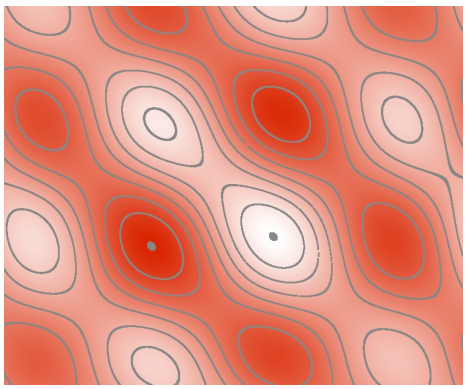
Maschinelles Lernen

Eine Einführung



Ehrhard Behrends, FU Berlin, WS 2016/17

Inhaltsverzeichnis

Einleitung

In den letzten Jahren ist immer öfter von „Big Data“ zu hören. Wenn es um mathematische Zusammenhänge geht, meint man damit das Problem, sinnvolle Informationen aus großen Datenmengen „automatisch“ zu extrahieren. Das können funktionale Zusammenhänge sein, Klassifizierungen, stochastische Abhängigkeiten usw. Hier einige typische Probleme:

- Welche Ziffern in $\{0, \ldots, 9\}$ sind gemeint? (Man soll zum Beispiel eine handgeschriebene Postleitzahl auswerten.)
- Es wurden viele Daten über Patienten gesammelt. Welche sind denn wesentlich, um eine spezielle Krankheit K vielleicht schon vor Ausbruch diagnostizieren zu können?
- Eine Bank hat viele Daten über einen Kunden zusammengetragen: Adresse, Einkommen, Alter, Schulden usw. Sollte man ihm einen Kredit von 10.000 Euro geben?
- Ein Surfer ruft immer wieder bestimmte Internetseiten auf. Handelt es sich vielleicht um einen Terroristen?
- Jemand kauft bei Amazon gewisse Bücher. Welche kann man ihm zusätzlich anbieten, bei welchen ist dann ein Kauf wahrscheinlich?

Gemeinsam ist allen Beispielen, dass es um riesige Datenmengen geht, dass ohne EDV also nichts zu machen sein wird. Klar ist auch, dass Stochastik eine Rolle spielen wird, denn Messungen sind in der Regel fehlerbehaftet.

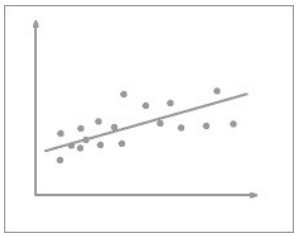
Wirklich kümmert sich die mathematische Statistik schon seit Jahrhunderten um derartige Fragen. Es sind aber viele neue Ideen dazugekommen, die zum Teil noch Gegenstand aktueller Forschung sind.

Durch „maschinelles Lernen“ sollen aber nicht nur praktische Probleme gelöst werden. Das Ziel ist viel ehrgeiziger: Man darf hoffen, durch das Modellieren von Lernprozessen auch zu verstehen, nach welchen Prinzipien „Lernen“ beim Menschen organisiert ist. Wie lernen Kinder ihre Muttersprache? Wie lernen wir ein Instrument? ...

In dieser Vorlesung soll es um einen Teilaspekt des Bereichs „Big Data“ gehen, den *mathematischen Hintergrund des maschinellen Lernens*. Wie der Name vermuten lässt, soll eine „Maschine“, d.h. ein Computer, in die Lage versetzt werden, mehr oder weniger selbständig aus vorgelegten Datensätzen etwas zu lernen. Die Kombination von zwei Ideen wird ein wichtige Rolle spielen:

- Ist „etwas Lineares“ zu lernen, so gibt es dafür viele wirksame Methoden. (Lineare Abhängigkeit des Ausgangs vom Eingang, Trennbarkeit von Daten durch einen Hyperebene usw.) Und manchmal ist es möglich, die Verfahren so zu beschreiben, dass nur die inneren Produkte der auftretenden Vektoren vorkommen.
- Wenn mit linearen Methoden nichts zu machen ist, so hat man viele Möglichkeiten, sie als Elemente eines Hilbertraumes aufzufassen und dort die „linearen“ Verfahren anzuwenden. De facto heißt das, dass – durch die richtige Wahl des Hilbertraums – auf diese Weise eine Fülle nichtlinearer Verfahren zur Verfügung stehen.

In Ansätzen kennt man diese Idee schon aus der „Elementaren Stochastik“.



Dort ist es doch oft sinnvoll, eine Gerade durch eine Punktwolke zu legen, und in der Abteilung „Lineare Regression“ lernt man, wie das optimal geht. Manchmal ist jedoch eher ein funktionaler Zusammenhang des Typs $x \mapsto ae^{bx}$ zu erwarten. Wie kann man a und b so finden, dass eine Punktwolke $(x_i, y_i), i = 1, \ldots, n$ möglichst gut approximiert wird, dass also stets $y_i \approx e^{bx_i}$ gilt? Nun ist $y = ae^{bx}$ gleichwertig zu $\log y = \log a + bx$. Deswegen approximiert man sinnvollerweise die Punktwolke der $(x_i, \log y_i)$ durch eine Gerade $\alpha + \beta x$ und kann dann die die gesuchten a, b aus $\log a = \alpha, b = \beta$ gewinnen.

Dieses Verfahren gehört zum Standard bei vielen biologischen Untersuchungen, früher hat man dazu „einfach logarithmisches Papier“ verwendet. Da war die eine Koordinate schon „gestaucht“. Man trug die (x_i, y_i) direkt ein und konnte dann (hoffentlich) sehen, dass sich die Punkte im Wesentlichen auf einer Geraden befinden[1)].

Man sollte noch erwähnen, dass sich längst nicht alle Aspekte des maschinelles Lernens mathematisch streng behandeln lassen. Viele Verfahren funktionieren hervorragend auch ohne exakte Analyse. Das lieben die Anwender, Mathematiker sind aber nicht wirklich zufrieden. Das hat das Thema mit einigen anderen Entwicklungen der Vergangenheit gemeinsam:

- Fuzzy Logik und Fuzzy Steuerung.
- Simulated Annealing.
- Data Mining.
- Neuronale Netze.
- Simulation durch schnell mischende Markovketten.
- ...

Das Skript ist wie folgt strukturiert. In *Kapitel 1* kümmern wir uns zunächst um ein klassiches Klassifikationsproblem: Kann man zwei Punktmengen linear trennen? Eine erste Antwort gibt der *Perzeptionsalgorithmus*, durch den eine trennende Hyperebene garantiert in einer vorher abschätzbaren Anzahl von Schritten gefunden werden kann, wenn die Mengen wirklich trennbar sind. Leider ist die Datenlage nicht immer so einfach, manchmal muss man einige Punkte

[1)] Doppelt logaritmisches Papier zum Eintragen der $(\log x_i, \log y_i)$ kam auch zum Einsatz, wenn ein Zusammenhang des Typs $y = ax^r$ vermutet wurde.

falsch klassifizieren. Doch wie sollen Fehler gewichtet werden? Im letzten Abschnitt besprechen wir erste Ideen, wie die Anwendbarkeit „linearer“ Verfahren stark erweitert werden kann.

Kapitel 2 ist Optimierungsproblemen gewidmet. Die Verfahren werden gebraucht, um konkrete Lösungen für Klassifikationsprobleme zu finden.

In *Kapitel 3* geht es dann um den funktionalanalytischen Hintergrund des heute als besonders erfolgversprechend angesehenen Ansatzes: Wir diskutieren *Hilberträume mit reproduzierendem Kern*: Was sind Kerne? Wie ist der Zusammenhang zu Hilberträumen? Wie kann man sich solche Räume vorstellen? Welche Verfahren werden dadurch möglich?

Danach, in *Kapitel 4*, wird es theoretischer: Es geht um eine Präsizierung der Problemstellung beim maschinellen Lernen. Man hat einen (unbekannten) Wahrscheinlichkeitraum vor sich, aus dem (theoretisch) beliebig viele Stichproben gezogen werden können. Wie kann man daraus auf optimale Weise auf funktionale Zusammenhänge oder Klassifzierungsmöglichkeiten schließen? Wesentlich wird die Festsetzung der Bewertung von Fehlern sein. Dazu führen wir die Begriffe „Verlustfunktion“ und „Risiko“ ein.

Am Ende stehen so genanne *Orakelungleichungen*: Wie oft muss man testen, um mit einer vorgegebenen Wahrscheinlichkeit (nahe bei 1) eine Funktion zu finden, die bis auf ε (klein!) nahe an dem optimal zu erreichenden Wert ist. Die Grundidee ist einfach, für die Präzisierung sind allerdings einige nichttriviale Ungleichungen aus der Stochastik vorzubereiten.

Eine letzte Bemerkung. Mehr noch als bei „Ethik-neutralen“ Gebieten wie etwa Topologie stellt sich natürlich die Frage, ob man sicher sein kann, dass mathematische Ergebnisse zum maschinellen Lernen nur zum Wohle der Menschheit eingesetzt werden können. Die Antwort ist ein klares Nein, und Beispiele sind auch schnell zu finden. Die Vergangenheit hat allerdings gezeigt, dass man nie sicher sein kann, welches Anwendungspotential ein Gebiet enthält. Ein berühmtes Beispiel einer angeblich „garantiert anwendungslosen“ Wissenschaft ist die Zahlentheorie, deren Ergebnisse seit einigen Jahrzehnten wegen ihrer Relevanz für die Kryptographie auch von Geheimdiensten verfolgt werden. Und selbst wenn man um die Gefahr weiß, überwiegen vielleicht die positiven Aspekte. Oder sollte man ein Brotmesser verbieten?

E. Behrends, Oktober 2016.

Bei der Vorbereitung dieser Vorlesung wurde die nachstehende **Literatur** verwendet:

E. Alpaydin: Introduction to Machine Learning.
MIT Press, 2014.

E. Behrends: An Introduction to Markov Chains with Special Emphasis on Rapid Mixing
Vieweg 1998.

E. Behrends: Elementare Stochastik.
Springer Spektrum, 2012.

E. Behrends: Mathematische Statistik
Skript zu einer Vorlesung an der FU.

N. Cristianini, J. Shawe-Taylor: An Introduction to Support Vector Machines.
Cambridge Univ. Press, 2000.

F. Cucker, St. Smale: On the Mathematical Foundations of Learning.
Bulletin of the AMS 39, 2001.

T. Hastie, R. Tibshirani, J. Friedman: The Elements of Statistical Learning.
Springer, 2009.

P. Huber, E. Ronchetti: Robust Statistics.
Wiley Series, 2009.

A. Klenke: Wahrscheinlichkeitstheorie.
Springer, 2005.

C. Rasmussen, C. Williams: Gaussian Processes for Machine Learning.
Vieweg, 2000.

R. Schapire, Y. Freund: Boosting.
Cambridge University Press, 2012.

I. Steinwart, A. Christmann: Support Vector Machines.
Springer, 2008.

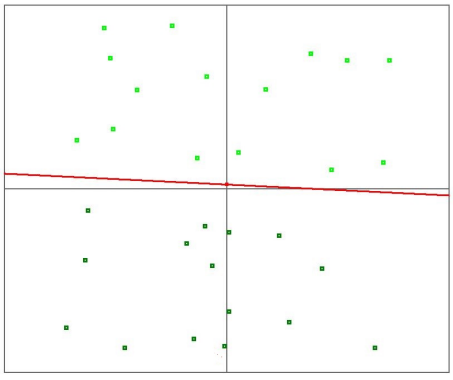
Kapitel 1

Maschinelles Lernen: Grundlagen

Beim „maschinellen Lernen“ geht es darum, aus umfangreichen Datenmengen Informationen zu destillieren. *Klassifikation* und *Regression* spielen dabei eine wichtige Rolle. Sehr wirkungsvoll wird dabei die Idee eingesetzt, durch Transformation lineare Methoden auf nichtlineare Situationen anwenden zu können. In diesem Kapitel sammeln wir erste Ergebnisse zu diesen Fragenkreisen.

1.1 Klassifikation (der Perceptron-Algorithmus)

Gegeben seien l Punkte im $\mathbb{R}^n$, die ein klassifizierendes Merkmal tragen, etwa wie die hell- und dunkelgrünen Punkte im nachstehenden Bild.



Die rote Gerade trennt die Punktmengen: Wenn sie (für einen geeigneten Vektor w und eine geeignete Zahl b) die Form $\phi_{w,b} : x \mapsto \langle w, x\rangle + b$ hat, so heißt das: Für hellgrüne Punkte ist $\phi_{w,b}$ größer als Null und für dunkelgrüne kleiner als Null. Die Hoffnung: Wenn ein neuer Punkt dazukommt, kann man

durch Auswertung von $\phi_{w,b}$ bei diesem Punkt prognostizieren, ob er hell- oder dunkelgrün ist.

Auch wenn die rote Gerade nicht eingezeichnet worden wäre, hätte man durch „scharfes Hinsehen“ eine trennende Gerade gefunden. Das ist in höheren Dimensionen leider nicht möglich. In der nachstehenden Tabelle sind Punkte aufgeführt, die für eine geeignete lineare Abbildung nach Einsetzen zu den Ergebnissen „größer“ und „kleiner“ führen. Diesmal gibt es keine anschauliche Möglicheit, die trennende Hyperebene zu raten. Mit den gleich zu besprechenden Methoden geht es aber doch!

kl	-0,71227	-0,50792	0,92064	-0,84912
kl	-0,49478	0,62308	0,09626	-0,85447
kl	0,62970	0,84365	0,39890	-0,29728
gr	-0,13983	-0,65838	-0,76784	0,66218
kl	0,00411	-0,30110	-0,27231	-0,70698
gr	0,96130	0,68525	-0,41067	0,88932
kl	0,57664	-0,10215	0,16574	0,50830
gr	-0,79397	-0,03476	0,34526	0,81979
kl	0,83874	-0,46545	0,36688	-0,46420
gr	0,12030	-0,23942	-0,85911	0,38891
gr	-0,88058	-0,48432	-0,82692	0,12827
kl	0,18284	-0,96098	-0,31381	0,50185
gr	-0,27811	0,24412	-0,96761	0,69388
gr	0,27432	0,40324	-0,00947	0,18600
gr	0,68536	0,42757	0,04189	0,51106
gr	0,75163	-0,53814	-0,78181	0,86825
kl	-0,82829	0,01214	0,61568	-0,93172
gr	0,14707	0,32560	-0,83115	-0,22261
kl	0,47720	-0,75907	0,77785	0,52896

Formalisiert liest sich das so: Es seien $x_1, \ldots, x_l \in \mathbb{R}^n$ und $y_1, \ldots, y_l \in \{-1, +1\}$. Ein Punkt x_i ist demnach je nach y_i als $+1$ bzw. -1 klassifiziert. Wir setzen voraus, dass es ein $w \in \mathbb{R}^n \setminus \{0\}$ und ein $b \in \mathbb{R}$ gibt, so dass $\langle w, x_i \rangle + b \geq 0$ (bzw. ≤ 0) für die i mit $y_i = 1$ (bzw. $y_i = -1$). Eleganter kann man das so schreiben:

$$y_i(\langle w, x_i \rangle + b) \geq 0 \text{ für } i = 1, \ldots, l.$$

Später soll ein x nach $+1$ oder -1 klassifiziert werden, je nachdem, wie das Vorzeichen von $\langle w, x_i \rangle + b$ ist. Das wird man sicher umso zuverlässiger machen können, je größer die Lücke zwischen den $+1$-Punkten und den -1-Punkten ist. Das motiviert, warum die Zahl

$$\gamma := \min_i y_i(\langle w, x_i \rangle + b)$$

eine wichtige Rolle spielen wird. Da die Hyperebene $\{w \mid \langle w, x_i \rangle + b = 0\}$ nur von der Richtung von w, nicht aber von der Länge abhängt, werden wir immer auf $||w|| = 1$ normieren. (Andernfalls könnte man γ beliebig klein machen.)

Wir setzen voraus: Es gibt $\hat{w}$ mit $||\hat{w}|| = 1$ und $\hat{b}$, so dass

$$\hat{\gamma} := \min_i y_i(\langle \hat{w}, x_i \rangle + \hat{b})$$

strikt positiv ist[1)]. Die Punkte mit $y_i = 1$ können also von den Punkten mit y_i strikt getrennt werden. Wir kennen $\hat{w}$ und $\hat{b}$ allerdings nicht, und das Problem

[1)] Übrigens ist $|\langle \hat{w}, x_i \rangle + b|$ der euklidische Abstand von x_i zu der durch w und b definierten Hyperebene.

besteht darin, etwas anderes, ebenfalls Trennendes zu finden. Die Lösung ist im folgenden Perzeptronalgorithmus (Novikoff, 1960) zu finden:

Satz 1.1.1. *Die Bezeichnungen und die Voraussetzung (lineare Trennbarkeit) seien wie vorstehend, und es sei* $R := \max_i ||x_i||$. *Betrachte das folgende Verfahren:*

- *Setze* $w_0 := 0 \in \mathbb{R}^n$, $b_0 := 0 \in \mathbb{R}$, $\hat{r} := 0$. *Wir nehmen an, dass* w_r, b_r *für* $r = 0, \ldots, \hat{r}$ *schon konstruiert sind.*
- *Überprüfe die Zahlen* $y_i(\langle x_i, w_{\hat{r}} \rangle + b_{\hat{r}})$ *für* $i = 1, \ldots, l$.
  - *Fall 1: Alle sind strikt positiv. Dann soll das Verfahren abbrechen.*
  - *Fall 2: Es gibt* j *mit* $y_j(\langle x_j, w_{\hat{r}} \rangle + b_{\hat{r}}) \leq 0$. *Mache dann ein update:*

$$w_{\hat{r}+1} := w_{\hat{r}} + y_j x_j, \ b_{\hat{r}+1} := b_{\hat{r}} + y_j, \ \hat{r} \mapsto \hat{r} + 1.$$

  Starte wieder beim zweiten Schritt, wobei $\hat{r}$ *um Eins erhöht wurde.*

Dann gilt: Nach spätestens $(2R/\hat{\gamma})^2$ *Schritten stoppt der Algorithmus mit einem* $w = w_{\hat{r}}$, *für das*

$$y_i(\langle w, x_i \rangle + b) > 0$$

für alle i *gilt.*

Beweis: Um die Idee herauszuarbeiten, lösen wir in einem Vorlauf ein etwas einfacheres Problem. Wir setzen voraus, dass es ein $\hat{w}$ mit $||\hat{w}|| = 1$ gibt, so dass

$$\hat{\gamma} := \min_i y_i \langle \hat{w}, x_i \rangle$$

strikt positiv ist. Die Punkte können also durch eine durch Null gehende Hyperebene strikt getrennt werden. Der neue Algorithmus sieht so aus:

- Setze $w_0 := 0 \in \mathbb{R}^n$, $k_0 := 0 \in \mathbb{N}_0$, $\hat{r} := 0$. Wir nehmen an, dass $w_{\hat{r}}, k_{\hat{r}}$ für $r = 0, \ldots, \hat{r}$ schon konstruiert sind.
- Überprüfe die Zahlen $y_i \langle w_{\hat{r}}, x_i \rangle$ für $i = 1, \ldots, l$.
  - Fall 1: Alle sind strikt positiv. Dann soll das Verfahren abbrechen.
  - Fall 2: Es gibt j mit $y_j(\langle x_j, w_{\hat{r}} \rangle) \leq 0$. Mache dann ein update:

$$w_{\hat{r}+1} := w_{\hat{r}} + y_j x_j, \ \hat{r} \mapsto \hat{r} + 1.$$

  Starte wieder beim zweiten Schritt, wobei $\hat{r}$ um Eins erhöht wurde.

Wie viele Schritte wird man höchstens brauchen? Dazu überlegen wir, was bei einem update passiert. Wie verändern sich die $\langle \hat{w}, w_{\hat{r}} \rangle$ nachdem ein x_j durch $w_{\hat{r}}$ falsch klassifiziert wurde?

$$\begin{aligned} \langle \hat{w}, w_{\hat{r}+1} \rangle &= \langle \hat{w}, w_{\hat{r}} \rangle + y_j \langle \hat{w}, x_j \rangle \\ &\geq \langle \hat{w}, w_{\hat{r}} \rangle + \hat{\gamma}. \end{aligned}$$

In jedem Korrekturschritt wird also $\langle \hat{w}, w_{\hat{r}} \rangle$ um $\hat{\gamma}$ größer, d.h.

$$\langle \hat{w}, w_r \rangle \geq r\hat{\gamma}.$$

Andererseits ist

$$\begin{aligned} ||w_{\hat{r}+1}||^2 &= \langle w_{\hat{r}} + y_j x_j, w_{\hat{r}} + y_j x_j \rangle \\ &= ||w_{\hat{r}}||^2 + 2y_j \langle x_j, w_{\hat{r}} \rangle + ||x_j||^2 \\ &\leq ||w_{\hat{r}}||^2 + R^2. \end{aligned}$$

Es folgt: $||w_r||^2 \leq rR^2$. Und daraus schließen wir mit der Cauchy-Schwarz-Ungleichung, dass

$$r\hat{\gamma} \leq \langle \hat{w}, w_r \rangle \leq ||\hat{w}||||w_r|| \leq \sqrt{r}R,$$

d.h $r \leq (R/\hat{\gamma})^2$. Das ist sogar um den Faktor 4 besser als die Behauptung, diese Verbesserung ergibt sich durch die Annahme $\hat{b} = 0$.

Der allgemeine Fall kann ähnlich behandelt werden, Die Grundidee: Trennen im $\mathbb{R}^n$ durch eine allgemeine Hyperebene erreicht man durch Trennen im $\mathbb{R}^{n+1}$ durch eine durch Null gehende Hyperebene.

Zunächst bemerken wir, dass wir o.E. $R = 1$ annehmen dürfen. Denn angenommen, wir haben den Satz für diesen Fall schon bewiesen. Ist dann eine beliebige Situation vorgelegt, so gehe von den x_i zu den x_i/R über. Dann muss das $\hat{\gamma}$ allerdings geändert werden: Aus $y_i(\langle x_i, \hat{w} \rangle + \hat{b}) \geq \hat{\gamma}$ folgt $y_i(\langle x_i/R, \hat{w} \rangle + \hat{b}/R) \geq \hat{\gamma}/R$. Es ist also $\hat{\gamma}$ durch $\hat{\gamma}/R$ zu ersetzen. Da die Norm der x_i/R höchstens Eins ist und der Satz dafür schon bewiesen sein soll, würde $r \leq \big(2/(\hat{\gamma}/R)\big)^2$ folgen, d.h. $r \leq (2R/\hat{\gamma})^2$ wie behauptet.

Es nehmen also $R = 1$ an. Wir stellen zunächst fest, dass dann $|\hat{b}| \leq 1$ gelten muss, denn ist etwa $y_1 = 1$ und $y_2 = -1$, so folgt

$$-\hat{b} \leq \langle x_1, \hat{w} \rangle - \hat{\gamma} \leq \langle x_1, \hat{w} \rangle \leq 1$$

und

$$\hat{b} \leq -\langle x_2, \hat{w} \rangle - \hat{\gamma} \leq -\langle x_2, \hat{w} \rangle \leq 1.$$

Der Trick besteht darin, die Situation in den $\mathbb{R}^{n+1}$ einzubetten. Wir gehen von den x_i, y_i (mit $x_i \in \mathbb{R}^n$) zu den $\tilde{x}_i, y_i$ über, wo $\tilde{x}_i := (x_i, 1) \in \mathbb{R}^{n+1}$. Dabei haben wir die Vektoren im $\mathbb{R}^{n+1}$ als (z, a) mit $z \in \mathbb{R}^n$ und $a \in \mathbb{R}$ geschrieben[2)]. Die inneren Produkte sind leicht auszurechnen: $\langle (z_1, a_1), (z_2, a_2) \rangle = \langle z_1, z_2 \rangle + a_1 a_2$. (Wir haben das gleiche Symbol $\langle \cdot, \cdot \rangle$ für das Skalarprodukt im $\mathbb{R}^n$ und im $\mathbb{R}^{n+1}$ verwendet.)

Wir setzen $\tilde{w} := (\hat{w}, \hat{b})$, es gilt dann $y_i \langle \tilde{x}_i, \tilde{w} \rangle = y_i(\langle x_i, \hat{w} \rangle + \hat{b}) \geq \hat{\gamma}$. Das machen wir auch für die im Algorithmus konstruierten Vektoren: $\tilde{w}_r := (w_r, b_r)$. Wir wissen aufgrund der Abschätzung $|\hat{b}| \leq 1$ schon, dass $||\tilde{w}|| \leq \sqrt{2}$. Nun verfolgen wir unseren Algorithmus im $\mathbb{R}^{n+1}$.

[2)] Wenn man Vektoren konsequent als Spalten schreibt, müsste es eigentlich $(z^\top, a)^\top$ heißen.

Angenommen, bis zur Nummer $\hat{r}$ st schon alles konstruiert, und es muss ein Update geben: Es gibt ein j mit $y_j(\langle x_j, w_{\hat{r}}\rangle + b_{\hat{r}}) \leq 0$, d.h. $y_j\langle \tilde{w}_{\tilde{r}}, \tilde{x}_j\rangle \leq 0$. Dann gehen wir zu $w_{\hat{r}+1}, b_{\hat{r}+1}$ über, und dabei ist die Definition so, dass $\tilde{w}_{\hat{r}+1} = \tilde{w}_{\hat{r}} + y_i\tilde{x}_i$. Es folgt

$$\begin{aligned}\langle \tilde{w}_{\hat{r}+1}, \tilde{w}\rangle &= \langle \tilde{w}_{\hat{r}} + y_i\tilde{x}_i, \tilde{w}\rangle \\ &= \langle \tilde{w}_{\hat{r}}, \tilde{w}\rangle + y_i\langle \tilde{x}_i, \tilde{w}\rangle \\ &\geq \langle \tilde{w}_{\hat{r}}, \tilde{w}\rangle + \hat{\gamma}.\end{aligned}$$

Wie im Fall $\hat{b} = 0$ folgt also $\langle \tilde{w}_r, \tilde{w}\rangle \geq r\hat{\gamma}$.
Für die Normen gilt (wegen $y_j\langle \tilde{w}_{\tilde{r}}, \tilde{x}_j\rangle \leq 0$):

$$\begin{aligned}||\tilde{w}_{\hat{r}+1}||^2 &= ||\tilde{w}_{\hat{r}}||^2 + 2y_j\langle \tilde{w}_{\tilde{r}}, \tilde{x}_j\rangle + ||\tilde{x}_j||^2 \\ &\leq ||\tilde{w}_{\hat{r}}||^2 + ||\tilde{x}_j||^2,\end{aligned}$$

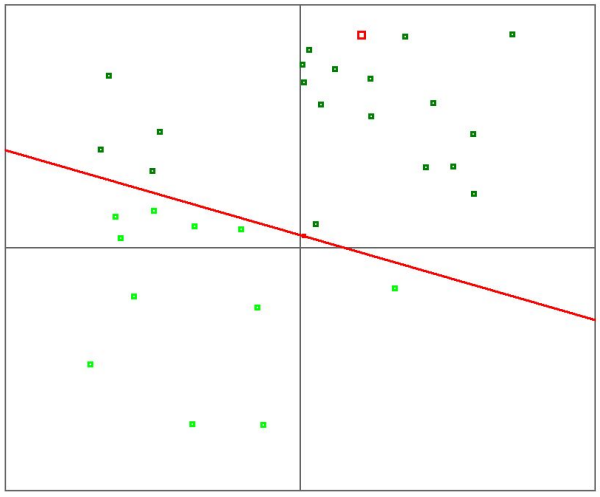
und das impliziert (wegen $||\tilde{x}_j||^2 \leq 2$) $||\tilde{w}_r||^2 \leq 2r$.
Wir kommen zum Finale:

$$\begin{aligned}r\hat{\gamma} &\leq \langle \tilde{w}_r, \tilde{w}\rangle \\ &\leq ||\tilde{w}_r||\,||\tilde{w}|| \\ &\leq \sqrt{2r}\sqrt{2},\end{aligned}$$

d.h. $r \leq (2/\hat{\gamma})^2$ wie behauptet. □

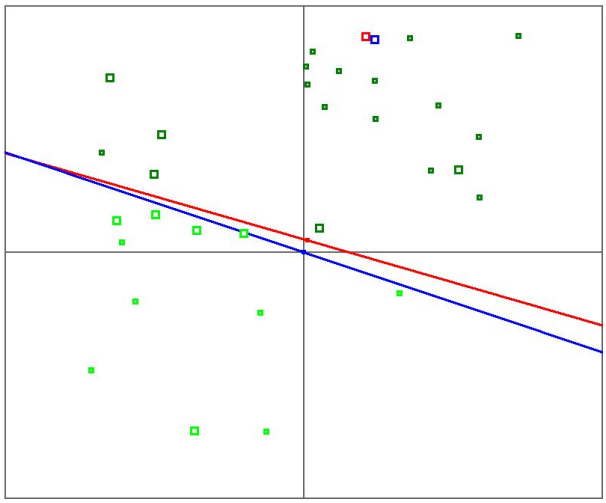
Bemerkung: Es ist sehr bemerkenswert, dass die Maximalzahl der updates nur von R und $\hat{\gamma}$ abhängt: Wie groß die Dimension des Raumes ist und wie viele Vektoren beteiligt sind, geht nicht explizit ein. Man kann den $\mathbb{R}^n$ sogar bei gleichem Beweis durch irgendeinen Raum mit Skalarprodukt ersetzen.

Hier ist ein *Beispiel*. Es wurden zunächst eine zufällige Hyperebene (rot) erzeugt (der zugehörige Einheitsvektor ist rot gekennzeichnet). Dann wurden 30 Punkte unter- und oberhalb generiert ($y = \pm 1$). Das sah dann so aus:



Eine Hyperebene und 30 Punkte

Es schloß sich ein Perceptron-Algorithmus an. Nach nur 4 Durchgängen (!) war eine trennende Hyperebene (blau) gefunden:



Der Perceptron-Algorithmus findet eine trennende Hyperebene

Das klappt auch dann, wenn man das Ergebnis nicht veranschaulichen kann. Hier ist der Ausschnitt aus einer Tabelle von 30 trennbaren Punkten im $\mathbb{R}^5$:

kl	-0,04584	0,34937	-0,37748	-0,81499	-0,96162
gr	-0,70454	-0,03991	0,82076	0,51706	0,99117
gr	0,96392	-0,46539	-0,09824	-0,29911	0,98866
gr	-0,23653	0,19490	0,02610	0,48841	0,51172
kl	-0,41213	0,27092	-0,78028	-0,75873	0,28612
kl	-0,55745	-0,38161	0,63064	-0,32009	-0,01992
gr	0,81890	0,54708	-0,42045	0,80569	0,28516
gr	0,07522	-0,26616	-0,20821	0,52128	0,94612
gr	0,35045	-0,76206	0,83127	0,21057	0,83700
gr	0,45132	0,59622	0,50010	0,48163	0,02443
gr	-0,96213	0,71981	-0,70830	0,90380	0,93724
gr	0,51647	-0,29047	-0,15651	0,27707	0,36579
kl	0,42378	-0,66288	0,27686	-0,37739	-0,00467
gr	0,62210	0,78924	-0,62226	0,51628	0,84996
gr	-0,53691	0,86796	-0,25526	0,24093	0,03385

30 trennbare Punkte im $\mathbb{R}^5$

Der Perceptron-Algorithmus braucht 7 Durchgänge, der zugehörige w-Vektor ist gleich $w = (0,64384, 0,36296, -0,20539, 0,47703, 0,42836)^t op$.

1.2 Maximaler Rand und Schlupfvariable

Zur Erinnerung: Gegeben sind Punkte, die zu den Klassen $y = 1$ und $y = -1$ gehören. Die sind durch eine unbekannte Hyperebene trennbar. (Die Menge dieser Punkte könnte man als Trainingsmenge interpretieren.) Und dann sucht man eine trennende Hyperebene in der Hoffnung, bei zukünftig vorgelegten Punkten die Klasse durch die Lage relativ zur Hyperebene ablesen zu können.

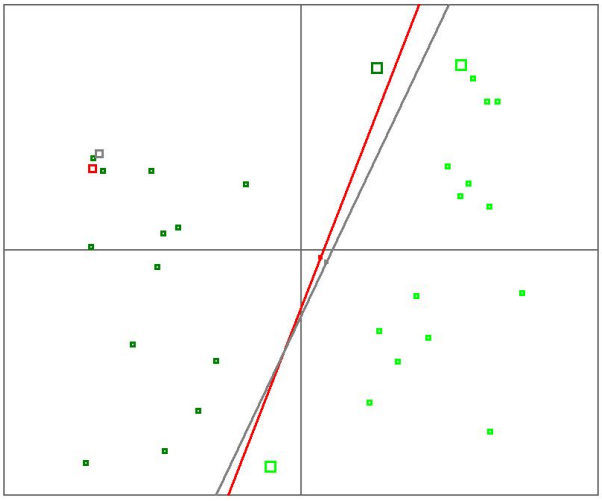
Dabei ist es im Interesse einer sicheren Klassifizierung sicher wünschenswert, den Minimalbstand beider Klassen zur trennenden Hyperebene so groß wie möglich zu haben. Nun gilt: Ist $w \in \mathbb{R}^n$ ein Einheitsvektor und $b \in \mathbb{R}$, so ist der Abstand von einem $x \in \mathbb{R}^n$ zur Hyperebene $H = \{\langle \cdot, w\rangle + b = 0\}$ gleich $|\langle x, w\rangle + b|$.

> Begründung: Der Vektor y bester Approximation in H erfüllt doch die Bedingung $x - y \in H^\perp = \mathbb{R}w$. Also ist $x - y = \alpha w$ für ein geeignetes α. D.h. $y = x - \alpha w$ und $y \in H$, es folgt $\langle x - \alpha w, w\rangle + b = 0$ oder $\alpha = \langle x, w\rangle + b$. Beachte noch $||x - y|| = ||\alpha w|| = |\alpha|$.

Man möchte also

$$\gamma := \min_i |\langle x_i, w\rangle + b|$$

so groß wie möglich machen, wenn $\{\langle \cdot, w\rangle + b = 0\}$ alle trennenden Hyperebenen (mit $||w|| = 1$) durchläuft. Im folgenden Bild ist die optimale Hyperebene grau eingezeichnet. Die Punkte, bei denen der minimale Abstand realisiert wird, sind hervorgehoben. Das sind die so genannten *Support-Vektoren*, die den support vector machines den Namen gaben.



Eine optimale Hyperebene (grau) und Supportvektoren

Für die Bestimmung der optimalen Hyperebene ist das folgende Lemma nützlich (der *hard margin classifiyer*):

Lemma 1.2.1. *Gegeben seien die linear trennbaren* $(x_i, y_i) \in \mathbb{R}^n \times \{-1, 1\}$ *für* $i = 1, \ldots, l$. *Betrachte das folgende Optimierungsproblem:*

- *Minimiere* $||w||^2$ *unter den Nebenbedingungen* $y_i(\langle x_i, w\rangle + b) \geq 1$, *wobei* $i = 1, \ldots, l$.

(Ausführlich heißt das: Betrachte im $\mathbb{R}^n \times \mathbb{R}$ die Menge aller (w, b), für die $y_i(\langle x_i, w\rangle + b) \geq 1$ für alle i gilt. Nach Voraussetzung ist diese Menge nicht leer. Minimiere darauf die Funktion $(w, b) \mapsto ||w||^2$.)

(i) Das Minimierungsproblem hat eine eindeutig bestimmte Lösung.

(ii) Wenn die Hyperebene $\{\langle \cdot, w\rangle + b = 0\}$ mit normiertem w und maximalem Rand γ trennt, so löst $(w/\gamma, b/\gamma)$ das Optimierungsproblem.

(iii) Sei umgekehrt (w, b) eine Lösung des Optimierungsproblems. Dann führt die Hyperebene $\{\langle \cdot, w/||w||\rangle + b/||w|| = 0\}$ zum maximalen Rand.

Beweis: (i) w_0, b_0 mögen strikt trennen, d.h. o.B.d.A. gilt $y_i(\langle x_i, w_0\rangle + b_0) \geq 1$ für $i = 1, \ldots, l$. Wir brauchen uns nur um die w mit $||w|| \leq ||w_0||$ zu kümmern.

Sei etwa $y_1 > 0$ und $y_2 < 0$, und w, b erfülle die Bedingugnen. Dann gilt (mit $M := \max ||x_i||$)

$$\langle x_1, w\rangle + b \geq 1, \; -(\langle x_2, w\rangle + b) \geq 1,$$

d.h.

$$b \geq 1 - \langle x_1, w\rangle \geq 1 - ||x_1||\,||w|| \geq 1 - M||w_0|| =: A$$

sowie

$$b \leq -1 - \langle x_2, w\rangle \leq -1 + ||x_2||\,||w|| \leq -1 + M||w_0|| =: B.$$

Es sind also nicht alle w, b in der Konkurrenz, sondern nur die mit $||w|| \leq ||w_0||$ und $b \in [\,A, B\,]$. Das ist eine kompakte Menge, und deswegen wird das Minimum angenommen.

Das Minimum ist auch eindeutig bestimmt. Angenommen, es wird bei w_1, b_1 und bei w_2, b_2 angenommen. Wegen der Konvexität des Problems dann auch bei $w := (w_1 + w_2)/2, b := (b_1 + b_2)/2$. Notwendig ist $w_1 = w_2$, denn andernfalls wäre (wegen der strikten Konvexität von $w \mapsto ||w||^2$) $||w|| < ||w_1||$.

Es muss auch $b_1 = b_2$ gelten. Angenommen, es wäre $b_1 < b_2$. Dann wäre $y_i(\langle x_i, w\rangle + b) > 1$ für $i = 1, \ldots, l$, und wir könnten unter Erhalt der Ungleichungen von w zu $w/(1 + \varepsilon)$ mit einem $\varepsilon > 0$ übergehen.

Die Gültigkeit von (ii) und (iii) liegt daran, dass $y_i(\langle x_i, w\rangle + b) \geq \gamma$ gleichwertig zur Ungleichung $y_i(\langle x_i, w/\gamma\rangle + b/\gamma) \geq 1$ ist. □

Ganz analog kann man zeigen:

Satz 1.2.2. *Die $(x_i, y_i) \in \mathbb{R}^n \times \{-1, +1\}, i = 1, \ldots, l$, seien durch eine durch Null gehende Hyperebene trennbar. Unter diesen Hyperebenen (also den Mengen $\{\langle \cdot, w\rangle = 0\}$) erhält man diejenige mit maximalem Rand durch Lösung des Optimierungsproblems*

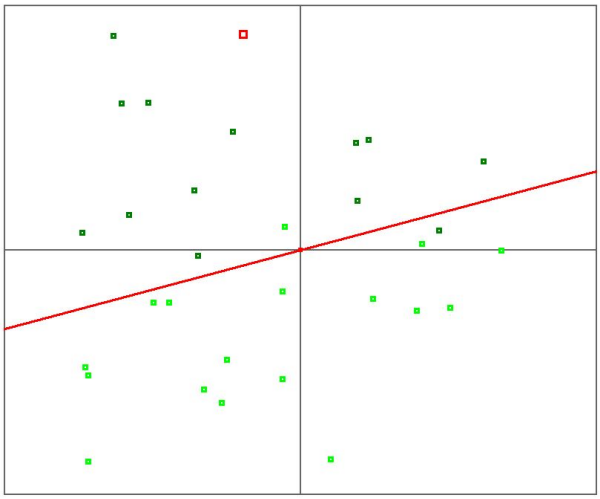
- $||w||^2 =$ *minimal unter den Nebenbedingungen $y_i\langle x_i, w\rangle \geq 1$ für alle i.*

Man beachte auch: $y_i\langle x_i, w\rangle \geq 1$ ist gleichwertig zu $\langle y_i x_i, w\rangle \geq 1$. Man sucht also einen Halbraum der Form $\{\langle \cdot, w\rangle\} \geq 1$ mit möglichst großem Abstand zur 0, so dass alle $y_i x_i$ in diesem Halbraum liegen. Deswegen ist es plausibel, dass

der optimale Vektor eine Linearkombination mit positiven Koponenten der $y_i x_i$ ist.

Angenommen, die $(x_i, y_i) \in \mathbb{R}^n$ sind vorgelegt. Es sind folgende Fälle zu berücksichtigen:

- Sie können linear getrennt werden, und das entsprechende γ ist beruhigend groß. Um diese Situation haben wir uns bisher gekümmert.
- Sie können zwar linear getrennt werden, aber das zugehörige γ ist – verursacht durch einen oder wenige „Ausreißer“ – sehr sehr klein. Wenn man die Ausreißer wegließe, wäre die Klassifikation viel überzeugender.
- Sie können nicht linear getrennt werden. Es gibt zwar eine „im Wesentlichen“ trennende Hyperebene, aber bei der liegen einige Punkte „ein bisschen“ auf der falschen Seite. (Siehe das nachstehende Bild.)



Die Hyperebene trennt nur „im Wesentlichen

Damit stellt sich das (auch in anderen Bereichen der Mathematik auftretende) Problem:

Wie geht man mit Fehlern um, wie wichtig nimmt man sie?

Anders ausgedrückt: Wenn die $z_i \in \mathbb{R}$ die richtigen Werte sind, man aber $z_i + \delta_i$ misst oder prognostiziert, wie sollte man dann die δ_i wichten[3)]? Die wichtigsten Antworten sind die folgenden:

[3)] Allgemeiner gibt es ähnliche Probleme immer, wenn man die „Größe“ von Vektoren vergleichen möchte.

- Mit der l^2-Norm: Maß für den Fehler ist $\left(\sum \delta_i^2\right)^{1/2}$. Dieses Fehlermaß ist in der Wahrscheinlichkeitstheorie und Statistik sehr verbreitet. Es hat den großen Vorteil, dass oft Hilbertraummethoden eingesetzt werden können.

  Inhaltlich bedeutet es, dass Fehler $\delta_i \in]-1,1[$ „abgemildert“ und Fehler $\delta_i \notin [-1,1]$ „besonders schwer gewichtet“ werden.

- Mit der l^1-Norm: Maß für den Fehler ist $\sum |\delta_i|$. Das ist eigentlich ein „faireres“ Fehlermaß, allerdings ist es strukturell viel schlechter zu behandeln[4)].

- Mit der l^∞-Norm: Maß für den Fehler ist $\max |\delta_i|$. Für manche Zwecke, zum Beispiel bei der Qualitätskontrolle, ist dieses Fehlermaß angemessen.

(Ganz allgemein kann man für $p > 0$ die Zahl $\left(\sum |\delta_i|^p\right)^{1/p}$ betrachten. Die vorstehenden Beispiele sind als Spezialfall enthalten.)

Zurück zum Klassifizierungsproblem. Im Idealfall hätte man gern die Ungleichung $y_i(\langle x_i, w\rangle + b) \geq 1$ (mit minimalem $||w||$), doch realistischer Weise wird man manchmal nur $y_i(\langle x_i, w\rangle + b) \geq 1 - \zeta_i$ mit hoffentlich kleinen ζ_i erreichen können. Das sind die *Schlupfvariablen* („slack variables“), sie messen, wie weit x_i „auf der falschen Seite“ der durch w, b definierten Hyperebene liegt.

Das führt zu *zwei Klassifizierungsansätzen*:

Soft margin classifier I: Die x_i, y_i seien wie bisher. Wir setzen aber nicht voraus, dass es eine trennende Hyperebene gibt. Bestimme dann $w \in \mathbb{R}^n$ und $b \in \mathbb{R}$ so, dass gilt:

- $y_i(\langle x_i, w\rangle + b) \geq 1 - \zeta_i$, wobei $\zeta_i \geq 0$ $(i = 1 \ldots, l)$.
- $||w||^2/2 + C \sum_i \zeta_i^2$ ist minimal.

Dabei ist $C > 0$ eine Konstante. Sie ist ein Maß dafür, wie wichtig wir Fehlklassifizierungen nehmen. (Der Faktor $1/2$ vor $||w||^2$ dient der Bequemlichkeit, denn bei der Lösung des Optimierungsproblems muss abgeleitet werden.)

Soft margin classifier II: Die x_i, y_i seien wie bisher. Wir setzen aber nicht voraus, dass es eine trennende Hyperebene gibt. Bestimme dann $w \in \mathbb{R}^n$ und $b \in \mathbb{R}$ so, dass gilt:

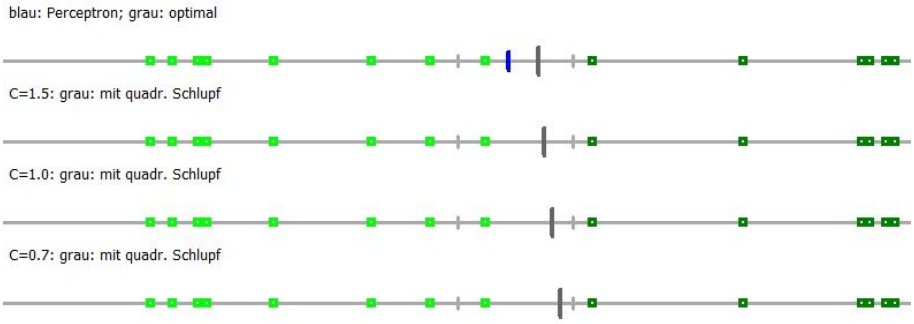
- $y_i(\langle x_i, w\rangle + b) \geq 1 - \zeta_i$, wobei $\zeta_i \geq 0$ $(i = 1 \ldots, l)$.
- $||w||^2/2 + C \sum_i |\zeta_i| = (||w||^2/2 + C \sum_i \zeta_i)$ ist minimal.

Dabei ist $C > 0$ eine Konstante. Sie ist ein Maß dafür, wie wichtig wir Fehlklassifizierungen nehmen.

Wie man diese Optimierungsprobleme behandeln kann, wird im nächsten Kapitel besprochen werden.

[4)] Man kann zum Beispiel den Median auf diese Weise einführen. Da sieht man schon die ersten Nachteile, denn Eindeutigkeit ist im Allgemeinen nicht mehr gegeben.

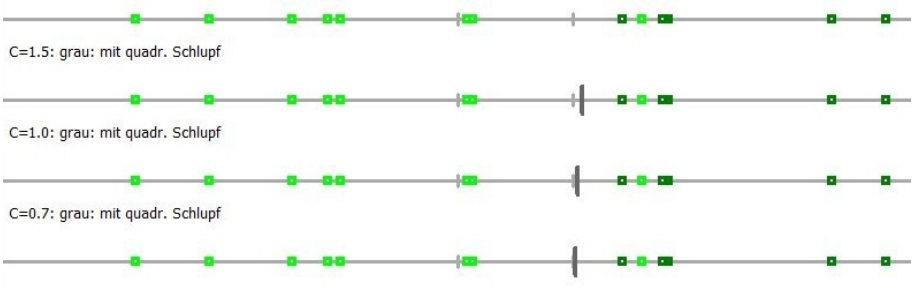
Zur Illustration gibt es jetzt noch zwei Beispiele zum soft margin classifier I. Es geht um 15 Punkte im $\mathbb{R}^1$, man muss also im $\mathbb{R}^{17}$ optimieren. Zunächst wurden 15 linear trennbare Punkte erzeugt:



Trennbare x_i: Schlupfvariable zu verschiedenen C.

- In jeder Zeile sind die x_i aufgeführt, die verschiedenen Klassen sind verschieden gefärbt.
- Die kurzen senkrechten Striche markieren die Werte 0 und 1
- Der Perzetron-Algorithmus hat im obersten Bild die blau gekennzeichnete Stelle gefunden, daneben ist in grau der Wert tingezeichnet, der optimal trennt.
- In den folgenden Zeilen sieht man die Lösungen (grau) für verschiedene Werte des Parameters C. Ist C groß, will man also Fehler vermeiden, so stimmt die Trennung mit dem optimalen Wert überein. Für kleinere C wandern die Werte nach links. Das ist auf die große Lücke zwischen dem ersten und zweiten Punkt von links der dunkelgrünen Werte zurückzuführen.

Und hier noch eine Situation, die nicht trennbar ist:



Nicht trennbare x_i: Schlupfvariable zu verschiedenen C.

- Mit kleiner werdendem C nimmt die Größe der ζ_i zu. Mehr und mehr wird der „falsch liegende“ Punkt einfach ignoriert.

1.3 Regression

In der Einleitung wurde schon daran erinnert, dass man manchmal eine „Punktwolke“ $(x_i, y_i)_{i=1,\dots,n}$ im $\mathbb{R}^2$ näherungsweise durch einen funktionalen Zusammenhang erklären möchte: Gesucht ist eine Funktion ϕ, so dass $\phi(x_i) \approx y_i$.

In vielen Fällen ist dabei ϕ eine affine Funktion. In der Theorie der *linearen Modelle* wird das wesentlich verallgemeinert. Da sind z_i $(i = 1, \dots, n)$ Vektoren im $\mathbb{R}^s$, die y_i sind Zahlen, und man sucht $\gamma \in \mathbb{R}^s$, so dass $y_i \approx \langle z_i, \gamma \rangle$.

> Diese Idee kann sehr vielfältig eingesetzt werden. Vermutet man zum Beispiel in der „Punktwolke“ $(x_i, y_i)_{i=1,\dots,n}$ des $\mathbb{R}^2$ einen Zusammenhang der Form $y_i \approx \gamma_0 + \gamma_1 x_i + \gamma_2 x_i^2$, so setze man $z_i := (1, x_i, x_i^2) \in \mathbb{R}^3$ und behandle das zugehörige lineare Modell.

Die Lösung sieht so aus. Definiere eine $n \times s$-Matrix A (die Designmatrix) als diejenige Matrix, deren Zeilen die z_i sind, und $y \in \mathbb{R}^n$ ist der Vektor $(y_1, \dots, y_n)^\top$. Das optimale γ ist dann durch

$$\hat{\gamma} := (A^\top A)^{-1} A^\top y$$

gegeben (Satz von Gauß-Markov). Einzelheiten findet man in Kapitel 4 meines Skripts zur mathematischen Statistik.

Auch diesen Ansatz wollen wir durch Transformation in einen neuen Raum verallgemeinern.

1.4 Übertragung ins Nichtlineare: die Featureabbildung

Nun wollen wir vom Linearen ins Nichtlineare gehen, die Idee soll am Klassifizierungsproblem erläutert werden. Gegeben ist eine Menge X und darin Punkte x_i, $i = 1, \dots, l$. Die x_i sollen zwei Klassen angehören, die durch -1 und 1 bezeichnet werden. Gegeben sind also $y_i \in \{-1, 1\}$. Gesucht ist eine Funktion $\phi : X \to \mathbb{R}$, für die $y_i \phi(x_i) > 0$ für alle i gilt. Das Ziel: Soll ein weiterer Punkt klassifiziert werden, so berechne $\phi(x)$, und je nach Vorzeichen dieser Zahl wird x als -1 oder $+1$ klassifiziert.

In Abschnitt 1.1 haben wir das für $X = \mathbb{R}^n$ und lineare Trennung durchgeführt. Jetzt wollen wir die x_i zunächst transformieren, um dann im Bildbereich die linearen Methoden anzuwenden[5)].

Genauer: Ist H ein Hilbertraum, so heißt eine Abbildung $\Phi : X \to H$ eine *Feature-Abbildung*. Man sucht dann $w \in H$ und $b \in \mathbb{R}$, so dass

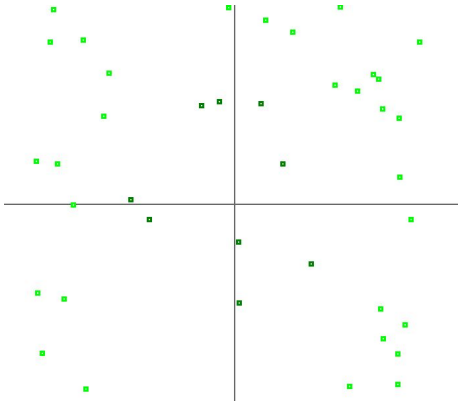
$$y_i\big(\langle \Phi(x_i), w \rangle + b\big) > 0$$

für alle i gilt. (Ziel: Sollen weitere $x \in X$ klassifiziert werden, so mache man das vom Vorzeichen von $\langle \Phi(x), w \rangle + b$ abhängig.) Im Grunde muss man also nur die Funktion $\langle \Phi(\cdot), w \rangle + b$ kennen.

[5)] Vergleiche auch die Beispiele aus der Einleitung.

Sehr bemerkenswert ist nun, dass man zur Bestimmung von w und b gar nicht wissen muss, was die $\Phi(x_i)$ eigentlich sind, denn im Verfahren von Kapitel 1 spielten nur die $\langle x_i, x_j \rangle$, jetzt also die $\langle \Phi(x_i), \Phi(x_j) \rangle$, eine Rolle. Diese Beobachtung werden wir in Kapitel 3 systematisch aufgreifen, wo wir uns mit Hilberträumen mit reproduzierendem Kern beschäftigen werden.

Als erstes Beispiel, wo diese Idee andgewendet werden kann, betrachten wir zwei Klassen des folgenden Typs:



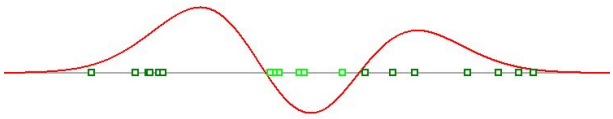
Es scheint so zu sein, dass man sie durch einen Kreis trennen kann. Damit liegt es nahe, als Feature-Abbildung die Abbildung $(x_1, x_2) \mapsto x_1^2 + x_2^2$ zu verwenden. Wirklich ergibt sich folgende Situation:

Da sind die Klassen leicht mit den bekannten Methoden trennbar, und alle anderen Ideen (beste Trennung, Schlupfvariable) lassen sich auch umsetzen. Ganz ähnlich kann man im $\mathbb{R}^3$ bei dafür geeigneten Situationen durch Kugelschalen trennen usw. Allgemein ist immer dann eine Feature-Abbildung in den $\mathbb{R}^1$ sinvoll, wenn die Höhenlinien einer einzigen Funktion $\phi : X \to \mathbb{R}$ zum Trennen ausreichen.

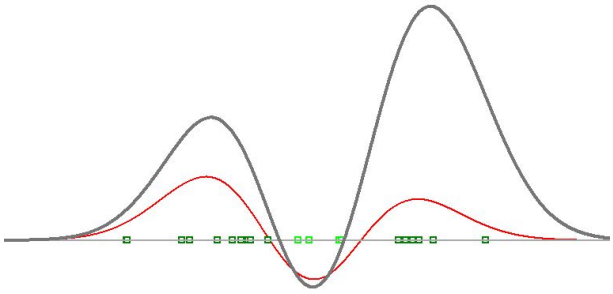
Hier ein weiteres Beispiel. X ist ein Intervall, darin sind zwei Klassen von Punkten gegeben (dunkel- und hellgrün). Sie wurden so erzeugt:

- Als Hilbertraum haben wir Funktionen gewählt, die „schnell abfallen“.
- Eine solche Funktion ϕ (rot im nächsten Bild) wurde durch Zufall ausgewählt, und dann wurden zufällige Punkte erzeugt, bei denen ϕ positiv oder negativ ist.

(Eigentlich kennen wir dieses ϕ aber nicht, es soll nur sichergestellt werden, dass man überhaupt eine Lösung finden kann.)



Zum vorstehenden ϕ wurden noch einmal 20 Punkte erzeugt, und dann wurde der Perceptron-Algorithmus angewendet. Recht schnell war eine Funktion gefunden, die die Punkte trennt (grau):



Wir werden das Thema viel ausführlicher in Kapitel 3 wieder aufgreifen.

Kapitel 2

Konvexe Optimierung

Im vorigen Kapitel tauchte an verschiedenen Stellen das Problem auf, die Funktion $w \mapsto ||w||^2$ auf recht speziellen Definitionsbereichen zu minimieren. Die Situation war stets die folgende:

- Es gibt einen Definitionsbereich $\Delta \subset \mathbb{R}^n$, der die Form $\bigcap_i \{g_i \leq 0\}$ für gewisse affine Funktionen $g_1, \ldots, g_k$ hat[1)].
- Man weiß, dass Δ nicht leer und kompakt ist.
- Die zu minimierende Funktion $f : \Delta \to \mathbb{R}$ ist differenzierbar und konvex oder sogar strikt konvex[2)].

Dann folgt sofort, dass ein Minimum in Δ existiert, das im Fall strikter Konvexität sogar eindeutig bestimmt ist.

> Die Voraussetzung ist für $w \mapsto ||w||^2$ erfüllt, denn wir haben den $\mathbb{R}^n$ mit der euklidischen Norm versehen, und damit gilt die Parallelogrammidentität
>
> $$||x+y||^2 + ||x-y||^2 = 2(||x||^2 + ||y||^2).$$
>
> Es folgt im Fall $x \neq y$
>
> $$\begin{aligned} ||\frac{x+y}{2}||^2 &= \frac{||x+y||^2}{4} \\ &= \frac{||x||^2 + ||y||^2}{2} - \frac{||x-y||^2}{4} \\ &< \frac{||x||^2 + ||y||^2}{2}. \end{aligned}$$

Doch wie kann man den Minimalwert finden? So etwas heißt ein *konvexes Optimierungsproblem*. In diesem Kapitel werden wir eine Lösungsstrategie vorstellen.

[1)] Affine Funktionen sind Funktionen der Form $x \mapsto \langle x, y \rangle + c$.

[2)] D.h. es ist stets $f\big((x+y)/2\big) < \big(f(x) + f(y)\big)/2$ für $x \neq y$.

2.1 Das Karush-Kuhn-Tucker-Theorem

Aus der Analysis weiß man: Ist $U \subset \mathbb{R}^n$ offen und $f : U \to \mathbb{R}$ differenzierbar, so ist für jedes lokale Minimum der Gradient an dieser Stelle gleich Null. Komplizierter war es schon, wenn f gewisse *Nebenbedingungen* $\phi_i(x) = 0$ (i=1,...,k) erfüllen soll. Ist dann x_0 ein lokaler Extremwert von f auf $\bigcap_i \{\phi_i = 0\}$, so ist notwendig der Gradient von f bei x_0 eine Linearkombination der Gradienten der ϕ_i an dieser Stelle. Heuristisch ist das klar, wenn man bedenkt, dass der Gradient die Richtung des stärksten Anstiegs angibt.

Formal geht man dann so vor. Man definiert die *Lagrangefunktion*

$$L(x, \lambda_1, \ldots, \lambda_k) := f + \lambda_1\phi_1 + \cdots + \lambda_k\phi_k,$$

und dann ist zu hoffen, dass aus den $n + k$ Gleichungen

$$\frac{\partial L}{\partial x}(x) = 0,\ \phi_1(x) = \cdots = \phi_k(x) = 0$$

die Unbekannten $x, \lambda_1, \ldots, \lambda_k$ ermittelt werden können. ($\partial L/\partial x$ steht für den Gradienten der Funktion L bei festgehaltenen λ's.)

Zur Illustration folgt ein *Beispiel*: An welchem Punkt des Einheitskreises $\{(x_1, x_2) \mid x_1^2 + x_2^2 = 1\}$ wird die Funktion $f(x_1, x_2) = x_1 + x_2^2$ extremal? Hier ist $k = 1$ und $\phi_1(x_1, x_2) = x_1^2 + x_2^2 - 1$.

Wegen $L(x_1, x_2, \lambda) = x_1 + x_2^2 + \lambda(x_1^2 + x_2^2 - 1)$ werden wir auf die folgenden Gleichungen geführt:

$$0 = \frac{\partial L}{\partial x_1} = 1 + 2\lambda x_1;$$

$$0 = \frac{\partial L}{\partial x_2} = 2x_2 + 2\lambda x_2;$$

$$x_1^2 + x_2^2 - 1 = 0.$$

Das sind 3 Gleichungen für die 3 Unbekannten x_1, x_2, λ. Als Lösungen ergeben sich die Punkte

$$(x_1, x_2) = (1, 0),\ (-1, 0),\ (0.5, \sqrt{3/4}),\ (0.5, -\sqrt{3/4}).$$

Die Zielfunktion f hat dort die Werte 1 bzw. -1 bzw. 5/4 bzw. 5/4. Bei $(-1, 0)$ bzw. bei $(0.5, \pm\sqrt{3/4})$ liegt also ein globales Minimum bzw. Maximum vor, und $(1, 0)$ ist ein lokales Minimum.

(In diesem Fall hätte man es auch einfacher haben können, indem man x_2^2 durch $1-x_1^2$ ersetzt. Dann ist nur noch die eindimensionale Funktion $x_1+(1-x_1^2)$ auf Extremwerte in $[-1, 1]$ zu untersuchen.)

Diese Ergebnisse reichen für unsere Zwecke leider nicht aus, denn wir haben nicht Nebenbedingungen des Typs $=$, sondern $\leq$.

Zur Analyse des Problems machen wir eine *Fallunterscheidung*. Dabei bezeichnen wir das eindeutig bestimmte Minimum von f in Δ mit w^*:

Fall 1: w^* liegt im Innern von Δ. Dann ist das Minimum mit Analysis-Methoden leicht zu finden: Wir müssen nur Punkte testen, bei denen der Gradient von f der Nullvektor ist.

Fall 2: w^* liegt auf dem Rand von Δ. Das bedeutet, dass gewisse $g_i(w^*)$ gleich 0 sind: Sei $I_0 := \{i \mid g_i(w^*) = 0\}$. ($w^*$ liegt also auf einer Kante oder in einer Ecke.) Wäre nun $-\operatorname{grad} f(w^*)$ *nicht* eine Linearkombination der $\operatorname{grad} g_i(w^*)$ $(i \in I_0)$ mit nichtnegativen Komponenten, so könnte man ein $\tilde{w} \in \Delta$ mit $f(\tilde{w}) < f(w^*)$ finden, ein Widerspruch zur Minimalität von $f(w^*)$. (Das Argument wird gleich vervollständigt.)

Um das einzusehen, muss man beachten, dass alle Abbildungen lokal linear sind und dann die folgenden Ergebnisse über lineare Abbildungen auf dem $\mathbb{R}^n$ anwenden:

Lemma 2.1.1. *(i) $V \subset \mathbb{R}^n$ sei ein Unterraum, der einen inneren Punkt e von $(\mathbb{R}^n)^+$ enthält. Weiter sei $g : V \to \mathbb{R}$ linear und positiv (aus $v \geq 0$ folgt $g(v) \geq 0$). Dann gibt es ein $w \in (\mathbb{R}^n)^+$ mit $g = \langle \cdot, w\rangle|_V$. Kurz: Positive Abbildungen können positiv fortgesetzt werden.*

(ii) $G, G_1, \ldots, G_k$ seien lineare Abbildungen auf dem $\mathbb{R}^n$. Es gebe ein v, an dem alle G_i strikt positiv sind. Dann ist G genau dann eine Linearkombination der G_i mit nichtnegativen Komponenten, wenn

$$\{G > 0\} \cap \bigcap_i \{G_i \leq 0\} = \emptyset.$$

Beweis: (i) Es reicht zu zeigen, dass man g positiv auf einen echt größeren Unterraum fortsetzen kann. Wähle $x_0 \notin V$ beliebig. Mit einem noch zu bestimmenden α wollen wir eine Fortsetzung von g durch $\tilde{g} : v + tx_0 \mapsto g(v) + t\alpha$ definieren. $\tilde{g}$ ist dann wohldefiniert und linear, und positiv ist diese Abbildung, wenn sie auf $D_+ := \{v \mid v \in V, x_0 + v \geq 0\}$ und $D_- := \{w \mid w \in V, -x_0 + w \geq 0\}$ positiv ist. Beide Mengen sind nicht leer, da $e \pm \varepsilon x_0 \geq 0$. Für $v \in D_-$ und $w \in D_+$ ist $v + w \geq 0$, also $g(v) + g(w) \geq 0$. Wir wollen $\alpha + g(v), -\alpha + g(w) \geq 0$, d.h. $-g(v) \leq \alpha \leq g(w)$ erreichen, und wegen $-g(v) \leq g(w)$ geht das. Kurz: Jedes α mit

$$\sup_{v \in D_+} -g(v) \leq \alpha \leq \inf_{w \in D_-} g(w)$$

führt zum Ziel.

(ii) Eine Richtung ist offensichtlich. Für die andere sei $\Phi : \mathbb{R}^n \to \mathbb{R}^k$ die Abbildung $x \mapsto (G_1(x), \ldots, G_k(x))$. Der Bildraum werde mit V bezeichnet. Definiere $h : V \to \mathbb{R}$ durch $\Phi(x) \mapsto G(x)$. Da die Voraussetzung insbesondere $\operatorname{kern} G \supset \bigcap \operatorname{kern} G_i$ impliziert, ist h wohldefiniert. Für V und h sind die Voraussetzungen von (i) erfüllt, und das verschafft uns nichtnegative $\alpha_1, \ldots, \alpha_k$ (die Komponenten des Vektors w in (i)) mit $G = \sum \alpha_i G_i$. □

Bemerkungen: 1. Teil (i) des Lemmas gilt auch ohne die Voraussetzung, dass V einen inneren Punkt des positiven Kegels enthält. Dann ist der Beweis allerdings viel schwieriger.

2. Teil (ii) kann auch mit einem Trennungssatz bewiesen werden. Wir hatten das Ergebnis so formuliert, wie wir es für den Beweis des KKT-Theorems brauchen. Es geht auch so (*Farkas-Lemma*): Sind $G, G_1, \ldots, G_k$ lineare Abbildungen auf einem $\mathbb{R}$-Vektorraum und gilt

$$\{G \geq 0\} \subset \bigcap_i \{G_i \geq 0\},$$

so gibt es $a_1, \ldots, a_k \geq 0$ mit $G = a_1 G_1 + \cdots + a_k G_k$. (Die Umkehrung gilt natürlich auch.)

(Beweisskizze für $V = \mathbb{R}^n$: Schreibe $G = \langle \cdot, g \rangle$ und $G_i = \langle \cdot, g_i \rangle$ für $i = 1, \ldots, k$. Setze $P := \{a_1 g_1 + \ldots, a_k g_k \mid a_i \geq 0\}$. Dann ist P konvex und abgeschlossen. Wäre $g \notin P$, gäbe es nach dem Trennungssatz von Hahn-Banach ein Funktional, das g von P trennt, also ein x, für das $\langle \cdot, x \rangle$ auf g strikt negativ und auf P – insbesondere den g_i – positiv ist. Es würde also $G_i(x) \geq 0$ und $G(x) < 0$ gelten. Widerspruch.)

Wir setzen nun die Diskussion von Fall 2 fort, ohne Einschränkung ist $w^* = 0$. Angenommen, $-\operatorname{grad} f(w^*)$ ist nicht eine Linearkombination der $\operatorname{grad} g_i(w^*)$ mit nichtnegativen Komponenten. Wir wenden Teil (ii) des vorigen Lemmas an mit $G := -\langle \cdot, \operatorname{grad} f(w^*) \rangle$ und $G_i = \langle \cdot, \operatorname{grad} g_i(w^*) \rangle$. (Es gibt wirklich ein v, wo alle G_i strikt positiv sind: Wähle ein w in der Nähe von w^* im Innern von Δ und setze $v := -w$.) Man findet also ein x mit $\langle x, \operatorname{grad} g_i(w^*) \rangle \leq 0$ (alle $i \in I$) und $\langle x, -\operatorname{grad} f(w^*) \rangle > 0$. Dann liegt εx für kleine ε in Δ, und $f(\varepsilon x) < f(0) = f(w^*)$. Widerspruch!

Das kann man so zusammenfassen:

Theorem 2.1.2. *(Karush-Kuhn-Tucker) Die g_i, f, Δ und w^* seien wie zu Beginn dieses Abschnitts. Dann gibt es Zahlen $\alpha_1^*, \ldots, \alpha_k^* \geq 0$, so dass gilt:*

- $(\operatorname{grad} f + \sum_i \alpha_i^* \operatorname{grad} g_i)(w^*) = 0$.
- $\alpha_i^* g_i(w^*) = 0$ *für alle i.*

Definiert man also die Lagrangefunktion *L als $L := f + \sum \alpha_i g_i$, so ist w^* unter den Punkten w^* mit $g_i(w^*) \leq 0$ und $\operatorname{grad} L(w^*) = 0$ zu finden.*

Die i mit $\alpha_i^* > 0$ (und folglich $g_i = 0$) heißen die i zu aktiven Randbedingungen.

Beweis: Liegt w^* im Innern von Δ, so sind alle $\alpha_i^* = 0$ und $\operatorname{grad} f(w^*) = 0$. Andernfalls ist $-\operatorname{grad} f$ nichtlineare Linearkombination gewisser $\operatorname{grad} g_i$. (Für die anderen i setzen wir $\alpha_i^* = 0$.) □

In unserem konkreten Fall (konvexe Zielfunktion) ist die vorstehende notwendige Bedingung auch hinreichend. Für w^* im Innern von Δ liegt es daran, dass die Hessematrix positiv semidefinit ist, und für Punkte am Rand muss man ausnutzen, dass $-\operatorname{grad} f$ nichtnegative Linearkombination gewisser $\operatorname{grad} g_i$

ist. Damit ist, von w^* aus gesehen, der Anstieg in jede Richtung in Δ positiv. (Beachte, dass die Gradienten der g_i aus Δ hinaus zeigen.)

Zur Illustration des Theorems betrachte man das folgende „Spielbeispiel“: $f(x) = (x-a)^2$ soll auf $\Delta := [\,-1, 1\,]$ minimiert werden. Wir wählen $g_1(x) = -1-x$ und $g_2(x) = x-1$. Es ist dann $L(x, \alpha_1, \alpha_2) = (x-a)^2 - \alpha_1(1+x) + \alpha_2(x-1)$. Der Gradient ist hier einfach die Ableitung $2(x-a) - \alpha_1 + \alpha_2$. Die vier Fälle α_1, α_2 gleich Null oder größer als Null führen zu $a < 0$, $a \in [\,0, 1\,]$, $a > 1$. ($\alpha_1, \alpha_2 > 0$ ist nicht möglich.)

2.2 Das duale Problem

Die Bedingungen $g_i(w) \leq 0$ sind für Rechnungen unbequem. Durch Übergang zu neuen Variablen kann man dieses Problem beheben.

f, die g_i und Δ seien so wie im vorigen Abschnitt. Insbesondere sind die g_i auf dem ganzen $\mathbb{R}^n$ definiert und affin, und f ist nicht nur auf Δ, sondern ebenfalls auf dem ganzen $\mathbb{R}^n$ definiert und dort differenzierbar und konvex. (Für unsere Probleme ist das ja erfüllt.)

Wieder definieren wir die Lagrangefunktion $L(w, \alpha) := f(w) + \sum_i \alpha_i g_i(w)$; dabei ist $\alpha = (\alpha_1, \ldots, \alpha_k)$. Die Lagrangefunktion ist also eine Funktion von $\mathbb{R}^{n+k}$ nach $\mathbb{R}$.

Für festes $\alpha \geq 0$ betrachten wir $L(\cdot, \alpha) : \mathbb{R}^n \to \mathbb{R}$. Das ist eine konvexe Funktion. Und dann definieren wir

$$\theta(\alpha) := \inf_{w \in \mathbb{R}^n} L(w, \alpha) \in \{-\infty\} \cup \mathbb{R}.$$

Auf Δ gilt (wegen $\alpha \geq 0$ und $g_i \leq 0$) $L(\cdot, \alpha) \leq f$, und deswegen ist $\theta(\alpha) \leq \inf f|_\Delta$; wir setzen voraus, dass das Infimum angenommen wird, also $\inf f|_\Delta = f(w^*)$ (mit $w^* \in \Delta$). Sehr bemerkenswert ist dann das folgende so genannte *starke Dualitätstheorem*:

Theorem 2.2.1. *Unter den vorstehenden Bedingungen und mit den vorstehenden Bezeichnungen gilt*

$$\sup_{\alpha \geq 0} \theta(\alpha) = f(w^*) (= \min f|_\Delta).$$

Beweis: Dass „$\leq$“ gilt, wurde schon gezeigt. Für den Beweis von „$\geq$“ unterscheiden wir zwei Fälle.

Fall 1: w^* liegt im Innern von Δ. Es ist $L(\cdot, \alpha) = f$ für $\alpha = 0$. Im vorliegenden Fall ist also $\theta(0) = f(w^*)$, und das zeigt $\sup_\alpha \theta(\alpha) \geq f(w^*)$.

Fall 2: w^* liegt auf dem Rand von Δ. Wir wählen $\alpha^* \geq 0$ wie im Karush-Kuhn-Tucker-Theorem. Die Funktion $L(\cdot, \alpha^*)$ ist konvex, und ihr Gradient ist bei w^* gleich Null. Das Minimum – das ist $\theta(\alpha^*)$ – ist also $L(w^*, \alpha^*)$. Und diese Zahl

ist gleich $f(w^*)$ aufgrund der Bedingungen $\alpha_i^* g_i(w^*) = 0$. Auch in diesem Fall gilt also $\sup\theta \geq f(w^*)$. □

Bemerkung: Betrachte noch einmal die Funktion $L = L(w, \alpha)$ auf $\mathbb{R}^n \times (\mathbb{R}^k)^+$. Bei festem α ist sie konvex, bei festem w linear. Man mache sich klar, dass (w^*, α^*) ein *Sattelpunkt* für L ist.

Für das vorstehend angegebene „Spielbeispiel“ kann man θ so berechnen:

- Die Ableitung nach x von $L(x, \alpha_1, \alpha_2) = (x-a)^2 - \alpha_1(1+x) + \alpha_2(x-1)$ wird Null bei $(\alpha_1 - \alpha_2)/2 + a$.
- Der Wert dieser Funktion an dieser Stelle ist
$$-\big(\frac{\alpha_2 - \alpha_1}{2}\big)^2 + a(\alpha_2 - \alpha_1) - \alpha_1 - \alpha_2.$$
- Das muss das Minimum sein, denn $L(\cdot, \alpha_1, \alpha_2)$ ist von der Form „quadratisch plus linear“. Die vorstehende Funktion ist also gleich θ.

Wo θ das Maximum im Bereich $\alpha \geq 0$ annimmt, wird von a abhängen. Hier einige Testläufe:

- $a = 0$. In diesem Fall ist das Minimum von f auf Δ gleich Null. Das sollte auch das Maximum von θ sein. Wirklich ist
$$\theta(\alpha_1, \alpha_2) = -\big(\frac{\alpha_2 - \alpha_1}{2}\big)^2 - \alpha_1 - \alpha_2,$$
und das Maximum im Bereich $\alpha \geq 0$ ist offensichtlich gleich Null.
- $a = 2$. Das Minimum von f auf Δ ist 1, das sollte auch das Maximum von θ sein. θ hat die konkrete Form
$$\theta(\alpha_1, \alpha_2) = -\big(\frac{\alpha_2 - \alpha_1}{2}\big)^2 - 3\alpha_1 + \alpha_2.$$
Fixiere ein $\alpha_1 \geq 0$. Für welches α_2 wird der Wert minimal? Die Ableitung nach α_2 wird Null bei $\alpha_2 = \alpha_1 + 2$, und θ hat an diesem α_1, α_2 den Wert $1 - 4\alpha_1$.
Daraus sieht man: Das Maximum wird bei $\alpha_1 = 0$ erreicht, es hat den Wert 1. Und α_2 ist dann gleich 2.

Die Moral: Will man f auf Δ minimieren, so sollte man zunächst θ als Funktion von α bestimmen und dann θ im Bereich $\alpha \geq 0$ maximieren. Dieses Maximum stimmt mit dem Minimum von f auf Δ überein. Der Hauptvorteil dabei ist, dass man statt der Nebenbedingungen $g_i \leq 0$ in der Regel einfachere Nebenbedingungen bzgl. der α_i zu berücksichtigen hat.

Hat man übrigens ein $\alpha^* \geq 0$ gefunden, wo θ maximal wird, so findet man w^* dadurch, dass man diejenige Stelle bestimmt, an der $L(\cdot, \alpha^*)$ minimal wird.

2.3 Konkrete Rechnungen

Wir werden bei den in Abschnitt 1.2 formulierten Problemen zu dualen Variablen $\alpha_1, \dots \alpha_k$ übergehen und das entspechende Maximierungsproblem lösen (siehe den vorigen Abschnitt). Das ist eigentlich eine weitere Baustelle. Bei nicht zu großen Problemen kann man es aber so machen:

- Starte mit irgendeinem erlaubten Satz $\alpha_1, \dots, \alpha_k$.
- Iteriere „sehr oft“ das folgende Verfahren:
  - Suche zwei zufällige Indizes i', j' (mit $i' \neq j'$).
  - Bestimme auf den erlaubten $\alpha_{i'}, \alpha_{j'}$ den maximalen Wert der Zielfunktion. Das ist das neue α.

Das zu den $\alpha' s$ gehörige w kann ja ausgerechnet und der Wert der zu minimierenden Funktion bestimmt werden. Man ist dann fertig, wenn der Unterschied dieser Zahl zum jeweiligen Wert der Zielfunktion für die α (der *duality gap*) klein genug ist; vgl. Theorem 2.2.1.

Hard margin classifier

Da ging es doch um folgendes Problem:

- Gegeben sind linear trennbare $(x_i, y_i) \in \mathbb{R}^n \times \{-1, 1\}$, $(i = 1, \dots, l)$.
- Gesucht sind $w \in \mathbb{R}^n$ und $b \in \mathbb{R}$, so dass $||w||^2/2$ minimal unter den Nebenbedingungen
$$y_i(\langle x_i, w\rangle + b) \geq 1$$
$(i = 1, \dots, l)$ ist.

Satz 2.3.1. *Das duale Problem lautet: Finde das Maximum der Funktion*

$$W(\alpha) := \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j \langle x_i, x_j \rangle$$

unter den Nebenbedingungen

$$\alpha_i \geq 0, \ \sum_i y_i \alpha_i = 0.$$

Beweis: Wir wollen die Ergebnisse des vorigen Abschnitts anwenden und betrachten zunächst die Lagrangefunktion:

$$L(w, b, \alpha) := \frac{1}{2}||w||^2 - \sum_{i=1}^{l} \alpha_i \big(y_i(\langle x_i, w\rangle + b) - 1\big);$$

beachte, dass Δ durch $g_i \leq 0$ definiert ist, die Bedingung $y_i(\langle x_i, w\rangle + b) \geq 1$ wird also in $g_i(w) = -\big(y_i(\langle x_i, w\rangle + b) - 1\big) \leq 0$ übersetzt. (L ist also eine Funktion in $n + 1 + l$ Variablen.)

Fixiere ein $\alpha \geq 0$ und minimiere $L(\cdot, \cdot, \alpha)$ auf $\mathbb{R}^{n+1}$. Dazu betrachten wir den Gradienten dieser Funktion:

- Gradient in Bezug auf w: $\partial L/\partial w = w - \sum y_i \alpha_i x_i$.
- Gradient in Bezug auf b: $\partial L/\partial b = \sum y_i \alpha_i$.

Dabei haben wir ausgenutzt, dass grad $||w||^2/2 = w$ und grad $\langle w, x_i \rangle = x_i$.

Fall 1: $\partial L/\partial b = \sum y_i \alpha_i \neq 0$.
In b-Richtung ist die Funkion also linear mit nichtverschwindender Steigung. Folglich ist das Minimum $-\infty$; solche α brauchen wir nicht zu berücksichtigen.

Fall 2: $\partial L/\partial b = \sum y_i \alpha_i = 0$.
In diesem Fall haben wir eine Chance, das Minimum in $\mathbb{R}$ zu finden. Das bestimmen wir, indem wir die Gleichung $\partial L/\partial w = 0$ lösen: Es folgt $w = \sum y_i \alpha_i x_i$. (Es muss ein Minimum sein, da die Zielfunktion konvex ist.)

Diesen Wert setzen wir in L ein, um den Wert des Minimums zu ermitteln. (Im vorigen Abschnitt hieß das Ergebnis $\theta(\alpha)$, hier soll es $W(\alpha)$ heißen.):

$$\begin{aligned} W(\alpha) &= L(w, b, \alpha) \\ &= \frac{1}{2}||w||^2 - \sum_{i=1}^{l} \alpha_i \big(y_i(\langle x_i, w\rangle + b) - 1\big) \\ &= \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \langle x_i, x_j\rangle - \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle x_i, x_j\rangle + \sum_i \alpha_i \\ &= \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \langle x_i, x_j\rangle. \end{aligned}$$

□

Daraus leiten wir die folgende *Handlungsanweisung* ab:

- Bestimme $\alpha^* = (\alpha_1^*, \ldots, \alpha_l^*) \geq 0$, für das $W(\cdot)$ unter den obigen Nebenbedingungen maximal wird. Im allgemeinen werden nur wenige $\alpha_i^* > 0$ sein. Die zugehörigen x_i heißen *Supportvektoren*.
- Setze $w^* := \sum y_i \alpha_i^* x_i$.
- Das zugehörige b^* finden wir durch folgende Überlegung. Die Hyperebene $\{\langle \cdot, w* \rangle + b*\}$ trennt doch die Punkte zu $y = 1$ optimal von denen zu $y = -1$. Folglich gilt

$$b^* = -\frac{\max_{i,y_i=-1}\langle x_i, w^*\rangle + \min_{i,y_i=1}\langle x_i, w^*\rangle}{2}.$$

Übrigens können während des Lösungsverfahrens für das Maximierungsproblem immer wieder feststellen, wie weit wir schon gekommen sind. Angenommen, irgendein α ist unser Kandidat, wo wir das Maximum vermuten. Setze $w := \sum y_i \alpha_i x_i$. Wenn dann $||w||^2/2$ „sehr nahe“ bei $W(\alpha)$ ist (der Unterschied heißt *duality gap*), so können wir aufhören und $\alpha^* := \alpha$ setzen.

Soft margin classifier, quadratische Wichtung

In Kapitel 1.3 hatten wir motiviert, warum es wichtig sein könnte, das Optimierungsproblem

- Minimiere $||w||^2 + C\sum_{i=1}^{l}\zeta_i^2$ unter den Nebenbedingungen

$$y_i(\langle x_i, w\rangle + b) \geq 1 - \zeta_i,\ \zeta_i \geq 0$$

in den $n + 1 + l$ Veränderlichen $w, b, \zeta_1, \ldots, \zeta_l$ zu lösen: Dadurch können auch eigentlich nicht trennbare Punktmengen behandelt werden.

Angenommen, wir finden eine Lösung für das Problem

- Minimiere $||w||^2 + C\sum_{i=1}^{l}\zeta_i^2$ unter den Nebenbedingungen

$$y_i(\langle x_i, w\rangle + b) \geq 1 - \zeta_i$$

(die Bedingungen $\zeta_i \geq 0$ sind also weggefallen). Wäre dann irgenein ζ_i negativ, so könnte man es durch 0 ersetzen. Denn dann ist die Ungleichung erst recht erfüllt, und $\sum_{i=1}^{l}\zeta_i^2$ ist noch kleiner geworden.
Kurz: Die Bedingungen $\zeta_i \geq 0$ können in der Problemstellung weggelassen werden.

Hier ist zunächst die Lagrangefunktion[3)]:

$$L(w, b, \zeta, \alpha) = \frac{||w||^2}{2} + \frac{C}{2}\sum_{i=1}^{l}\zeta_i^2 - \sum_i \alpha_i\big(y_i(\langle x_i, w\rangle + b) - 1 + \zeta_i\big);$$

Das ist eine Funktion in $n + 1 + 2l$ Veränderlichen.

Fixiere ein $\alpha \geq 0$, wir wollen $W(\alpha) := \inf_{w,b,\zeta} L(w, b, \zeta)$ bestimmen. Dann wissen wir, dass $\max_{\alpha\geq 0} W_\alpha$ Lösung des Minimierungsproblems ist.

Zunächst berechnen wir die partiellen Ableitungen:

Der Gradient in Richtung w ist $w - \sum_{i=1}^{l} y_i\alpha_i x_i$.

Der Gradient in Richtung ζ ist $C\zeta - \alpha$.

Der Gradient in Richtung b ist $\sum y_i\alpha_i$.

Fall 1: $\sum y_i\alpha_i \neq 0$.

Dann wird L in b-Richtung beliebig klein, d.h. $W(\alpha) = -\infty$. Solche α müssen wir nicht berücksichtigen.

Fall 2: $\sum y_i\alpha_i = 0$.

Dann ist die Extremstelle leicht durch Nullsetzen des Gradienten zu ermitteln, und die muss ein Minimum sein, denn L ist konvex:

$$w = \sum_{i=1}^{l} y_i\alpha_i x_i;\ \zeta = \alpha/C.$$

[3)] Dabei wurde die Zielfunktion aus Bequemlichkeitsgründen mit dem Faktor 0.5 versehen.

Um $W(\alpha)$ zu ermitteln, setzen wir diese Werte in L ein. Wir erhalten

$$W(\alpha) = \sum \alpha_i - \frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j \big(\langle x_i, x_j \rangle + \frac{1}{C} \delta_{i,j} \big).$$

($\delta_{i,j}$ bezeichnet dabei das Kroneckersymbol.) Wir fassen zusammen:

Satz 2.3.2. *Die x_i, y_i und $C > 0$ seien gegeben. Um dann w, b, ζ mit möglichst kleinem $||w||^2 + C \sum \zeta^2$ und $y_i(\langle x_i, w \rangle + b) \geq 1 - \zeta_i,\ \zeta_i \geq 0$ (alle i) zu finden, verfahre wie folgt:*

- *Finde $\alpha^* \geq 0$, für das $W(\alpha*)$ maximal ist. Dabei ist W die Funktion*

$$W(\alpha) = \sum \alpha_i - \frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j \big(\langle x_i, x_j \rangle + \frac{1}{C} \delta_{i,j} \big),$$

  das Minimum wird im Bereich $\alpha \geq 0, \sum y_i \alpha_i = 0$ gesucht.

- *Setze $w^* := \sum_i \alpha_i^* y_i x_i$.*

- *Suche ein i' mit $\alpha_{i'}^* > 0$. Ermittle b^* aus der Gleichung*

$$y_{i'} \big(\langle x_{i'}, w^* \rangle + b^* \big) = 1 - \alpha_{i'}^* / C.$$

Es ist dann

$$\{ \sum_i \alpha_i^* y_i \langle \cdot, x_i \rangle + b^* = 1 \}$$

die gesuchte Hyperebene.

Für die Norm von w^* gilt: $||w^*||^2 = \sum_i \alpha_i^* - \langle \alpha^*, \alpha^* \rangle / C$. Der normalisierte Rand ist also

$$\frac{1}{||w^*||} = \frac{1}{\sqrt{\sum_i \alpha_i^* - \langle \alpha^*, \alpha^* \rangle / C}}.$$

Nur für die Supportvektoren (also die x_i mit $\alpha_i^* > 0$) sind die Schlupfvariablen ζ_i^* von Null verschieden: Es gilt $\zeta^* = \alpha^* / C$.

Beweis: Aufgrund des Karush-Kuhn-Tucker-Theorems gilt

$$\alpha_i^* \big(y_i (\langle x_i, w \rangle + b^*) - 1 + \zeta_i^* \big) = 0$$

für alle i. Aus $\alpha_{i'}^* > 0$ folgt also $y_{i'}(\langle x_{i'}, w \rangle + b^*) = 1 - \zeta_{i'}^*$, und $\zeta_{i'}^*$ darf durch $\alpha_{i'}^* / C$ ersetzt werden.

Es fehlt noch die Rechnung

$$\begin{aligned}\langle w^*, w^*\rangle &= \langle \sum_i \alpha_i^* y_i x_i, \sum_j \alpha_j^* y_j x_j\rangle \\ &= \sum_j \alpha_j^* y_j \sum_i \alpha_i^* y_i \langle x_i, x_j\rangle \\ &= \sum_j \alpha_j^* (1 - \zeta_j^* - y_j b^*) \\ &= \sum_j \alpha_j^* - \sum_j \alpha_j^* \zeta_j^* \\ &= \sum_j \alpha_j^* - \frac{1}{C}\langle \alpha^*, \alpha^*\rangle.\end{aligned}$$

□

Soft margin classifier, lineare Wichtung

Und wie können wir den Fall behandeln, wenn wir die Schlupfvariablen ζ_i durch $C\sum\zeta_i$ wichten? Es soll doch $||w||^2/2 + C\sum\zeta_i$ unter den Nebenbedingungen

$$y_i(\langle x_i, w\rangle + b) \geq 1 - \zeta_i,\ \zeta_i \geq 0$$

minimiert werden. Anders als im vorstehenden Fall können die Bedingungen $\zeta_i \geq 0$ nicht ignoriert werden. Zusätzlich zu den α_i kommen also weitere Variable $r_i \geq 0$ dazu. Die Lagrangefunktion (mit $n+1+3l$ Variablen) hat damit die Form

$$L(w, b, \zeta, \alpha, r) = \frac{||w||^2}{2} + C\sum\zeta_i - \sum\alpha_i\big(y_i(\langle w, x_i\rangle + b) - 1 + \zeta_i\big) - \sum r_i\zeta_i.$$

Hier sind die Komponenten des Gradienten, wenn wir α und r festlassen:

Der Gradient in Richtung w ist $w - \sum_{i=1}^{l} y_i\alpha_i x_i$.

Der Gradient in Richtung ζ_i ist $C - \alpha_i - r_i$.

Der Gradient in Richtung b ist $\sum\alpha_i y_i$.

Fall 1: $\sum y_i\alpha_i \neq 0$ oder $C - \alpha_i - r_i \neq 0$.

Dann wird L in b-Richtung oder in ζ_i-Richtung beliebig klein, d.h. $W(\alpha) = -\infty$. Solche α müssen wir nicht berücksichtigen.

Fall 2: $\sum y_i\alpha_i = 0$ und $C - \alpha_i - r_i = 0$.

In diesem Fall setzen wir auch die restlichen Gradientengleichungen Null und setzen das Ergebnis in L ein, um $W(\alpha, r) := \inf_{w,b,\zeta} L(w, b, \zeta, \alpha, r)$ zu ermitteln. Da fällt viel weg, wir erhalten

$$W(\alpha, r) = \sum\alpha_i - \frac{1}{2}\sum_{i,j} y_i y_j \alpha_i \alpha_j \langle x_i, x_j\rangle.$$

Bemerkenswerter Weise ist das die gleiche Funktion wie beim ersten unserer Probleme. Der Definitionsbereich ist allerdings anders: Das muss unter den Nebenbedingungen

$$\alpha_i, r_i \geq 0,\ \alpha_i + r_i = C,\ \sum \alpha_i y_i = 0$$

maximiert werden. Dabei kann man die r_i elimineren, indem man die Bedingungen $\alpha_i, r_i \geq 0$, $\alpha_i + r_i = C$ durch $0 \leq \alpha_i \leq C$ ersetzt (*box constraints*). Wir fassen zusammen:

Satz 2.3.3. *Die x_i, y_i und $C > 0$ seien gegeben. Um dann w, b, ζ mit möglichst kleinem $||w||^2 + C\sum \zeta$ und $y_i(\langle x_i, w\rangle + b) \geq 1 - \zeta_i,\ \zeta_i \geq 0$ (alle i) zu finden, verfahre wie folgt:*

- *Finde $\alpha^* \geq 0$, für das $W(\alpha^*)$ maximal ist. Dabei ist W die Funktion*

$$W(\alpha) = \sum \alpha_i - \frac{1}{2}\sum_{i,j} y_i y_j \alpha_i \alpha_j \langle x_i, x_j\rangle,$$

  das Minimum wird im Bereich $0 \leq \alpha \leq C,\ \sum y_i\alpha_i = 0$ gesucht.

- *Setze $w^* := \sum_i \alpha_i^* y_i x_i$.*

- *Suche ein i' mit $0 < \alpha_{i'}^* < C$. Ermittle b^* aus der Gleichung*

$$y_{i'}\big(\langle x_{i'}, w^*\rangle + b^*\big) = 1.$$

Es ist dann

$$\{\sum_i \alpha_i^* y_i \langle \cdot, x_i\rangle + b^* = 1\}$$

die gesuchte Hyperebene.

Für die Norm von w^* gilt: $||w^*||^2 = \sum_{i,j} \alpha_i^*\alpha_j^* y_i y_j \langle x_i, x_j\rangle$. Der normalisierte Rand ist also

$$\frac{1}{||w^*||} = \frac{1}{\sqrt{\sum_{i,j} \alpha_i^*\alpha_j^* y_i y_j \langle x_i, x_j\rangle}}.$$

Supportvektoren sind jetzt die x_i, für die $\alpha_i^* > 0$ oder $r_i^* > 0$ (also $\alpha_i^* \in \,]\,0, C\,[$) gilt.

Beweis: Nur $||w^*||^2 = \sum_{i,j} \alpha_i^*\alpha_j^* y_i y_j \langle x_i, x_j\rangle$ ist noch nachzutragen, doch das ist wegen $||w^*||^2 = \langle w^*, w^*\rangle$ klar. □

Regression

Wir hatten das Regressionsproblem so eingeführt: Gesucht ist eine affine Funktion $f = \langle \cdot, w\rangle + b$, so dass

$$\sum_i |y_i - f(x_i)|^2$$

minimal wird. Zwei Varianten spielen eine Rolle:

- Statt $|y_i - f(x_i)|^2$ kann man beliebige Verlustfunktionen zulassen.
- Später soll das Ganze in beliebigen Hilberträumen nachgemacht werden. Wenn die viele Funktionen enthalten, wird oft eine „sehr gute“ Approximation mit „sehr komplizierten“ Funktionen möglich sein. (Man spricht von *overfitting*.) Da macht man sich zunutze, dass „komplizierte“ Funktionen oft auch eine „große“ Norm haben. Und deswegen versucht man, einen Kompromiss zu finden. Man wählt ein $C > 0$, und dann soll die Summe $||w||^2 + C \sum_i |y_i - f(x_i)|^2$ minimiert werden.

Es folgt eine Kurzfassung des Ansatzes von Christiani-Taylor (Abschnitt 6.2). Da führt man zwei Familien von Schlupfvariablen ein. Die einen messen Abweichungen nach oben, die anderen nach unten. Genauer geht es, bei gegebenen x_i, y_i, um das folgende Extremalproblem:

- Minimiere $|||w||^2 + C \sum_i (\zeta_i^2 + \hat{\zeta}_i^2)$.
- Nebenbedingungen sind dabei:

  $\langle x_i, w\rangle + b \le y_i + \zeta_i$ für alle i.

  $\langle x_i, w\rangle + b \ge y_i - \hat{\zeta}_i$ für alle i.

  $\zeta_i, \hat{\zeta}_i \ge 0$ für alle i.

  (Das sind $4l$ Nebenbedingungen.)

In diesem Fall hängt die Lagrangefunktion von $n{+}1{+}4l$ Variablen $w, b, \alpha_i, \hat{\alpha}_i, r_i, \hat{r}_i$ ab (n für w, eine für b, $4l$ für die Nebenbedingungen.) Man berechnet die θ-Funktion auf die übliche Weise und zieht daraus Folgerungen. Zum Beispiel soll die partielle Ableitung nach b verschwinden, man erhält $\sum \alpha_i - \hat{\alpha}_i = 0$. Und fasst man die partiellen Ableitungen nach den Komponenten von w zusammen, so ergibt sich $w = \sum(\alpha_i - \hat{\alpha}_i)x_i$. Das Ergebnis setzt man in L ein (falls nicht θ gleich $-\infty$ ist), berücksichtigt, dass stets $\zeta_i\hat{\zeta}_i = 0$ gilt, tauft die oft auftretende Differenz $\hat{\alpha}_i - \alpha_i$ in β_i um und langt so schließlich bei dem folgenden Maximierungsproblem an:

- Maximiere
$$\sum_i y_i\beta_i - 0.5 \sum \beta_i\beta_j(\langle x_i, x_j\rangle + \delta_{ij}/C)$$
unter der Nebenbedingung $\sum_i \beta_i = 0$.

Kapitel 3

Hilberträume mit reproduzierendem Kern

Wenn man eine „lineare“ Methode entwickelt hat, in der nur Auswertungen von Skalarprodukten in einem Hilbertraum H vorkommen, so kann man die manchmal auch für gewisse Mengen X anwenden, indem man X durch eine Abbildung Φ nach H abbildet. So ein $\Phi : X \to H$ heißt im Englischen *feature map*, wir werden von einer *Feature-Abbildung* sprechen.

Jedes derartige Φ induziert eine Abbildung $k : X \times X \to \mathbb{R}$ (oder nach $\mathbb{C}$ für $\mathbb{C}$-Hilberträume) durch $k : (x, x') \mapsto \langle \Phi(x), \Phi(x) \rangle$. (Achtung: Es ist kein Tippfehler, rechts sind x, x' vertauscht. Für reelle Räume ist das natürlich belanglos.) Im vorliegenden Kapitel wollen wir den Zusammenhang zwischen X, k, Φ und H klären.

3.1 Hilberträume

Es ist sicher sinnvoll, die wichtigsten Fakten zu Hilberträumen noch einmal zusammenzustellen.

Es sei $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ und H ein linearer $\mathbb{K}$-Vektorraum. Eine Anbbildung $\langle \cdot, \cdot \rangle : H \times H \to \mathbb{K}$ heißt *inneres Produkt* (oder *Skalarprodukt*), wenn gilt:

- Stets ist $\langle x, x \rangle$ reell und nichtnegativ, und aus $\langle x, x \rangle = 0$ folgt $x = 0$.
- $\langle ax + by, z \rangle = a\langle x, z \rangle + b\langle y, z \rangle$.
- $\langle z, ax + by \rangle = \overline{a}\langle z, x \rangle + \overline{b}\langle z, y \rangle$.
- $\langle x, y \rangle = \overline{\langle y, x \rangle}$.

Dann wird durch $||x|| := \sqrt{\langle x, x \rangle}$ wirklich eine Norm definiert, im Beweis wird wesentlich von der *Cauchy-Schwarzschen Ungleichung* $|\langle x, y \rangle| \leq ||x||\,||y||$ Gebrauch gemacht.

Wenn jede Cauchy-Folge in $(H, ||\cdot||)$ konvergent ist, wenn H bezüglich $||\cdot||$ also vollständig ist, spricht man von einem *Hilbertraum*.

Standardbeispiele sind:

- Der $\mathbb{K}^n$ mit dem euklidischen Skalarprodukt

$$\langle x, y\rangle := \sum_{i=1}^{n} x_i \overline{y_i}.$$

- Der Folgenraum $l^2 := \{(x_n) \mid \sum_n |x_n|^2 < \infty\}$ mit dem Skalarprodukt

$$\langle x, y\rangle := \sum_{i=1}^{\infty} x_i \overline{y_i}.$$

- Der Raum $L^2(\mathbb{R}) := \{f \mid \int_{\mathbb{R}} |f|^2 < \infty\}$ der quadratintegrablen Funktionen (modulo Funktionen, die fast überall Null sind) mit dem Skalarprodukt

$$\langle f, g\rangle := \int_{\mathbb{R}} f\overline{g}.$$

Hilberträume sind deswegen so wichtig, weil die Eigenschaften in vielen Fällen so sind wie im Endlichdimensionalen. Aus der Cauchy-Schwarz-Ungleichung folgt zum Beispiel, dass man im Fall $\mathbb{K} = \mathbb{R}$ den Winkel α zwischen zwei von Null verschiedenen Vektoren x, y durch $\cos\alpha := \langle x, y\rangle/(||x||\,||y||)$ definieren kann. Konsequenterweise heißen dann x, y (für beliebiges $\mathbb{K}$) *orthogonal*, wenn $\langle x, y\rangle$=0 gilt. Schreibweise: $x \perp y$.

Wir werden besonders die folgenden Ergebnisse benötigen:

Satz 3.1.1. *Es sei H ein Hilbertraum.*

(i) Für jedes y ist $x \mapsto \langle x, y\rangle$ eine stetige lineare Abbildung von H nach $\mathbb{K}$.

(ii) Das gilt auch umgekehrt: Zu jeder stetigen linearen Abbildung $\phi : H \to \mathbb{K}$ gibt es ein eindeutig bestimmtes y mit $\phi = \langle \cdot, y\rangle$.

(iii) $K \subset H$ sei nicht leer, abgeschlossen und konvex. Dann gibt es zu jedem $x_0 \in H$ ein eindeutig bestimmtes $x \in K$ mit minimalem Abstand zu x_0:

$$||x - x_0|| = \min_{y \in K} ||y - x_0||.$$

(Zu diesem Ergebnis gibt es Kommentare am Ende des Abschnitts.)

(iv) Ist insbesondere $K = N$ ein abgeschlossener Unterraum, so versteht man unter $N^\perp$ die Menge der $x \in H$, für die die beste Approximation gleich Null ist. Ein x liegt genau dann in $N^\perp$, wenn x zu allen $y \in N$ orthogonal ist. Allgemeiner ist $y' \in N$ die beste Approximation an ein $x \in H$, wenn $x - y' \perp y$ für alle $y \in N$ gilt.

(v) H ist die direkte Summe aus N und $N^\perp$: Zu jedem $y \in H$ gibt es eindeutig bestimmte $x \in N$, $x' \in N^\perp$ mit $y = x + x'$. Es ist dann $||y||^2 = ||x||^2 + ||x'||^2$.

(vi) $(N^\perp)^\perp = N$ für alle abgeschlossenen Unterräume.

(vii) Sei H ein endlich-dimensionaler Hilbertraum. Dann gibt es Vektoren $e_1, \ldots, \ldots, e_n$ mit $||e_i|| = 1$ und $\langle e_i, e_j \rangle = 0$ (für $i \neq j$), eine so genannte Orthonormalbasis*, so dass jedes $x \in H$ mit eindeutig bestimmten $a_1, \ldots, a_n \in \mathbb{K}$ als*

$$x = \sum_{i=1}^{n} a_i e_i$$

geschrieben werden kann. Es ist dann $||x||^2 = \sum_{i=1}^{n} |a_i|^2$.

(viii) Sei H ein unendlich-dimensionaler Hilbertraum. Er soll aber „nicht zu groß" sein: Man soll eine dichte abzählbare Teilmenge auswählen können.

Dann gibt es Vektoren $e_1, e_2, \ldots$ mit $||e_i|| = 1$ und $\langle e_i, e_j \rangle = 0$ (für $i \neq j$), eine abzählbare Orthonormalbasis*, so dass jedes $x \in H$ mit eindeutig bestimmten $a_1, a_2, \ldots \in \mathbb{K}$ als*

$$x = \sum_{i=1}^{\infty} a_i e_i$$

geschrieben werden kann. Diese Reihen sind jeweils unbedingt konvergent*, und es gilt $||x||^2 = \sum_{i=1}^{\infty} |a_i|^2$.*

Etwas komplizierter wird es, wenn wir uns nicht mehr auf „nicht zu große" Hilberträume beschränken und alle zulassen. Dann gilt immer noch: Es gibt eine (evtl. überabzählbare) Indexmenge I und eine Familie $(e_i)_{i \in I}$ mit folgenden Eigenschaften:

- $||e_i|| = 1$ für alle i und $\langle e_i, e_j \rangle = 0$ für $i \neq j$.
- Jedes $x \in H$ kann mit eindeutig bestimmten $a_i,\ i \in I$ aus $\mathbb{K}$ als

$$x = \sum_{i \in I} a_i e_i$$

  geschrieben werden kann. Diese Reihen sind jeweils *unbedingt konvergent*, und es gilt $||x||^2 = \sum_{i \in I} |a_i|^2$.

Dabei ist zunächst nicht offensichtlich, was die hier auftretenden Summen $\sum_{i \in I} y_i$ für $y_i \in \mathbb{K}$ oder allgemeiner $y_i \in Y$ (ein Banachraum) bedeuten sollen. Es folgen einige Informationen zu diesem Problem.

1. Ist I leer, so sei $\sum_{i \in I} y_i := 0$.

2. Ist I endlich und nicht leer, so schreibe I als $\{i_1, \ldots, i_n\}$. Wir setzen dann

$$\sum_{i \in I} y_i := \sum_{k=1}^{n} y_{i_k}.$$

Es ist noch zu bemerken, dass das wohldefiniert ist: Wählt man eine andere Aufzählung von i, so kommt das Gleiche heraus. Das liegt daran, dass die Addition in jedem Banachraum kommutativ und assoziativ ist.
Es gelten dann die üblichen Rechenregeln, etwa $\sum_{i\in I} \alpha b_i = \alpha \sum_{i\in I} b_i$.
Sind alle $b_i \geq 0$, so kann man schnell sehen, dass

$$\sum_{i\in I} b_i = \sup_{J\subset I} \sum_{i\in J} b_j.$$

3. Ist I beliebig und sind alle $b_i \geq 0$, so kann man die vorstehende Beobachtung zur Definition erheben:

$$\sum_{i\in I} b_i := \sup_{J\subset I,\ J \text{ endlich}} \sum_{j\in J} b_j.$$

Das liefert ein wohldefiniertes Ergebnis in $[0,\infty]$, und im Fall einer endlichen Reihensumme gelten wieder die üblichen Rechenregeln: $\sum_{i\in I} \alpha b_i = \alpha \sum_{i\in I} b_i$ für $\alpha \geq 0$, $\sum_{i\in I}(b_i + c_i) = \sum_{i\in I} b_i + \sum_{i\in I} c_i$ usw.

Hier gibt es übrigens ein erstes nichttriviales Resultat: Ist $\sum_{i\in I} b_i < \infty$, so ist $\{i \mid b_i > 0\}$ höchstens abzählbar. Das liegt daran, dass dann alle Mengen $\{i \mid b_i > 1/k\}$ endlich sein müssen und man $\{i \mid b_i > 0\}$ als $\bigcup_k \{i \mid b_i > 1/k\}$ schreiben kann.

Kurz: $\sum_{i\in I} b_i$ kann als „gewöhnliche“ Reihe aufgefasst werden. (Wobei allerdings zu beachten ist, dass sie unbedingt konvergent ist.)

Die Tatsache, dass Reihen mit positiven Gliedern unbedingt konvergent sind, spielt übrigens auch in der Wahrscheinlicheitstheorie eine Rolle. Wenn man ein Ereignis E disjunkt in $E_1, E_2, \ldots$ zerlegt hat, so soll doch $\mathbb{P}(E) = \sum_i \mathbb{P}(E_i)$ gelten. Und das wäre nicht wohldefiniert, wenn die Reihensumme von der Reihenfolge der E_i abhinge.

4. Es bleibt noch der für uns interessante Fall $\sum_{i\in I} b_i$ mit $b_i \in \mathbb{K}$ (oder allgemeiner: $\sum_{i\in I} y_i$, wobei die y_i Elemente eines Banachraums Y sind) zu behandeln. Man geht so vor: Für ein $y \in Y$ soll $\sum_{i\in I} y_i = y$ bedeuten, dass

- $I_0 := \{i \mid y_i \neq 0\}$ ist höchstens abzählbar.
- Schreibt man I_0 auf irgendeine Weise als $I_0 = \{i_1, i_2, \ldots\}$, so ist $\sum_{k=1}^{\infty} y_{i_k} = y$, d.h. $\lim_{k\to\infty} y_{i_1} + y_{i_2} + \cdots + y_{i_k} = y$.

Kurz: $\sum_{i\in I} y_i$ steht für eine unbedingt konvergente Reihe in Y. Wissenswert ist dann noch:

- Aus $\sum_i ||y_i|| < \infty$ folgt, dass $\sum_i y_i$ existiert: absolute Konvergenz impliziert Konvergenz.
- Im Endlichdimensionalen gilt auch die Umkehrung, aber in jedem unendlichdimensionalen Raum gibt es Gegenbeispiele dafür (Satz von Dvoretzky-Rogers).

Der vorstehende Struktursatz für beliebige Hilberträume kann auch so interpretiert werden, dass man ein H, das eine Orthonormalbasis $(e_i)_{i\in I}$ besitzt, mit dem $l^2(I)$ identifizieren kann. Das soll der Raum aller Tupel $(x_i)_{i\in I}$ (mit $x_i \in \mathbb{K}$) sein, für die $\sum_{i\in I} |x_i|^2$ endlich ist.

Wir schließen diesen Abschnitt mit einer elementaren Beobachtung, die für unsere Untersuchungen eine wichtige Rolle spielt:

Lemma 3.1.2. *Sei H ein Hilbertraum, und $x_1, x_2, \ldots \in H$. Gesucht ist ein $w \in H$ mit minimaler Norm, für das gewisse Bedingungen bezüglich der x_i erfüllt sind, wobei in diesen Bedingungen nur die Zahlen $\langle x_i, w\rangle$ auftreten. Wenn es dann so ein w gibt, so liegt es im Abschluss der linearen Hülle der x_i.*

Beweis: Sei U der Abschluss der linearen Hülle der x_i, und für $w_1 \in H$ werde das Problem gelöst (Bedingungen erfüllt, minimale Norm). Schreibe $w_1 = w_0 + w'$ mit $w_0 \in U$ und $w' \in U^\perp$. Dann erfüllt auch $w_0 = w_1 - w'$ alle Bedingungen, da $\langle \cdot, w'\rangle = 0$ auf U. Auch gilt $||w_1||^2 = ||w_0||^2 + ||w'||^2$, und da $||w_1||$ minimal war, muss $w' = 0$ sein. □

Oft geht es übrigens nur um endlich viele x_i, und dann ist die lineare Hülle der x_i schon abgeschlossen.

Es folgen noch einige Bemerkungen zu Teil (iii) von Satz 3.1.1. Dadurch soll deutlich werden, dess es sich um eine besondere Eigenschaft von Hilberträumen handelt und dass die Konvexität wesentlich für das Ergebnis ist.

1. Im allgemeinen ist der Satz in beliebigen Banachräumen falsch.

Sei f ein Funktional auf einem Raum X, das seine Norm nicht annimmt. Sei etwa $||f|| = 1 = \sup_{||y||\leq 1} |f(y)|$, aber es soll kein y mit $||y|| \leq 1$ geben, für das $|f(y)] = 1$ gilt. Solche Funktionale existieren auf allen nichtreflexiven Banachräumen (Satz von James). Als konkretes Beispiel denke man an $(x_n) \mapsto \sum(1 - 1/n)x_n$ auf l^1.

Sei nun $x \in X$ mit $f(x) = 1$ gegeben. Wir behaupten, dass es an $K := \ker f$ keine beste Approximation gibt. Angenommen, das wäre doch der Fall. Ohne Einschränkung soll 0 die beste Approximation sein, d.h. $||x||$ ist der Abstand zu K. Wir behaupten, dass $||x|| = 1$ gilt, f würde also doch die Norm annehmen. Widerspruch!

Sei $y \in K$. Dann ist $f(x-y) = 1 \leq ||x-y||$. Also ist $d(x, K) \geq 1$. Es gilt auch $d(x, K) \leq 1$. Sei dazu $\varepsilon > 0$ und y mit $||y|| = 1$ so gewählt, dass $f(y) \geq 1 - \varepsilon$. Für $\lambda = 1/f(y)$ liegt $x - \lambda y$ in K, also

$$d(x, K) \leq ||x - (x - \lambda y)|| = |\lambda| = \frac{1}{1-\varepsilon}.$$

Damit gilt $d(x, K) \leq 1$, insgesamt also $d(x, K) = 1$.

2. Ist K in 3.1.1(iii) nicht konvex, so ist Eindeutigkeit nicht garantiert.
Man muss nur an einen Kreisring im $\mathbb{R}^2$ und an $x_0 = 0$ denken.

3. Ist K in 3.1.1(iii) nicht konvex, so ist die Existenz nicht garantiert.
Sei $x_0 = 0$ im l^2 und $K = \{(1 + 1/n)e_n \mid n \in \mathbb{N}\}$, wobei e_n für den n-ten Einheitsvektor steht. Dann ist K abgeschlossen, $d(x_0, K) = 1$, aber der Abstand wird in K nicht realisiert.

4. Schon im $\mathbb{R}^2$ mit der Maximumsnorm gibt es Gegenbeispiele zur Eindeutigkeit. Ist K die Einheitskugel, so hat $(2, 0)$ unendlich viele beste Approximationen.

3.2 Kerne

Es sei X eine Menge, H ein $\mathbb{K}$-Hilbertraum (mit $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$) und $\Phi : X \to H$. Wie schon erwähnt, werden wir Φ eine *Feature-Abbildung* nennen[1)].
Die Abbildung

$$k : X \times X \to \mathbb{K}, \ (x, x') \mapsto \langle \Phi(x'), \Phi(x) \rangle$$

heißt dann ein *Kern* auf X.

Es ist offensichtlich, dass es auf X im Allgemeinen eine unübersehbare Fülle von Kernen geben wird. Wir werden sehen, dass es im Wesentlichen reicht, sich auf Hilberträume zu beschränken, deren Elemente Funktionen auf X sind.

Beispiele: 0. Sei $\alpha \geq 0$. Für jedes X ist die konstante Abbildung $(x, x') \mapsto \alpha$ ein Kern.

1. X sei Teilmenge eines Hilbertraums, etwa $X \subset \mathbb{K}^n$. Wählt man Φ als identische Einbettung, so ist $k = \overline{\langle \cdot, \cdot \rangle}$, also $k(x, x') = \overline{\langle x, x' \rangle} = \langle x', x \rangle$.

2. Gegeben seien beliebige Funktionen $f_i : X \to \mathbb{K}$, $i = 1, \ldots, n$. Definiert man $\Phi : X \to \mathbb{K}^n$ durch $x \mapsto \big(f_i(x)\big)_{i=1,\ldots,n}$, so ist der zugehörige Kern durch $k(x, x') = \sum_{i=1}^n f_i(x')\overline{f_i(x)}$ gegeben.

Wenn man im Fall unendlich vieler f_n verlangt, dass $\big(f_n(x)\big)_{n=1,\ldots}$ stets zum l^2 gehört, so kann man $\mathbb{K}^n$ durch den Hlbertraum l^2 ersetzen.

Als konkrete Anwendung betrachten wir *Cosinuskerne*. Es seien $a_0, a_1, \ldots$ nichtegativ, und $\sum_n a_n < \infty$. Setze $f(t) := \sum_{n=0}^\infty a_n \cos(nt)$ für $t \in [-\pi, \pi]$. Wir behaupten, dass $k(x, x') := f(x - x')$ ein Kern ist.
Wirklich ist

$$k(x, x') = a_0 + \sum_{n=1}^\infty a_n \sin(nx) \sin(nx') + \sum_{n=1}^\infty a_n \cos(nx) \cos(nx').$$

Wenn wir also $f_n(x) = \sqrt{a_n} \sin(nx)$ bzw. $f_n(x) = \sqrt{a_n} \cos(nx)$ setzen, so führen beide Summen zu Kernen. Und wir werden gleich zeigen, dass Summen von Kernen wieder Kerne sind.

[1)] Das Wort „Feature“ hat im Englischen viele Bedeutungen. Die Internetseite `leo.org` bietet 212 Übersetzungsmöglichkeiten an. So beginnt die Liste: Eigenschaft, Besonderheit, Gesichtszug, Merkmal, Feuilleton, äußere Erscheinung, Charakteristikum, Charakterzug, Wesensmerkmal, ...

3. Sei H ein beliebiger Hilbertraum und $(e_i)_{i\in I}$ eine Orthonormalbasis. Wähle Abbildungen $f_i : X \to \mathbb{K}$, so dass für jedes x die Reihensumme $\sum_{i\in I} f_i(x)e_i$ existiert. (Es sind also höchstens abzählbar viele $f_i(x) \neq 0$, und die Reihe ist unbedingt konvergent.) Definiere $\Phi : X \to H$ durch $x \mapsto \sum_{i\in I} f_i(x)e_i$. Der zugehörige Kern ist dann $k(x, x') = \sum_{i\in I} f_i(x')\overline{f_i(x)}$.

Jeder Kern entsteht auf diese Weise. Sei nämlich $\Phi : X \to H$ eine beliebige Feature-Abbildung. Wir wählen irgendeine Orthonormalbasis $(e_i)_{i\in I}$ und definieren $f_i : X \to \mathbb{K}$ so: Für jedes i_0 und jedes x soll $f_{i_0}(x)$ der Koeffizient bei e_{i_0} in der Darstellung von $\Phi(x)$ als Reihe $\sum a_i e_i$ sein.

4. Sei $k : X \times X \to \mathbb{K}$ ein Kern auf Y. Für $Y \subset X$ ist dann die Einschränkung von k auf $Y \times Y$ ein Kern auf Y. (Klar: Man betrachte den gleichen Raum H und die Einschränkung von Φ auf Y).

Allgemeiner: Ist Y beliebig und $\tau : Y \to X$ eine Abbildung, so ist $(y, y') \mapsto k\big(\tau(y), \tau(y')\big)$ ein Kern auf Y.

Lemma 3.2.1. *Mit k, k_1, k_2 sind auch $k_1 + k_2$ und ak für jedes $a \geq 0$ Kerne auf X.*

Beweis: k_i werde durch Φ_i auf H_i erzeugt $(i = 1, 2)$. $H := H_1 \times H_2$ sei der Produkt-Hilbertraum. ($\langle(x_1, x_2), (y_1, y_2)\rangle := \langle x_1, y_1\rangle + \langle x_2, y_2\rangle$.) Definiert man noch $\Phi : x \mapsto (\Phi_1(x), \Phi_2(x))$, so ist das zu Φ gehörige k gleich $k_1 + k_2$.

Um ak als Kern zu erkennen, wähle man das gleiche H und die Feature-Abbildung $\Phi'(x) = \sqrt{a}\Phi(x)$.

($a < 0$ ist im Allgemeinen nicht zulässig, denn für einen Kern gilt sicher stets $k(x, x) \geq 0$.) □

Etwas aufwändiger ist es einzusehen, dass auch Produkte von Kernen wieder Kerne sind. Wir beginnen mit einer vergleichsweise einfachen Situation. Dazu sei $X = \mathbb{K}^n$ und k der schon oben eingeführte Kern $k(x, x') := \langle x', x\rangle$. Wir behaupten, dass $\tilde{k}(x, x') := \langle x', x\rangle^2$ auch ein Kern ist. Dazu betrachten wir die Abbildung $\tilde{\Phi}$ von X in den Hilbertraum $\mathbb{K}^{n^2}$, die durch $x \mapsto (x_i x_j)_{i,j=1,\ldots,n}$ definiert ist. Dann gilt

$$\begin{aligned}
\tilde{k}(x, x') &= \langle x', x\rangle^2 \\
&= (x'_1\overline{x_1} + \cdots + x'_n\overline{x_n})(x'_1\overline{x_1} + \cdots + x'_n\overline{x_n}) \\
&= \sum_{i,j} x'_i\overline{x_i}x'_j\overline{x_j} \\
&= \langle (x'_i x'_j)_{i,j}, (\overline{x_i x_j})_{i,j}\rangle \\
&= \langle \tilde{\Phi}(x'), \tilde{\Phi}(x)\rangle.
\end{aligned}$$

Ganz analog lässt sich zeigen, dass alle $\langle x', x\rangle^k$ für $k \in \mathbb{N}$ Kerne auf $\mathbb{K}^n$ sind. Es gilt aber viel allgemeiner:

Lemma 3.2.2. *Mit $k, \tilde{k}$ ist auch $k \cdot \tilde{k}$ ein Kern auf X.*

Beweis: Nach Voraussetzung gibt es Hilberträume $H, \tilde{H}$ und Abbildungen $\Phi, \tilde{\Phi}$, so dass stets $k(x,x') = \langle \Phi(x'), \Phi(x)\rangle$ sowie $\tilde{k}(x,x') = \langle \tilde{\Phi}(x'), \tilde{\Phi}(x)\rangle$ gilt[2)]. Wir wählen geeignete Mengen $I, \tilde{I}$ und geeignete Familien $(f_i)_{i\in I}$ und $(\tilde{f}_{\tilde{i}})_{\tilde{i}\in\tilde{I}}$ wie vorstehend in Beispiel 3.

Dann betrachten wir den Hilbertraum $l^2(I\times J)$. Es ist dann leicht (aber etwas langwierig) einzusehen, dass die Abbildung $x \mapsto \big(f_i(x)\tilde{f}_{\tilde{i}}(x)\big)_{(i,\tilde{i})\in I\times\tilde{I}}$ eine Feature-Abbildung von X nach $l^2(I\times J)$ ist, die zum Kern $k\cdot\tilde{k}$ gehört. □

Kann man einer Abbildung $k : X \times X \to \mathbb{K}$ ansehen, ob sie ein Kern ist? Kerne erfüllen sicher zwei Bedingungen

- Stets ist $k(x,x') = \overline{k(x',x)}$, ein Kern ist also *symmetrisch*. Das folgt aus der entsprechenden Eigenschaft für Skalarprodukte.
- Kerne sind *positiv semidefinit*, d.h.

  $$\sum_{i,j=1,\dots,n} a_i\overline{a_j}k(x_j,x_i) \geq 0$$

  für beliebige $x_1,\dots,x_n \in X$ und $a_1,\dots,a_n \in \mathbb{K}$.

  (Setze $v := \sum_{i=1,\dots,n} a_i\Phi(x_i)$. Dann ist

  $$\begin{aligned} 0 &\leq \langle v, v\rangle \\ &= \langle \sum_{i=1,\dots,n} a_i\Phi(x_i), \sum_{j=1,\dots,n} a_j\Phi(x_j)\rangle \\ &= \sum_{i,j=1,\dots,n} a_i\overline{a_j}k(x_j,x_i). \end{aligned}$$

  Das beweist die Behauptung.)

Bemerkenswerter Weise sind Kerne durch diese Bedingungen schon charakterisiert:

Satz 3.2.3. *Es sei $k : X \times X \to \mathbb{K}$ symmetrisch und positiv semidefinit. Dann ist k ein Kern.*

Beweis: Wir müssen einen Hilbertraum „aus dem Nichts“ erschaffen. Die Grundidee besteht darin, die Featureabbildung als $x \mapsto k(\cdot,x)$ zu definieren. Allerdings fehlt noch der Hilbertraum, der diese Funktionen enthält ...

Als linearen Raum betrachten wir den $\mathbb{K}$-Vektorraum V aller Abbildungen von X nach $\mathbb{K}$ mit der üblichen punktweisen Linearstruktur. Die $k(\cdot,x)$ sind spezielle Elemente von V, mit H_0 bezeichnen wir die lineare Hülle:

$$H_0 := \{ \sum_{i=1\dots,n} a_i k(\cdot,x_i) \mid n\in\mathbb{N},\ x_i \in X,\ a_i \in \mathbb{K}\}.$$

[2)] Wir bezeichnen die Skalarprodukte in $H, \tilde{H}$ mit dem gleichen Symbol.

H_0 ist sicher ein Vektorraum.

Nun zur Definition des inneren Produkts. Später soll doch stets $k(x,x') = \langle k(\cdot,x'), k(\cdot,x)\rangle$ erfüllt sein, damit $x \mapsto k(\cdot,x)$ Feature-Abbildung zu k ist. Da das innere Produkt auch bilinear sein soll, haben wir praktisch keine andere Wahl, als $\langle\cdot,\cdot\rangle$ so zu definieren:

$$\langle \sum_{i=1\ldots,n} a_i k(\cdot,x_i), \sum_{j=1\ldots,m} b_j k(\cdot,x'_j)\rangle := \sum_{i=1,\ldots,n,j=1,\ldots,m} a_i\overline{b_j}k(x'_j,x_i).$$

Dann gilt:

$\langle\cdot,\cdot\rangle$ *ist wohldefiniert*[3)]

Beweis dazu: Wir müssen zeigen: haben $f,g \in H_0$ zwei Darstellungen, also

$$f = \sum_{i=1\ldots,n} a_i k(\cdot,x_i) = f = \sum_{i=1}^{\tilde{n}} \tilde{a}_i k(\cdot,\tilde{x}_i),$$

$$g = \sum_{j=1\ldots,m} b_j k(\cdot,x'_j) = \sum_{j=1}^{\tilde{m}} \tilde{b}_j k(\cdot,\tilde{x}'_i),$$

so ist

$$\sum_{i=1,\ldots,n,j=1,\ldots,m} a_i\overline{b_j}k(x'_j,x_i) = \sum_{i=1,\ldots,\tilde{n},j=1,\ldots,\tilde{m}} \tilde{a}_i\overline{\tilde{b}_j}k(\tilde{x}'_j,\tilde{x}_i).$$

Wirklich ist

$$\begin{aligned}
\sum_{i=1,\ldots,n,j=1,\ldots,m} a_i\overline{b_j}k(x'_j,x_i) &= \sum_j \overline{b_j} f(x'_j)\\
&= \sum_j \overline{b_j}\sum_{i=1}^{\tilde{n}} \tilde{a}_i k(x'_j,\tilde{x}_i)\\
&= \sum_j \overline{b_j}\sum_{i=1}^{\tilde{n}} \tilde{a}_i \overline{k(\tilde{x}_i,x'_j)}\\
&= \sum_i \tilde{a}_i \overline{g(\tilde{x}_i)}\\
&= \sum_i \tilde{a}_i \overline{\sum_{j=1}^{\tilde{m}} \tilde{b}_j k(\tilde{x}_i,\tilde{x}'_i)}\\
&= \sum_{i,j} \tilde{a}_i\overline{\tilde{b}_j}\overline{k(\tilde{x}_i,\tilde{x}'_i)}\\
&= \sum_{i,j} \tilde{a}_i\overline{\tilde{b}_j}k(\tilde{x}'_i,\tilde{x}_i).
\end{aligned}$$

Für alle $f,g \in H_0$ *gilt* $\langle f,g\rangle = \overline{\langle g,f\rangle}$

[3)] Die Definition hängt also nicht von der zufälligen Darstellung der Elemente von H_0 ab.

Das folgt aus $k(x, x') = \overline{k(x', x)}$.

$\langle \cdot, \cdot \rangle$ *ist linear in der ersten und konjugiert linear in der zweiten Komponente.*
Das ist aufgrund der Definiton klar.

$\langle f, f \rangle \geq 0$ *für alle* f.
Das ist eine Umformulierung der positiven Definitheit.

Aus $\langle f, f \rangle = 0$ *folgt* $f = 0$.
Bisher wissen wir nur, dass $\langle \cdot, \cdot \rangle$ ein so genanntes Semi-Skalarprodukt ist. Das reicht aber schon, um die Cauchy-Schwarzsche Ungleichung zu beweisen: $|\langle f, g \rangle|^2 \leq \langle f, f \rangle \langle g, g \rangle$.

Nun sei $\langle f, f \rangle = 0$. Wendet man die Ungleichung speziell für $g = k(\cdot, x)$ an, so folgt

$$|f(x)|^2 = |\langle f, g \rangle| \leq \langle f, f \rangle \langle g, g \rangle = 0.$$

f ist also wirklich die Nullfunktion.

H_0 ist also mit einem Skalarprodukt versehen worden. In Bezug auf die induzierte Norm $||f|| = \sqrt{\langle f, f \rangle}$ muss H_0 allerdings nicht vollständig sein, es könnte nur ein *Prähilbertraum* (= Raum mit innerem Produkt) sein. Doch dafür gibt es bewährte Methoden, wir gehen zur *Vervollständigung* H über, dabei werden wir H_0 als Unterraum von H aufassen

Die Feature-Abbildung definieren wir wie geplant als $\Phi : x \mapsto k(\cdot, x) \in H$. Dann gilt stets

$$\langle \Phi(x'), \Phi(x) \rangle = \langle k(\cdot, x'), k(\cdot, x) \rangle = k(x, x').$$

□

Zusatz: Es ist doch $\langle f, k(\cdot, x) \rangle = f(x)$ für $f \in H_0$. Aus der Cauchy-Schwarzschen Ungleichung folgt daraus, dass $f \mapsto f(x)$ gleichmäßig stetig ist. Wenn dann (f_n) eine Cauchy-Folge in H_0 ist, ist auch $(f_n(x))$ eine Cauchy-Folge (in $\mathbb{R}$) und damit konvergent. Anders ausgedrückt: Man kann die $f \in H$ mit Elementen aus $\mathrm{Abb}(X, \mathbb{K})$ identifizieren.

Es folgt übrigens noch einmal, dass Summen und positive Vielfache von Kernen wieder Kerne sind[4)].

Wenn man die vorstehenden Ergebnisse kombiniert, kann man schon eine Vielzahl von Kernen konstruieren. Wichtig wird noch das folgende Korollar werden:

Korollar 3.2.4. *$k_1, k_2, \ldots$ seien Kerne auf X, so dass für alle x, x' die Reihe $k(x, x') := \sum_{i=1}^{\infty} k_n(x, x')$ konvergent ist. Dann ist auch k ein Kern. Ebenso: Wenn alle Folgen $(k_n(x, x'))_n$ konvergent sind, so ist die durch $k(x, x') := \lim_n k_n(x, x')$ definierte Funktion ein Kern.*

[4)] Für Produkte lässt sich das nicht so ohne Weiteres übertragen.

Beweis: Das folgt unmittelbar aus dem vorigen Charakterisierungssatz. □

Hier noch einige wichtige Beispiele:
1. $X = \mathbb{K}^n$, und $c \geq 0$. Dann ist $k(x, x') := (c + \langle x', x\rangle)^l$ für jedes $l \in \mathbb{N}$ ein Kern.
2. Es sei $X = \mathbb{K}^n$ und $\sum_m a_m z^m$ eine Potenzreihe mit unendlichem Konvergenzradius, so dass $a_m \geq 0$ für alle m gilt. Dann ist $k(x, x') := \sum_m a_m \langle x', x\rangle^m$ ein Kern auf $\mathbb{K}^n$. (Das verallgemeinert übrigens das vorstehende Beispiel.) Insbesondere ist $\exp(c\langle x', x\rangle)$ ein Kern für jedes $c \geq 0$.
3. Ein wichtiges Beispiel sind die *Gaußkerne zum Parameter* $\gamma > 0$ auf dem $\mathbb{R}^n$:

$$k_\gamma(x, x') := \exp\left(-\frac{||x - x'||^2}{\gamma^2}\right).$$

Das ist wirklich ein Kern. Die Symmetrie ist klar. Zum Nachweis der Definitheit schreibe $k_\gamma(x, x')$ als

$$\frac{\exp(2\langle x', x\rangle/\gamma^2)}{\big(\exp(||x||^2)/\gamma^2\big)\big(\exp(||x'||^2/\gamma^2)\big)}.$$

Sind dann $a_1, \ldots, a_r \in \mathbb{R}$ und $x_1, \ldots, x_r \in \mathbb{R}^n$ beliebig, so ist

$$\sum_{i,j} a_i a_j k_\gamma(x_j, x_i) = \sum_{i,j} \tilde{a}_i \tilde{a}_j \tilde{k}(x_j, x_i);$$

dabei ist $\tilde{k}(x, x') := \exp(2\langle x', x\rangle/\gamma^2)$ und $\tilde{a}_i := a_i \exp(||x_i||^2)/\gamma^2$.
Die Summe ist damit nichtnegativ wegen Beispiel 2.

Man mache sich qualitativ klar, wie k_γ von γ abhängt. Angenommen, γ ist klein. Dann wird $k_\gamma(x, x')$ schon für nahe beieinander liegende x, x' klein sein, die $\Phi(x), \Phi(x')$ werden also „beinahe orthogonal“ sein.

Und nun sei γ „groß“. Wenn x, x' nahe beieinander liegen, sind dann $\Phi(x), \Phi(x')$ „fast parallel“.

Zum Abschluss dieses Abschnitts wollen wir Kerne auf endlichen Mengen explizit beschreiben. Es sei $X = \{1, \ldots, n\}$; wie sieht der allgemeinste Kern k auf X aus? Schreibe k als Matrix K. Dann ist K symmetrisch und positiv semidefinit, kann also nach Koordinatentransformation als Diagonalmatrix mit nichtnegativen Einträgen λ_i geschrieben werden. Macht man die Transformation wieder rückgängig, so heißt das: Es gibt orthonormale $\psi_1, \ldots, \psi_n \in \mathbb{K}^n$, so dass stets

$$k(i, j) = \sum_{l=1}^{n} \lambda_l \psi_l(i) \overline{\psi_l(j)}$$

gilt. Das ist ein Spezialfall eines Ergebnisses von Mercer, durch das ein Kern auf einem tolologischen Raum durch die Eigenfunktionen eines durch k definierten Integraloperators dargestellt werden kann.

3.3 Hilberträume mit reproduzierendem Kern

Nach Definition gehört zu jedem Kern auf X eine Feature-Abbildung Φ von X in einen Hilbertraum H: Es soll stets $k(x,x') = \langle \Phi(x'), \Phi(x)\rangle$ gelten. Stets kann man beliebig viele H und Φ finden, die das gleiche k liefern. Zum Beispiel könnte man H durch H^r (mit dem Produkt-Skalarprodukt) und Φ durch $x \mapsto (1/\sqrt{r})\big(\Phi(x),\dots,\Phi(x)\big)$ ersetzen, wobei $r \in \mathbb{N}$ beliebig ist.

Falls man nur an k interessiert ist, wäre es wünschenswert, unter den vielen möglichen H und Φ kanonische Kandidaten auszuzeichnen. Einen möglichen Ansatz liefert Satz 3.2.4, da wurde zunächst ein Prähilbertraum aus Funktionen auf X konstruiert. Diese Idee soll nun etwas ausführlicher entwickelt werden.

Definition 3.3.1. *Es sei $X \neq \emptyset$ eine Menge. Wir bezeichnen mit $Abb(X,\mathbb{K})$ den $\mathbb{K}$-Vektorraum aller Abbildungen von X nach $\mathbb{K}$. Ein Unterraum $H \subset Abb(X,\mathbb{K})$ sei bezüglich eines Skalarprodukts $\langle\cdot,\cdot\rangle$ ein Hilbertraum.*

Ist k ein Kern auf X, so heißt k ein reproduzierender Kern *zu H, wenn die folgenden Bedingungen erfüllt sind:*

(i) k ist der zur Feature-Abbildung $x \mapsto k(\cdot,x)$ gehörige Kern.

(ii) H ist „nicht größer als erforderlich“: die lineare Hülle der $k(\cdot,x)$ soll dicht in H liegen.

(iii) Weitergehend als (i) soll gelten: Für alle $f \in H$ und alle x ist $\langle f, k(\cdot,x)\rangle = f(x)$. ((i) besagt, dass das für die Funktionen $k(\cdot,x')$ gilt. Daraus folgt, dass die Gleichung auch auf der linearen Hülle gilt, aber es ist nicht klar, ob sie für Funktionen im Abschluss auch richtig ist.)

Hat ein Hilbertraum H von Funktionen auf X einen reproduzierenden Kern, so soll er Hilbertraum mit reproduzierendem Kern (reproducing kernel Hilbert space, RKHS) *genannt werden*

Man kann einem Hilbertraum $H \subset \mathrm{Abb}(X,\mathbb{K})$ schnell ansehen, ob er zu einem reproduzierenden Kern gehört:

Satz 3.3.2. *Für einen Hilbertraum $H \subset Abb(X,\mathbb{K})$ sind die folgenden Aussagen äquivalent:*

(i) Es gibt einen Kern k, so dass k ein reproduzierender Kern zu H ist.

(ii) Die Abbildungen $f \mapsto f(x)$ (von H nach $\mathbb{K}$) sind für alle x stetig.

Beweis: Es gelte zunächst (i). Fixiere x und setze $g := k(\cdot,x)$. Aus der Cauchy-Schwarzschen Ungleichung folgt

$$|f(x)| = |\langle f,g\rangle| \le ||f||\,||g||.$$

Da $f \mapsto f(x)$ linear ist, folgt die Stetigkeit.

Nun setzen wir (ii) voraus, und wir geben ein x vor. $f \mapsto f(x)$ ist linear und stetig und folglich wegen Satz 3.1.1(ii) von der Form $f \mapsto \langle f, g_x\rangle$ für ein geeignetes $g_x \in H$. Definiere $k(\cdot,x) := g_x$. Auf diese Weise wird eine Abbildung $k : X \times X \to \mathbb{K}$ erzeugt.

Klar ist, dass alle $k(\cdot, x)$ zu H gehören, und setzt man speziell $f = k(\cdot, x')$, so folgt wirklich $k(x, x') = \langle k(\cdot, x'), k(\cdot, x)\rangle$: Es ist also $x \mapsto k(\cdot, x)$ wirklich eine Feature-Abbildung zu k.

Es fehlt nur noch der Nachweis der zweiten Eigenschaft aus Definition 3.3.1. Bezeichne mit G den Abschluss der linearen Hülle der $k(\cdot, x)$ in H. Wir wollen $H = G$ zeigen. Dazu reicht es zu beweisen, dass der Orthogonalraum trivial ist. Sei also $f \in H$ senkrecht zu allen $k(\cdot, x)$. Das heißt doch nach Definition von k, dass $f(x) = 0$ für alle x gilt. Es ist also $f = 0$, und damit ist alles gezeigt[5)]. □

Korollar 3.3.3. *H sei ein Hilbertraum von Funktionen auf X. Sind k_1, k_2 reproduzierende Kerne zu H, so gilt $k_1 = k_2$.*

Beweis: Im vorstehenden Satz haben wir k aus der Stetigkeitsbedingung konstruiert. □

Es ist zu betonen, dass die Stetigkeitsbedingung in (ii) für Hilberträume von Funktionen untypisch ist. Betrachten wir etwa den Hilbertraum der quadratintegrablen Funktionen auf $\mathbb{R}$, so würde die Bedingung besagen dass kleine $\int f^2\, d\lambda$ auf kleine $f(x)$ schließen lassen. Das stimmt aber nicht! Das bedeutet, dass wir Hilberträume quadratintegrablen Funktionen höchstens im diskreten Fall antreffen werden[6)].

Zu jedem Kern können viele geeignete Hilberträume gehören. Die Beziehung zwischen Kernen und Hilberträumen mit reproduzierendem Kern ist aber eineindeutig:

Satz 3.3.4. *Es sei k ein Kern auf X, und $\Phi : X \to \hat{H}$ sei eine Feature-Abildung. O.B.d.A. sei $\hat{H}$ der Abschluss der linearen Hülle der $\Phi(x), x \in X$. (Denn das ist auch ein Hilbertraum, und außerhalb liegende Punkte spielen für Φ keine Rolle.) Dann gibt es einen eindeutig bestimmten RKHS H zum Kern k, so dass $\hat{H}$ und H als Hilberträume isomorph sind.*

Es reicht also, sich auf RKHS zu beschränken.

Beweis: Definiere $H := \{\langle w, \Phi(\cdot)\rangle, w \in \hat{H}\}$. Das ist sicher ein Unterraum von $\mathrm{Abb}(X, \mathbb{K})$, und die Abbildung $\Psi : w \mapsto \langle w, \Phi(\cdot)\rangle$ von $\hat{H}$ nach H ist nach Definition surjektiv und offensichtlich linear. Sie ist auch injektiv: Ist $\langle w, \Phi(\cdot)\rangle$ die Nullabbildung, so steht w senkrecht auf allen $\Phi(x)$, also auch senkrecht auf der abgeschlossenen linearen Hülle dieser Vektoren (das ist $\hat{H}$). Das impliziert $w = 0$.

Ψ ist also bijektiv. Wir machen H durch die Definition $\langle \Psi(w), \Psi(w')\rangle := \langle w, w'\rangle$ zu einem zu $\hat{H}$ isomorphen Hilbertraum. (Achtung: In der Definition steht links das zu definierende Skalarprodukt, rechts das Skalarprodukt in $\hat{H}$).

[5)] Dieses Argument zeigt übrigens, dass die zweite Bedingung in Definition 3.3.1 aus der dritten folgt. Es hätte also gereicht, nur Bedingung (iii) zu fordern.

[6)] Allerdings geht es im allgemeinen Fall auch nicht wirklich um Funktionen, sondern um Klassen von Funktionen.

Wir müssen noch zeigen, dass H ein RKHS zum Kern k ist. Für $x \in X$ ist $k(x', x) = \langle \Phi(x), \Phi(x') \rangle$, d.h. $k(\cdot, x) = \Psi\big(\Phi(x)\big) \in H$. Ist $f := \Psi(w)$ beliebig in H, so soll $\langle f, k(\cdot, x) \rangle = f(x)$ für alle x gelten. Das stimmt, denn

$$\langle \Psi(w), \Psi\big(\Phi(x)\big) \rangle = \langle w, \Phi(x) \rangle = \big(\Psi(w)\big)(x) = f(x).$$

Die Eindeutigkeit von H ist klar, da H der Abschluss der linearen Hülle der $k(\cdot, x)$ ist. □

Es ist nicht leicht, Beispiele von Hilbert-Funktionenräumen zu finden, die kein RKHS sind. Der Grund: Auf vollständigen Räumen sind lineare Abbildungen fast immer stetig; als Faustregel: Alles, was man konkret hinschreiben kann, ist auch stetig. Das liegt am Closed-graph-Theorem.

Trotzdem gibt es Gegenbeispiele. Für deren Konstruktion ist es allerdings notwendig, das Zornsche Lemma einzusetzen. Hier ist eine Konstruktion von Halmos aus einer Arbeit von Alpey und Mills[7] ("A family of Hilbert spaces which are not reproducing kernel Hilbert spaces." J. Anal. Appl. 1, No. 2, 107-111 (2003)).

- Wähle irgendein X, auf dem es einen unendlichdimensionalen RKHS H gibt. (Wir haben einige Beispiele kennen gelernt.)
- Wähle ein lineares unstetiges Funktional $\phi : H \to \mathbb{K}$. Das geht, auf jedem unendlichdimensionalen normierten Raum gibt es so etwas. (*Hier* geht das Zornsche Lemma ein. Man braucht es, um eine Basis zu finden. Manchmal spricht man von einer *Hamelbasis*; der Name erinnert an den Berliner Profesosr Georg Hamel, 1877 – 1954.)
- Definiere $\hat{X} := X \cup \{x_0\}$, wobei x_0 irgendein Punkt ist, der nicht zu X gehört.
- Für $f \in H$ sei $\hat{f} : \hat{X} \to \mathbb{K}$ die Funktion, die auf X gleich f und bei x_0 gleich $\phi(x_0)$ ist. $\hat{H}$ sei die Menge aller $\hat{f}$. Das ist offensichtlich ein Funktionenraum.
- Auf $\hat{H}$ erklären wir ein Skalarprodukt durch $\langle \hat{f}, \hat{g} \rangle := \langle f, g \rangle$. Das macht $\hat{f}$ zu einem Hilbertraum.
- Die Abbildung $\hat{f} \mapsto \hat{f}(x_0)$ ist nicht stetig.

Folglich ist $\hat{H}$ kein RKHS.

(Viel einfacher lässt sich ein Beispiel eines Prähilbertraumes von Funktionen auf X finden, bei dem nicht alle Auswertungen stetig sind:

- $X := \mathbb{N}_0$, und

  $H := \{(x_i)_{i=0,1,\dots} \mid \text{ höchstens endlich viele } x_i \neq 0,\ x_0 = x_1 + x_2 + \cdots\}$.

[7] Dank an Dirk Werner für diesen Tipp.

- $\langle (x_i), (y_i) \rangle := \sum_{i=1}^{\infty} x_i y_i / n$.

Das ist ein Skalarprodukt, und $(x_i) \mapsto x_0$ ist unstetig.)

3.4 RKHS: Beispiele

Alle Kerne aus Abschnitt 3.2 liefern Beispiele. Das ist recht theoretisch, doch wie kann man RKHS visualisieren?

Eindimensionale Beispiele

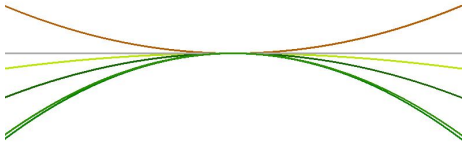
Wir nehmen zunächst an, dass X eindimensional (etwa $X = [0,1]$ oder $X = \mathbb{R}$) und $\mathbb{K} = \mathbb{R}$ ist. Ist k ein zum RKHS H gehöriger Kern, so ist doch H der Abschluss der linearen Hülle der $k(\cdot, x)$ mit $x \in X$. „Typische“ Elemente in H sind also Funktionen der Form $\sum_{i=1}^{n} a_i k(\cdot, x_i)$.

Beispiel 0: $k(x', x) := c > 0$. Das ist nicht wirklich interessant, H besteht aus allen konstanten Funktionen.

Beispiel 1: $k(x', x) := xx'$. Diesmal besteht H aus den Funktionen $x \mapsto xa$, mit $a \in \mathbb{R}$. Das sind die Einschränkungen auf X aller linearen Abbildungen auf $\mathbb{R}$.

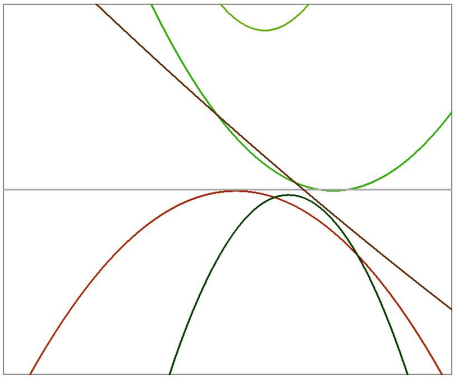
Beispiel 1': $k(x', x) := xx' + c$ (wobei c eine positive Konstante ist). Diesmal ergeben sich die Einschränkungen auf X der $x \mapsto ax + b$.

Beispiel 2: $k(x', x) := (xx')^2$. Auch das ist noch ein recht kleiner Hilbertraum: die $x \mapsto \alpha x^2$. Hier sieht man einige Beispiele (sie sind in verschiedenen Farben skizziert):



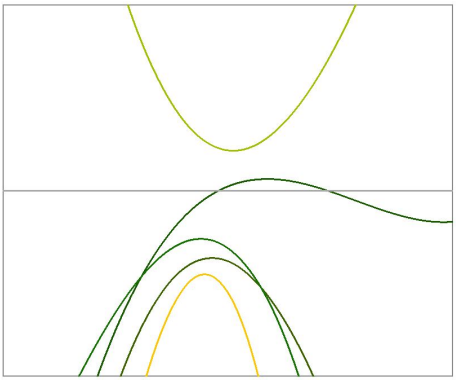
Funktionen im RKHS zu $k(x', x) := (xx')^2$

Beispiel 3: $k(x', x) := (1 + xx')^2$. Diesmal ergeben sich alle quadratischen Funktionen (d.h., die Parabeln):



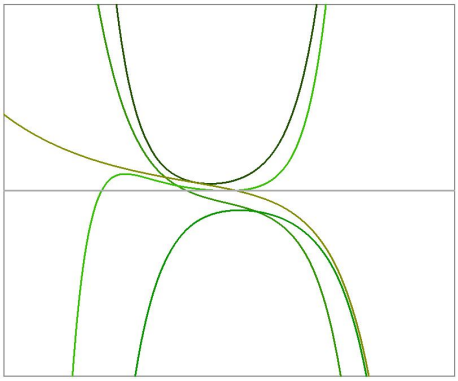
Funktionen im RKHS zu $k(x', x) := (1 + xx')^2$

Beispiel 4: $k(x', x) := (1 + xx')^3$. Und jetzt erhalten wir alle Funktionen höchstens dritten Grades:



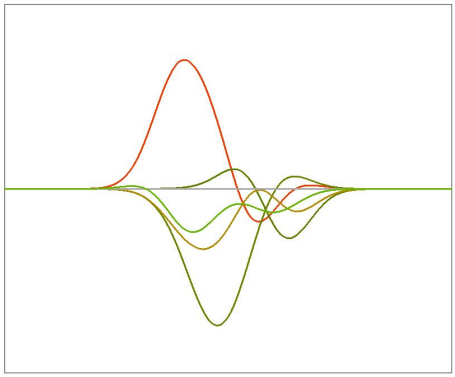
Funktionen im RKHS zu $k(x', x) := (1 + xx')^3$

Beispiel 5: $k(x', x) := \exp(xx')$, ein Exponentialkern. Erwartungsgemäß erhalten wir Funktionen, die wie die Exponentialfunktionen sowie Sinus und Cosinus hyperbolicus aussehen:

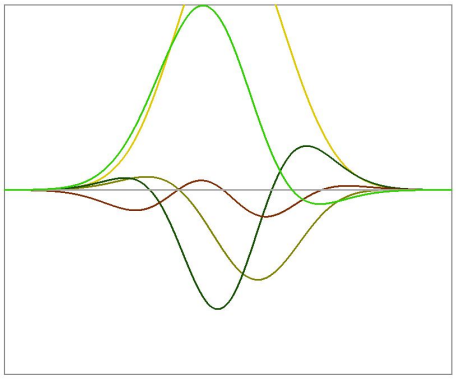


Funktionen im RKHS zu $k(x', x) := \exp(xx')$

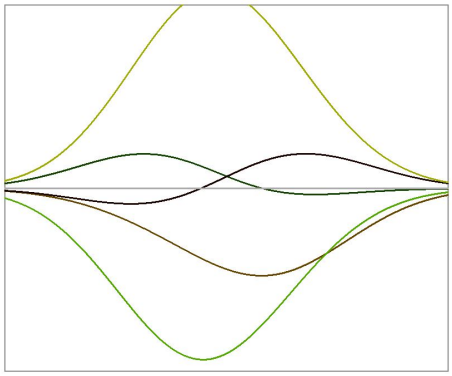
Beispiel 6: Und nun betrachten wir Gaußkerne $k_\gamma(x', x) := \exp(-||x - x'||/\gamma^2)$ für ein kleines, ein mittleres und ein großes γ. Typische Funktionen sehen dann so aus:



Gaußkern, γ klein

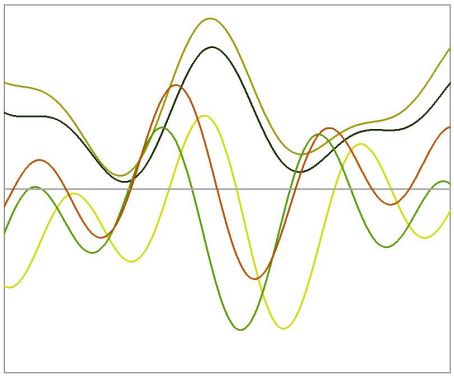


Gaußkern, γ mittel



Gaußkern, γ groß

Beispiel 7: $k(x', x) := \cos(x - x') + \cos\big(3(x - x')\big)$, ein Cosinuskern.

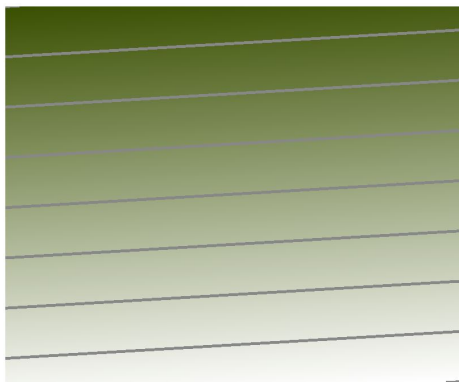


Funktionen im RKHS zu $k(x', x) := \cos(x - x') + \cos\big(3(x - x')\big)$

Etwas schwieriger ist es schon, sich *RKHS in höheren Dimensionen* vorzustellen. Wenigstens für Funktionen, die auf einer Teilmenge des $\mathbb{R}^2$ definiert sind, wollen wir es versuchen. Wir visualisieren wie folgt:

> Sei $\Delta \subset \mathbb{R}^2$, $\phi : \Delta \to \mathbb{R}$ und F eine Farbe. Mit A bzw. B bezeichnen wir das Minimum bzw. Maximum von ϕ. Ein Punkt $x \in \Delta$ wird dann so eingefärbt, dass die Farbe – in Abhängigkeit von $\phi(x)$ – linear zwischen F (bei A) und weiß (bei B) interpoliert. Zusätzlich sind einige Höhenlinien $\{\phi = c\}$ eingezeichnet.
>
> Anders ausgedrückt: Die „hohen“ Werte sind weiß, die „tieferen“ dunkler eingefärbt[8)].

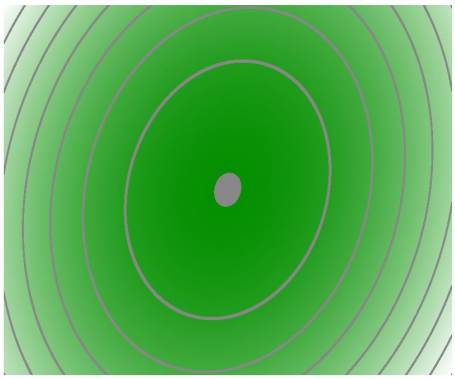
Beispiel 8: $k(x', x) := \langle x, x' \rangle$ auf $[-1, 1]^2$. Der RKHS besteht aus linearen Abbildungen.



Funktionen im RKHS zu $k(x', x) := \langle x, x' \rangle$ auf $[-1, 1]^2$

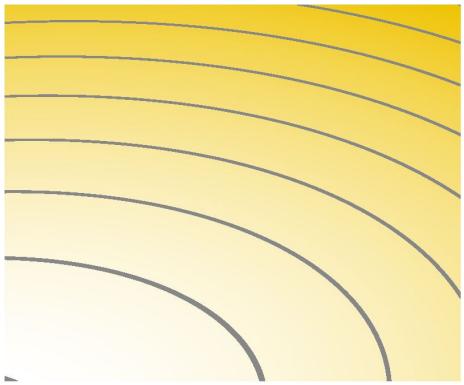
[8)] So ist zum Beispiel das Titelbild entstanden.

Beispiel 9: $k(x', x) := \langle x, x' \rangle^2$ auf $[\,-1, 1\,]^2$. Der RKHS besteht aus quadratischen Abbildungen, die bei 0 verschwinden.



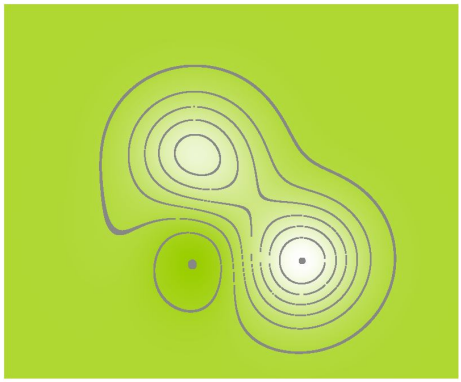
Funktionen im RKHS zu $k(x', x) := \langle x, x' \rangle^2$ auf $[\,-1, 1\,]^2$

Beispiel 10: $k(x', x) := (1 + \langle x, x' \rangle)^2$ auf $[\,-1, 1\,]^2$. Der RKHS besteht aus quadratischen Abbildungen.



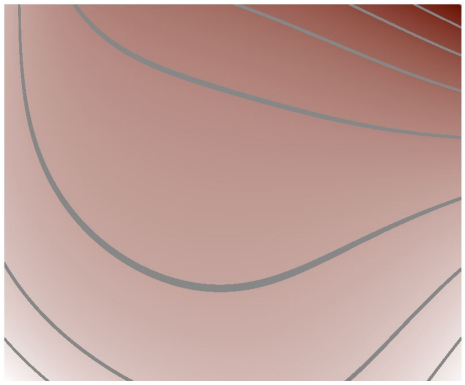
Funktionen im RKHS zu $k(x', x) := (1 + \langle x, x' \rangle)^2$ auf $[\,-1, 1\,]^2$

Beispiel 11: $k(x', x) := \exp(-||x - x'||^2/0.1)$ auf $[\,-1, 1\,]^2$. Der RKHS besteht aus schnell abfallenden Abbildungen.



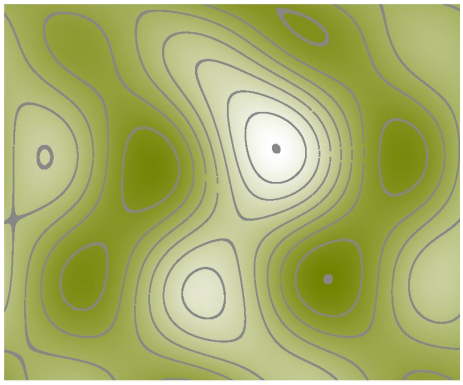
Funktionen im RKHS zu $k(x', x) := \exp(-||x - x'||^2/0.1)$ auf $[\,-1, 1\,]^2$

Beispiel 12: $k(x', x) := \exp(\langle x, x' \rangle)$ auf $[\,-1, 1\,]^2$. Der RKHS besteht aus Linearkombinationen von Exponentialfunktionen.



Funktionen im RKHS zu $k(x', x) := \exp(\langle x, x' \rangle)$ auf $[\,-1, 1\,]^2$

Beispiel 13: $k(x', x) := \cos\langle x, x' \rangle + \cos(3\langle x, x' \rangle)$ auf $[\,-1, 1\,]^2$. Der RKHS besteht aus Linearkombinationen von trigonometrischen Funktionen[9)].



Funktionen im RKHS zu $k(x', x) := \cos\langle x, x' \rangle + \cos(3\langle x, x' \rangle)$ auf $[\,-1, 1\,]^2$

Hier noch ein Beispiel, das zu analytischen Funktionen führt:

Beispiel 14: Sei $X := \,]-1, 1\,[$. Für $a = (a_n)_{n=0,1,\ldots} \in l^2$ definieren wir eine Funktion f_a durch $f_a(x) := \sum a_n x^n$. Da die a_n beschränkt sind, ist f_a eine analytische Funktion auf X, und man kann die a_n aus f_a rekonstruieren. H soll die Menge der f_a bezeichnen.

[9)] Man sieht eine ähnliche Visualisierung auch auf dem Titelbild.

Setze $\langle f_a, f_b\rangle := \langle a, b\rangle = \sum a_n b_n$, das macht H zu einem Hilbertraum von Funktionen auf X. Ist es ein RKHS? Die Antwort ist positiv, denn

$$\begin{aligned} |f_a(x)| &= |\sum a_n x^n| \\ &= |\langle (a_n), (x^n)\rangle| \\ &\leq ||f_a||\,||(x^n)||, \end{aligned}$$

wobei $||(x^n)||$ im l^2 zu berechnen ist:

$$||(x^n)||^2 = 1 + x^2 + x^4 + \cdots = \frac{1}{1-x^2}.$$

Und wie sieht der zugehörige Kern aus? Es soll doch $\langle f_a, k(\cdot, x)\rangle = f_a(x)$ sein. Dabei ist $k(\cdot, x) = f_b$ für ein geeignetes, noch unbekanntes b. Kurz: Wir suchen ein $b \in l^2$, so dass $f_a(x) = \sum a_n x^n = \sum a_n b_n$ gilt. Damit ist klar: $k(\cdot, x) = f_{(x^n)}$, also

$$k(x', x) = \sum x'^n x^n = \frac{1}{1 - xx'}.$$

Ganz analog kann man das auch im Komplexen mit analytischen Funktionen auf $\{z \mid |z| < 1\}$ machen. Das führt dann zum *Hardy-Raum* $H^2(\mathbb{D})$.

Beispiel 15: Hier ist noch ein theoretisches Beispiel. Sei X irgendeine Menge, und darauf sei eine $\mathbb{R}$-wertige Funktion ψ erklärt. Dann wird, wie leicht zu sehen, durch

$$k(x', x) := \psi(x)\psi(x')$$

ein Kern definiert. Wenn ψ nicht gerade die Nullfunktion ist, ist der zugehörige RKHS gleich $\mathbb{R}\psi$, und $\langle \psi, \psi\rangle = 1$

Damit sind leicht Beispiele für „pathologische“ Kerne anzugeben. So ein k wird, zum Beispiel, im Allgemeinen weder messbar noch stetig sein.

3.5 Der Kern bestimmt die Eigenschaften des RKHS

Aus Eigenschaften von k lassen sich Eigenschaften der Funktionen im zugehörigen RKHS ablesen. Wir beschränken uns auf einige typische Beispiele.

Beschränktheit

Sei H der RKHS zu k. Dann gilt:

$$\begin{aligned} k(x, x) &= \langle k(\cdot, x), k(\cdot, x)\rangle_H \\ &= ||k(\cdot, x)||_H^2. \end{aligned}$$

Satz 3.5.1. *$H \subset Abb(X, \mathbb{K})$ sei der zum Kern $k : X \times X \to \mathbb{K}$ gehörige RKHS. H besteht genau dann aus beschränkten Funktionen, wenn $x \mapsto k(x, x)$ beschränkt ist.*

Beweis: Die Cauchy-Schwarzsche Ungleichung liefert

$$\begin{aligned} |k(x,x')|^2 &= |\langle k(\cdot,x), k(\cdot,x')\rangle|^2 \\ &\leq ||k(\cdot,x)||_H^2 ||k(\cdot,x')||_H^2 \\ &= k(x,x)k(x',x'). \end{aligned}$$

Sei zunächst $M := \sup_x k(x,x) < \infty$. Für $f \in H$ und beliebiges x ist

$$|f(x)| = |\langle f, k(\cdot,x)\rangle| \leq ||f||_H ||k(\cdot,x)|| \leq ||f||_H \sqrt{M}.$$

f ist also beschränkt.

Schwieriger ist die Umkehrung einzusehen.

Beweis I:

Alle f seien beschränkt. Wir betrachten dann die Abbildung $\Psi : f \mapsto (f(x))_x$ von H in den Banachraum $l^\infty(X)$, die nach Voraussetzung wohldefiniert und sicher auch linear ist.

Wir behaupten, dass sie abgeschlossen ist. Sei dazu (f_n) eine gegen ein $f_0 \in H$ konvergente Folge, und $(\Psi(f_n))$ möge gegen ein $g \in l^\infty$ konvergieren. Da alle Abbildungen $f \mapsto f(x)$ stetig sind, folgt $\Psi f_0 = g$, d.h., Ψ ist wegen des Satzes vom abgeschlossenen Graphen stetig. Es gibt also ein M, so dass stets $||\Psi f||_\infty \leq M||f||_H$. Insbesondere ist für jedes x

$$|k(x,x)| \leq ||k(\cdot,x)||_\infty = ||\Psi\big(k(\cdot,x)\big)||_\infty \leq M||k(\cdot,x)||_H \leq M\sqrt{k(x,x)}.$$

Damit ist $k(x,x) \leq M^2$.

Beweis II: Man kann auch ein anderes tiefliegendes Resultat der Funktionalanalysis verwenden, den Satz von der gleichmäßigen Beschränktheit: Eine Familie von stetigen Funktionalen auf einem Banachraum, die punktweise beschränkt ist, ist normbeschränkt.

> Formal findet eine Vertauschung von Quantoren „für alle“ und „es existiert“ statt. Das gibt es auch bei Ergebnissen im Zusammenhang mit Kompaktheit.

Alle $f \in H$ seien beschränkt, und diesmal betrachten wir die Funktionale $\delta_x : f \to \langle f, k(\cdot,x)\rangle = f(x)$ auf H. Die sind stetig mit $||\delta_x|| = ||k(\cdot,x)||_H$. (Denn

$$|f(x)| = |\langle f, k(\cdot,x)\rangle \leq ||f||||k(\cdot,x)||$$

sowie

$$|\delta_x(k(\cdot,x))| = k(x,x) = ||k(\cdot,x)||^2.)$$

Für jedes f sind die Zahlen $\delta_x(f)$ beschränkt, deswegen müssen nach dem zitierten Satz auch die Normen beschränkt sein. Es gibt ein R, so dass $||k(\cdot,x)||_H \leq R$ für alle x.

Weiter geht es wie oben:

$$\begin{aligned} k(x,x')^2 &= \langle k(\cdot,x), k(\cdot x')\rangle^2 \\ &\leq ||k(\cdot,x)||_H^2 ||k(\cdot,x')||_H^2 \\ &\leq R^4. \end{aligned}$$

Ein Gegenbeispiel:

Die Tatsache, dass tiefliegende Ergebnisse wichtig waren, lässt vermuten, dass das Ergebnis für nicht vollständige H falsch ist. Wirklich gilt: Es gibt X und $k : X \times X \to \mathbb{R}$ und einen aus Funktionen auf X bestehenden Prä-Hilbertraum H, so dass gilt:

- Alle $k(\cdot,x)$ gehören zu H, und $x \mapsto k(\cdot,x)$ ist eine Feature-Abbildung. (D.h.,
$$\langle k(\cdot,x), k(\cdot,x')\rangle = k(x',x).)$$
- Alle $k(\cdot,x)$ sind beschränkt.
- $x \mapsto k(x,x)$ ist unbeschränkt.

Man kann zum Beispiel $X = \mathbb{N}$ und die Feature-Abildung $n \mapsto (1,1,\ldots,1,0,0,\ldots)$ in den l^2 betrachten. Dann ist $k(m,n) := \min\{m,n\}$, und H soll die lineare Hülle der $k(\cdot,n)$ sein. □

Stetigkeit

X sei nicht nur eine Menge, sondern ein topologischer Raum. Der Zusammenhang zwischen der Stetigkeit von k und der Stetigkeit der Funktionen im zugehörigen RKHS H ist komplizierter als bei der entsprechenden Frage nach der Beschränktheit. Klar ist nur:

- Sind alle $f \in H$ stetig, so sind alle $k(x',\cdot)$ und alle $k(\cdot,x)$ stetig. (Denn $k(\cdot,x) \in H$, und $k(x',\cdot) = \overline{k(\cdot,x')}$.)

Satz 3.5.2. *Äquivalent sind:*

(i) k ist separat stetig und beschränkt auf X^2.

(ii) Alle $f \in H$ sind stetig und beschränkt.

Beweis: Es gelte (i). Dann wissen wir schon, dass alle $f \in H$ beschränkt sind. Stetig sind dann auch sicher alle f in der linearen Hülle H_0 der $k(\cdot,x)$, einem dichten Unterraum.

Zu $f \in H$ wählen wir eine Folge f_n in H_0 mit $||f_n - f||_H \to 0$. Mit den Ergebnissen von Satz 3.5.1 folgt, dass dann auch f_n gleichmäßig gegen f geht und folglich stetig ist.

Die Umkehrung ist, wieder wegen Satz 3.5.1, klar. □

Die Stetigkeit von k (also nicht nur die separate Stetigkeit) ist dann noch nicht garantiert:

Satz 3.5.3. *Es gibt einen RKHS auf einem topologischen Raum X, für den alle $f \in H$ beschränkt und stetig sind und für den k unstetig ist.*

Beweis: (Das Ergebnis geht auf Lehto zurück, die hier vorgestellte Konstruktion scheint neu zu sein.) Es sei $X := \{0\} \cup \{1/n \mid n \in \mathbb{N}\}$, versehen mit der Spurtopologie von $\mathbb{R}$. Wir betten l^2 in $\mathrm{Abb}(X, \mathbb{K})$ wie folgt ein: Für $a \in l^2$ sei $\Lambda(a)$ die Abbildung, die bei Null verschwindet und bei $1/n$ den Wert a_n hat. Das innere Produkt wird auf H, dem Bild von Λ, übernommen. (H besteht also aus den Funktionen auf X, deren Funktionswerte quadratsummierbar sind und die bei 0 verschwinden.)

Die Feature-Abbildung bildet 0 auf Null und $1/n$ auf die Indikatorfunktion von $\{1/n\}$ ab. Dann gilt:

- Alle $f \in H$ sind stetig und beschränkt.
- H ist der RKHS zu k.
- $k(0,0) = 0$ und $k(1/n, 1/n) = 1$.

k ist also nicht stetig bei $(0,0)$. □

Für unsere Zwecke ist das folgende Ergebnis am Wichtigsten:

Satz 3.5.4. *k sei stetig. Dann sind alle $f \in H$ stetig.*

Beweis: Für beliebige x, y gilt

$$\begin{aligned} ||k(\cdot,x) - k(\cdot,y)||_H^2 &= \langle k(\cdot,x) - k(\cdot,y), k(\cdot,x) - k(\cdot,y)\rangle \\ &= k(x,x) - 2k(x,y) + k(yy), \end{aligned}$$

also

$$||k(\cdot,x) - k(\cdot,y)||_H = \sqrt{k(x,x) - 2k(x,y) + k(y,y)}.$$

Damit weiß man: Ist k stetig, so folgt aus $x_i \to x$, dass $k(\cdot, x_i) \to k(\cdot, x)$ in H. (Eigentlich wurde hier nur gebraucht, dass alle $k(\cdot, x)$ und $x \mapsto k(x,x)$ stetig sind.)

Ist nun f beliebig, so muss man nur noch die Beziehung $\langle f, k(\cdot, y)\rangle = f(y)$ ausnutzen, um $f(x_i) \to f(x)$ garantieren zu können, falls $x_i \to x$. Folglich ist jedes f bei jedem x stetig. □

Messbarkeit

X sei eine Menge, und darauf sei $\mathcal{A}$ eine σ-Algebra. Gegeben seien ein Kern k auf $X \times X$ und der zugehörige RKHS H.

Satz 3.5.5. *Äquivalent sind:*

(i) Alle $k(\cdot, x)$ sind $\mathcal{A}$-messbar.

(ii) Alle $f \in H$ sind $\mathcal{A}$-messbar.

Beweis: (i) folgt aus (ii), da die $k(\cdot, x)$ zu H gehören. Wenn (i) gilt, sind alle f in H_0, der linearen Hülle der $k(\cdot, x)$, messbar. Ist $f \in H$ beliebig, so gibt es eine Folge (f_n) in H_0 mit $||f_n - f||_H \to 0$. Die H-Konvergenz impliziert aber die punktweise Konvergenz, d.h. f ist als punktweiser Limes messbarer Funktionen ebenfalls messbar. □

3.6 Konkrete Rechnungen

Nun kann das Programm verwirklicht werden, lineare Techniken im Nichtlinearen anzuwenden, falls in dem jeweiligen Algorithmus nur innere Produkte auftauchen. In Abschnitt 1.4 wurde das schon für den Perceptron-Argorithmus angewandt, jetzt wollen wir auch die Ergebnisse aus Abschnitt 2.3 übertragen. Im Grunde ist es ganz einfach, man muss nur die inneren Produkte $\langle x_i, x_j \rangle$ an allen Stellen durch $k(x_i, x_j)$ ersetzen.

Als exemplarisches Beispiel betrachten wir das *Problem der optimalen Klassifikation trennbarer Punktmengen*. Man hat also x_i, y_i gegeben, und es ist ein RKHS H auf X mit reproduzierendem Kern k vorgelegt, von dem man weiß, dass es ein $f \in H$ und ein $b \in \mathbb{R}$ mit

$$y_i(f(x_i) + b) > 0,\ i = 1, \ldots, l$$

gibt. Wenn man optimal durch ein Element in $f^* \in H$ und ein $b^* \in \mathbb{R}$ trennen möchte, muss man das folgende Problem lösen:

- Finde das Maximum der Funktion

$$W(\alpha) := \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j k(x_i, x_j)$$

  unter den Nebenbedingungen

$$\alpha_i \geq 0,\ \sum_i y_i \alpha_i = 0.$$

Mal angenommen, die α_i^* lösen das Problem (nur wenige dieser Zahlen werden von Null verschieden sein). Setze $f^* := \sum y_i \alpha_i^* k(\cdot, x_i)$, und b^* wird als

$$b^* := -\frac{\max_{i, y_i = -1} f^*(x_i) + \min_{i, y_i = 1} f^*(x_i)}{2}$$

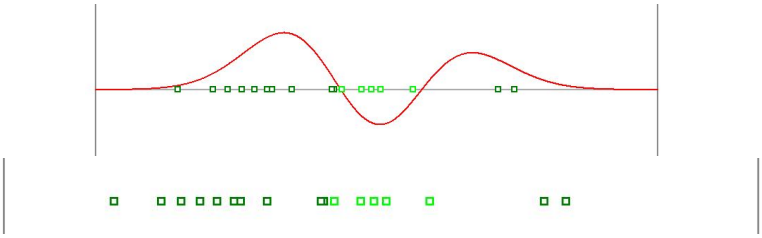
definiert.

Ganz ähnlich geht man vor, falls mit Schlupfvariablen gearbeitet werden soll oder muss oder wenn es um Regressionsprobleme geht.

Nachstehend werden als typisches Beispiel mehrere Datensätze mit dem Perceptron-Algorithmus behandelt. Der Ablauf ist stets so:

- Fixiere einen Kern k auf $X = [\,-1,1\,]$ oder auf $X = [\,-1,1\,] \times [\,-1,1\,]$.
- Erzeuge eine zufällige Linearkombination f_0 der $k(\cdot, x)$. also ein Element des zugehörigen RKHS.
- Suche eine Zahl b_0 und erzeuge Zufallspunkte x mit $f_0(x) < b$ und Zufallspunkte x mit $f_0(x) > b$. Wir haben also Paare $(x_i, y_i) \in X \times \{-1, 1\}$ für $i = 1, \ldots, l$. Die x_i gehören zu zwei Klassen, und die sind im $RKHS$ zu k trennbar.
- Vergiss f_0 und b_0 und trenne die Klassen durch den Perceptronalgorithmus.

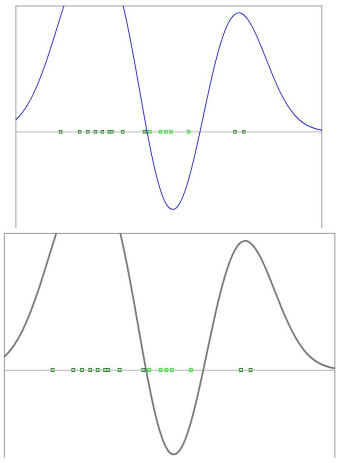
Beispiel 1: Wir arbeiten mit einem Gaußkern auf $[\,-1,1\,]$. Hier sind f_0 (rot) und die Zufallspunkte. (f_0 ist so skaliert, dass $b_0 = 0$.)



Perceptron für einen Gaußkern

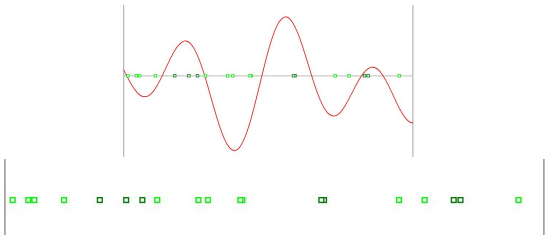
Oben ist f_0 mit den Punkten abgebildet, darunter sind nur die zwei Klassen zu sehen. *Damit* (und mit der Kenntnis von k) soll getrennt werden.

Der Perceptron-Algorithmus (blau) braucht nur einen Durchgang, eine klassifizierende Funktion ist grau eingezeichnet:



Perceptron für einen Gaußkern

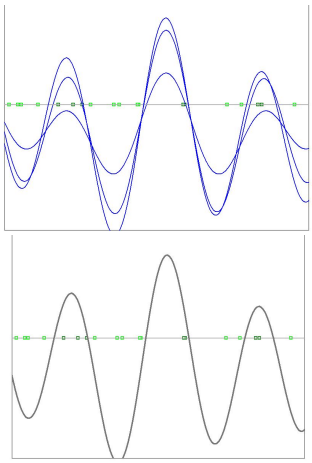
Beispiel 2: Wir arbeiten mit einem Cosinuskern auf $[\,-1,1\,]$. Hier sind f_0 (rot) und die Zufallspunkte. (f_0 ist so skaliert, dass $b_0 = 0$.)



Perceptron für einen Cosinuskern

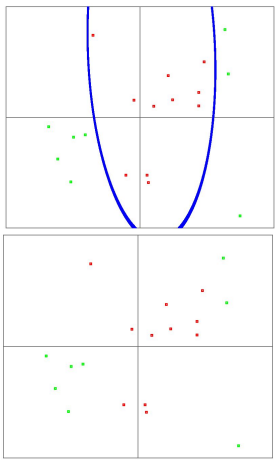
Oben ist f_0 mit den Punkten abgebildet, darunter sind nur die zwei Klassen zu sehen. *Damit* (und mit der Kenntnis von k) soll getrennt werden.

Der Perceptron-Algorithmus (blau) braucht nur viele Durchgänge, eine klassifizierende Funktion ist grau eingezeichnet:



Perceptron für einen Cosinuskern

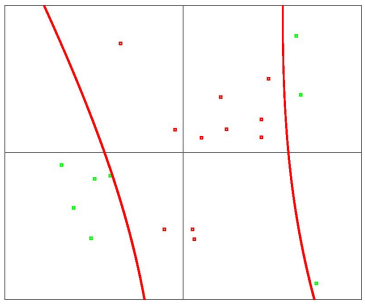
Beispiel 3: Wir arbeiten mit einem Polynomkern auf $[-1,1] \times [-1,1]$. Hier sieht man die Menge $\{f_0 = b_0\}$ (rot) und die Zufallspunkte.



Perceptron für einen Polynomkern auf $[-1,1] \times [-1,1]$

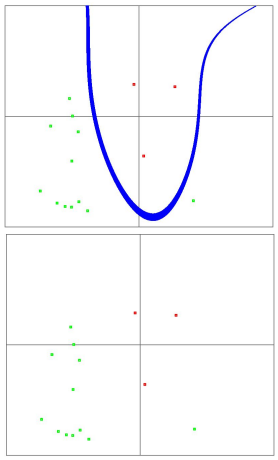
Nur mit Hilfe der zwei Klassen (und mit der Kenntnis von k) soll getrennt werden.

Der Perceptron-Algorithmus findet nach einigen Durchgängen eine klassifizierende Funktion f und ein b. Die trennende Menge $\{f = b\}$ ist rot eingezeichnet:



Perceptron für einen Polynomkern auf $[-1,1] \times [-1,1]$

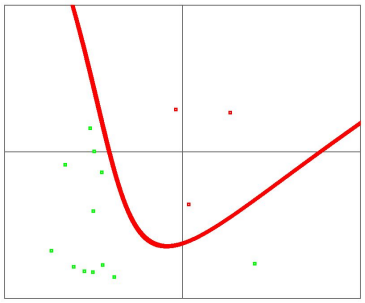
Beispiel 4: Wir arbeiten mit einem Exponentialkern auf $[-1,1] \times [-1,1]$. Hier sieht man die Menge $\{f_0 = b_0\}$ (rot) und die Zufallspunkte.



Perceptron für einen Exponentialkern auf $[-1,1] \times [-1,1]$

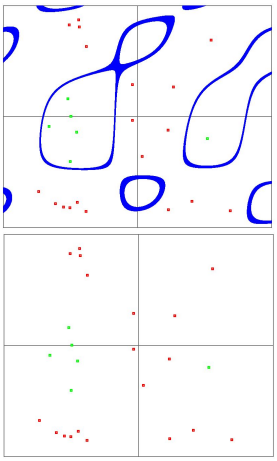
Nur mit Hilfe der zwei Klassen (und mit der Kenntnis von k) soll getrennt werden.

Der Perceptron-Algorithmus findet nach einigen Durchgängen eine klassifizierende Funktion f und ein b. Die trennende Menge $\{f = b\}$ ist rot eingezeichnet:



Perceptron für einen Exponentialkern auf $[-1,1] \times [-1,1]$

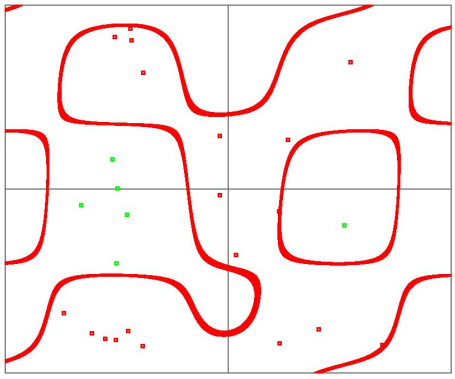
Beispiel 5: Wir arbeiten mit einem Cosinuskern auf $[-1,1] \times [-1,1]$. Hier sieht man die Menge $\{f_0 = b_0\}$ (rot) und die Zufallspunkte.



Perceptron für einen Cosinuskern auf $[-1,1] \times [-1,1]$

Nur mit Hilfe der zwei Klassen (und mit der Kenntnis von k) soll getrennt werden.

Der Perceptron-Algorithmus findet nach einigen Durchgängen eine klassifizierende Funktion f und ein b. Die trennende Menge $\{f = b\}$ ist rot eingezeichnet:



Perceptron für einen Cosinuskern auf $[-1,1] \times [-1,1]$

Kapitel 4

Der theoretische Hintergrund

Wir haben in den letzten Kapiteln einige Algorithmen kennen gelernt, durch die Klassifizierungs- und Regressionsprobleme konkret behandelt werden können. Jetzt wollen wir das Thema etwas theoretischer angehen: Was ist das Modell, und wie kann man das Ziel formulieren?

Dazu wird es erforderlich sein, Begriffe und Fakten aus der Stochastik zu verwenden.

4.1 Stochastik I: Erinnerungen

Es wird im Folgenden vorausgesetzt, dass die folgenden Sachverhalte bekannt sind:

Wahrscheinlichkeitsräume

- Eine *σ-Algebra* $\mathcal{E}$ auf einer Menge Ω ist eine Teilmenge der Potenzmenge, die unter allen Mengenoperationen stabil ist, bei denen höchstens abzählbar viele Elemente von $\mathcal{E}$ beteilgt sind.
- Sei $\mathcal{E}$ eine σ-Algebra auf Ω. Eine Abbildung $\mathbb{P} : \mathcal{E} \to [0,1]$ heißt ein *Wahrscheinlichkeitsmaß*, wenn $\mathbb{P}(\Omega) = 1$ ist und
$$\mathbb{P}\big(\bigcup_n E_n\big) = \sum_n \mathbb{P}(E_n)$$
für jede Folge (E_n) von paarweise disjunkten Mengen in $\mathcal{E}$ gilt.
- Ein *Wahrscheinlichkeitsraum* ist ein Tripel $(\Omega, \mathcal{E}, \mathbb{P})$; dabei ist Ω eine Menge, $\mathcal{E}$ eine σ-Algebra auf Ω und $\mathbb{P}$ ein Wahrscheinlichkeitsmaß auf $(\Omega, \mathcal{E})$.
- Die σ-Algebra der *Borelmengen* auf dem $\mathbb{R}^n$ ist die kleinste σ-Algebra, die alle offenen Teilmengen enthält. Faustregel: *Jede* Teilmenge, die in den Anwendungen jemals vorkommen kann, ist eine Borelmenge.

Wichtige Beispiele für Wahrscheinlichkeitsräume

- Ist Ω endlich oder höchstens abzählbar, so ist $\mathcal{E}$ in der Regel die Potenzmenge. Ein Wahrscheinlichkeitsmaß ist dann durch die Angabe der Zahlen $\mathbb{P}(\{\omega\})$ definiert. (Diese Zahlen müssen nichtnegativ sein und sich zu Eins summieren.)

- Die wichtigsten Beispiele dazu sind

  - *Laplaceräume*: Da ist Ω endlich, und alle Elementarereignisse haben die gleiche Wahrscheinlichkeit.
  - *Bernoulliräume*. Hier ist $\Omega = \{0, 1\}$, und es reicht die Angabe der Zahl $p = \mathbb{P}(\{1\})$ („Wahrscheinlichkeit für Erfolg“), um das Wahrscheinlichkeitsmaß festzulegen.
  - Abgeleitet von Bernoulliräumen sind die *geometrische Verteilung* (warten auf den ersten Erfolg), die *Binomialverteilung* (k Erfolge in n Versuchen), die *hypergeometrische Verteilung* (Ziehen ohne Zurücklegen) und die *Poissonverteilung* (Grenzwert von Binomialverteilungen).

- Sei zunächst Ω eine „einfache“ Teilmenge von $\mathbb{R}$ (etwa ein Intervall) und $f : \Omega \to \mathbb{R}$ eine „gutartige“ (etwa eine stetige) nichtnegative Funktion mit Integral Eins. Dann wird dadurch ein Wahrscheinlichkeitsraum durch die Festsetzung
$$\mathbb{P}(E) := \int_E f(x)\,dx$$
definiert. Dabei kann E eine beliebige Borelmenge sein. Für die Anwendungen reicht es aber so gut wie immer, sich für E ein Teilintervall von Ω vorzustellen. f heißt die *Dichtefunktion* zu dem so definierten Wahrscheinlichkeitsmaß.

- Die wichtigsten Beispiele sind

  - Die *Gleichverteilung* auf $[a, b]$; da ist $f(x) := 1/(b - a)$.
  - Die *Exponentialverteilung* zum Parameter $\lambda > 0$; sie ist durch die Dichtefunktion
  $$f(x) := \lambda \cdot e^{-\lambda x}$$
  auf $\mathbb{R}^+$ definiert. Durch die Exponentialverteilung kann gedächtnisloses Warten beschrieben werden.
  - Die *Normalverteilungen* $N(\mu, \sigma^2)$ auf $\mathbb{R}$. Sie haben – für $\mu \in \mathbb{R}$ und $\sigma > 0$ – die Dichtefunktion
  $$f(x) := \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}.$$
  Sie spielen in der Statistik eine ganz besonders wichtige Rolle.

- Die gleiche Idee kann in allen Situationen ausgenutzt werden, in denen ein Integral zur Verfügung steht. Wer also auf $\mathbb{R}$ das Lebesgue-Integral kennen gelernt hat, kann integrierbare Dichten zulassen, wer die Integration im $\mathbb{R}^n$ beherrscht, kann leicht Wahrscheinlichkeitsmaße auf den Borelmengen dieses Raumes angeben usw.

Wahrscheinlichkeitstheorie: Grundbegriffe

- Bedingte Wahrscheinlichkeit.
- Was bedeutet „Unabhängigkeit“ für zwei, endlich viele bzw. beliebig viele Ereignisse?
- Zufallsvariable.
- Erwartungswert und Streuung.
- Unabhängigkeit für Zufallsvariable.

Grenzwertsätze

Die Grenzwertsätze besagen, „dass der Zufallseinfluss verschwindet“, wenn sich „viele“ Zufallseinflüsse unabhängig überlagern. Man sollte kennen:

- Die Definitionen „Konvergenz in Wahrscheinlichkeit“, „Konvergenz in Verteilung“, „Fast sichere Konvergenz“.
- Das Wurzel-n-Gesetz.
- Die Lemmata von Borel-Cantelli.
- Die Tschebyscheff-Ungleichung und die Markov-Ungleichung.
- Das schwache Gesetz der großen Zahlen.
- Das starke Gesetz der großen Zahlen.
- Den zentralen Grenzwertsatz.

4.2 Ein stochastisches Modell des maschinellen Lernens

Beim Klassifizierungs- und beim Regressionsproblem war die Situation doch die folgende:

- Es gibt eine Menge X von „Datensätzen“ und eine Menge $Y \subset \mathbb{R}$ von „Ergebnissen“: Beim Klassifizieren war $Y = \{-1, +1\}$ und bei der Regression $Y = \mathbb{R}$.
- Man hat einen Satz $(x_i, y_i) \in X \times Y$ $(i = 1, \ldots, l)$ „zum Trainieren“ zur Verfügung. Damit soll man eine Funktion $f : X \to Y$ vorschlagen.

- Die Hoffnung: Ist ein weiteres $x \in X$ gegeben, so ist $f(x)$ ein vernünftiger Vorschlag für die Klassifizierung bzw. diejenige Funktion, die durch die Regressionsfunktion approximiert werden sollte.

$X \times Y$ als Wahrscheinlichkeitsraum

Um zu einem abstrakten Modell zu kommen, stellen wir uns vor, dass X ein Wahrscheinlichkeitsraum ist: Das Maß soll μ_X heißen. Die x_i, die man zum Trainieren zur Verfügung hat, sind durch unabhängige Abfragen dieses Wahrscheinlichkeitsraums entstanden.

Ist irgendein x gewählt, so entsteht das zugehörige $y \in Y$ ebenfalls zufällig. Y trägt nämlich für jedes x ein Wahrscheinlichkeitsmaß $\mu(\cdot, x)$, und das y wird gemäß $\mu(\cdot, x)$ ausgesucht. $\mu(\cdot, x)$ könnte zum Beispiel die Gleichverteilung auf $[\, f_0(x) - 0.1, f_0(x) + 0.1\,]$ sein, wobei $f_0 : X \to \mathbb{R}$ eine feste Funktion ist. (Durch dieses Maß wird eine fehlerbehaftete f_0-Messung modelliert.)

Satz 4.2.1. *(i) Durch μ_X und die $\mu(\cdot, x)$ wird ein Wahrscheinlichkeitsmaß μ auf $X \times Y$ induziert.*

(ii) Umgekehrt gilt das (für nicht zu pathologische Räume X) auch: Für jedes Maß μ auf $X \times Y$ kann man μ_X und $\mu(\cdot, x)$ (alle $x \in X$) finden, so dass μ daraus wie in (i) entsteht.

Beweis: (Auf Feinheiten wie Messbarkeits- und Integrierbarkeitsvorausetzungen gehen wir hier nicht ein.)
(i) Für $A \subset X \times Y$ und $x \in X$ definiere $A_x := \{y \mid (x, y) \in A\}$. Setze dann

$$\mu(A) := \int_X \mu(A_x|x)\, d\mu_X.$$

(ii) Das ist viel schwieriger. Wer es ganz genau wissen möchte, sollte (zum Beispiel) Kapitel 14 im Buch von Klenke konsultieren (oder den Anhang, insbesondere Lemma A.3.16, im Buch von Steinwart-Christmann). Inhaltlich sind die $\mu(\cdot, x)$ bedingte Wahrscheinlichkeiten. (x ist bekannt, was kann man denn für y prognostizieren?)

Die Definition von μ_X ist unproblematisch: $\mu_X(A) := \mu(A \times Y)$. Wir beweisen die Existenz der $\mu(\cdot, x)$ nur für zwei Spezialfälle:

Spezialfall 1: X ist endlich. Für $\mu_X(\{x\}) = 0$ sei $\mu(\cdot, x)$ ein beliebiges Wahrscheinlichkeitsmaß auf Y, und andernfalls ist es durch

$$\mu(A|x) := \mu(\{x\} \times A)/\mu_X(\{x\})$$

definiert.

Spezialfall 2: Y ist endlich. Wir illustrieren die Idee am Fall $Y = \{-1, +1\}$ (die Klassifikationssituation). Die Einschränkungen von μ auf die Teilmengen $X \times \{-1\}$ und $X \times \{+1\}$ können als Maße $\mu_{\pm 1}$ auf X aufgefasst werden. Dann

ist μ_1 absolutstetig gegen $\mu_{-1} + \mu_1$, es gibt also eine Radon-Nikodym-Dichte f mit

$$d\mu_1 = f\, d(\mu_{-1} + \mu_1).$$

Dabei ist f fast sicher $[\,0,1\,]$-wertig.

Wähle einen Vertreter, der *überall* $[\,0,1\,]$-wertig ist und definiere dann $\mu(\cdot|x)$ für $x \in X$ auf Y durch „$f(x)$ mal Diracmaß auf $+1$ plus $1 - f(x)$ mal Diracmaß auf -1“. □

Die Verlustfunktion

Es ist also $X \times Y$ ein Wahrscheinlichkeitsraum, und wir suchen – unter Verwendung der (x_i, y_i) – eine Funktion $f : X \to Y$, so dass für später zu erzeugende (x, y) die Zahl $f(x)$ „möglichst nahe“ bei y liegt. Dazu muss man sich natürlich darauf verständigen, wie man den Unterschied zwischen diesen beiden Zahlen bewertet.

Formal sieht es so aus, dass man eine *Verlustfunktion* $L : \mathbb{R}^2 \to \mathbb{R}^+$ gegeben hat; $L\big(f(x), y\big)$ ist dann der „Verlust“, wenn $f(x)$ vorausgesagt wird, aber y eintritt.

Hier sind einige Beispiele für Verlustfunktionen:

1. $L(z, y) := |z - y|^p$ für irgendeine $p \geq 1$. Wir werden überwiegend mit diesem L und $p = 2$ arbeiten (*quadratischer Verlust*). Das ist angemessen für Regressionsaufgaben.

2. $L(z, y) := 1$ für $z \neq y$, und $L(z, y) := 0$ für $z = y$. Das ist die naheliegende Verlustfunktion bei Klassifizierungsproblemen.

3. $L(z, y) := 0$ für $|z - y| \leq \varepsilon$, und $L(z, y) := (|z - y| - \varepsilon)^p$ sonst: Kleine Fehler werden ignoriert.

4. Angenommen, es ist $Y = \{-1, 1\}$, wir wollen also klassifizieren. Manchmal ist es günstig, den Verlust dann so zu bewerten:

Für $y = 1$ sei $L(z, y) := \max\{0, 1 - z\}$, und für $y = -1$ setzt man $L(z, y) := \max\{0, 1 + z\}$. Der Hauptvorteil dieses L („*hinge loss*“) besteht darin, dass man es beim Optimieren mit konvexen Problemen zu tun hat.

Im Buch von Steinwart-Christmann ist Verlustfunktionen ein ganzes Kapitel gewidmet. Viele Aspekte können eine Rolle spielen:

- Ist die Verlustfunktion symmetrisch, gilt also $L(z, y) = L(y, z)$? Das ist nicht in allen Fällen angemessen.

- Ist sie stetig? differenzbar? beschränkt? integrierbar? ist $L(\cdot, y)$ konvex oder sogar strikt konvex?

Wir bleiben bei der quadratischen Verlustfunktion.

Risiko!

X, Y und die Verlustfunktion sind bekannt, μ ist aber unbekannt. Wenn man sich für eine Funktion f entschieden hat, so ist doch die Zahl

$$\mathcal{R}_\mu(f) := \int_{X\times Y} (f(x)-y)^2 \, d\mu = \int_X \int_Y (f(x)-y)^2 \, d\mu(\cdot|x) \, d\mu_X$$

der im Mittel zu erwartende Verlust, also das *Risiko*, und es wäre sicher wünschenswert, das f so zu wählen, dass $\mathcal{R}_\mu(f)$ minimal ist. Wir setzen also

$$\mathcal{R}^*_\mu := \inf_{f:X\to\mathbb{R}} \mathcal{R}_\mu(f),$$

und wir hoffen, mit einem geeigneten f zu erreichen, dass $\mathcal{R}^*_\mu \approx \mathcal{R}_\mu(f)$.

Was ist denn im besten Fall zu erwarten? Aus der elementaren Stochastik weiß man:

> G sei eine reelle Zufallsvariable. An der Stelle $x = \mathbb{E}(G)$ wird die Funktion $x \mapsto \mathbb{E}(G-x)^2$ minimal.

Deswegen sollte man eine Funktion f_μ punktweise durch $f_\mu(x) := \mathbb{E}_{\mu(\cdot,x)}(y) = \int_Y y \, d\mu(\cdot|x)$ definieren. Dann gilt sicher $\mathcal{R}_\mu(f_\mu) = \mathcal{R}^*_\mu$. Doch leider kennen wir f_μ nicht, da μ unbekannt ist.

Die Abweichung zwischen $\mu(\cdot|x)$ und $f_\mu(x)$ kann durch die Streuung der Zufallsvariablen y auf $(Y, \mu(\cdot|x))$ gemessen werden. Die entsprechende Varianz (also $\int_Y (y-f_\mu(x))^2 \, d\mu(\cdot|x)$) soll $\sigma^2_\mu(x)$ heißen, und wir definieren

$$\sigma^2_\mu := \int_X \sigma^2_\mu(x) \, d\mu_X.$$

Das ist gerade die Streuung von $(x,y) \mapsto y$, es ist eine natürliche untere Schranke dafür, was bei diesem Problem im besten Fall zu erreichen sein wird.

Hier noch eine weitere Erinnerung aus der elementaren Stochastik:

> Ist G eine Zufallsvariable, so gilt für jedes α:
>
> $$\mathbb{E}(\alpha - G)^2 = (\alpha - \mathbb{E}(G))^2 + \mathbb{E}\big(G - \mathbb{E}(G)\big)^2.$$

Das kann man direkt ausrechnen. (Es ist auch als Variante des Satzes von Pythagoras in einem gewissen L^2-Raum aufzufassen, denn $\alpha - \mathbb{E}(G) \perp \mathbb{E}(G) - G$.)

Mit $G = y$ auf $(Y, \mu(\cdot|x))$ und $\alpha = f(x)$ bedeutet es

$$\int \big(f(x)-y\big)^2 \, d\mu(\cdot|x) = \big(f(x) - f_\mu(x)\big)^2 + \sigma^2_\mu(x),$$

und wenn wir diese Identität gegen μ_X aufintegrieren, folgt

$$\mathcal{R}_\mu(f) = \int_X \big(f(x) - f_\mu(x)\big)^2 \, d\mu_X + \sigma^2_\mu.$$

Das bedeutet, dass das Problem in zwei Teile zerfällt. Einen, wo man versucht, den L^2-Unterschied zwischen f und f_μ klein zu bekommen, und einen (das σ_μ^2), wo man nichts machen kann.

Experimente

Das einzige, was man zur Verfügung hat, sind Experimente: Man fragt $(X \times Y, \mu)$ „sehr oft“ ab (unabhängige Abfragen gemäß μ) und erhält so eine „Testmenge“ $T = (x_i, y_i)_{i=1,\dots,l}$, die man optimal zum „Lernen“ von f_μ nutzen soll. Formal heißt das, dass ein Algorithmus anzugeben ist, der jedem T ein $f_T : X \to Y$ (oder von X nach $\mathbb{R}$) zuordnet, und zwar so, dass man sich mit größer werdendem l immer mehr darauf verlassen kann, dass $\mathcal{R}_\mu(f_T)$ nahe bei $\mathcal{R}_\mu^*$ liegt. (Es soll natürlich auch so sein, dass f_T mit vertretbarem Aufwand zu bestimmen ist.)

4.3 Stochastik II: Ungleichungen

Durch die Gesetze der großen Zahlen wird in der elementaren Stochastik präzisiert, inwiefern größere Abweichungen vom Erwartungswert unwahrscheinlich sind, wenn man Mittelwerte unabhängiger Abfragen betrachtet. Was aber lässt sich für eine einzelne Zufallsvariable aussagen? Ein typisches Beispiel ist die Tschebyscheff-Ungleichung: Für jede Zufallsvariable gilt

$$\mathbb{P}\big(|Z - \mathbb{E}(Z)| \geq t\big) \leq \frac{\sigma^2(Z)}{t^2},$$

oder

$$\mathbb{P}\big(|Z - \mathbb{E}(Z)| < t\big) \geq 1 - \frac{\sigma^2(Z)}{t^2}.$$

Das wird für unsere Zwecke nicht ausreichen, wir benötigen die *Hoeffding-ungleichung* (s.u.).

Sei X eine reellwertige Zufallsvariable. Wir wollen ihr eine im folgenden wichtige Funktion $C_X : \mathbb{R} \to \mathbb{R}$ zuordnen, die *Kumulantenerzeugende Funktion*:

Definition 4.3.1. *$C_X : \mathbb{R} \to \mathbb{R}$ ist durch*

$$C_X(\theta) := \log \mathbb{E}(e^{\theta X})$$

definiert. (Wir werden stets voraussetzen, dass die hier betrachteten Erwartungswerte existieren).

Bemerkungen:

1. Ist X gleich der Konstanten r, so ist $C_X(\theta) = \theta r$.

2. Für $N(0,1)$-verteilte X kann man beweisen: $C_X(\theta) = \log(e^{\theta^2/2}) = \theta^2/2$.

3. Sind X, Y unabhängig, so ist $C_{X+Y} = C_X + C_Y$. Dabei haben wir ausgenutzt, dass mit X, Y auch $e^{\theta X}, e^{\theta Y}$ unabhängig sind und dass deswegen der Erwartungswert mit Produkten vertauscht.

4. Klar ist, dass stets $C_{\alpha X}(\theta) = C_X(\alpha\theta)$ gilt.

5. Fixiere ein $\alpha > 0$, wir definieren X als diejenige Zufallsvariable, für die $\mathbb{P}_X = (\delta_{-\alpha} + \delta_\alpha)/2$ gilt: Mit jeweils Wahrscheinlichkeit 0.5 wird α oder $-\alpha$ erzeugt. Es ist dann $\mathbb{E}(e^{\theta X}) = \cosh(\alpha\theta)$, für „große“ positive θ ist also $C_X(\theta) \approx \theta\alpha - \log 2$. Vergleicht man nur Zufallsvariable mit Erwartungswert 0, so scheint der Anstieg von C_X ein Maß für die Streuung zu sein.

Hier ist der Satz von Cramér:

Satz 4.3.2. *$X, X_1, \ldots, X_n$ seien Zufallsvariable, und $X_1, \ldots, X_n$ seien unabhängig. Für $t > 0$ gilt dann*

(i) $\mathbb{P}(X \geq t) \leq \exp\big(\inf_{\theta>0} -\theta t + C_X(\theta)\big)$.

(ii) $\mathbb{P}\big(\sum_{i=1,\ldots,n} X_i \geq t\big) \leq \exp\big(\inf_{\theta>0} -\theta t + \sum_i C_{X_i}(\theta)\big)$.

Beweis: (i) Fixiere ein $\theta > 0$. Aufgrund der *Markovungleichung* gilt für jede positive Zufallsvariable Y:

$$\mathbb{P}(Y \geq s) \leq \frac{\mathbb{E}(Y)}{s}.$$

Das nutzen wir wie folgt aus:

$$\begin{aligned} \mathbb{P}(X \geq t) &= \mathbb{P}\big(\exp(\theta X) \geq \exp(\theta t)\big) \\ &\leq e^{-\theta t}\mathbb{E}\big(\exp(\theta X)\big) \\ &= \exp(-\theta t + C_X(\theta)). \end{aligned}$$

Nun muss man nur noch das Infimum über alle θ bilden.

(ii) Hier ist nur (i) mit Bemerkung 3 zu kombinieren. □

Wir testen die Cramér-Ungleichung an einigen Beispielen.

1. X sei konstant gleich r, also $C_X(\theta) = \theta r$. Auf der rechten Seite der Cramérungleichung steht dann

$$\inf_{\theta>0} \theta(r - t),$$

und das ist 0 für $t \leq r$ und $-\infty$ für $t > r$. Die Cramérungleichung sagt also richtig voraus:

- $\mathbb{P}(X \geq t) = 1$ für $t \leq r$.
- $\mathbb{P}(X \geq t) = 0$ für $t > r$.

2. X sei nun $N(0,1)$-verteilt. Dann ist $C_X(\theta) = \theta^2/2$. Für $t > 0$ ist

$$\inf_{\theta>0} -\theta t + \theta^2/2 = -t^2/2$$

(mit elementarer Analysis), und damit liefert die Ungleichung

$$\mathbb{P}(X \geq t) \leq e^{-t^2/2}.$$

(In Wirklichkeit gilt sogar $\mathbb{P}|X| \geq t) \leq e^{-t^2/2}$, hier kommt nur $\mathbb{P}(|X| \geq t) \leq 2e^{-t^2/2}$ heraus.)

Die Cramérungleichung ist etwas unhandlich. Deswegen zeigen wir noch etwas leichter anwendbare Varianten. Wir beginnen mit einer eher technischen Vorbereitung.

Lemma 4.3.3. *(i)* $\cosh x \leq e^{x^2/2}$ *für alle* x.
(ii) X *sei eine beschränkte Zufallsvariable:* $|X| \leq M$ *fast sicher. Es sei auch* $\mathbb{E}X = 0$. *Dann ist* $C_X(\theta) \leq M^2\theta^2/2$.
(iii) Seien $t, \alpha > 0$. *Dann ist* $\min_{\theta>0} -\theta t + \alpha\theta^2/2 = t^2/(2\alpha)$.

Beweis: (i) Die Potenzreihenentwicklungen von $\cosh x$ und $e^{x^2/2}$ sind

$$\frac{1}{2}\big((1 + x + \frac{x^2}{2!} \pm \cdots) + (1 - x + \frac{x^2}{2!} \pm \cdots)\big) = 1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \cdots$$

und

$$1 + \frac{x^2}{2} + \frac{x^4}{2^2 2! + \cdots}.$$

Der typische Summand in der der ersten bzw. zweiten Reihe lautet

$$\frac{x^{2n}}{(2n)!} \text{ bzw. } \frac{x^{2n}}{2^n n!}.$$

Die Behauptung folgt dann aus der offensichtlichen Ungleichung $(2n)! \geq 2^n n!$.
(ii) Schreibe X als Konvexkombination von $\pm M$, wobei die Parameter variieren:

$$X = Y(-M) + (1 - Y)M, \text{ mit } Y := \frac{M - X}{2M} \in [\,0, 1\,].$$

Die Funktion $f : x \mapsto e^{\theta x}$ ist konvex, für jedes $\lambda \in [\,0, 1\,]$ ist also

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y).$$

In unserem Fall heißt das: punktweise gilt

$$e^{\theta X} \leq Y e^{-\theta M} + (1 - Y)e^{\theta M}.$$

Wir integrieren diese Ungleichung, setzen die konkrete Form von Y ein und nutzen aus, dass $\mathbb{E}(X) = 0$ gilt. So folgt

$$\mathbb{E}(e^{\theta X}) \leq \frac{M}{2M} e^{-\theta M} + \frac{M}{2M} e^{\theta M} = \cosh(\theta M).$$

Wegen (i) kann das mit $\exp\big((\theta M)^2/2\big)$ weiter abgeschätzt werden, und das beweist (wegen der Monotonie des Logarithmus) die Behauptung.
(iii) Das ist mit elementarer Analysis klar. □

Wir zeigen nun die *Hoeffdingungleichung*:

Satz 4.3.4. *$X_1, \dots, X_n$ seien beschränkte und unabhängige Zufallsvariable mit Erwartungswert 0, und zwar sei jeweils $|X_l| \le M_l$. Für $t > 0$ ist dann*
(i) $\mathbb{P}\big(\sum_l X_l \ge t\big) \le \exp\big(-t^2/(2\sum_l M_l^2)\big)$.
(ii) $\mathbb{P}\big(|\sum_l X_l| \ge t\big) \le 2\exp\big(-t^2/(2\sum_l M_l^2)\big)$.

Beweis: (i) Es ist

$$C_{X_1+\cdots+X_n} = C_{X_1} + \cdots + C_{X_n} \le \frac{\theta^2}{2}(M_1^2 + \cdots + M_n^2)$$

aufgrund der obigen Bemerkung 3 und Lemma 4.3.3 (ii). Die Aussage folgt nun aus der Cramérschen Ungleichung und 4.3.3 (iii) (mit $\alpha = \sum_l M_l^2$).
(ii) Man wende (i) für die $-X_l$ an, beachte, dass $\{|Y| \ge t\} = \{Y \ge t\} \cup \{-Y \ge t\}$ und erinnere sich an $\mathbb{P}(E \cup F) \le \mathbb{P}(E) + \mathbb{P}(F)$. □

Die Hoeffdingungleichung hat ein interessantes Korollar. Man stelle sich vor, dass eine Münze geworfen wird, die mit ± 1 beschriftet ist. Unabhängige Abfragen sollen durch Zufallsvariable $\varepsilon_1, \dots, \varepsilon_n$ modelliert werden (eine so genannte *Rademacherfolge*). Wir geben uns Zahlen $a_1, \dots, a_n$ vor und fragen, wie groß

$$a_1\varepsilon_1 + a_2\varepsilon_2 + \cdots + a_n\varepsilon_n$$

wohl sein wird. Hier ist die Antwort:

Korollar 4.3.5. *Für jedes $u > 0$ ist*

$$\mathbb{P}\left(|\sum_l a_l\varepsilon_l| \ge u\sqrt{\sum_l a_l^2}\right) \le 2e^{-u^2/2}.$$

Beweis: Wir wenden die Hoeffdingungleichung mit $X_l := a_l\varepsilon_l$ an. Es ist dann $M_l^2 = a_l^2$, und man muss nur noch $t = u\sqrt{\sum_l a_l^2}$ setzen. □

Überraschenderweise spielt hier also die l^2-Norm des Vektors (a_l) eine Rolle. Als Spezialfall betrachten wir $a_1 = \cdots = a_n = 1$. Dann erhalten wir das folgende Ergebnis:

> $|\varepsilon_1 + \cdots + \varepsilon_n|$ wird nicht wesentlich größer als $\sqrt{n}$. Genauer:
>
> $$\mathbb{P}\big(|\varepsilon_1 + \cdots + \varepsilon_n| \ge u\sqrt{n}\big) \le 2e^{-u^2/2}.$$
>
> So kommt es zum Beispiel nur mit einer Wahrscheinlichkeit von höchstens $e^{-3^2/2} \approx 0.11\ldots$ vor, dass $\varepsilon_1 + \cdots + \varepsilon_{10000}$ außerhalb von $[-300, 300]$ liegt.

Hier noch ein weiteres Korollar, durch das mögliche Abweichungen des Stichprobenmittels vom Erwartungswert kontrolliert werden können:

Korollar 4.3.6. *$X_1, \dots, X_n$ seien beschränkte und unabhängige Zufallsvariable mit Erwartungswert 0, und zwar sei jeweils $|X_l| \le M$ auf Ω. Für $\varepsilon > 0$ ist dann*
(i) $\mathbb{P}\big(\sum_l X_l/n \ge \varepsilon\big) \le \exp\big(-n\varepsilon^2/(2M^2)\big)$.
(ii) $\mathbb{P}\big(|\sum_l X_l|/n \ge \varepsilon\big) \le 2\exp\big(-n\varepsilon^2/(2M^2)\big)$.

Beweis: Man muss nur beachten, dass $\sum \cdots /n \ge \varepsilon$ gleichwertig zu $\sum \cdots \ge n\varepsilon$ ist und dann mit $t = n\varepsilon$ arbeiten. □

Bemerkung: Zum Vergleich stellen wir noch einmal *zwei Abschätzungen für n Abfragen* einer Zufallsvariablen X mit $|X| \le M$ und $\mathbb{E}(X) = 0$ gegenüber:

$$\text{Hoeffding:}\ \ \mathbb{P}\big(|\sum_l X_l|/n \ge \varepsilon\big) \le 2\exp\big(-n\varepsilon^2/(2M^2)\big).$$

$$\text{Tschebyscheff:}\ \ \mathbb{P}\big(|\sum_l X_l|/n \ge \varepsilon\big) \le \frac{M^2}{n\varepsilon^2}.$$

Für kleine n ist die Tschebyscheff-Ungleichung besser, für große die Hoeffding-Ungleichung. Hier ist ein Beispiel zum Vergleich für $M = 1$ und $\varepsilon = 0.1$:

n	Hoeffding	Tschebyscheff
100	1.21	1
300	0.446	0.333
500	0.164	0.2
1000	0.014	0.1
1500	0.001	0.066
2000	0.00009	0.05

(Natürlich sind Aussagen der Form $\mathbb{P}(A) \le 1$ völlig wertfrei.)

Das nächste Ziel ist – der Vollständigkeit halber – der Beweis der *Bernstein-Ungleichung*. Einige Vorbereitungen sammeln wir in

Lemma 4.3.7. *(i) X sei eine Zufallsvariable mit $\mathbb{E}X = 0$. Für geeignete Zahlen R, σ soll*

$$\mathbb{E}\big(|X|^n\big) \le n! R^{n-2}\sigma^2/2$$

für alle $n \in N$ gelten (wobei meist $\sigma^2 := \mathbb{E}X^2$). Für die $\theta > 0$ mit $R\theta < 1$ gilt dann

$$C_X(\theta) \le \frac{\theta^2\sigma^2}{2}\frac{1}{1-R\theta}.$$

(ii) Für $t, \alpha > 0$ gilt

$$\inf_{0<R\theta<1}\left(-\theta t + \frac{\theta^2\alpha^2}{2(1-R\theta)}\right) \le \frac{-t^2/2}{\alpha^2 + Rt}$$

Beweis: (i) Wir beginnen mit der Berechnung des Erwartungswerts von $e^{\theta X}$. Wir nutzen aus, dass $\mathbb{E}(X) = 0$ und dass (wegen $X^n \leq |X|^n$) $\mathbb{E}(X^n) \leq \mathbb{E}(|X|^n)$ gilt.

$$\begin{aligned}
\mathbb{E}e^{\theta X} &= 1 + \theta\mathbb{E}X + \frac{\theta^2\sigma^2}{2}\big(\sum_{n\geq 2}\frac{\theta^{n-2}}{n!(\sigma^2/2)}\mathbb{E}(X^n)\big) \\
&= 1 + \frac{\theta^2\sigma^2}{2}\big(\sum_{n\geq 2}\frac{\theta^{n-2}}{n!(\sigma^2/2)}\mathbb{E}(X^n)\big) \\
&\leq 1 + \frac{\theta^2\sigma^2}{2}\big(\sum_{n\geq 2}\frac{\theta^{n-2}}{n!(\sigma^2/2)}\mathbb{E}(|X|^n)\big) \\
&\leq 1 + \frac{\theta^2\sigma^2}{2}\sum_{n\geq 2}(R\theta)^{n-2} \\
&= 1 + \frac{\theta^2\sigma^2}{2}\frac{1}{1-R\theta} \\
&\leq \exp\big(\frac{\theta^2\sigma^2}{2}\frac{1}{1-R\theta}\big).
\end{aligned}$$

Durch Logarithmieren folgt die Behauptung.
(ii) Setze $\theta_0 := t/(\alpha^2 + Rt)$. Dann ist $R\theta_0 \in \,]\,0,1\,[$, also

$$\begin{aligned}
\inf_{0<R\theta<1}\frac{\theta^2\alpha^2}{2(1-R\theta)} &\leq \frac{\theta_0^2\alpha^2}{2(1-R\theta_0)} \\
&= \frac{t^2\alpha^2}{2(\alpha^2+Rt)^2}\frac{1}{1-\frac{Rt}{\alpha^2+Rt}} - \frac{t^2}{\alpha^2+Rt} \\
&= \frac{-t^2/2}{\alpha^2+Rt}.
\end{aligned}$$

□

Satz 4.3.8. *(Bernsteinungleichung) $X_1, \ldots, X_n$ seien unabhängige Zufallsvariable mit $\mathbb{E}(X_l) = 0$ und $\sigma_l := \mathbb{E}(X_l)^2$. Es gebe $R > 0$, so dass für alle m*

$$\mathbb{E}\big(|X_l|^m\big) \leq m!R^{m-2}\sigma_l^2/2$$

gilt. Wir setzen $\sigma^2 := \sigma_1^2 + \cdots + \sigma_n^2$. (Das ist die Varianz von $X_1 + \cdots + X_n$). Für $t > 0$ ist dann
(i)

$$\mathbb{P}\big(\sum_l X_l \geq t\big) \leq \exp\big(\frac{-t^2/2}{\sigma^2+Rt}\big).$$

(ii)

$$\mathbb{P}\big(|\sum_l X_l| \geq t\big) \leq 2\exp\big(\frac{-t^2/2}{\sigma^2+Rt}\big).$$

Beweis: (i) Durch Kombination von Bemerkung 3 zu Beginn dieses Abschnitts und Lemma 4.3.7 folgt

$$C_{\sum X_l}(\theta) \leq \frac{\theta^2\sigma^2}{2}\frac{1}{1-R\theta}.$$

Damit schließen wir aus der Cramérschen Ungleichung:

$$\begin{aligned}
\mathbb{P}\big(\sum_l X_l \geq t\big) &\leq \inf_{\theta>0} \exp\big(-\theta t + C_{\sum_l X_l}(\theta)\big) \\
&\leq \inf_{0<R\theta<1} \exp\big(-\theta t + C_{\sum_l X_l}(\theta)\big) \\
&\leq \inf_{0<R\theta<1} \exp\big(-\theta t + \frac{\theta^2\sigma^2}{2}\frac{1}{1-R\theta}\big) \\
&\leq \exp\big(-\frac{t^2/2}{\sigma^2 + Rt}\big)
\end{aligned}$$

(ii) Das folgt sofort aus (i), wenn man (i) auch noch für die $-X_l$ ausnutzt. □

4.4 Orakelungleichungen

Nach den vorstehenden Vorbereitungen können nun Ergebnisse bewiesen werden, die so etwas wie eine theoretische Rechtfertigung der hier beschriebenen konkreten Verfahren darstellen. Wir erinnern noch einmal an das Problem:

- Es gibt eine Menge $X \times Y$, die ein uns unbekanntes Wahrscheinlichkeitsmaß μ trägt. Die bei gegebenem x und variierendem y entstehenden „lokalen“ Erwartungswerte erzeugen die Funktion $x \mapsto f_\mu(x)$ (siehe Abschnitt 4.1). *Die* würden wir gern identifizieren.

- f_μ hat eine problemimmanente Ungenauigkeit σ_μ^2

- Wir dürfen eine Stichprobe $S = \{(x_i, y_i) \mid i = 1, \ldots, l\}$ vom Umfang l gemäß μ ziehen (unabhängig). Aufgrund von S soll eine Funktion $f_S = X \to \mathbb{R}$ erzeugt werden, und $\mathcal{R}_\mu(f_S)$ soll möglichst nahe bei σ_μ^2. liegen. Wegen
$$\mathcal{R}_\mu(f) = \int_X \big(f(x) - f_\mu(x)\big)^2 \, d\mu_X + \sigma_\mu^2.$$
heißt das, dass $\int_X \big(f_S(x) - f_\mu(x)\big)^2 \, d\mu_X$ möglichst klein sein soll.

Das wollen wir nun für ein spezielles Verfahren analysieren.

Das Lernverfahren

Vorgegeben sei ein RKHS H auf X. Er soll so reichhaltig sein, dass wir $f_\mu \in H$ annehmen dürfen: Wir suchen nur noch in H. Ist dann $S = \{(x_i, y_i) \mid i = 1, \ldots, l\}$ eine Stichprobe, so suche ein $f_S \in H$, so dass

$$\frac{1}{l} \sum_{i=1}^{l} |f(x_i) - y_i|^2$$

(der mittlere quadratische Fehler bei der Arbeit mit f) unter den $f \in H$ minimiert wird:

$$\frac{1}{l} \sum_{i=1}^{l} |f_S(x_i) - y_i|^2 = \min_{f \in H} \frac{1}{l} \sum_{i=1}^{l} |f(x_i) - y_i|^2.$$

Sofort stellen sich einige Fragen:

- Gibt es so ein f_S ? (Ja, unter geeigneten Kompaktheitsforderungen.)
- Ist f_S eindeutig bestimmt? (Nein.)
- Kann man das Fehlermaß $(x - y)^2$ durch eine andere Verlustfunktion ersetzen? (Ja, aber dann lässt sich f_S nur unter gewissen Konvexitätsbedingungen für L in zumutbarer Zeit bestimmen.)
- Ist f_S im vorliegenden Fall gut bestimmbar? (Ja, mit den Methoden von Kapitel 2, es ist ja ein quadratisches Minimierungsproblem.)

Die einfache Idee

Die Grundidee ist einfach. Betrache für ein $f \in H$ die Funktion f_Y, die durch $(x, y) \mapsto \big(f(x) - y\big)^2$ definiert ist. Dann ist $\mathcal{R}_\mu(f)$ der Erwartungswert von f_Y, und der Mittelwert einer zufällig gezogenen Stichprobe ist $\frac{1}{l} \sum_{i=1}^{l} |f(x_i) - y_i|^2$. Ist l groß genug, so sollte nach den Ergebnissen des vorigen Kapitels

$$\frac{1}{l} \sum_{i=1}^{l} |f(x_i) - y_i|^2 \approx \mathcal{R}_\mu(f)$$

gelten. Insbesondere wäre also auch

$$\frac{1}{l} \sum_{i=1}^{l} |f_S(x_i) - y_i|^2 \approx \mathcal{R}_\mu(f_S).$$

Andererseits ist, wegen $f_\mu \in H$,

$$\frac{1}{l} \sum_{i=1}^{l} |f_S(x_i) - y_i|^2 \leq \frac{1}{l} \sum_{i=1}^{l} |f_\mu(x_i) - y_i|^2,$$

und es folgt

$$\begin{aligned}\sigma_\mu^2 &\leq \mathcal{R}_\mu(f_S)\\ &\approx \sum_{i=1}^{l} |f_S(x_i) - y_i|^2\\ &\leq \sum_{i=1}^{l} |f_\mu(x_i) - y_i|^2\\ &\approx \sigma_\mu^2,\end{aligned}$$

und daraus kann man $\mathcal{R}_\mu(f_S) \approx \sigma_\mu^2$ schließen. Genauer: Wenn man wüsste, dass

$$\left| \frac{1}{l} \sum_{i=1}^{l} |f(x_i) - y_i|^2 - \mathcal{R}_\mu(f) \right| \leq \varepsilon$$

für $f = f_S$ und $f = f_\mu$ gilt (beachte hier, dass $\mathcal{R}_\mu(f_\mu) = \sigma_\mu^2$), so folgte

$$\begin{aligned}\sigma_\mu^2 &\leq \mathcal{R}_\mu(f_S)\\ &\leq \sum_{i=1}^{l} |f_S(x_i) - y_i|^2 + \varepsilon\\ &\leq \sum_{i=1}^{l} |f_\mu(x_i) - y_i|^2 + \varepsilon\\ &\leq \sigma_\mu^2 + 2\varepsilon,\end{aligned}$$

also $|\sigma_\mu^2 - \mathcal{R}_\mu(f_S)| \leq 2\varepsilon$.

Leider stimmt $|\frac{1}{l}\sum_{i=1}^{l} |f(x_i) - y_i|^2 - \mathcal{R}_\mu(f)| \leq \varepsilon$ für ein einzelnes f nur mit einer gewissen Wahrscheinlichkeit, und deswegen ist noch einige Arbeit zu leisten, um die Idee zu verwirklichen.

Wir behandeln zunächst in den nächsten drei Unterabschnitten Vorbereitungen:

- Mit welcher Wahrscheinlichkeit ist $|\frac{1}{l}\sum_{i=1}^{l} |f(x_i) - y_i|^2 - \mathcal{R}_\mu(f)| \leq \varepsilon$ zu erwarten?
- Wenn f_1 „nahe bei“ f_2 liegt, so ist $\mathcal{R}_\mu(f_1) \approx \mathcal{R}_\mu(f_2)$.
- Mit dem Begriff „Entropie“ kann man die „metrische Komplexität“ eines metrischen Raumes quantifizieren.

Und danach führen wir alles zusammen.

Wie nahe ist $\frac{1}{l}\sum_{i=1}^{l} |f(x_i) - y_i|^2$ *bei* $\mathcal{R}_\mu(f)$ *?*

Wir fixieren ein f und betrachten die auf dem Wahrscheinlichkeitsraum $X \times Y$ definierte Zufallsvariable $Z' := |f(x) - y|^2$. Wir ziehen den Erwartungswert $\mathcal{R}_\mu(f)$ ab und erhalten so $Z := Z' - \mathcal{R}_\mu(f)$, eine Zufallsvariable mit Erwartungswert Null. Wir wollen weiter annehmen, dass es eine Zahl M gibt, so dass $|Z'|$ durch M beschränkt ist; dann ist auch $|Z|$ durch M beschränkt.

Wenn wir nun l zufällige Abfragen (x_i, y_i) aus $X \times Y$ ziehen, so ist die Differenz $\frac{1}{l}\sum_{i=1}^{l} |f(x_i) - y_i|^2 - \mathcal{R}_\mu(f)$ ein Ausdruck der Form $\sum Z_i/l$, wobei die Z_i unabhängige Kopien von Z sind. Das Korollar 4.3.6 liefert dann

$$\mathbb{P}\big(\big|\frac{1}{l}\sum_{i=1}^{l} |f(x_i) - y_i|^2 - \mathcal{R}_\mu(f)\big| \geq \varepsilon\big) \leq 2\exp(-l\varepsilon^2/(2M)^2).$$

Bei vorgelegtem ε und bekannten M kann man also die Wahrscheinlichkeit für „Versager“ (zu großer Abstand) beliebig klein machen, wenn man l nur groß genug wählt.

Eine Abschätzung für den Fehler beim Risikoschätzen

Wir kürzen den Fehler, den man beim Ersetzen von $\mathcal{R}_\mu(f)$ durch den Stichprobenfehler $\sum(f(x_i) - y_i)^2/l$ macht, mit $L_S(f)$ ab:

$$L_S(f) := \mathcal{R}_\mu(f) - \frac{1}{l}\sum_{i=1}^{l}(f(x_i) - y_i)^2.$$

Lemma 4.4.1. *$f_1, f_2 : X \to \mathbb{R}$ seien beschränkt, und $|f_j(x) - y| \leq M$ für alle $(x, y) \in X \times Y$ und $j = 1, 2$. Dann gilt*

$$|L_S(f_1) - L_S(f_2)| \leq 4M||f_1 - f_2||_\infty.$$

Beweis: Wichtig für den Beweis ist die elementare Gleichung

$$(a - c)^2 - (b - c)^2 = (a - b)(a + b - 2c).$$

Zunächst folgt

$$\begin{aligned}
|\mathcal{R}_\mu(f_1) - \mathcal{R}_\mu(f_2)| &= |\int_{X\times Y} \big(f_1(x) - y\big)^2 - \big(f_2(x) - y\big)^2 \, d\mu| \\
&= |\int_{X\times Y} \big(f_1(x) - f_2(x)\big)\big(f_1(x) + f_2(x) - 2y\big) \, d\mu| \\
&\leq ||f_1 - f_2||_\infty |\int_{X\times Y} \big(f_1(x) - y + f_2(x) - y\big) \, d\mu| \\
&\leq ||f_1 - f_2||_\infty \big(\int_{X\times Y} |f_1(x) - y| \, d\mu + \int_{X\times Y} |f_1(x) - y| \, d\mu\big) \\
&\leq 2M \, ||f_1 - f_2||_\infty.
\end{aligned}$$

Ganz analog ergibt sich

$$|\frac{1}{l}\sum_{i=1}^{l}(f_1(x_i)-y_i)^2-\frac{1}{l}\sum_{i=1}^{l}(f_2(x_i)-y_i)^2|\leq 2M\,||f_1-f_2||_\infty,$$

und daraus folgt sofort die Behauptung. □

Überdeckungszahlen

Definition 4.4.2. *Sei D eine Teilmenge eines metrischen Raumes (M,d), und η sei positiv. Ein η-*Netz für *D ist dann eine Teilmenge Δ von M, so dass zu jedem $x\in D$ ein $y\in\Delta$ mit $d(x,y)\leq\eta$ existiert. Mit $\mathcal{N}(D,\eta)$ bezeichnen wir die minimale Anzahl eines η-Netzes; das ist die zu η gehörige Überdeckungszahl.*

Hier ein Beispiel. Sei $\eta>0$. Wie viele Punkte muss man sich in der Einheitskugel eines n-dimensionalen Raumes aussuchen, um ein η-Netz zu erhalten?

Lemma 4.4.3. *Sei X ein s-dimensionaler normierter Raum, die Einheitskugel werde mit B bezeichnet. Zu $\eta>0$ gibt es dann ein η-Netz in X für B mit höchstens $(1+2/\eta)^s$ Elementen.*

Beweis: Für die Kugeln $B(x,\alpha)$ (um x, mit Radius α) ist das euklidische Volumen das α^s-fache des Volumens V von B. Wir wählen eine maximale Teilmenge $x_1,\ldots,x_r$ von B mit der Eigenschaft, dass $||x_i-x_j||>\eta$ für $i\neq j$ ist. Dann gilt:

- $\{x_1,\ldots,x_r\}$ ist ein η-Netz für B. (Das ist klar.)
- Die r Kugeln $B(x_i,\eta/2)$ sind disjunkt und liegen in $B(0,1+\eta/2)$. Deswegen ist das Volumen von $\bigcup_i B(x_i,\eta/2)$ höchstens gleich dem Volumen von $B(0,1+\eta/2)$.

  Aufgrund der Vorbemerkung heißt das:

$$r\left(\frac{\eta}{2}\right)^s V\leq\left(1+\frac{\eta}{2}\right)^s V.$$

  Und daraus folgt sofort $r\leq(1+2/\eta)^s$. □

Anders ausgedrückt heißt das: $\mathcal{N}(B,\eta)\leq(1+2/\eta)^s$.

Finale

Nun kombinieren wir die Vorbereitungen. X,Y und μ sind wie bisher, und H sei ein RKHS auf X. Ein Zahl M sei so wählbar, dass alle auftretenden Funktonen durch M beschränkt sind. Für jede Trainingsmenge S soll f_S in H existieren. Weiter geben wir uns ein $\varepsilon>0$ und ein $\delta>0$ vor.

> Das Ziel: Es soll l, der Umfang der Trainingsmenge, so bestimmt werden, dass mit Wahrscheinlichkeit $1-\delta$ die Zahl $\mathcal{R}_\mu(f)$ um höchstens 2ε vom optimal zu erreichenden Wert σ_μ^2 entfernt liegt.

Wir haben schon begründet:

> Wenn $|\frac{1}{l}\sum_{i=1}^{l}|f(x_i)-y_i|^2-\mathcal{R}_\mu(f)|\le\varepsilon$ für $f=f_S$ und $f=f_\mu$ gilt, so ist $|\sigma_\mu^2-\mathcal{R}_\mu(f_S)|\le 2\varepsilon$.

Damit erreichen wir das Ziel wie folgt:

Schritt 1: Wir erinnern daran, dass

$$\mathbb{P}\big(|L_S(f)|\ge\varepsilon\big)=\mathbb{P}\big(\big|\frac{1}{l}\sum_{i=1}^{l}|f(x_i)-y_i|^2-\mathcal{R}_\mu(f)\big|\ge\varepsilon\big)\le 2\exp(-l\varepsilon^2/(2M)^2)$$

für jedes einzelne der hier zu betrachtenden f.

Schritt 2: Sei $n:=\mathcal{N}_{\varepsilon/(4M)}$ die zur Menge der hier relevanten f gehörige Überdeckungszahl (relativ zur Supremumsnorm) für den Wert $\eta=\varepsilon/(4M)$. Es gibt also $f_1,\dots,f_n$ mit der Eigenschaft: Für jedes hier relevanten f gibt es ein i mit der Eigenschaft $||f-f_i||_\infty\le\varepsilon/(4M)$. Wegen Lemma 4.4.1 ist dann $|L_S(f)-L_S(f_i)|\le\varepsilon$.

Wenn also für ein f die Ungleichung $|L_S(f)|\ge 2\varepsilon$ gilt, so muss es auch ein f_i mit $|L_S(f_i)|\ge\varepsilon$ geben. Kurz: Die Teilmenge

$$\{S\mid \text{ es existiert } f \text{ mit } |L_S|\ge 2\varepsilon\}$$

von $(X\times Y)^l$ liegt in

$$\bigcup_{i=1}^{n}\{S\mid |L_S(f_i)|\ge\varepsilon\}.$$

Folglich ist wegen Schritt 1 und $\mathbb{P}(\bigcup A_i)\le\sum\mathbb{P}(A_i)$:

$$\mathbb{P}\big(\sup_f|L_S(f)|)\big|\ge 2\varepsilon\big)\le 2\mathcal{N}_{\varepsilon/(4M)}\exp(-l\varepsilon^2/(2M)^2).$$

Das kann man so zusammenfassen:

Satz 4.4.4: *(Orakelungleichung) Mit Wahrscheinlichkeit*

$$1-2\mathcal{N}_{\varepsilon/(8M)}\exp(-l\varepsilon^2/(4M)^2)$$

kann man garantieren, dass unser Verfahren (wähle f_S zu S) zu einer zufällig vorgelegten Trainingsmenge S eine Funktion liefert, deren Risiko höchstens 2ε vom optimalen Wert entfernt ist.

Beachte, dass diese Wahrscheinlichkeit für große l beliebig nahe bei 1 liegt.

Beweis: Es ist nur in der vor wenigen Zeilen bewiesenen Ungleichung ε durch $\varepsilon/2$ zu ersetzen:

$$\mathbb{P}\big(\sup_f|L_S(f)|)\big|\ge\varepsilon\big)\le 2\mathcal{N}_{\varepsilon/(8M)}\exp(-l\varepsilon^2/(4M)^2).$$

Beachte noch

$$\mathbb{P}\big(\sup_f |L_S(f)|)\big| \geq \varepsilon\big) \leq 1 - \mathbb{P}\big(\sup_f |L_S(f)|)\big| \leq \varepsilon\big).$$

Mit der angegebenen Wahrscheinlichkeit sind also $|L_S(f_S)|, |L_S(f_\mu)| \leq \varepsilon$, und das impliziert, wie schon begründet, die Behauptung. □