Methoden der Statistik – Teil 2
Gliederung zur Vorlesung im Wintersemester 2019/20

Martin Wahl
Humboldt-Universität zu Berlin
martin.wahl@math.hu-berlin.de

24. Februar 2020

Inhaltsverzeichnis

Literatur

1) M. Trabs, M. Jirak, K. Krenz und M. Reiß. *Methoden der Statistik und des maschinellen Lernens: Eine mathematische Einführung.* Buchprojekt.

2) C. Giraud. *Introduction to high-dimensional statistics.* CRC Press, Boca Raton, FL, 2015.

3) S. Shalev-Shwartz and S. Ben-David: Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, 2014.

4) M. Wainwright. *High-dimensional statistics. A non-asymptotic viewpoint.* Cambridge University Press, Cambridge, 2019

5) T. Hastie, R. Tibshirani, J. Friedman. *The elements of statistical learning. Data mining, inference, and prediction.* Springer, New York, 2009.

6) G. James, D. Witten, T. Hastie, R. Tibshirani. *An introduction to statistical learning. With applications in R.* Springer, New York, 2013.

7) I. Steinwart und A. Christmann. *Support vector machines.* Springer, New York, 2008.

8) J. Shawe-Taylor und N. Cristianini. *Kernel Methods for Pattern Analysis.* Cambridge University Press, USA, 2004.

1 Klassifizierung

Häufig müssen in der Statistik Entscheidungen zwischen zwei oder mehr Alternativen aufgrund komplexer Daten getroffen werden. Dies führt auf sogenannte Klassifikationsprobleme. Beispiele sind Spam-Filter (Klassifikationen zwischen ‘mit Sicherheit Spam’, ‘mit Sicherheit kein Spam’ und Klassen dazwischen), medizinische Datenanalyse (wie EKG weist auf Krankheit hin oder nicht) oder Schrifterkennung (ASCII-Code auf der Basis von Pixelmuster). Hier werden wir nur binäre Klassifikation mit Klassen (Labels) 0 und 1 betrachten. Verallgemeinerungen mit mehr als 2 Klassen werden teilweise in den Übungen besprochen.

1.1 Bayes-Klassifizierer

Sei (X, Y) ein Paar von Zufallsvariablen wobei X Werte in $(\mathcal{X}, \mathscr{F})$ und Y Werte in $\{0, 1\}$ annimmt. X ist der gegebene Featurevektor, Y das gewünschte Label. Sei $\eta : \mathcal{X} \to [0, 1]$ gegeben durch

$$\eta(x) = \mathbb{P}(Y = 1 | X = x).$$

Wir gehen kurz auf die Definition und Bedeutung von η ein. Ist $\mathcal{X}$ diskret, so kann man $\mathbb{P}(Y = 1|X = x) = \mathbb{P}(Y = 1, X = x)/\mathbb{P}(X = x)$ setzen. Für allgemeines X definieren wir $\mathbb{P}(Y = 1|X = x) = \mathbb{E}(Y|X = x)$ als bedingte Erwartung von Y gegeben $X = x$, wobei wir ausnutzen, dass Y nur die Werte 0 und 1 annimmt und somit $Y = \mathbb{1}_{\{Y=1\}}$ gilt. Wir bemerken außerdem, dass die Verteilung von (X, Y) durch η und der Verteilung P^X von X eindeutig bestimmt ist: für eine integrierbare Funktion g gilt

$$\mathbb{E}g(X, Y) = \int_{\mathcal{X}} \left(g(x, 1)\eta(x) + g(x, 0)(1 - \eta(x))\right) dP^X(x).$$

Man kann (X, Y) also wie folgt erzeugen: erst $X \sim P^X$, dann $Y \sim \mathrm{Ber}(\eta(X))$.

1.1 Definition. Eine messbare Funktion $h : \mathcal{X} \to \{0, 1\}$ heißt Klassifizierer. Für einen Klassifizierer h heißt

$$R(h) = \mathbb{P}(Y \neq h(X)) = \int \mathbb{1}_{\{y \neq h(x)\}} \, dP^{X,Y}(x, y)$$

Klassifizierungsfehler von h. Dabei bezeichnet $P^{X,Y}$ die Verteilung von (X, Y). Der Klassifizierer h^* definiert durch

$$h^*(x) = \begin{cases} 1, & \text{falls } \eta(x) > 1/2 \\ 0, & \text{falls } \eta(x) \leq 1/2 \end{cases}$$

heißt Bayes- oder MAP-Klassifizierer.

1.2 Satz. *Es gilt*

$$R(h^*) = \min_h R(h) = \mathbb{E} \min(\eta(X), 1 - \eta(X)),$$

wobei das Minimum über alle Klassifizierer genommen wird.

Beweis. Zunächst gilt

$$R(h) = \int_{\mathcal{X}} \mathbb{P}(Y \neq h(x)|X = x) \, dP^X(x).$$

(Ist $\mathcal{X}$ diskret, so folgt dies gerade aus der Formel von der totalen Wahrscheinlichkeit, ansonsten verwende $R(h) = \mathbb{E}\mathbb{1}_{\{Y \neq h(X)\}} = \mathbb{E}\mathbb{E}(\mathbb{1}_{\{Y \neq h(X)\}}|X)$).

Weiter gilt (für das a-posteriori Risiko)

$$\begin{aligned}
&\mathbb{P}(Y \neq h(x)|X = x) \\
&= \mathbb{P}(Y \neq h(x), Y = 1|X = x) + \mathbb{P}(Y \neq h(x), Y = 0|X = x) \\
&= \mathbb{P}(0 = h(x), Y = 1|X = x) + \mathbb{P}(1 = h(x), Y = 0|X = x) \\
&= \mathbb{1}_{\{h(x)=0\}} \mathbb{P}(Y = 1|X = x) + \mathbb{1}_{\{h(x)=1\}} \mathbb{P}(Y = 0|X = x) \\
&= \mathbb{1}_{\{h(x)=0\}} \eta(x) + \mathbb{1}_{\{h(x)=1\}} (1 - \eta(x)). \qquad (1.1)
\end{aligned}$$

Dieser Ausdruck wird minimal für

$$h(x) = \begin{cases} 1, & \text{falls } \eta(x) > 1 - \eta(x) \\ 0, & \text{falls } \eta(x) \leq 1 - \eta(x) \end{cases}.$$

Dies ist gerade der Bayes-Klassifizierer und die Behauptung folgt. □

Man nennt

$$R^* = R(h^*) = \min_h R(h)$$

auch das Bayes-Risiko. Der Bayes-Klassifizierer besitzt also minimales Risiko. Die Verteilung von (X, Y) ist unbekannt, daher kann der Bayes-Klassifizierer in der Realität nicht verwendet werden. Stattdessen beobachten wir Trainingsdaten

$$(X_1, Y_1), \ldots, (X_n, Y_n) \qquad \text{mit} \qquad (X_1, Y_1), \ldots, (X_n, Y_n), (X, Y) \text{ i.i.d.}$$

Unser Ziel ist es einen Klassifizierer

$$\hat{h}_n = \hat{h}_n(X_1, Y_1, \ldots, X_n, Y_n) : \mathcal{X} \to \{0, 1\}$$

zu konstruieren, so dass der sogenannte Generalisierungsfehler von $\hat{h}_n$

$$R(\hat{h}_n)$$

möglichst nah an R^* ist. Beachte dabei, dass $R(\hat{h}_n)$ eine Zufallsvariable ist, der Erwartungswert wird nur bezüglich (X, Y) genommen (betrachte R als Funktion definiert auf der Menge aller Klassifizierer):

$$R(\hat{h}_n) = \int \mathbb{1}_{\{y \neq \hat{h}_n(x)\}} \, dP^{X,Y}(x, y).$$

Dabei interpretieren wir $R(\hat{h}_n)$ als den mittleren Fehler den wir machen wenn wir $\hat{h}_n$ zum Klassifizieren einer neuen Beobachtung (X, Y) verwenden.

1.2 k-NN-Klassifizierer

In diesem Kapitel sei $\mathcal{X} = \mathbb{R}^p$ versehen mit dem Standardskalarprodukt $\langle x, y\rangle = \sum_{j=1}^p x_j y_j$ und zugehöriger Euklidischer Norm $\|x\| = \sqrt{\langle x, x\rangle}$, $x, y \in \mathbb{R}^p$. Für $x \in \mathbb{R}^p$ sei

$$X_{(1)}(x), \ldots, X_{(n)}(x)$$

eine Umordnung von $X_1, \ldots, X_n$ mit der Eigenschaft, dass

$$\|X_{(j)}(x) - x\| \leq \|X_{(j+1)}(x) - x\|, \qquad \forall j = 1, \ldots, n-1.$$

Dabei heißt $X_{(j)}(x)$ j-nächste Nachbar von x. Des Weiteren sei

$$N_k(x) = \{X_{(1)}(x), \ldots, X_{(k)}(x)\}.$$

1.3 Methode. Der k-NN-Klassifizierer $\hat{h}_n^{k\text{-NN}} : \mathbb{R}^p \to \{0, 1\}$ ist definiert durch

$$\hat{h}_n^{k\text{-NN}}(x) = \begin{cases} 1, & \sum\limits_{i: X_i \in N_k(x)} \mathbb{1}_{\{Y_i = 1\}} > \sum\limits_{i: X_i \in N_k(x)} \mathbb{1}_{\{Y_i = 0\}} \\ 0, & \text{sonst} \end{cases}.$$

Während obige Definition auf einer Mehrheitsregel basiert, kann der k-NN-Klassifizierer auch als Plug-in-Methode verstanden werden:

$$\hat{h}_n^{k\text{-NN}}(x) = \begin{cases} 1, & \hat{\eta}^{k\text{-NN}}(x) > 1/2 \\ 0, & \text{sonst} \end{cases}$$

mit Schätzer $\hat{\eta}$ von η definiert durch

$$\hat{\eta}^{k\text{-NN}}(x) = \frac{1}{k} \sum_{i: X_i \in N_k(x)} Y_i = \sum_{i=1}^n w_i(x) Y_i, \qquad w_i(x) = \begin{cases} 1/k, & X_i \in N_k(x) \\ 0, & \text{sonst} \end{cases}.$$

Zum Beispiel gilt $h_n^{1\text{-NN}}(x) = Y_i$ mit $X_i = X_{(1)}(x)$. Der 1-NN-Klassifizierer ist also eine interpolierende Klassifikationsmethode.

1.4 Satz. *Die Zufallsvariable X nehme Werte in $[-M, M]^p$ an und η sei L-Lipschitz-stetig (d.h. $|\eta(x) - \eta(x')| \leq L\|x - x'\|$ für alle $x, x' \in \mathbb{R}^p$). Dann gilt*

$$\mathbb{E}R(h_n^{1\text{-}NN}) \leq 2R^* + Cn^{-\frac{1}{p+1}}$$

mit einer Konstanten $C = C(L, M, p) > 0$.

1.5 Bemerkungen.

1) Für $n \to \infty$ konvergiert der erwartete Generalisierungsfehler gegen das doppelte Bayesrisiko. Der 1-NN-Klassifizierer ist also nicht konsistent. Man kann zeigen, dass es für großes p sogar interpolierende Klassifikationsmethoden die annährend konsistent sind (siehe zum Beispiel [1]). Solche Resultate sind von Interesse, da viele aktuelle Methoden des maschinellen Lernens stark überfitted werden und trotzdem noch gute Generalisierungseigenschaften besitzen.

2) Der Term $n^{-1/(p+1)}$ kann mit etwas mehr Aufwand durch $n^{-1/p}$ ersetzt werden (vgl. Aufgabe 1.7 unten).

3) Fluch der Dimension: Für $n^{-1/p} \leq \epsilon$ benötigen wir $n \geq \epsilon^{-p}$ Datenpunkte, für kleineres n sind die Punkte $X_1, \dots, X_n$ in der Regel zu isoliert. In hoher Dimensions wird daher oft erst eine Dimensionsreduktion durchgeführt.

Beweis von Satz 1.4. Es gilt (vergleiche (1.1))

$$R(h_n^{\text{1-NN}}) = \int_{\mathbb{R}^p} \left(\mathbb{1}_{\{h_n^{\text{1-NN}}(x)=0\}} \eta(x) + \mathbb{1}_{\{h_n^{\text{1-NN}}(x)=1\}} (1-\eta(x)) \right) dP^X(x).$$

Des Weiteren gilt

$$\mathbb{P}(h_n^{\text{1-NN}}(x) = 1 | X_1, \dots, X_n) = \eta(X_{(1)}(x))$$

und somit

$$\mathbb{P}(h_n^{\text{1-NN}}(x) = 1) = \mathbb{E}\eta(X_{(1)}(X)).$$

Wenden wir außerdem die Lipschitz-Stetigkeit von η an, so folgt

$$\begin{aligned}
\mathbb{E}R(h_n^{\text{1-NN}}) &= \mathbb{E}\Big[(1-\eta(X_{(1)}(X)))\eta(X) + \eta(X_{(1)}(X))(1-\eta(X))\Big] \\
&= \mathbb{E}\Big[((\eta(X)-\eta(X_{(1)}(X)))(2\eta(X)-1) + 2\eta(X)(1-\eta(X))\Big] \\
&\leq L\mathbb{E}[\|X - X_{(1)}(X)\|] + 2\mathbb{E}[\eta(X)(1-\eta(X))].
\end{aligned}$$

Da

$$\mathbb{E}[2\eta(X)(1-\eta(X))] \leq \mathbb{E}[\min(\eta(X), 1-\eta(X))] = R^*,$$

folgt die Behauptung aus folgendem Lemma:

1.6 Lemma. *Unter den Voraussetzungen aus Satz 1.4 gilt*

$$\mathbb{E}[\|X - X_{(1)}(X)\|] \leq C n^{-\frac{1}{p+1}}$$

mit einer Konstanten $C = C(M,p) > 0$.

Wir müssen noch Lemma 1.6 zeigen. Sei hierfür o.B.d.A. $M = 1$, so dass $X \in [-1,1]^p$. Für $N \in \mathbb{N}$ sei $Q_1, \dots, Q_{(2N)^p}$ eine Überdeckung von $[-1,1]^p$ aus Quadern mit Duchmesser $\sqrt{p}N^{-1}$. Expliziter betrachten wir also die Quader $\prod_{j=1}^p [(a_j - 1)/N, a_j/N]$ mit $a_j \in \{-N+1, \dots, N\}$, $j = 1, \dots, N$. Es gilt

$$\begin{aligned}
&\mathbb{E}[\|X - X_{(1)}(X)\| \mathbb{1}_{\{X \in Q_j\}}] \\
&\leq \mathbb{E}[\|X - X_{(1)}(X)\| \mathbb{1}_{\{X \in Q_j\}} \mathbb{1}_{\{\exists i : X_i \in Q_j\}}] \\
&+ \mathbb{E}[\|X - X_{(1)}(X)\| \mathbb{1}_{\{X \in Q_j\}} \mathbb{1}_{\{\forall i : X_i \notin Q_j\}}] \\
&\leq \sqrt{p} N^{-1} \mathbb{P}(X \in Q_j) + 2\sqrt{p} \mathbb{P}(X \in Q_j)(1 - \mathbb{P}(X \in Q_j))^n.
\end{aligned}$$

Setzen wir $q_j = \mathbb{P}(X \in Q_j)$, so folgt

$$\begin{aligned}
\mathbb{E}[\|X - X_{(1)}(X)\|] &\leq \sum_{j=1}^{(2N)^p} \mathbb{E}[\|X - X_{(1)}(X)\| \mathbb{1}_{\{X \in Q_j\}}] \\
&\leq \sqrt{p} N^{-1} + 2\sqrt{p} \sum_{j=1}^{(2N)^p} q_j (1 - q_j)^n \\
&\leq \sqrt{p} N^{-1} + 2\sqrt{p} (2N)^p \max_{q \in [0,1]} q(1-q)^n.
\end{aligned}$$

Setzen wir

$$q(1-q)^n \leq q e^{-nq} = \frac{1}{n} nq e^{-nq} \leq \frac{1}{ne}$$

(verwende hierfür $1 + x \leq e^x$, $x \in \mathbb{R}$, und $x e^{-x} \leq e^{-1}$, $x \geq 0$) ein, so folgt

$$\mathbb{E}[\|X - X_{(1)}(X)\|] \leq 2\sqrt{p}\Big(N^{-1} + \frac{2^p N^p}{ne}\Big) \leq \sqrt{p} n^{-\frac{1}{p+1}} + 2^{2p+1} \sqrt{p} e^{-1} n^{\frac{p}{p+1} - 1},$$

wobei wir $n^{1/(p+1)} \leq N < n^{1/(p+1)} + 1 \leq 2n^{1/(p+1)}$ gewählt haben. □

1.7 Aufgabe. *Es gelten die Voraussetzungen aus Satz 1.4 mit $p \geq 2$. Dann gilt sogar*

$$\mathbb{E}[\|X - X_{(1)}(X)\|] \leq C n^{-\frac{1}{p}}$$

mit einer Konstanten $C = C(M, p) > 0$.

Beweis. Wir geben hier nur die Hauptschritte im Beweis und überlassen es dem Leser die Details auszuarbeiten. Als erstes zeigt man, dass $\mathbb{P}(\|X - X_{(1)}(X)\| > \delta) = \mathbb{E}(1 - \mu(S_{X,\delta}))^n$ für alle $\delta > 0$ gilt, wobei μ die Verteilung von X ist und $S_{x,\delta} = \{y \in \mathbb{R}^p : \|x - y\| \leq \delta\}$. Als nächstes zeigt man mit Hilfe eines ähnlichen Überdeckungsargumentes wie im Beweis von Lemmas 1.6, dass $\mathbb{E}(1 - \mu(S_{X,\delta}))^n \leq C\delta^{-p}/n$. Die Behauptung folgt dann aus diesen beiden Schritte in Kombination mit der Formel $\mathbb{E}Z = \int_0^\infty \mathbb{P}(Z \geq z)\, dz \leq a + \int_a^\infty \mathbb{P}(Z \geq z)\, dz$, gültig für alle Zufallsvariablen $Z \geq 0$ und reelle Zahlen $a \geq 0$. □

1.8 Satz. *Es gelten die Voraussetzungen aus Satz 1.4 mit $k \leq n/2$ und $p \geq 2$. Dann gilt*

$$\mathbb{E}R(h_n^{k\text{-}NN}) - R^* \leq \frac{2}{\sqrt{k}} + C\Big(\frac{k}{n}\Big)^{\frac{1}{p}}$$

mit einer Konstanten $C = C(L, M, p) > 0$.

Der Beweis verwendet die Interpretation des k-NN-Klassifizierer als Plug-in-Methode und basiert auf folgender Aufgabe:

1.9 Aufgabe. *Sei $h : \mathcal{X} \to \{0, 1\}$ ein Klassifizierer. Zeige:*

(a) Es gilt

$$R(h) = \mathbb{E}\big[\eta(X) + \mathbb{1}_{\{h(X)=1\}}(1 - 2\eta(X))\big].$$

(b) Ist h^* der Bayes-Klassifizierer, so gilt

$$R(h) - R(h^*) = 2\mathbb{E}\big[|\eta(X) - 1/2|\mathbb{1}_{\{h(X)\neq h^*(X)\}}\big].$$

(c) Für eine (feste) Funktion $\hat{\eta} : \mathcal{X} \to [0, 1]$ betrachte nun den Plug-in-Klassifizierer $h(x) = \mathbb{1}_{\{\hat{\eta}(x)>1/2\}}$. Dann gilt

$$R(h) - R(h^*) \leq 2\mathbb{E}\big[|\eta(X) - \hat{\eta}(X)|\big].$$

1.10 Bemerkungen.

1) Für $n \to \infty$ und $k \to \infty$ mit $k/n \to 0$ konvergiert der erwartete Generalisierungsfehler gegen das Bayesrisiko. In diesem Fall ist der k-NN-Klassifizierer also konsistent.

2) Die Wahl von k hat starken Einfluss auf die Güte. Ein k von der Größenordnung $n^{-2/(2+p)}$ liefert eine Schranke von der Größenordnung $n^{-1/(2+p)}$. In der Vorlesung Nichtparametrische Statistik wird gezeigt, dass $n^{-1/(2+p)}$ die optimale Konvergenzrate für das Schätzen von η unter Lipschitz-Bedingungen ist.

3) Im Fall $k = 1$ liefert Satz 1.8 ein schwächeres Resultat als Satz 1.4. Dies liegt daran, dass der Beweis von Satz 1.8 auf Aufgabe 1.9 beruht, das Schätzen von $\mathbb{1}_{\{\eta(x)>1/2\}}$ im Allgemeinen jedoch einfacher ist als das Schätzen von η.

(a) Für Beispiele siehe Figures 2.13, 2.15, 2.16 in [5].

Beweis von Satz 1.8. Mit $w_i(x) = (1/k)\mathbb{1}_{\{X_i \in N_k(x)\}}$ gilt

$$\sum_{i=1}^{n} w_i(x) = 1 \qquad \text{und} \qquad \hat{\eta}^{k\text{-NN}}(x) = \sum_{i=1}^{n} w_i(x) Y_i.$$

Verwenden wir außerdem Aufgabe 1.9, so gilt

$$\begin{aligned}
&\mathbb{E}R(\hat{h}_n^{k\text{-NN}}) - R^* \\
&\leq 2\mathbb{E}|\hat{\eta}^{k\text{-NN}}(X) - \eta(X)| \\
&= 2\mathbb{E}\Big|\sum_{i=1}^n w_i(X)(Y_i - \eta(X))\Big| \\
&\leq 2\mathbb{E}\Big|\sum_{i=1}^n w_i(X)(Y_i - \eta(X_i))\Big| + 2\mathbb{E}\Big|\sum_{i=1}^n w_i(X)(\eta(X_i) - \eta(X))\Big| =: 2(S+D).
\end{aligned}$$

Der Term S ist ein stochastischer Fehler, der Term D ein Approximationsfehler. Es gilt nun

$$\begin{aligned}
S^2 &\leq \mathbb{E}\sum_{i,j=1}^n w_i(X)w_j(X)(Y_i - \eta(X_i))(Y_j - \eta(X_j)) \\
&= \mathbb{E}\sum_{i}^n w_i^2(X)(Y_i - \eta(X_i))^2,
\end{aligned}$$

wobei wir in der ersten Ungleichung die Cauchy-Schwarz-Ungleichung, und in der zweiten Ungleichung die Unabhängigkeit von $(X_1, Y_1), \ldots, (X_n, Y_n), (X, Y)$ und die Identitäten $\mathbb{E}(Y_i - \eta(X_i)) = \mathbb{E}(Y_i - \mathbb{E}(Y_i|X)) = 0$ verwendet haben. Benutzen wir außerdem, dass die nicht-negativen Gewichte $w_i(X)$ beschränkt sind durch $1/k$ und sich aufsummieren zu 1, so folgt

$$S^2 \leq \mathbb{E}\max_i w_i(X)\sum_{i=1}^n w_i(X) \leq 1/k.$$

Für den Approximationsfehler gilt

$$D \leq \mathbb{E}\sum_{i=1}^n w_i(X)|\eta(X_i) - \eta(X)| \leq \mathbb{E}\frac{L}{k}\sum_{j=1}^k \|X_{(j)}(X) - X\|.$$

Diese Ausdrücke können mit einem Trick auf 1-nächste Nachbarn zurückgeführt werden. Zerlege hierfür die Daten $S = \{X_1, \ldots, X_n\}$ in $\leq k+1$ Segmente

$$S_j = \Big\{X_i : i = (j-1)\Big\lfloor\frac{n}{k}\Big\rfloor + 1, \ldots, j\Big\lfloor\frac{n}{k}\Big\rfloor\Big\}, \quad 1 \leq j \leq k$$

und sofern n/k keine natürliche Zahl ist zusätzlich noch

$$S_{k+1} = \Big\{X_i : i = k\Big\lfloor\frac{n}{k}\Big\rfloor + 1, \ldots, n\Big\}.$$

Definieren wir nun $X_{(1,j)}(X)$ als 1-nächste Nachbar von X in S_j, $j \leq k$, so gilt

$$\frac{1}{k}\sum_{j=1}^{k}\|X_{(j)}(X)-X\| \leq \frac{1}{k}\sum_{j=1}^{k}\|X_{(1,j)}(X)-X\|$$

und somit

$$\begin{aligned}\mathbb{E}\frac{1}{k}\sum_{j=1}^{k}\|X_{(j)}(X)-X\| &\leq \mathbb{E}\frac{1}{k}\sum_{j=1}^{k}\|X_{(1,j)}(X)-X\| \\ &= \mathbb{E}\|X_{(1,1)}(X)-X\| \leq C\Big\lfloor\frac{n}{k}\Big\rfloor^{-\frac{1}{p}} \leq C2^{p+1}\Big(\frac{k}{n}\Big)^{\frac{1}{p}},\end{aligned}$$

wobei wir in der vorletzten Ungleichung Aufgabe 1.7 angewendet haben (mit $\lfloor n/k \rfloor$ anstatt n Beobachtungen) und in der letzten Ungleichung $\lfloor n/k \rfloor \geq n/k - 1 \geq n/(2k)$ verwendet haben. □

Obwohl das k-NN Verfahren einen einfachen Ansatz verfolgt, liefert es oft sehr gute Ergebnisse. Des Weiteren bemerken wir noch, dass das k-NN Verfahren nicht dem ERM-Prinzip folgt, sondern (zumindest für $k \gg 1$) auf einem (nichtparametrischen) Plug-in-Prinzip beruht.

1.3 ERM-Prinzip

1.11 Definition. Für einen Klassifizierer h heißt

$$R_n(h) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{Y_i \neq h(X_i)\}}$$

empirischer Klassifizierungsfehler von h. Ist $\mathcal{H}$ eine Menge von Klassifizierern, so heißt ein Klassifizierer $\hat{h}_n$ ERM-Klassifizierer, falls

$$\hat{h}_n \in \operatorname*{argmin}_{h\in\mathcal{H}} R_n(h).$$

1.12 Bemerkung. Ein Minimierer existiert immer, da $R_n(h)$ nur die Werte $0, 1/n, \ldots, 1$ annehmen kann (ist i.A. jedoch nicht eindeutig).

Die Wahl von $\mathcal{H}$ spielt eine entscheidende Rolle. Das Analogon zur Bias-Varianz-Zerlegung ist:

$$R(\hat{h}_n) - R^* = \underbrace{R(\hat{h}_n) - \inf_{h\in\mathcal{H}} R(h)}_{(1)} + \underbrace{\inf_{h\in\mathcal{H}} R(h) - R^*}_{(2)}.$$

Dabei ist (2) ein (deterministischer) Bias-Term. Er wird kleiner je größer wir $\mathcal{H}$ wählen. Dieser Term wird entweder ignoriert oder mit Hilfe von Modellannahmen in den Griff bekommen. Andererseits ist (1) ein stochastischer Fehler. Er misst den Fehler, den wir machen wenn wir beim Minimieren R durch R_n ersetzen.

1.13 Lemma. *Sei $\mathcal{H}$ eine Menge von Klassifizierern und $\hat{h}_n$ ein zugehöriger ERM-Klassifizierer. Dann gilt*

$$\mathbb{E}R(\hat{h}_n) - \inf_{h\in\mathcal{H}} R(h) \le \mathbb{E}\sup_{h\in\mathcal{H}}\{R_n(h) - R(h)\}.$$

Beweis. Für ein $\epsilon > 0$ sei $h_\epsilon \in \mathcal{H}$ mit $R(h_\epsilon) \le \inf_{h\in\mathcal{H}} R(h) + \epsilon$. Dann gilt

$$\begin{aligned} R(\hat{h}_n) - \inf_{h\in\mathcal{H}} R(h) &\le R(\hat{h}_n) - R(h_\epsilon) + \epsilon \\ &= R(\hat{h}_n) - R_n(\hat{h}_n) + R_n(\hat{h}_n) - R(h_\epsilon) + \epsilon \\ &\le R(\hat{h}_n) - R_n(\hat{h}_n) + R_n(h_\epsilon) - R(h_\epsilon) + \epsilon \\ &\le \sup_{h\in\mathcal{H}}\{R(h) - R_n(h)\} + R_n(h_\epsilon) - R(h_\epsilon) + \epsilon, \end{aligned}$$

wobei wir in der zweiten Ungleichung die Definition von $\hat{h}_n$ verwendet haben. Nehmen wir den Erwartungswert, so folgt

$$\mathbb{E}R(\hat{h}_n) - \inf_{h\in\mathcal{H}} R(h) \le \mathbb{E}\sup_{h\in\mathcal{H}}\{R(h) - R_n(h)\} + \epsilon.$$

Da $\epsilon > 0$ beliebig war, folgt die Behauptung. □

$\mathcal{H}$ endlich

1.14 Aufgabe. *Sei $\mathcal{H} = \{h_1, \dots, h_M\}$ eine endliche Menge von Klassifizierern und $\hat{h}_n$ ein zugehörige ERM-Klassifizierer. Dann gilt*

$$\mathbb{E}R(\hat{h}_n) \le \min_{j\le M} R(h_j) + 2\sqrt{\frac{2\log M}{n}}.$$

1.15 Aufgabe. *Sei $\mathcal{H} = \{h_1, \dots, h_M\}$ eine endliche Menge von Klassifizierern mit der Eigenschaft, dass $R(h_j) = 0$ für ein $j \le M$. Des Weiteren sei $\hat{h}_n$ ein zugehöriger ERM-Klassifizierer. Beweise schrittweise, dass*

$$\mathbb{E}R(\hat{h}_n) \le \frac{1 + \log M}{n}.$$

(a) Es gilt $R_n(\hat{h}_n) = 0$ fast sicher.

(b) Es folgt $\mathbb{P}(R(\hat{h}_n) > \epsilon) \le \sum_{k:R(h_k)>\epsilon} \mathbb{P}(R_n(h_k) = 0) \le M(1-\epsilon)^n$.

(c) Die Behauptung folgt aus (b) und $\mathbb{E}R(\hat{h}_n) = \int_0^\infty \mathbb{P}(R(\hat{h}_n) > \epsilon)\, d\epsilon$.

Man spricht von einer schnellen Rate (fast rate).

Affine Klassifikation

Ab jetzt nehmen wir an, dass $(X_1, Y_1), \ldots, (X_1, Y_n), (X, Y)$ Werte in $\mathbb{R}^p \times \{\pm 1\}$ annehmen, d.h. wir ersetzen die Labels $0, 1$ durch $-1, +1$.

Wir beginnen damit einige Fakten aus der Vektorgeometrie aufzulisten. Für $x, y \in \mathbb{R}^p$ seien das Standardskalarprodukt und die zugehörige Euklidische Norm definiert durch $\langle x, y\rangle = \sum_{j=1}^p x_j y_j$ und $\|x\| = \sqrt{\langle x, x\rangle}$. Es gilt:

1) Die Menge $H = \{x : \langle w, x\rangle + b = 0\}$ ist eine Hyperebene in $\mathbb{R}^p$.

2) Sind $x_1, x_2 \in H$, so gilt $\langle w, x_1 - x_2\rangle = 0$, d.h. $w, x_1 - x_2$ sind orthogonal. Der Vektor $w^* = w/\|w\|$ heißt auch Normalenvektor von H.

3) Ist $x \in \mathbb{R}^p$, so ist der Abstand zwischen x und H gegeben durch

$$\frac{1}{\|w\|}|\langle w, x\rangle + b)|, \qquad x_0 \in H. \tag{1.2}$$

4) Ist $x \in \mathbb{R}^p$, so ist $\|w\|^{-1}(\langle w, x\rangle + b) = \langle w^*, x - x_0\rangle$ der Abstand mit Vorzeichen.

Wir betrachten im Folgenden also alle Klassifizierer $h : \mathbb{R}^p \to \{\pm 1\}$ der Form

$$h(x) = \operatorname{sign}(\langle w, x\rangle + b) := \begin{cases} +1, & \text{falls } \langle w, x\rangle + b > 0, \\ -1, & \text{falls } \langle w, x\rangle + b \le 0, \end{cases}$$

mit $w \in \mathbb{R}^p$ und $b \in \mathbb{R}$.

1.16 Aufgabe. *Sei $\mathcal{H} = \{h(x) = \operatorname{sign}(\langle w, x\rangle + b) : w \in \mathbb{R}^p, b \in \mathbb{R}\}$ die Menge aller affinen Klassifizierer und $\hat{h}_n$ ein zugehörige ERM-Klassifizierer. Dann gilt*

$$\mathbb{E}R(\hat{h}_n) \le \min_{h \in \mathcal{H}} R(h) + 2\sqrt{\frac{2(p+1)\log(n+1)}{n}}.$$

Allerdings ist die ERM-Methode im affinen Fall für großes p nicht mehr implementierbar. Als Illustration betrachte den Fall $p \ge n$ und $X_1, \ldots, X_n$ linear unabhängig. Dann lässt sich jede Teilmenge von $\{X_1, \ldots, X_n\}$ und deren Komplement durch eine Hyperebene separieren. Bei der Berechnung des ERM-Klasifizierers müssen also im schlimmsten Falle alle 2^n Möglichkeiten überprüft werden, was schon im Fall $n \approx 40 - 80$ zu Berechenbarkeitsproblemen führt (NP-schweres Problem). Eine Ausnahme bildet dabei der Fall, dass das Trainingssample durch eine Hyperebene trennbar ist, d.h. es existiert $(w, b) \in \mathbb{R}^p \times \mathbb{R}$, so dass $Y_i(\langle w, X_i\rangle + b) > 0$ für alle $i = 1, \ldots, n$.

1.17 Aufgabe. *Betrachte Rosenblatts Perzeptron-Algorithmus:*

input: *A training set* $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^p \times \{\pm\}$

initialize: $w^{(1)} = (0, \ldots, 0)$

for $t = 1, 2, \ldots$

if $(\exists i \text{ s.t. } y_i \langle w^{(t)}, x_i \rangle \leq 0)$ *then* $w^{(t+1)} = w^{(t)} + y_i x_i$

else output $w^{(t)}$

Wir nehmen an, dass es ein $w \in \mathbb{R}^p$ *gibt mit* $y_i \langle w, x_i \rangle > 0$ *für alle* $i = 1, \ldots, n$. *Sei* $M = \max_{i \leq n} \|x_i\|$, $B = \min\{\|w\| : \forall i, y_i \langle w, x_i \rangle \geq 1\}$ *und* $w^* \in \mathbb{R}^p$ *ein Vektor der das Minimum in der Definition von* B *annimmt. Zeige:*

(a) Läuft der Perzeptron-Algorithmus für T *Schritte, so gilt*

$$\|w^{(T+1)}\|^2 \leq TM^2 \quad \textit{und} \quad \langle w^*, w^{(T+1)} \rangle \geq T.$$

(b) Schließen Sie mit Hilfe der Cauchy-Schwarz-Ungleichung, dass

$$T \leq B^2 M^2.$$

Der Perzeptron-Algorithmus stoppt also nach höchstens B^2M^2 *Schritten und liefert einen linearen Klassifizierer* $h(x) = \text{sign}(\langle w, x \rangle)$ *welcher alle Beobachtungen aus der Trainingsmenge richtig klassifiziert.*

1.4 Support vector machines (SVM)

Motivation

Sind $(x_1, y_1), \ldots, (x_n, y_n)$ durch eine Hyperebene separierbar, d.h. existieren $w \in \mathbb{R}^p$ und $b \in \mathbb{R}$ mit $y_i(\langle w, x_i \rangle + b) > 0$ für alle $i = 1, \ldots, n$, so stellt sich die Frage welche separierende Hyperebene man wählen soll. Eine Idee besteht darin den Abstand (margin) zwischen der Hyperebene und den Daten zu maximieren. Dies führt zu einem Optimierungsproblem mit Nebenbedingungen:

$$\begin{array}{lll} \text{maximiere} & \min\limits_{i \leq n} |\langle w, x_i \rangle + b| & \text{über } w \in \mathbb{R}^p, \|w\| = 1, b \in \mathbb{R} \\ \text{unter der NB} & y_i(\langle w, x_i \rangle + b) > 0 & i = 1, \ldots, n. \end{array} \tag{1.3}$$

Eine äquivalente Formulierung von (1.3) ist wie folgt:

$$\begin{array}{lll} \text{minimiere} & \|w\| & \text{über } w \in \mathbb{R}^p, b \in \mathbb{R} \\ \text{unter der NB} & y_i(\langle w, x_i \rangle + b) \geq 1 & i = 1, \ldots, n. \end{array} \tag{1.4}$$

Ist $(\hat{w}, \hat{b})$ eine Lösung von (1.4) und H die Hyperebene definiert durch $H = \{x : \langle \hat{w}, x \rangle + \hat{b} = 0\}$, so ist der Abstand zwischen H und den Datenpunkten gerade gegeben durch (vergleiche (1.2))

$$\min_{i=1,\ldots,n} \frac{1}{\|\hat{w}\|} |\langle \hat{w}, x_i \rangle + \hat{b}| = \frac{1}{\|\hat{w}\|}.$$

Es wird also wieder gerade der Abstand maximiert unter der Nebenbedingung, dass alle Datenpunkte richtig klassifiziert werden. Daher folgt, dass $(\hat{w}/|\hat{w}|, \hat{b}/|\hat{w}|)$ eine Lösung von (1.3) ist. Die äquivalenten Methoden (1.3) (1.4) werden oft auch Hard-SVM genannt.

Nun können Klassen in der Regel nicht durch eine Hyperebene separiert werden. Wollen wir diese Bedingung verhindern, so müssen wir die Nebenbedinungen in (1.4) abschwächen. Eine Idee ist es nur $y_i(\langle w, x_i\rangle + b) \geq 1 - \xi_i$ zu fordern für sogenannte Schlupfvariablen (slack variables) $\xi_i \geq 0$. Allerdings müssen wir diese zusätzlich bestrafen, sonst sind die Nebenbedingungen trivialerweise erfüllt. Für $\lambda > 0$ betrachten wir also das Optimierungsproblem

$$\begin{array}{lll} \text{minimiere} & \lambda\|w\|^2 + \frac{1}{n}\sum_{i=1}^n \xi_i & \text{über } w \in \mathbb{R}^p, b \in \mathbb{R}, \xi \in \mathbb{R}^n \\ \text{unter der NB} & y_i(\langle w, x_i\rangle + b) \geq 1 - \xi_i & i = 1, \dots, n \\ & \xi_i \geq 0 & i = 1, \dots, n. \end{array} \tag{1.5}$$

Ist $(\hat{w}, \hat{b}, \hat{\xi})$ eine Lösung von (1.6), so gilt $\hat{\xi}_i \geq \max(1 - y_i(\langle \hat{w}, x_i\rangle + \hat{b}), 0)$. Es gilt sogar Gleichheit, da sonst eine Widersprch zum Minimum entsteht. (1.5), ist also äquivalent zu

1.18 Methode. Für $\lambda > 0$ betrachte

$$(\hat{w}, \hat{b}) \in \operatorname*{argmin}_{w \in \mathbb{R}^p, b \in \mathbb{R}} \frac{1}{n}\sum_{i=1}^n \max(1 - y_i(\langle w, x_i\rangle + b), 0) + \lambda\|w\|^2. \tag{1.6}$$

Dann heißt der Klassifizierer

$$\hat{h}_n^{\mathrm{SVM}}(x) = \operatorname{sign}(\langle w, x_i\rangle + b)$$

Stützvektor-Klassifizierer (SVM-Klassifizierer).

Die äquivalenten Methoden (1.5) und (1.6) werden oft auch Soft-SVM genannt. Die folgende Aufgabe lässt sich mit elementaren Methoden, oder aber leichter mit den Rechenregeln des Subdifferentials aus Appendix A.2 lösen.

1.19 Aufgabe. *Zeige, dass die Lösung $\hat{w}$ des Optimierungsproblems* (1.6) *von der Form $\hat{w} = \sum_{i=1}^n \hat{\alpha}_i y_i x_i$ ist, wobei*

$$\begin{cases} \hat{\alpha}_i = 0, & \textit{falls } y_i(\langle \hat{w}, x_i\rangle + b) > 1, \\ \hat{\alpha}_i = 1/(2\lambda n), & \textit{falls } y_i(\langle \hat{w}, x_i\rangle + b) < 1, \\ \hat{\alpha}_i \in [0, 1/(2\lambda n)], & \textit{falls } y_i(\langle \hat{w}, x_i\rangle + b) = 1. \end{cases}$$

Vektoren x_i mit $\hat{\alpha}_i \neq 0$ heißen auch Stützvektoren (support vectors).

Der Generalisierungsfehler des SVM-Klassifizierers

Seien nun wieder $(X_1, Y_1), \dots, (X_n, Y_n), (X, Y)$ i.i.d. mit Werten in $\mathbb{R}^p \times \{\pm 1\}$. Wir beginnen mit einer weiteren Interpretation des SVM-Klassifizierers als konvexe Relaxation des affinen ERM-Klassifizierers. Für $\mathcal{H} = \{h(x) = \operatorname{sign}(\langle w, x\rangle) : w \in \mathbb{R}^p, \|w\| = 1\}$ ist der zugehörige ERM-Klassifizierer definiert durch

$$\hat{h}_n^{\mathrm{ERM}} \in \operatorname*{argmin}_{h\in\mathcal{H}} R_n(h) = \operatorname*{argmin}_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{-Y_i h(X_i)>0\}},$$

wobei wir in der Gleichheit verwendet haben, dass $Y_i h(X_i)$ nur die Werte ± 1 annimmt. Der Einfachheit halber betrachten wir hier nur affine Klassifizierer und setzen $b = 0$. Der allgemeine kann auf den linearen Fall zurückgeführt werden indem man $X_i \mapsto (X_i, 1) \in \mathbb{R}^{p+1}$ betrachtet und $\langle w, x\rangle + b = \langle (w, b), (x, 1)\rangle$ schreibt. In den letzten Kapiteln haben wir gesehen, dass der ERM-Klassifizierer gute statistische Eigenschaften besitzt. Allerdings kann er in der Praxis oft nicht verwendet werden, da er nur schwer zu berechnen ist (sowohl $\mathcal{H}$ als auch die Indikatorfunktion sind nicht konvex). Ähnlich wie im Fall der Modellwahl (Lasso-Schätzer) ist es unser Ziel eine konvexe Relaxation zu finden.
1. Schritt. Wir ersetzen $\mathbb{1}_{\{-z>0\}}$ durch das hinge loss $(1-z)_+ = \max(1-z, 0)$ für alle $z \in \mathbb{R}$.
2. Schritt. Wir ersetzen $\mathcal{H}$ durch $\{h(x) = \langle w, x\rangle : \|w\| \le \lambda'\}$. Ein h kann zum Klassifizeren verwendet werden indem wir $\operatorname{sign}(h)$ betrachten.

Das resultierende Minimierungsproblem kann immer noch als ERM-Problem interpretiert werden indem wir

$$R^{\mathrm{hinge}}(w) = \mathbb{E}(1 - Y\langle w, X\rangle)_+ \quad \text{und} \quad R_n^{\mathrm{hinge}}(w) = \frac{1}{n}\sum_{i=1}^{n}(1 - Y_i\langle w, X_i\rangle)_+$$

setzen.

1.20 Methode. Für $\lambda' > 0$ betrachte

$$\hat{w}_n^{\mathrm{SVM}} \in \operatorname*{argmin}_{w\in\mathbb{R}^p:\|w\|\le\lambda'} R_n^{\mathrm{hinge}}(w). \tag{1.7}$$

Dann heißt der Klassifizierer

$$\hat{h}_n^{\mathrm{SVM}}(x) = \operatorname{sign}(\langle w_n^{\mathrm{SVM}}, x\rangle)$$

Stützvektor-Klassifizierer (SVM-Klassifizierer).

1.21 Satz. *Es gelte $X \le M$ fast sicher. Dann gilt für den SVM-Klassifizierer aus* (1.7)*, dass*

$$\mathbb{E}R(\hat{h}_n^{\mathrm{SVM}}) \le \min_{\|w\|\le\lambda'} R^{hinge}(w) + \frac{2\lambda' M}{\sqrt{n}}.$$

Insbesondere, falls $\mathbb{P}(Y\langle w^*, X\rangle \geq 1) = 1$ *für ein Vektor* $w^* \in \mathbb{R}^p$*, so folgt mit der Wahl* $\lambda' = \|w^*\|$*, dass*

$$\mathbb{E}R(\hat{h}_n^{\mathrm{SVM}}) \leq \frac{2\|w^*\|M}{\sqrt{n}}.$$

1.22 Bemerkungen.

1) Die Rate $n^{-1/2}$ heißt auch langsame Rate. Unter der Zusatzannahme $\mathbb{P}(Y\langle w^*, X\rangle \geq 1) = 1$ für ein Vektor $w^* \in \mathbb{R}^p$ kann man sogar eine sogenannte schnelle n^{-1}-Rate beweisen, allerdings mit schwierigerem Beweis mittels lokalen Rademacher-Komplexitäten.

2) Die zweite Schranke in Satz 3.28 hängt nur von M und $\|w^*\|$ ab. Sie ist insbesondere Dimensionsunabhängig (führt zu kernel SVM).

3) In der Praxis muss λ' adaptiv gewählt werden.

Das folgende Lemma ist eine Verallgemeinerung von Lemma 1.13.

1.23 Lemma. *Der SVM-Klassifizierer aus Methode 1.20 erfüllt*

$$\mathbb{E}R(\hat{h}_n^{\mathrm{SVM}}) \leq \min_{\|w\|\leq\lambda'} R^{hinge}(w) + \mathbb{E}\sup_{\|w\|\leq\lambda'}\{R^{hinge}(w) - R_n^{hinge}(w)\}.$$

Beweis. Zunächst gilt mit $\hat{h}_n^{\mathrm{SVM}}(x) = \mathrm{sign}(\langle \hat{w}_n^{\mathrm{SVM}}, x\rangle)$,

$$\begin{aligned} R(\hat{h}_n^{\mathrm{SVM}}) &= \int \mathbb{1}_{\{-yh_n^{\mathrm{SVM}}(x)>0\}}\, dP^{X,Y}(x,y) \\ &\leq \int \mathbb{1}_{\{-y\langle w_n^{\mathrm{SVM}}, x\rangle \geq 0\}}\, dP^{X,Y}(x,y) \\ &\leq \int (1 - y\langle w_n^{\mathrm{SVM}}, x\rangle)_+\, dP^{X,Y}(x,y) = R^{\mathrm{hinge}}(\hat{w}_n^{\mathrm{SVM}}). \end{aligned}$$

Es folgt, dass

$$\mathbb{E}R(\hat{h}_n^{\mathrm{SVM}}) \leq \min_{\|w\|\leq\lambda'} R^{\mathrm{hinge}}(w) + \mathbb{E}R^{\mathrm{hinge}}(\hat{w}_n^{\mathrm{SVM}}) - \min_{\|w\|\leq\lambda'} R^{\mathrm{hinge}}(w). \quad (1.8)$$

Sei nun $w' \in \mathbb{R}^p$ mit $\|w'\| \leq \lambda'$ und

$$R^{\mathrm{hinge}}(w') = \min_{\|w\|\leq\lambda'} R^{\mathrm{hinge}}(w).$$

Dann gilt

$$\begin{aligned} &R^{\mathrm{hinge}}(\hat{w}_n^{\mathrm{SVM}}) - R^{\mathrm{hinge}}(w') \\ &= R^{\mathrm{hinge}}(\hat{w}_n^{\mathrm{SVM}}) - R_n^{\mathrm{hinge}}(\hat{w}_n^{\mathrm{SVM}}) + R_n^{\mathrm{hinge}}(\hat{w}_n^{\mathrm{SVM}}) - R^{\mathrm{hinge}}(w') \\ &\leq R^{\mathrm{hinge}}(\hat{w}_n^{\mathrm{SVM}}) - R_n^{\mathrm{hinge}}(\hat{w}_n^{\mathrm{SVM}}) + R_n^{\mathrm{hinge}}(w') - R^{\mathrm{hinge}}(w') \\ &\leq \sup_{\|w\|\leq\lambda'}\{R^{\mathrm{hinge}}(w) - R_n^{\mathrm{hinge}}(w)\} + R_n^{\mathrm{hinge}}(w') - R^{\mathrm{hinge}}(w'), \end{aligned}$$

wobei wir in der ersten Ungleichung die Definition von $\hat{w}_n^{\mathrm{SVM}}$ verwendet haben. Nehmen wir den Erwartungswert, so folgt

$$\mathbb{E}R^{\mathrm{hinge}}(\hat{w}_n^{\mathrm{SVM}}) - \min_{\|w\|\leq\lambda'} R^{\mathrm{hinge}}(w) \leq \mathbb{E} \sup_{\|w\|\leq\lambda'} \{R^{\mathrm{hinge}}(w) - R_n^{\mathrm{hinge}}(w)\}.$$

Einsetze in (1.8) liefert die Behauptung. □

Proof of Satz 3.28. Verwenden wir Lemma 1.23, so bleibt zu zeigen, dass

$$\mathbb{E} \sup_{\|w\|\leq\lambda'} \{R^{\mathrm{hinge}}(w) - R_n^{\mathrm{hinge}}(w)\} \leq \frac{2\lambda' M}{\sqrt{n}}.$$

Mit Hilfe des Symmetrisierungstrick aus Lemma A.1 erhalten wir

$$\begin{aligned}
&\mathbb{E} \sup_{\|w\|\leq\lambda'} \{R^{\mathrm{hinge}}(w) - R_n^{\mathrm{hinge}}(w)\} \\
&= \mathbb{E} \sup_{\|w\|\leq\lambda'} \Big\{\frac{1}{n}\sum_{i=1}^n -(1 - Y_i\langle w, X_i\rangle)_+ + \mathbb{E}(1 - Y_i\langle w, X_i\rangle)_+\Big\} \\
&\leq 2\,\mathbb{E} \sup_{\|w\|\leq\lambda'} \frac{1}{n}\sum_{i=1}^n \epsilon_i(1 - Y_i\langle w, X_i\rangle)_+ \\
&\leq 2 \max_{y\in\{0,1\}^n} \sup_{x\in B_M(0)^n} \mathbf{E} \sup_{\|w\|\leq\lambda'} \frac{1}{n}\sum_{i=1}^n \epsilon_i(1 - y_i\langle w, x_i\rangle)_+,
\end{aligned}$$

mit $\epsilon_1, \ldots, \epsilon_n$ i.i.d. Rademacher Zufallsvariablen und $B_M(0) = \{x \in \mathbb{R}^p : \|x\| \leq M\}$. Die Funktion $(1 - z)_+$ ist 1-Lipschitz stetig. Das Konstraktionsprinzip aus Lemma A.3 liefert nun

$$\begin{aligned}
&2\mathbf{E}_\epsilon \sup_{\|w\|\leq\lambda'} \frac{1}{n}\sum_{i=1}^n \epsilon_i(1 - y_i\langle w, x_i\rangle)_+ \\
&\leq 2\mathbf{E}_\epsilon \sup_{\|w\|\leq\lambda'} \frac{1}{n}\sum_{i=1}^n -\epsilon_i y_i\langle w, x_i\rangle \\
&= 2\mathbf{E}_\epsilon \sup_{\|w\|\leq\lambda'} \frac{1}{n}\sum_{i=1}^n \epsilon_i\langle w, x_i\rangle,
\end{aligned}$$

wobei wir in der letzten Gleichheit verwendet haben, dass die $-\epsilon_i y_i$ die gleiche (gemeinsame) Verteilung haben wie die ϵ_i. Es gilt nun für $\|w\| \leq \lambda'$ und $x \in (B_M(0))^n$, dass

$$\frac{1}{n}\sum_{i=1}^n \epsilon_i\langle w, x_i\rangle = \langle w, \frac{1}{n}\sum_{i=1}^n \epsilon_i x_i\rangle \leq \|w\|\Big\|\frac{1}{n}\sum_{i=1}^n \epsilon_i x_i\Big\| \leq \lambda'\Big\|\frac{1}{n}\sum_{i=1}^n \epsilon_i x_i\Big\|$$

wobei wir die Cauchy-Schwarz-Ungleichung angewendet haben. Es folgt für $x \in B_M(0)^n$, dass

$$\begin{aligned}\mathbf{E}_\epsilon \sup_{\|w\|\le\lambda'} \frac{1}{n}\sum_{i=1}^n \epsilon_i\langle w, x_i\rangle &\le 2\lambda'\mathbb{E}\Big\|\frac{1}{n}\sum_{i=1}^n \epsilon_i x_i\Big\| \\ &\le \lambda'\Big(\mathbb{E}\Big\|\frac{1}{n}\sum_{i=1}^n \epsilon_i x_i\Big\|^2\Big)^{1/2} \\ &= \lambda'\Big(\mathbb{E}\Big(\frac{1}{n^2}\sum_{i,j=1}^n \epsilon_i\epsilon_j\langle x_i, x_j\rangle\Big)^{1/2} \\ &\le \lambda'\Big(\mathbb{E}\Big(\frac{1}{n^2}\sum_{i=1}^n \|x_i\|^2\Big)^{1/2} \le \frac{\lambda' M}{\sqrt{n}}.\end{aligned}$$

Einsetzen liefert die Behauptung. □

1.5 LDA und logistische Regression

1.24 Aufgabe. *Sei (X, Y) ein Paar von Zufallsvariablen mit Werten in $\mathbb{R}^p \times \{0, 1\}$ und Verteilung*

$$\mathbb{P}(Y = k) = \pi_k \quad \text{und} \quad \mathbb{P}(X \in dx | Y = k) = f_k(x)\,dx, \quad k \in \{0, 1\}, x \in \mathbb{R}^p,$$

wobei $\pi_0 + \pi_1 = 1$ und f_0, f_1 zwei Dichten in $\mathbb{R}^p$ sind. Zeige:

(a) Der Bayes-Klassifizierer h^* ist gegeben durch $h^*(x) = \mathbb{1}_{\{\pi_1 f_1(x) > \pi_0 f_0(x)\}}$ und das zugehörige Bayes-Risiko erfüllt

$$R^* = \int_{\mathbb{R}^p} \min(\pi_0 f_0(x), \pi_1 f_1(x))\,dx.$$

Seien f_0 und f_1 nun gegeben durch die Normalverteilungsdichte

$$f_k(x) = (2\pi)^{-p/2}\sqrt{\det(\Sigma_k^{-1})}\exp\Big(-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\Big), \quad k = 0, 1,$$

mit invertierbaren Kovarianzmatrizen $\Sigma_k \in \mathbb{R}^{p\times p}$ und Mittelwerten $\mu_k \in \mathbb{R}^p$.

(b) Gilt $\Sigma_0 = \Sigma_1 = \Sigma$ und $\mu_0 \neq \mu_1$, so ist die Bedingung $\pi_1 f_1(x) > \pi_0 f_0(x)$ äquivalent zu

$$(\mu_1 - \mu_0)^T\Sigma^{-1}\Big(x - \frac{\mu_0 + \mu_1}{2}\Big) > \log(\pi_0/\pi_1).$$

1.25 Methode. Seien $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^p \times \{0, 1\}$ Daten. Für $k = 0, 1$ setze $n_k = |\{i : Y_i = k\}|$ sowie

$$\hat\pi_k = \frac{n_k}{n}, \quad \hat\mu_k = \frac{1}{n_k}\sum_{i:Y_i=k} X_i, \quad \hat\Sigma = \frac{1}{n-2}\sum_{k=0,1}\sum_{i:Y_i=k}(x-\hat\mu_k)(x-\hat\mu_k)^T$$

Dann heißt

$$\hat{h}_n^{\mathrm{LDA}}(x) = \begin{cases} 1 & (\mu_1 - \mu_0)^T \Sigma^{-1}\Big(x - \frac{\mu_0+\mu_1}{2}\Big) > \log(\pi_0/\pi_1) \\ 0 & \text{sonst} \end{cases}$$

LDA-Klassifizierer.

Im Modell der logistischen Regression nimmt man an, dass

$$\log \frac{\mathbb{P}(Y=1|X=x)}{\mathbb{P}(Y=0|X=x)} = \alpha + \beta^T x, \qquad \alpha \in \mathbb{R}, \beta \in \mathbb{R}^p. \tag{1.9}$$

Dies ist äquivalent zu

$$\mathbb{P}(Y=1|X=x) = \frac{e^{\alpha+\beta^T x}}{1+e^{\alpha+\beta^T x}}, \quad \mathbb{P}(Y=0|X=x) = \frac{1}{1+e^{\alpha+\beta^T x}}.$$

1.26 Methode. Seien $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^p \times \{0,1\}$ Daten. Für

$$(\hat{\alpha}, \hat{\beta}) \in \operatorname*{argmax}_{\alpha\in\mathbb{R}, \beta\in\mathbb{R}^p} \prod_{i=1}^n \frac{e^{Y_i(\alpha+\beta^T X_i)}}{1+e^{\alpha+\beta^T X_i}}$$

heißt

$$\hat{h}_n^{\mathrm{LR}}(x) = \begin{cases} 1 & \frac{e^{\hat{\alpha}+\hat{\beta}^T x}}{1+e^{\hat{\alpha}+\hat{\beta}^T x}} > 1/2 \\ 0 & \text{sonst} \end{cases} = \begin{cases} 1 & \hat{\alpha}+\hat{\beta}^T x > 0 \\ 0 & \text{sonst} \end{cases}$$

Logistische-Regression-Klassifizierer.

Sowohl unter dem Modell der logistischen Regression als auch unter dem Modell der LDA aus Aufgabe 1.24(b) gilt (1.9). Beide Klassifizierer sind zudem affin. Der Unterschied liegt darin wie α unf β geschätzt werden, LDA verfolgt einen parametrischen Ansatz, logistischen Regression einen semiparamterischen Ansatz.

2 Hauptkomponentenanalyse (PCA)

2.1 Motivation: Die erste Hauptkomponente

Seien $X_1, \ldots, X_n \in \mathbb{R}^p$ Daten. Wir nehmen an, dass diese in einem ersten Schritt schon zentriert wurden, d.h. es gilt $\sum_{i=1}^n X_i = 0$. Um die Daten möglichst gut durch eine Gerade durch 0 zu beschreiben, betrachten wir das Optimierungsproblem

$$\text{minimiere} \quad \frac{1}{n}\sum_{i=1}^n \|X_i - \langle v, X_i\rangle v\|^2 \qquad \text{über } v \in \mathbb{R}^p, \|v\| = 1. \tag{2.1}$$

Es gilt

$$\frac{1}{n}\sum_{i=1}^{n}\|X_i - \langle v, X_i\rangle v\|^2 = \frac{1}{n}\sum_{i=1}^{n}(\|X_i\|^2 - \langle v, X_i\rangle^2)$$
$$= \frac{1}{n}\sum_{i=1}^{n}(\|X_i\|^2 - v^T X_i X_i^T v_i).$$

Mit empirischer Kovarianzmatrix $\hat{\Sigma} = n^{-1}\sum_{i=1}^{n} X_i X_i^T$ ist (2.1) also äquivalent zu

$$\text{maximiere} \quad v^T\hat{\Sigma}v = \langle\hat{\Sigma}v, v\rangle \qquad \text{über } v \in \mathbb{R}^p, \|v\| = 1. \tag{2.2}$$

Wir suchen also gleichzeitig die Richtung mit maximaler empirischer Varianz. Sei nun im Folgenden $\hat{u}_1$ eine Lösung von (2.1) bzw. (2.2).

2.1 Definition. Der Vektor $(\langle X_1, \hat{u}_1\rangle, \ldots, \langle X_n, \hat{u}_1\rangle)^T \in \mathbb{R}^n$ heißt erste Hauptkomponente (PC).

Möglicherweise werden die Daten nicht gut genug durch die erste PC repräsentiert. Wir betrachten daher im Folgenden Approximationen durch d-dimensionale lineare Unterräume V von $\mathbb{R}^p$.

Bevor wir das tun, wollen wir noch sehen, dass $\hat{u}_1$ eine Eigenvektor von $\hat{\Sigma}$ ist mit Eigenwert $\hat{\lambda}_1$, d.h. es gilt

$$\hat{\Sigma}\hat{u}_1 = \hat{\lambda}_1\hat{u}_1 \quad \text{mit} \quad \hat{\lambda}_1 = \langle\hat{\Sigma}\hat{u}_1, \hat{u}_1\rangle. \tag{2.3}$$

Verwenden wir die Definitionen von $\hat{u}_1$ und $\hat{\lambda}_1$, so erhalten wir

$$\langle\hat{\Sigma}(\hat{u}_1 + \alpha h), \hat{u}_1 + \alpha h\rangle \leq \lambda_1\langle\hat{u}_1 + \alpha h, \hat{u}_1 + \alpha h\rangle \qquad \forall\alpha \in \mathbb{R}, h \in \mathbb{R}^p$$

und somit durch Umformungen wegen $\hat{\Sigma} = \hat{\Sigma}^T$,

$$\alpha\langle\hat{\Sigma}\hat{u}_1 - \hat{\lambda}_1\hat{u}_1, h\rangle + \alpha^2\langle\hat{\Sigma}h - \hat{\lambda}_1 h, h\rangle \leq 0 \qquad \forall\alpha \in \mathbb{R}, h \in \mathbb{R}^p.$$

Setzen wir $\alpha = \epsilon\langle\hat{\Sigma}\hat{u}_1 - \hat{\lambda}_1\hat{u}_1, h\rangle$ mit $\epsilon > 0$ beliebig klein, so folgt

$$\langle\hat{\Sigma}\hat{u}_1 - \hat{\lambda}_1\hat{u}_1, h\rangle = 0 \qquad \forall h \in \mathbb{R}^p,$$

und wir erhalten (2.3). Mathematisch haben wir es bei PCA also mit der Diagonalisierung der empirischen Kovarianzmatrix zu tun.

2.2 Elementare Eigenschaften von PCA

Notationen und Definition von PCA

Für $x, y \in \mathbb{R}^p$ sei

$$\langle x, y\rangle = \sum_{j=1}^{p} x_j y_j, \qquad \|x\| = \sqrt{\langle x, x\rangle}.$$

Sei $\mathbb{R}^{p\times p}$ die Menge aller $p \times p$-Matrizen. Schreibweise $A = (a_{jk})$. Für $A, B \in \mathbb{R}^{p\times p}$ sind die Spur, das Hilbert-Schmidt-Skalarprodukt sowie die Frobenius- bzw. Hilbert-Schmidt-Norm definiert durch

$$\operatorname{tr}(A) = \sum_{j=1}^{p} a_{jj},$$

$$\langle A, B\rangle_{\mathrm{HS}} = \operatorname{tr}(A^T B) = \sum_{j=1}^{p}\sum_{k=1}^{p} a_{jk}b_{jk}, \qquad \|A\|_{\mathrm{HS}} = \sqrt{\langle A, A\rangle_{\mathrm{HS}}}.$$

$A \in \mathbb{R}^{p\times p}$ heißt positiv semi-definit (Schreibweise $A \geq 0$), falls $\langle Ax, x\rangle \geq 0$ für alle $x \in \mathbb{R}^p$. A heißt symmetrisch, falls $A = A^T$.

Ist V ein linearer Unterraum von $\mathbb{R}^p$ (Schreibweise $V \leq \mathbb{R}^p$), so bezeichnet $P_V \in \mathbb{R}^{p\times p}$ die Orthogonalprojektion auf V. Ist $v_1, \ldots, v_d$ eine Orthonormalbasis (ONB) von V, so gilt

$$P_V = \sum_{j=1}^{d} v_j v_j^T = (v_1 \cdots v_d)(v_1 \cdots v_d)^T.$$

Allgemeiner erfüllt P_V die Eigenschaften $P_V^2 = P_V$, $P_V^T = P_V$, $\operatorname{Bild} P_V = V$ und ist zudem durch diese Eigenschaften eindeutig bestimmt (Übung 11.2).

2.2 Definition. Für $V \leq \mathbb{R}^p$ heißt

$$R_n(V) = \frac{1}{n}\sum_{i=1}^{n} \|X_i - P_V X_i\|^2$$

empirischer Rekonstruktionsfehler von V.

2.3 Methode. Für $d \leq p$ berechnet PCA

$$\hat{V}_d \in \operatorname*{argmin}_{\dim V = d} R_n(V). \tag{2.4}$$

Charakterisierung von $\hat{V}_d$

Für $V \leq \mathbb{R}^p$ gilt

$$\begin{aligned}
R_n(V) &= \frac{1}{n}\sum_{i=1}^{n} \|X_i - P_V X_i\|^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} \|X_i\|^2 - \langle X_i, P_V X_i\rangle - \langle P_V X_i, X_i\rangle + \|P_V X_i\|^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} \|X_i\|^2 - \|P_V X_i\|^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i^T X_i - X_i^T P_V X_i \\
&= \frac{1}{n}\sum_{i=1}^{n} \operatorname{tr}(X_i X_i^T - X_i X_i^T P_V) = \operatorname{tr}(\hat{\Sigma}(I - P_V))
\end{aligned}$$

mit empirischer Kovarianzmatrix

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T.$$

Somit ist (2.4) äquivalent zu

$$\text{maximiere} \quad \operatorname{tr}(\hat{\Sigma} P_V) \qquad \text{über } V \leq \mathbb{R}^p, \dim V = d \tag{2.5}$$

oder konkreter äquivalent zu

$$\text{maximiere} \quad \sum_{j=1}^{d} \langle \hat{\Sigma} v_j, v_j \rangle \qquad \text{über } v_1, \dots, v_d \text{ ONS in } \mathbb{R}^p. \tag{2.6}$$

Wir erhalten also zwei Interpretationen von PCA: (2.4) bedeutet minimaler Rekonstruktionsfehler, (2.5) bedeutet maximale Variabilität bzw. Streuung in den projizierten Daten.

Da $\hat{\Sigma}$ symmetrisch und positiv semi-definit ist, folgt aus dem Spektralsatz, dass es reelle Zahlen $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p \geq 0$, sowie eine ONB $\hat{u}_1, \dots, \hat{u}_p$ von $\mathbb{R}^p$ gibt, so dass $\hat{\Sigma}\hat{u}_j = \hat{\lambda}_j \hat{u}_j$ für alle $j = 1, \dots, p$, oder kompakter

$$\hat{\Sigma} = \sum_{j=1}^{p} \hat{\lambda}_j \hat{u}_j \hat{u}_j^T = (\hat{u}_1 \cdots \hat{u}_p) \operatorname{diag}(\hat{\lambda}_1 \cdots \hat{\lambda}_p)(\hat{u}_1 \cdots \hat{u}_p)^T.$$

2.4 Satz. *Das Minimierungsproblem in* (2.4) *besitzt die Lösung*

$$\hat{V}_d = \operatorname{span}(\hat{u}_1, \dots, \hat{u}_d) \quad \textit{und es gilt} \quad R_n(\hat{V}_d) = \sum_{j>d} \hat{\lambda}_j.$$

2.5 Bemerkung. Die Lösung $\hat{V}_d$ ist eindeutig bestimmt falls $\hat{\lambda}_d > \hat{\lambda}_{d+1}$.

Beweis. Wegen (2.6) reicht es zu zeigen, dass

$$\hat{\lambda}_1 + \cdots + \hat{\lambda}_d = \max_{v_1,\dots,v_d \text{ ONS in } \mathbb{R}^p} \sum_{j=1}^{d} \langle \hat{\Sigma} v_j, v_j \rangle$$

und das Maximum wird angenommen für $\hat{u}_1, \dots, \hat{u}_d$. Dies ist eine Variationscharakterisierung für Teilspuren, welche die Charakterisierung des maximalen Eigenwertes aus Kapitel 2.1 erweitert.

" $\leq$ ". Setzen wir $\hat{u}_1, \dots, \hat{u}_d$ ein, so folgt

$$\sum_{j=1}^{d} \langle \hat{\Sigma} v_j, v_j \rangle = \hat{\lambda}_1 + \cdots + \hat{\lambda}_d.$$

" $\geq$ ". Ist nun $v_1, \dots, v_d$ Orthonormalsystem (ONS) in $\mathbb{R}^p$, so gilt

$$\sum_{j=1}^{d} \langle \hat{\Sigma} v_j, v_j \rangle = \sum_{j=1}^{d} \sum_{k=1}^{p} \hat{\lambda}_k \langle \hat{u}_k, v_j \rangle^2 = \sum_{k=1}^{p} \hat{\lambda}_k \Big(\sum_{j=1}^{d} \langle \hat{u}_k, v_j \rangle^2 \Big).$$

Es gilt nun

$$\sum_{j=1}^{d} \langle \hat{u}_k, v_j \rangle^2 \leq \|\hat{u}_k\|^2 = 1$$

und andererseits

$$\sum_{k=1}^{p} \sum_{j=1}^{d} \langle \hat{u}_k, v_j \rangle^2 = \sum_{j=1}^{d} \|v_j\|^2 = d.$$

Es folgt

$$\sum_{j=1}^{d} \langle \hat{\Sigma} v_j, v_j \rangle \leq \hat{\lambda}_1 + \cdots + \hat{\lambda}_d.$$

Somit gilt die erste Behauptung. Die zweite Behauptung folgt aus $R_n(\hat{V}_d) = \operatorname{tr}(\hat{\Sigma}) - \operatorname{tr}(\hat{\Sigma} P_{\hat{V}_d}) = \sum_{j>d} \hat{\lambda}_j$. □

2.6 Satz. *Für $k \geq 1$ gilt*

$$\hat{u}_k \in \operatorname*{argmax}_{\|v\|=1, v \perp \hat{u}_1, \ldots \hat{u}_{k-1}} \langle \hat{\Sigma} v, v \rangle = \operatorname*{argmax}_{\|v\|=1, v \perp \hat{u}_1, \ldots \hat{u}_{k-1}} \frac{1}{n} \sum_{i=1}^{n} \langle X_i, v \rangle^2.$$

Dies ist eine weitere Variationscharakterisierung für Eigenvektoren bzw. Eigenwerte. Die $\hat{u}_k$ sind also die sukzessiven Richtungen mit maximaler empirischer Varianz.

Beweis. Für $\|v\| = 1$ mit $v \perp \hat{u}_1, \ldots \hat{u}_{k-1}$ gilt $v = \sum_{j \geq k} \langle v, \hat{u}_j \rangle \hat{u}_j$ und es folgt

$$\langle \hat{\Sigma} v, v \rangle = \sum_{j=k}^{p} \hat{\lambda}_j \langle v, \hat{u}_j \rangle^2 \leq \hat{\lambda}_k \sum_{j=k}^{p} \langle v, \hat{u}_j \rangle^2 = \hat{\lambda}_k \|v\|^2 = \hat{\lambda}_k$$

mit Gleichheit für $v = \hat{u}_k$. □

2.7 Definition. Der Vektor $(\langle X_1, \hat{u}_k \rangle, \ldots, \langle X_n, \hat{u}_k \rangle)^T \in \mathbb{R}^n$ heißt k-te Hauptkomponente (PC).

2.8 Bemerkungen.

1) Im Allgemeinen (sofern nicht $\mathbb{E} X_i = 0$ für alle i) sollte man PCA auf die zentrierten Daten $X_i - \bar{X} = X_i - n^{-1} \sum_{i=1}^{n} X_i$ anwenden.

2) Die Dimension d wird häufig so gewählt, dass ein gewisses Perzentil (z.B. 50%, 75%, 90%) der erklärten Varianz $\sum_{j=1}^{d} \hat{\lambda}_j$ der gesamten Varianz $\sum_{j=1}^{p} \hat{\lambda}_j$ erreicht wird.

Berechung von PCA

In vielen Anwendungen ist p viel größer als n (kernel PCA, Gesichtserkennung $p = 1024^2$). Dann ist die Diagonalisierung von $\hat{\Sigma}$ viel zu aufwendig. Es bleibt die Frage, ob und wie die PCs in diesem Fall berechnet werden können. Setze hierfür

$$\mathbf{X} = \begin{pmatrix} X_1^T \\ \vdots \\ X_n^T \end{pmatrix} \in \mathbb{R}^{n \times p}.$$

Dann gilt

$$\mathbf{X}^T\mathbf{X} = n\hat{\Sigma} \in \mathbb{R}^{p \times p}$$

und

$$\mathbf{X}\mathbf{X}^T = (\langle X_i, X_j \rangle)_{i,j=1}^n \in \mathbb{R}^{n \times n}.$$

Die Matrix $\mathbf{X}\mathbf{X}^T$ heißt auch Gramsche Matrix. Betrachte nun die Diagonalisierung

$$\mathbf{X}\mathbf{X}^T = \sum_{j=1}^{r} \hat{\sigma}_j^2 \hat{v}_j \hat{v}_j^T$$

mit $\hat{\sigma}_1 \geq \cdots \geq \hat{\sigma}_r > 0$, $\hat{v}_1, \ldots, \hat{v}_r$ ONS in $\mathbb{R}^n$ und

$$r = \mathrm{Rang}(\mathbf{X}\mathbf{X}^T) = \mathrm{Rang}(\mathbf{X}^T\mathbf{X}) = \mathrm{Rang}(\hat{\Sigma}).$$

2.9 Proposition. *Es gilt* $\hat{u}_j = \sigma_j^{-1}\mathbf{X}^T\hat{v}_j$, $j = 1, \ldots, r$.

Beweis. Zunächst bilden $\hat{\sigma}_1^{-1}\mathbf{X}^T\hat{v}_1, \ldots, \hat{\sigma}_r^{-1}\mathbf{X}^T\hat{v}_r$ ein ONS in $\mathbb{R}^p$:

$$\langle \hat{\sigma}_j^{-1}\mathbf{X}^T\hat{v}_j, \hat{\sigma}_k^{-1}\mathbf{X}^T\hat{v}_k \rangle = \hat{\sigma}_j^{-1}\hat{\sigma}_k^{-1}\langle \mathbf{X}\mathbf{X}^T\hat{v}_j, \hat{v}_k \rangle = \begin{cases} 1, & j = k, \\ 0, & j \neq k. \end{cases}$$

Des Weiteren gilt

$$\mathbf{X}^T\mathbf{X}\hat{\sigma}_j^{-1}\mathbf{X}^T\hat{v}_j = \hat{\sigma}_j^{-1}\mathbf{X}^T\mathbf{X}\mathbf{X}^T\hat{v}_j = \hat{\sigma}_j^2\hat{\sigma}_j^{-1}\mathbf{X}^T\hat{v}_j, \quad j = 1, \ldots, r.$$

Es folgt also $\hat{\lambda}_j = \hat{\sigma}_j^2/n$ und $\hat{u}_j = \hat{\sigma}_j^{-1}\mathbf{X}^T\hat{v}_j$, $j = 1, \ldots, r$. □

2.10 Korollar. *Die k-te PC ist gegeben durch $\hat{\sigma}_k\hat{v}_k$.*

Beweis. Es gilt

$$(\langle X_1, \hat{u}_k \rangle, \ldots, \langle X_n, \hat{u}_k \rangle)^T = \mathbf{X}\hat{u}_k = \hat{\sigma}_k^{-1}\mathbf{X}\mathbf{X}^T\hat{v}_k = \hat{\sigma}_k\hat{v}_k.$$

□

2.11 Aufgabe. *Seien $X_1, \dots, X_n \in \mathcal{X}$ und $\Phi : \mathcal{X} \to \mathbb{R}^p$ eine Abbildung. Zeige:*

(a) Wenden wir PCA auf $\Phi(X_1), \dots, \Phi(X_n)$ an, so müssen wir zur Berechnung der ersten k Hauptkomponenten nur die k größten Eigenwerte und die dazugehörigen Eigenvektoren von

$$K = (\langle \Phi(X_i), \Phi(X_j) \rangle)_{i,j=1}^n \in \mathbb{R}^{n \times n}$$

berechnen.

(b) Wenden wir PCA auf die zentrierten Daten $\Phi(X_i) - n^{-1} \sum_{j=1}^n \Phi(X_j)$, $i = 1, \dots, n$ an, so müssen wir zur Berechnung der ersten k Hauptkomponenten nur die k größten Eigenwerte und die dazugehörigen Eigenvektoren von

$$K - \mathbf{1}_n K - K \mathbf{1}_n + \mathbf{1}_n K \mathbf{1}_n$$

berechnen, wobei $(\mathbf{1}_n)_{ij} = 1/n$ für alle $i, j = 1, \dots, n$.

PCA als beste affine Approximation der Daten

2.12 Aufgabe. *Für $X_1, \dots, X_n \in \mathbb{R}^p$ und $d \leq p$, betrachte das Optimierungsproblem*

$$\min_{\mu, (z_i), V} \sum_{i=1}^n \|X_i - \mu - V z_i\|^2 \tag{1}$$

über $\mu \in \mathbb{R}^p$, $z_1, \dots, z_n \in \mathbb{R}^d$ mit $\sum_{i=1}^n z_i = 0$ und $V = (v_1 \cdots v_d) \in \mathbb{R}^{p \times d}$ mit $v_1, \dots, v_d$ ONS in $\mathbb{R}^p$. Zeige:

Eine Lösung $(\hat{\mu}, (\hat{z}_i), \hat{V})$ ist von der Form $\hat{\mu} = \bar{X} = n^{-1} \sum_{i=1}^n X_i$, $\hat{z}_i = \hat{V}^T (X_i - \bar{X})$ und $\hat{V}$ ist eine Lösung des Optimierungsproblems

$$\min_V \sum_{i=1}^n \|X_i - \bar{X} - V V^T (X_i - \bar{X})\|^2$$

über $V = (v_1 \cdots v_d) \in \mathbb{R}^{p \times d}$ mit $v_1, \dots, v_d \in \mathbb{R}^p$ ONS.

2.3 Der Generalisierungsfehler von PCA

PCA als empirische Risikominimierung

Seien $X_1, \dots, X_n, X$ unabhängige und identisch verteile Zufallsvariablen mit Werten in $\mathbb{R}^p$. Der Einfachheit halber nehmen wir im Folgenden an, dass $\mathbb{E}X = 0$. Außerdem gelte $\mathbb{E}\|X\|^2 < \infty$. Für $V \leq \mathbb{R}^p$ heißt

$$R(V) = \mathbb{E}\,\|X - P_V X\|^2$$

Rekonstruktionsfehler von V. Sei $d \in \mathbb{N}$. Setze

$$V_d \in \operatorname*{argmin}_{\dim V = d} R(V). \tag{2.7}$$

Dann ist V_d ein Unterraum der Dimension d welcher X am besten beschreibt. In der Realität können wir V_d nicht berechnen da die Verteilung von X (und daher auch $R(\cdot)$) unbekannt ist. Stattdessen betrachten wir wie in Kapitel 2.2 den empirischen Rekonstruktionsfehler

$$R_n(V) = \frac{1}{n} \sum_{i=1}^{n} \|X_i - P_V X_i\|^2$$

und berechnen

$$\hat{V}_d \in \operatorname*{argmin}_{\dim V = d} R_n(V).$$

Unser Ziel ist es obere Schranken für $R(\hat{V}_d)$, den Rekonstruktionsfehler von PCA, zu finden. Genauer gesagt wollen wir obere Schranken für das sogennante Exzessrisiko (excess risk) von $\hat{V}_d$

$$\mathcal{E}(\hat{V}_d) = R(\hat{V}_d) - R(V_d)$$

beweisen. Beachte dabei, dass $R(\hat{V}_d)$ eine Zufallsvariable ist (betrachte R als Funktion definiert auf der Menge aller Unterräume), der Erwartungswert wird nur bezüglich X genommen:

$$R(\hat{V}_d) = \mathbb{E}_X \|X - P_{\hat{V}_d} X\|^2.$$

Dabei interpretieren wir $R(\hat{V}_d)$ als den Fehler den wir machen wenn wir eine neue Beobachtung $X = X_{n+1}$ auf $\hat{V}_d$ projizieren.

Charakterisierung von V_d

Wie im Fall des empirischen Rekonstruktionsfehlers gilt für $V \leq \mathbb{R}^p$

$$R(V) = \mathbb{E}\|X - P_V X\|^2 = \mathbb{E}(\|X\|^2 - \|P_V X\|^2) = \operatorname{tr}(\Sigma(I - P_V)).$$

mit Kovarianzmatrix $\Sigma = \mathbb{E}XX^T$ von X. Somit ist (2.7) äquivalent zu

$$\text{maximiere} \quad \operatorname{tr}(\Sigma P_V) \qquad \text{über } V \leq \mathbb{R}^p, \dim V = d \tag{2.8}$$

oder konkreter äquivalent zu

$$\text{maximiere} \quad \sum_{j=1}^{d} \langle \Sigma v_j, v_j \rangle \qquad \text{über } v_1, \ldots, v_d \text{ ONS in } \mathbb{R}^p. \tag{2.9}$$

Da Σ symmetrisch und positiv semi-definit ist, folgt aus dem Spektralsatz, dass es reelle Zahlen $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ und eine ONB $u_1, \dots u_p$ von $\mathbb{R}^p$ gibt, so dass $\Sigma u_j = \lambda_j u_j$ für alle $j = 1, \dots, p$, oder kompakter

$$\Sigma = \sum_{j=1}^{p} \lambda_j u_j u_j^T.$$

Setzen wir die Spektralzerlegung in (2.9) ein, so erhalten wir mit dem gleichen Argument wie im Beweis von Satz 2.4 die folgende Beschreibung von V_d.

2.13 Satz. *Das Minimierungsproblem in* (2.7) *besitzt die Lösung*

$$V_d = \operatorname{span}(u_1, \dots, u_d) \quad \textit{und es gilt} \quad R(V_d) = \sum_{j>d} \lambda_j.$$

Der Satz von Davis-Kahan

Sei im Folgenden

$$\Sigma = \sum_{j=1}^{p} \lambda_j u_j u_j^T, \qquad \hat{\Sigma} = \sum_{j=1}^{p} \hat{\lambda}_j \hat{u}_j \hat{u}_j^T$$

mit $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ und $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p \geq 0$ reelle Zahlen und $u_1, \dots, u_p$ und $\hat{u}_1, \dots, \hat{u}_p$ Orthonormalbasen in $\mathbb{R}^p$, und

$$V_d = \operatorname{span}(u_1, \dots, u_d), \qquad \hat{V}_d = \operatorname{span}(\hat{u}_1, \dots, \hat{u}_d).$$

Wir betrachten $\hat{\Sigma} = \Sigma + (\hat{\Sigma} - \Sigma)$ als Störung von Σ. Wir wollen nun die folgende Frage beantworten. Wie verändern sich die Eigenräume bei einer solchen Störung? Wie nah liegt $\hat{V}_d$ an V_d?

2.14 Satz (Davis-Kahan, $\sin\Theta$-Satz)**.** *Es gilt*

$$\|P_{\hat{V}_d} - P_{V_d}\|_{\mathrm{HS}}^2 \leq \min\left(2d, \frac{4\|\hat{\Sigma} - \Sigma\|_{\mathrm{HS}}^2}{(\lambda_d - \lambda_{d+1})^2}\right).$$

2.15 Bemerkungen.

1) Zur Erinnerung: $\langle A, B\rangle_{\mathrm{HS}} = \operatorname{tr}(A^T B)$ und $\|A\|_{HS}^2 = \langle A, A\rangle_{\mathrm{HS}} = \sum_{j=1}^{p}\sum_{k=1}^{p} a_{jk}^2$

2) Im Fall $d = 1$ gilt

   $$\|P_{\hat{V}_1} - P_{V_1}\|_{\mathrm{HS}}^2 = \|\hat{u}_1\hat{u}_1^T - u_1 u_1^T\|_{\mathrm{HS}}^2 = 2 - 2\langle \hat{u}_1, u_1\rangle^2 = 2\sin^2(\measuredangle(\hat{u}_1, u_1)).$$

   In diesem Fall kann der Satz von Davis-Kahan also umgeschrieben werden als

   $$\sin(\measuredangle(\hat{u}_1, u_1)) \ \leq \frac{\sqrt{2}\|\hat{\Sigma} - \Sigma\|_{\mathrm{HS}}}{\lambda_d - \lambda_{d+1}}.$$

   Eine analoge Interpretation von $\|P_{\hat{V}_d} - P_{V_d}\|_{\mathrm{HS}}^2$ wird in Aufgabe 2.16 auch für den Fall $d \geq 2$ erarbeitet.

3) Man kann anhand von Beispielen sehen, dass der Satz von Davis-Kahan bis auf Konstanten scharf ist. Sei hierfür

$$\Sigma = \begin{pmatrix} 1+\epsilon & 0 \\ 0 & 1-\epsilon \end{pmatrix} = (1+\epsilon)\begin{pmatrix}1\\0\end{pmatrix}\begin{pmatrix}1\\0\end{pmatrix}^T + (1-\epsilon)\begin{pmatrix}0\\1\end{pmatrix}\begin{pmatrix}0\\1\end{pmatrix}^T$$

und

$$\hat{\Sigma} = \begin{pmatrix} 1 & \epsilon \\ \epsilon & 1 \end{pmatrix} = (1+\epsilon)\begin{pmatrix}1/\sqrt{2}\\1/\sqrt{2}\end{pmatrix}\begin{pmatrix}1/\sqrt{2}\\1/\sqrt{2}\end{pmatrix}^T + (1-\epsilon)\begin{pmatrix}1/\sqrt{2}\\-1/\sqrt{2}\end{pmatrix}\begin{pmatrix}1/\sqrt{2}\\-1/\sqrt{2}\end{pmatrix}^T.$$

Daher folgt

$$\sin(\measuredangle(\hat{u}_1, u_1)) = \frac{1}{\sqrt{2}} = \frac{1}{\sqrt{2}}\frac{2\epsilon}{2\epsilon} = \frac{1}{\sqrt{2}}\frac{\|\hat{\Sigma}-\Sigma\|_{\mathrm{HS}}}{\lambda_1-\lambda_2}.$$

Beweis. Setze

$$P = P_{V_d} = \sum_{j=1}^{d} u_j u_j^T \quad \text{und} \quad \hat{P} = P_{\hat{V}_d} = \sum_{j=1}^{d} \hat{u}_j \hat{u}_j^T.$$

Dann gilt

$$\|P-\hat{P}\|_{\mathrm{HS}}^2 = \langle P-\hat{P}, P-\hat{P}\rangle_{\mathrm{HS}} = \|P\|_{\mathrm{HS}}^2 + \|\hat{P}\|_{\mathrm{HS}}^2 - 2\langle P, \hat{P}\rangle_{\mathrm{HS}}.$$

Setzen wir die Identitäten

$$\|P\|_{\mathrm{HS}}^2 = \sum_{j=1}^{d}\|u_j\|^2 = d, \qquad \|\hat{P}\|_{\mathrm{HS}}^2 = d$$

ein, so erhalten wir

$$\|P-\hat{P}\|_{\mathrm{HS}}^2 = 2d - 2\underbrace{\langle P, \hat{P}\rangle_{\mathrm{HS}}}_{\geq 0} \leq 2d,$$

und der erste Teil der Ungleichung folgt. Weiter gilt

$$P-\hat{P} = (I-\hat{P}+\hat{P})P - \hat{P}(I-P+P) = (I-\hat{P})P - \hat{P}(I-P)$$

mit

$$I-P = \sum_{j=d+1}^{p} u_j u_j^T$$

Orthogonalprojektion auf $V_d^{\perp}$. Daher folgt

$$\begin{aligned}
\langle \Sigma, P-\hat{P}\rangle_{\mathrm{HS}} &= \langle \Sigma, (I-\hat{P})P\rangle_{\mathrm{HS}} - \langle \Sigma, \hat{P}(I-P)\rangle_{\mathrm{HS}} \\
&= \sum_{j=1}^{d} \lambda_j \underbrace{\langle u_j, (I-\hat{P})u_j\rangle}_{\geq 0} - \sum_{k=d+1}^{p} \lambda_k \underbrace{\langle u_k, \hat{P}u_k\rangle}_{\geq 0} \\
&\geq \lambda_d \sum_{j=1}^{d}\langle u_j, (I-\hat{P})u_j\rangle - \lambda_{d+1}\sum_{k=d+1}^{p}\langle u_k, \hat{P}u_k\rangle \\
&= \lambda_d(d - \langle P, \hat{P}\rangle_{\mathrm{HS}}) - \lambda_{d+1}(d - \langle P, \hat{P}\rangle_{\mathrm{HS}}).
\end{aligned}$$

Es folgt

$$\begin{aligned}\|P-\hat{P}\|_{\mathrm{HS}}^2 = 2d - 2\langle P,\hat{P}\rangle_{\mathrm{HS}} &\le \frac{2\langle \Sigma, P-\hat{P}\rangle_{\mathrm{HS}}}{\lambda_d-\lambda_{d+1}}\\ &\le \frac{2\langle \Sigma-\hat{\Sigma}, P-\hat{P}\rangle_{\mathrm{HS}}}{\lambda_d-\lambda_{d+1}}\\ &\le \frac{2\|\Sigma-\hat{\Sigma}\|_{\mathrm{HS}}\|P-\hat{P}\|_{\mathrm{HS}}}{\lambda_d-\lambda_{d+1}},\end{aligned}$$

wobei wir die Cauchy-Schwarz-Ungleichung und die Tatsache benutzt haben, dass $\langle\hat{\Sigma},\hat{P}\rangle_{\mathrm{HS}} \ge \langle\hat{\Sigma},P\rangle_{\mathrm{HS}}$ ($\hat{P}$ maximiert den Term $\langle\hat{\Sigma},P_V\rangle_{\mathrm{HS}}$ für V mit $\dim V = d$). Teilen durch $\|P-\hat{P}\|_{\mathrm{HS}}$ und quadrieren liefert den zweiten Teil der Ungleichung. □

2.16 Aufgabe. *Sei $A = (\langle u_j,\hat{u}_k\rangle)_{j,k=1}^d \in \mathbb{R}^{d\times d}$ mit Singulärwertzerlegung $A = \sum_{j=1}^d \sigma_j\psi_j\varphi_j^T$, wobei $1 \ge \sigma_1 \ge \cdots \ge \sigma_d \ge 0$ und $\psi_1,\dots,\psi_d$ und $\varphi_1,\dots,\varphi_d$ Orthonormalbasen von $\mathbb{R}^d$. Dann heißt*

$$\vartheta_j := \vartheta_j(V_d,\hat{V}_d) := \arccos(\sigma_j) \in [0,\pi/2]$$

der j-te Hauptwinkel (principal angle) zwischen V_d und $\hat{V}_d$. Zeige:

(a) Es gilt

$$\cos(\vartheta_1) = \max_{v\in V_d, w\in\hat{V}_d} \frac{|\langle v,w\rangle|}{\|v\|\|w\|}.$$

Formuliere und beweise eine analoge Formel für $\cos(\vartheta_j)$, $j = 2,\dots,d$.

(b) Es gilt

$$\|P_{\hat{V}_d} - P_{V_d}\|_{\mathrm{HS}}^2 = 2\sum_{j=1}^d \sin^2(\vartheta_j).$$

Exzessrisiko von PCA

2.17 Satz. *Für $d \le p$ gilt*

$$\mathbb{E}R(\hat{V}_d) - R(V_d) \le \min\Big(\sqrt{\frac{2d}{n}(\mathbb{E}\|X\|^4 - \|\Sigma\|_{\mathrm{HS}}^2)}, \frac{2}{n}\frac{\mathbb{E}\|X\|^4 - \|\Sigma\|_{\mathrm{HS}}^2}{\lambda_d-\lambda_{d+1}}\Big).$$

Man spricht von einer langsamen $n^{-1/2}$-Rate und einer schnellen n^{-1}-Rate. Allerdings hängt die schnelle Rate von der spectral gap $\lambda_d - \lambda_{d+1}$ ab, die prinzipiell sogar gleich Null sein kann. Es können also beide Teile der Ungleichung dominieren.

2.18 Beispiel. Im Fall $X \sim \mathcal{N}(0,\Sigma)$ gilt $\mathbb{E}\|X\|^4 = \operatorname{tr}^2(\Sigma) + 2\|\Sigma\|_{\mathrm{HS}}^2 \le 2\operatorname{tr}^2(\Sigma) + \|\Sigma\|_{\mathrm{HS}}^2$ und wir erhalten

$$\mathbb{E}R(\hat{V}_d) - R(V_d) \le \min\Big(\sqrt{\frac{2d}{n}}\operatorname{tr}(\Sigma), \frac{2}{n}\frac{\operatorname{tr}^2(\Sigma)}{\lambda_d - \lambda_{d+1}}\Big).$$

Beweis. Es gilt

$$\begin{aligned} R(\hat{V}_d) - R(V_d) &= \operatorname{tr}(\Sigma(I - P_{\hat{V}_d})) - \operatorname{tr}(\Sigma(I - P_{V_d})) \\ &= \operatorname{tr}(\Sigma(P_{V_d} - P_{\hat{V}_d})) = \langle \Sigma, P_{V_d} - P_{\hat{V}_d}\rangle_{\mathrm{HS}}. \end{aligned}$$

Es folgt, dass

$$R(\hat{V}_d) - R(V_d) \le \langle \Sigma - \hat{\Sigma}, P_{V_d} - P_{\hat{V}_d})\rangle_{\mathrm{HS}} \le \|\Sigma - \hat{\Sigma}\|_{\mathrm{HS}}\|P_{V_d} - P_{\hat{V}_d}\|_{\mathrm{HS}}.$$

Setzen wir nun Satz 2.14 ein, so erhalten wir

$$R(\hat{V}_d) - R(V_d) \le \min\left(\sqrt{2d}\|\hat{\Sigma} - \Sigma\|_{\mathrm{HS}}, \frac{2\|\hat{\Sigma} - \Sigma\|_{\mathrm{HS}}^2}{\lambda_d - \lambda_{d+1}}\right).$$

Nehmen wir den Erwartungswert und wenden die Cauchy-Schwarz-Ungleichung an, so folgt

$$\mathbb{E}R(\hat{V}_d) - R(V_d) \le \min\left(\sqrt{2d\mathbb{E}\|\hat{\Sigma} - \Sigma\|_{\mathrm{HS}}^2}, \frac{2\mathbb{E}\|\hat{\Sigma} - \Sigma\|_{\mathrm{HS}}^2}{\lambda_d - \lambda_{d+1}}\right).$$

Wir müssen also noch $\mathbb{E}\|\hat{\Sigma} - \Sigma\|_{\mathrm{HS}}^2$ berechnen:

$$\begin{aligned} \mathbb{E}\|\hat{\Sigma} - \Sigma\|_{\mathrm{HS}}^2 &= \mathbb{E}\Big\|\frac{1}{n}\sum_{i=1}^n X_i X_i^T - \Sigma\Big\|_{\mathrm{HS}}^2 \\ &= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \mathbb{E}\operatorname{tr}((X_iX_i^T - \Sigma)(X_jX_j^T - \Sigma)) \\ &= \frac{1}{n}\mathbb{E}\operatorname{tr}((XX^T - \Sigma)(XX^T - \Sigma)). \end{aligned}$$

Der letzte Ausdruck ist gleich mit $\mathbb{E}\|X\|^4 - \|\Sigma\|_{\mathrm{HS}}^2$ und die Behauptung folgt. □

Das Exzessrisiko mittels Rademacher-Komplexitäten

2.19 Aufgabe. *Es gelte $\|X\| \le M$ fast sicher. Zeige:*

(a) Es gilt

$$\begin{aligned} \mathbb{E}R(\hat{V}_1) - R(V_1) &\le \mathbb{E}\sup_{\dim V = 1}(R(V) - R_n(V)) \\ &= \mathbb{E}\sup_{\|v\|=1}\frac{1}{n}\sum_{i=1}^n(\langle v, X_i\rangle^2 - \mathbb{E}\langle v, X_i\rangle^2). \end{aligned}$$

(b) Seien $\epsilon_1, \dots, \epsilon_n$ i.i.d. Rademacher Zufallsvariablen, unabhängig von $X_1, \dots, X_n$. Erkläre die folgenden Rechenschritte:

$$\mathbb{E} \sup_{\|v\|=1} \frac{1}{n} \sum_{i=1}^{n} (\langle v, X_i \rangle^2 - \mathbb{E}\langle v, X_i \rangle^2) \leq 2\mathbb{E} \sup_{\|v\|=1} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \langle v, X_i \rangle^2$$
$$\leq 4M\mathbb{E} \sup_{\|v\|=1} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \langle v, X_i \rangle.$$

(c) Es folgt $\mathbb{E}R(\hat{V}_1) - R(V_1) \leq \frac{4M^2}{\sqrt{n}}$.

3 Kern-Methoden

Unser Ziel in diesem Kapitel ist es die Ausdruckstärke von linearen Verfahren wie SVM und PCA zu vergrößern indem wir die Daten mittels Featureabbildung (feature map) in einen höher-dimensionalen Featureraum (feature space) abbilden. Als motivierendes Beispiel betrachte (x_i, y_i) mit $i = -10, -9, \dots, 0, 1, \dots, 9, 10$, $x_i = i$ und $y_i = 1$ falls $|i| \leq 2$ und $y_i = -1$ falls $|i| > 2$. Des Weiteren sei $\Phi : \mathbb{R} \to \mathbb{R}^2$ gegeben durch $\Phi(x) = (x, x^2)$. Dann können die eingebetteten Daten $(\Phi(x_i), y_i)$ durch die Hyperebene $\langle (0,1)^T, x \rangle - 5 = 0$ separiert werden.

3.1 SVM, Ridge-Regression und PCA mit Featureabbildung

SVM

Seien $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathcal{X} \times \{\pm 1\}$ Daten und $\Phi : \mathcal{X} \to H$ eine Abbildung mit $(H, \langle \cdot, \cdot \rangle, \|\cdot\|)$ (möglicherweise unendlichdimensionaler) Hilbertraum (zum Beispiel $H = \mathbb{R}^p$). Wir wollen SVM auf die eingebetteten Daten $(\Phi(X_i), Y_i)_{i=1}^n$ anwenden. Betrachte also für $\lambda > 0$

$$\hat{w}_n^{\mathrm{SVM}} \in \operatorname*{argmin}_{w \in H} \frac{1}{n} \sum_{i=1}^{n} (1 - Y_i \langle w, \Phi(X_i) \rangle)_+ + \lambda \|w\|^2 \tag{3.1}$$

und

$$\hat{h}_n^{\mathrm{SVM}}(x) = \mathrm{sign}(\langle \hat{w}_n^{\mathrm{SVM}}, \Phi(x) \rangle).$$

3.1 Lemma (Darsteller-Formel)**.** *Es gilt*

$$\hat{w}_n^{SVM} = \sum_{j=1}^{n} \hat{\alpha}_j \Phi(X_j) \tag{3.2}$$

mit

$$\hat{\alpha} \in \operatorname*{argmin}_{\alpha \in \mathbb{R}^n} \left(\frac{1}{n} \sum_{i=1}^{n} \Big(1 - Y_i (K\alpha)_i\Big)_+ + \lambda \alpha^T K \alpha \right),$$

wobei

$$K = (\langle \Phi(X_i), \Phi(X_j) \rangle)_{i,j=1}^n.$$

Des Weiteren gilt

$$\hat{h}_n^{SVM}(x) = \begin{cases} 1 & \sum_{j=1}^n \hat{\alpha}_j \langle \Phi(x), \Phi(X_j) \rangle > 0 \\ 0 & sonst \end{cases}.$$

Wir müssen bei der Berechnung von $\hat{w}_n^{\mathrm{SVM}}$ und $\hat{h}_n^{\mathrm{SVM}}(x)$ also nicht unbedingt Φ berechnen, es reicht die Funktion $\langle \Phi(\cdot), \Phi(\cdot) \rangle : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (an den Beobachtungen und x) zu kennen.

Beweis. Sei $V = \mathrm{span}(\Phi(X_1), \ldots, \Phi(X_n))$ und $V^\perp$ das orthogonale Komplement von V in H. Ist $w \in H$, so schreibe $w = w_V + w_{V^\perp}$ mit $w_V \in V$ und $w_{V^\perp} \in V^\perp$. Dann gilt $\|w\|^2 = \|w_V\|^2 + \|w_{V^\perp}\|^2$ und $\langle w, \Phi(X_i) \rangle = \langle w_V, \Phi(X_i) \rangle$ für alle $i = 1, \ldots, n$. Es folgt, dass

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^n (1 - Y_i(\langle w, \Phi(X_i) \rangle)_+ + \lambda \|w\|^2 \\ &= \frac{1}{n} \sum_{i=1}^n (1 - Y_i(\langle w_V, \Phi(X_i) \rangle)_+ + \lambda \|w_V\|^2 + \lambda \|w_{V^\perp}\|^2. \end{aligned}$$

Ist daher w eine Lösung von (3.1), so muss $w_{V^\perp} = 0$ gelten und w ist von der Form (3.2). Für $w = \sum_{j=1}^n \alpha_j \Phi(X_j)$ gilt

$$\langle w_V, \Phi(X_i) \rangle = \sum_{j=1}^n K_{ij} \alpha_j = (K\alpha)_i, \qquad \|w\|^2 = \alpha^T K \alpha.$$

Setzen wir dies in (3.1), so folgt die letzte Behauptung. □

Wir merken noch an, dass (3.1) eine eindeutige Lösung besitzt. Existenz folgt dabei mit Hilfe eines Kompaktheitsargument, Eindeutigkeit folgt aus der Konvexität des hinge losses in Kombination mit der strengen Konvexität des Strafterms:

$$\left\| \frac{w + w'}{2} \right\|^2 = \frac{1}{4}(2\|w\|^2 + 2\|w'\|^2 - \|w - w'\|^2) < \frac{1}{2}(\|w\|^2 + \|w'\|^2)$$

für alle $w, w' \in H$ mit $w \neq w'$.

Ridge-Regression

Seien $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \mathbb{R}$ Daten und $\Phi : \mathcal{X} \to H$ eine Abbildung mit Hilbertraum H. Betrachte für $\lambda > 0$

$$\hat{w}_n^{\mathrm{Ridge}} \in \operatorname*{argmin}_{w \in H} \frac{1}{n} \sum_{i=1}^n (Y_i - \langle w, \Phi(X_i) \rangle)^2 + \lambda \|w\|^2. \tag{3.3}$$

3.2 Lemma. *Es gilt*

$$\hat{w}_n^{Ridge} = \sum_{j=1}^{n} \hat{\alpha}_j \Phi(X_j) \tag{3.4}$$

mit $\hat{\alpha} \in \mathbb{R}^n$ *gegeben durch*

$$\hat{\alpha} = (K + n\lambda I)^{-1} Y,$$

wobei $K = (\langle \Phi(X_i), \Phi(X_j) \rangle)_{i,j=1}^n$ *und* $Y = (Y_1, \ldots, Y_n)^T$.

Beweis. Die erste Behauptung folgt mit dem gleichen Argument wie für den SVM-Klassifizierer. Für $w = \sum_{j=1}^n \alpha_j \Phi(X_j)$ gilt

$$\langle w, \Phi(X_i) \rangle = (K\alpha)_i \quad \text{und} \quad \|w\|^2 = \alpha^T K \alpha.$$

Setzen wir dies in (3.4) ein so folgt, dass

$$\hat{\alpha} \in \operatorname*{argmin}_{\alpha \in \mathbb{R}^n} \|Y - K\alpha\|^2 + n\lambda \alpha^T K \alpha.$$

Bilden wir den Gradienten, so erhalten wir, dass $\hat{\alpha}$ die folgende Gleichung erfüllen muss

$$-2K(Y - K\hat{\alpha}) + 2n\lambda K\hat{\alpha} = 0,$$

d.h.

$$K(-Y + (K + n\lambda I)\hat{\alpha}) = 0.$$

Also ist $\hat{\alpha}$ von der Form $(K + n\lambda I)^{-1}(Y + \beta)$ mit $K\beta = 0$. Ein solches β erfüllt $\sum_{j=1}^n \beta_j \Phi(X_j) = 0$ (die Norm zum Quadrat ist gleich $\beta K \beta = 0$) und hat somit keinen Einfluss auf (3.4). □

PCA

Seien $X_1, \ldots, X_n \in \mathcal{X}$ Daten und $\Phi : \mathcal{X} \to H$ mit Hilbertraum H. Betrachte

$$\hat{u}_1 \in \operatorname*{argmax}_{w \in H, \|w\|=1} \frac{1}{n} \sum_{i=1}^{n} \langle w, \Phi(X_i) \rangle^2 \tag{3.5}$$

und iterativ für $j = 2, \ldots, n$,

$$\hat{u}_j \in \operatorname*{argmax}_{\|w\|=1, w \perp \hat{u}_1, \ldots, \hat{u}_{j-1}} \frac{1}{n} \sum_{i=1}^{n} \langle w, \Phi(X_i) \rangle^2$$

3.3 Lemma. *Es gilt*

$$\hat{u}_1 = \sum_{j=1}^{n} \hat{\alpha}_j \Phi(X_j) \tag{3.6}$$

mit $\hat{\alpha} \in \mathbb{R}^n$ *gegeben durch*

$$\hat{\alpha} = \hat{\sigma}_1^{-1} \hat{v}_1,$$

wobei $\hat{\sigma}_1^2$ *maximaler Eigenwert von* $K = (\langle \Phi(X_i), \Phi(X_j) \rangle)_{i,j=1}^n$ *und* $\hat{v}_1$ *zugehöriger Eigenvektor ist. Insebsondere ist die erste Hauptkomponente* $(\langle \hat{u}_1, \Phi(X_1) \rangle, \ldots, \langle \hat{u}_1, \Phi(X_n) \rangle)^T$ *gegeben durch* $\hat{\sigma}_1 \hat{v}_1$.

Beweis. Ist $w \in H$, so schreibe wieder $w = w_V + w_{V^\perp}$ mit $w_V \in V = \operatorname{span}(\Phi(X_1), \ldots, \varphi(X_n))$ und $w_{V^\perp} \in V^\perp$. Dann gilt

$$\frac{1}{n} \sum_{i=1}^{n} \langle w, \Phi(X_i) \rangle^2 = \frac{1}{n} \sum_{i=1}^{n} \langle w_V, \Phi(X_i) \rangle^2 \leq \frac{1}{n} \sum_{i=1}^{n} \langle w_V / \|w_V\|, \Phi(X_i) \rangle^2.$$

Daher ist eine Lösung von (3.5) von der Form (3.6). Für $w = \sum_{j=1}^n \alpha_j \Phi(X_j)$ gilt

$$\frac{1}{n} \sum_{i=1}^{n} \langle w, \Phi(X_i) \rangle^2 = \frac{1}{n} \alpha^T K^2 \alpha \quad \text{und} \quad \|w\|^2 = \alpha^T K \alpha.$$

Setzen wir dies in (3.5) so folgt, dass

$$\hat{\alpha} \in \operatorname*{argmax}_{\alpha^T K \alpha = 1} \alpha^T K^2 \alpha.$$

Es folgt, dass $K^{1/2} \hat{\alpha} = \hat{v}_1$ und somit $\hat{\alpha} = \hat{\sigma}_1^{-1} \hat{v}_1 + \beta$ mit $K\beta = 0$. Ein solches β erfüllt $\sum_{j=1}^n \beta_j \Phi(X_j) = 0$ und hat somit keinen Einfluss auf (3.5). □

In all diesen Beispielen (SVM, Ridge, PCA) mussten wir Φ nicht unbedingt ausrechnen, es reichte die Funktion

$$\langle \Phi(\cdot), \Phi(\cdot) \rangle : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$

(ausgewertet an den Beobachtungen) zu kennen.

Wir wollen nun einen umgekehrten Zugang finden. Anstatt Φ zu konstruieren und dann die inneren Produkte zu berechnen, wollen wir mit einem sogenannten Kern $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ starten der unsere inneren Produkte in einem (möglicherweise unbekannten) Featureraum berechnet. Man spricht vom sogenannten Kern-Trick (kernel trick).

3.4 Definition. Eine Funktion $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ heißt Kern, falls es einen (möglicherweise unendlichdimensionalen) Hilbertraum $(H, \langle \cdot, \cdot \rangle)$ und eine Abbildung $\Phi : \mathcal{X} \to H$ gibt, so dass

$$k(x, y) = \langle \Phi(x), \Phi(y) \rangle, \qquad \forall x, y \in \mathcal{X}.$$

Φ heißt Featureabbildung (feature map), H Featureraum (feature space).

Wir wollen im Folgenden unter anderem die beiden Fragen beantworten: Wann ist k ein Kern? Wie können Kerne konstruiert werden?

3.2 Charakterisierung von Kernen

3.5 Definition. Eine Abbildung $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ heißt positiv definit, falls für alle $m \geq 1$, $x_1, \ldots, x_m \in \mathcal{X}$ und $a_1, \ldots, a_m \in \mathbb{R}$

$$\sum_{j=1}^{m} \sum_{k=1}^{m} a_j a_k k(x_j, x_k) \geq 0.$$

k heißt symmetrisch, falls $k(x, y) = k(y, x)$ für alle $x, y \in \mathcal{X}$.

3.6 Satz. *Eine Abbildung $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ ist ein Kern genau dann, wenn sie symmetrisch und positiv definit ist.*

Beweis. Wir zeigen hier nur die einfachere Hinrichtung, die Rückrichtung befindet sich nach Satz 3.16. Ist k ein Kern, so folgt aus der Darstellung $k(x, y) = \langle \Phi(x), \Phi(y) \rangle$, dass k symmetrisch ist. Des Weiteren folgt

$$\sum_{j=1}^{m} \sum_{k=1}^{m} a_j a_k k(x_j, x_k) = \sum_{j=1}^{m} \sum_{k=1}^{m} a_j a_k \langle \Phi(x_j), \Phi(x_k) \rangle = \Big\| \sum_{j=1}^{m} a_j \Phi(x_j) \Big\|^2 \geq 0.$$

Also ist k symmetrisch und positiv definit. □

3.3 Konstruktion von Kernen

3.7 Lemma (Summe von Kernen)**.** *Sind $k_1, k_2 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ zwei Kerne, so ist auch $k_1 + k_2$ ein Kern.*

Beweis. Wegen Satz 3.6 sind k_1, k_2 symmetrisch und positiv definit. Also folgt, dass $k_1 + k_2$ symmetrisch ist und es gilt $\sum_{j=1}^{m} \sum_{k=1}^{m} a_j a_k (k_1(x_j, x_k) + (k_2(x_j, x_k)) \geq 0$. Also folgt die Behauptung aus Satz 3.6. Alternativ kann man für Featureabbildungen $\Phi_j : \mathcal{X} \to H_j$, $j = 1, 2$, auch $\Phi_1 \oplus \Phi_2 : \mathcal{X} \to H_1 \oplus H_2, x \mapsto (\Phi_1(x), \Phi_2(x))$ betrachten. □

3.8 Lemma (Produkt von Kernen)**.** *Sind $k_1, k_2 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ zwei Kerne, so ist auch $k_1 \cdot k_2$ ein Kern.*

Beweis. Wegen Satz 3.6 sind k_1, k_2 symmetrisch und positiv definit. Es folgt, dass $k_1 \cdot k_2$ symmetrisch ist. Sei nun $m \geq 1$, $x_1, \ldots, x_m \in \mathcal{X}$ und $a_1, \ldots, a_m \in \mathbb{R}$. Dann ist $K_1 = (k_1(x_j, x_k))_{j,k=1}^{m}$ symmetrisch und positiv semi-definit, d.h. wir können $K_1 = FF^T$ schreiben mit $F = (f_{jk}) \in \mathbb{R}^{m \times m}$. Hieraus folgt $k(x_j, x_k) = \sum_{i=1}^{m} f_{ji} f_{ki}$ und somit

$$\sum_{j=1}^{m} \sum_{k=1}^{m} a_j a_k k_1(x_j, x_k) k_2(x_j, x_k) = \sum_{i=1}^{m} \Big(\sum_{j=1}^{m} \sum_{k=1}^{m} a_j f_{ji} a_k f_{ki} k_2(x_j, x_k) \Big) \geq 0,$$

wobei wir in der Ungleichung verwendet haben, dass k_2 positiv definit ist. Also folgt die Behauptung aus Satz 3.6. Alternativ kann man auch mit Tensorprodukten und deren Vervollständigung argumentieren. □

3.9 Lemma (Polynomialer Kern)**.** *Für natürliche Zahlen $m \geq 0$ und $d \geq 1$ und eine relle Zahl $c \geq 0$ ist die Funktion $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ gegeben durch $k(x, y) = (\langle x, y\rangle + c)^m$ ein Kern.*

Beweis. Offensichtlich sind $\langle \cdot, \cdot \rangle$ und die konstante Funktion $k(x, y) = c$ für $c \geq 0$ Kerne. Die Behauptung folgt nun durch Anwendung von Lemma 3.7 und Lemma 3.8. □

3.10 Aufgabe. *Zeige, dass sich der Kern aus Lemma 3.9 mit $c = 1$ durch eine Featureabbildung $\Phi : \mathbb{R}^d \to \mathbb{R}^D$ mit $D = \binom{d+m}{m}$ realisiert werden kann.*

3.11 Lemma (Limiten von Kernen)**.** *Sei $k_n : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $n \geq 1$, eine Folge von Kernen die punktweise gegen eine Funktion $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ konvergiert (d.h. $\lim_{n\to\infty} k_n(x, y) = k(x, y)$). Dann ist k ein Kern.*

Beweis. Wegen Satz 3.6 sind die k_n symmetrisch und positiv definit. Symmetrie von k folgt aus $k(x, y) = \lim_{n\to\infty} k_n(x, y) = \lim_{n\to\infty} k_n(y, x) = k(y, x)$. Positive Definitheit folgt analog aus

$$\sum_{j=1}^{m}\sum_{k=1}^{m} a_j a_k k(x_j, x_k) = \lim_{n\to\infty} \sum_{j=1}^{m}\sum_{k=1}^{m} a_j a_k k_n(x_j, x_k) \geq 0.$$

Also folgt die Behauptung aus Satz 3.6. □

3.12 Lemma (Exponential-Kern)**.** *Für eine natürliche Zahl $d \geq 1$ und eine reelle Zahl $\sigma > 0$ ist die Funktion $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ gegeben durch $k(x, y) = \exp(\langle x, y\rangle/\sigma^2)$ ein Kern.*

Beweis. Aus Lemma 3.7 und Lemma 3.8 folgt, dass $\sum_{k\leq n} \frac{\langle x,y\rangle^k}{k!\sigma^{2k}}$ für alle $n \geq 1$ einen Kern definiert. Da dieser Ausdruck für $n \to \infty$ punktweise gegen $\exp(\langle x, y\rangle/\sigma^2)$ konvergiert, folgt die Behauptung aus Lemma 3.11. □

3.13 Aufgabe. *Sei $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ ein Kern. Zeige, dass die Abbildung $k' : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ definiert durch $k'(x, y) = k(x, y)/\sqrt{k(x, x)k(y, y)}$, falls der Nenner ungleich Null ist und $k'(x, y) = 0$, sonst, auch eine Kern ist.*

Kombinieren wir diese Aufgabe mit Lemma 3.12, so erhalten wir den wichtigen Gauß-Kern (RBF kernel):

3.14 Lemma (Gauß-Kern)**.** *Für eine natürliche Zahl $d \geq 1$ und eine relle Zahl $\sigma > 0$ ist die Funktion $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ gegeben durch $k(x, y) = \exp(-\|x - y\|^2/(2\sigma^2))$ ein Kern.*

3.15 Lemma (Histogramm-Kern)**.** *Die Funktion $k : [0, 1] \times [0, 1] \to \mathbb{R}$ gegeben durch $k(x, y) = x \wedge y = \min(x, y)$ ist ein Kern.*

Beweis. k ist symmetrisch und es gilt

$$k(x,y) = x \wedge y = \int_0^1 \mathbb{1}_{[0,x]}(t)\mathbb{1}_{[0,y]}(t)\,dt$$

Aus dieser Darstellung folgt leicht die positive Definitheit:

$$\sum_{j=1}^m \sum_{k=1}^m a_j a_k k(x_j,x_k) = \int_0^1 \Big(\sum_{j=1}^m a_j \mathbb{1}_{[0,x_j]}(t)\Big)^2 dt \geq 0.$$

Also folgt die Behauptung aus Satz 3.6. □

3.4 Reproducing kernel Hilbert spaces (RKHS)

3.16 Satz. *Sei $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ eine symmetrische und positiv definite Abbildung. Dann existiert genau ein Hilbertraum H_k von Funktionen $f : \mathcal{X} \to \mathbb{R}$ mit*

(a) $k(x,\cdot) \in H_k$ für alle $x \in \mathcal{X}$,

(b) $f(x) = \langle f, k(x,\cdot)\rangle_{H_k}$ für alle $x \in \mathcal{X}$ und alle $f \in H_k$.

H_k heißt auch der zu k gehörige RKHS (reproducing kernel Hilbert space). Man nennt k oft auch reproduzierender Kern (reproducing kernel) und (b) die reproduzierende Eigenschaft (reproducing property).

Wir zeigen zunächst wie man mit Hilfe von Satz 3.16 die Rückrichtung aus Satz 3.6 zeigen kann. Ist $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ eine symmetrische und positiv definite Abbildung und H_k der zugehörige RKHS, so kann man

$$\Phi : \mathcal{X} \to H, x \mapsto k(x,\cdot)$$

betrachten. Dann gilt

$$\langle \Phi(x), \Phi(y)\rangle_{H_k} = \langle k(x,\cdot), k(y,\cdot)\rangle_{H_k} = k(x,y),$$

wobei wir in der zweiten Gleichheit die reproduzierende Eigenschaft verwendet haben. Wir erhalten also, dass k ein Kern ist. Man nennt Φ auch die kanonische Featureabbildung.

3.17 Beispiel. Sei $\mathcal{X} = \mathbb{R}^d$ und $k(x,y) = \langle x,y\rangle = \sum_{j=1}^d x_j y_j$. Dann ist der zu k gehörige RKHS gegeben durch den Raum aller Linearformen $H_k = \{f_w = \langle w,\cdot\rangle : w \in \mathbb{R}^d\}$ versehen mit dem Skalarprodukt $\langle f_w, f_v\rangle_{H_k} = \langle w,v\rangle$, $v,w \in \mathbb{R}^d$. Offensichtlich ist $(H_k, \langle\cdot,\cdot\rangle_{H_k})$ ein endlich-dimensionaler Innenproduktraum und somit ein Hilbertraum. Des Weiteren gilt $k(x,\cdot) = \langle x,\cdot\rangle = f_x \in H_k$ für alle $x \in \mathbb{R}^d$ und

$$\langle f_w, k(x,\cdot)\rangle_{H_k} = \langle f_w, f_x\rangle_{H_k} = \langle w,x\rangle = f_w(x), \qquad \forall x,w \in \mathbb{R}^d.$$

3.18 Beispiel. Sei $\mathcal{X} = [0,1]$ und $k(x,y) = x \wedge y$. Dann ist der Raum

$$H_k = \Big\{ f : [0,1] \to \mathbb{R} : f(0) = 0, f \text{ absolut stetig mit } f' \in L^2[0,1] \Big\}$$

versehen mit dem Skalarprodukt $\langle f, g \rangle_{H_k} = \int_0^1 f'(t) g'(t)\, dt$, der zu k gehörige RKHS. Man zeigt zunächst, dass $(H_k, \langle \cdot, \cdot \rangle_{H_k})$ ein Hilbertraum ist (die Vollständigkeit erhält man dabei aus der Vollständigkeit von $L^2[0,1]$, angewendet auf die Ableitungen). Wir setzen nun $R_x : [0,1] \to \mathbb{R}, y \mapsto x \wedge y = k(x,y)$. Dann gilt $R_x' = \mathbb{1}_{[0,x]}$ da $\int_0^y \mathbb{1}_{[0,x]}(t)\, dt = x \wedge y$. Es folgt $k(x,\cdot) = R_x \in H_k$ und

$$\langle f, k(x,\cdot) \rangle_{H_k} = \int_0^1 f'(t) \mathbb{1}_{[0,x]}(t)\, dt = \int_0^x f'(t)\, dt = f(t).$$

Wir sehen an diesem Beispiel, dass die RKHS-Norm stark mit Glattheitseigenschaften der Funktionen verbunden ist. Ein analoges Resultat gilt für Sobolevräume höherer Ordnung.

3.19 Beispiel. Ist $(H_k, \langle \cdot, \cdot \rangle_{H_k})$ ein RKHS mit reproduzierendem Kern k, so gilt für alle $x \in \mathcal{X}$ und $f \in H_k$

$$|f(x)| = |\langle f, k(x,\cdot) \rangle_{H_k}| \leq \|f\|_{H_k} \|k(x,\cdot)\|_{H_k} = \sqrt{k(x,x)} \|f\|_{H_k},$$

wobei wir in den Gleichheiten die reproduzierende Eigenschaft und in der Ungleichung die Cauchy-Schwarz-Ungleichung verwendet haben. Die Punktevaluationen $L_x : H \to \mathbb{R}, f \mapsto f(x)$ sind also alle beschränkt (d.h. für alle $x \in \mathcal{X}$ existiert ein $M > 0$ mit $|L_x f| \leq M \|f\|$ für alle $f \in H_k$). Konvergiert nun (f_n) in H_k Norm gegen f, so gilt

$$|f(x) - f_n(x)| \leq \sqrt{k(x,x)} \|f - f_n\|_{H_k} \xrightarrow{n \to \infty} 0, \qquad \forall x \in \mathcal{X},$$

d.h. Konvergenz in Norm impliziert punktweise Konvergenz. Insbesondere kann der Hilbertraum $(L^2[0,1], \langle \cdot, \cdot \rangle_{L^2})$, wobei $\langle f, g \rangle_{L^2} = \int_0^1 f(x) g(x)\, dx$ kein RKHS sein kann.

Umgekehrt kann man sehen, dass man für jeden Hilbertraum H von Funktionen $f : \mathcal{X} \to \mathbb{R}$, für den die Punktevaluationen $L_x : \mathcal{H} \to \mathbb{R}, f \mapsto f(x)$ beschränkt sind, eine symmetrische, positiv definite Abbildung $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ konstruieren kann, so dass (a) und (b) aus Satz 3.16 erfüllt sind. In der Tat erhält man mit Hilfe des Rieszschen Darstellungssatzes (siehe Satz A.18) für alle $x \in \mathcal{X}$ ein $R_x \in H$ mit $L_x = \langle R_x, \cdot \rangle_H$. Setze nun $k(x,y) = \langle R_x, R_y \rangle = R_x(y) = R_y(x)$. Dann gilt $k(x,\cdot) \in H$ für alle $x \in \mathcal{X}$ und $\langle f, k(x,\cdot) \rangle_H = \langle f, R_x \rangle_H = L_x f = f(x)$ für alle $x \in \mathcal{X}$ und $f \in H$.

Beweis von Satz 3.16

Wir setzen

$$H_0 = \Big\{ f : \mathcal{X} \to \mathbb{R} : f = \sum_{i=1}^{m} a_i k(x_i, \cdot), m \in \mathbb{N}, x_i \in \mathcal{X}, a_i \in \mathbb{R}, i = 1, \ldots, m \Big\}.$$

Sind $f = \sum_{i=1}^{m} a_i k(x_i, \cdot)$ und $g = \sum_{j=1}^{n} b_j k(y_j, \cdot)$ aus H_0, so setzen wir

$$\langle f, g \rangle_{H_0} = \sum_{i=1}^{m} \sum_{j=1}^{n} a_i b_j k(x_i, y_j).$$

3.20 Lemma. *$(H_0, \langle \cdot, \cdot \rangle_{H_0})$ ist ein Innenproduktraum mit $f(x) = \langle f, k(x, \cdot) \rangle_{H_0}$ für alle $x \in \mathcal{X}$ und $f \in H_0$.*

Beweis. Es ist klar, dass H_0 ein Vektorraum ist. Desweiteren ist $\langle \cdot, \cdot \rangle_{H_0}$ wohldefiniert, da

$$\langle f, g \rangle_{H_0} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i b_j k(x_i, y_j) = \sum_{i=1}^{m} a_i g(x_i) = \sum_{j=1}^{m} b_j f(y_j). \tag{3.7}$$

Hieraus folgt, dass

$$\langle f, k(x, \cdot) \rangle_{H_0} = 1 \cdot f(x) = f(x)$$

für alle $x \in \mathcal{X}$ und $f \in H_0$. Es ist klar, dass $\langle \cdot, \cdot \rangle_{H_0}$ bilinear und symmetrisch ist und dass $\|f\|_{H_0} = 0$ falls $f = 0$. Insbesondere gilt die Cauchy-Schwarz-Ungleichung: $|\langle f, g \rangle_{H_0}| \leq \|f\|_{H_0} \|g\|_{H_0}$ für alle $f, g \in H_0$ (siehe Satz A.15). Setzen wir $g = k(x, \cdot)$, so folgt

$$|f(x)| = |\langle f, k(x, \cdot) \rangle_{H_0}| \leq \|f\|_{H_0} \|k(x, \cdot)\|_{H_0} = \sqrt{k(x, x)} \|f\|_{H_0}. \tag{3.8}$$

Somit impliziert $\|f\|_{H_0} = 0$, dass $f = 0$. □

Ein Innenproduktraum bzw. Prähilbertraum kann zu einem Hilbertraumk vervollständigt werden indem man H_k als den Raum aller Cauchyfolgen modulo aller Nullfolgen setzt. Wir wollen allerdings H_k als Raum von Funktionen $f : \mathcal{X} \to \mathbb{R}$ realisieren und müssen deshalb ein wenig Vorarbeit leisten.

3.21 Lemma. *Sei (f_n) eine Cauchyfolge aus H_0. Dann konvergiert die Folge $(f_n(x))$ für alle $x \in \mathcal{X}$.*

Beweis. Aus (3.8) folgt, dass für alle $n, m \in \mathbb{N}$

$$|f_n(x) - f_m(x)| \leq \sqrt{k(x, x)} \|f_n - f_m\|_{H_0}.$$

Daher ist für beliebiges x die Zahlenfolge $(f_n(x))$ eine Cauchyfolge und die Behauptung folgt aus der Vollständigkeit von $\mathbb{R}$. □

3.22 Lemma. *Eine Folge (f_n) ist eine Cauchyfolge in H_0 die punktweise gegen 0 konvergiert genau dann, wenn $\|f_n\|_{H_0} \to 0$ für $n \to \infty$.*

Beweis. Da Cauchyfolgen beschränkt sind, existiert ein $M > 0$ mit $\|f_n\|_{H_0} \leq M$ für alle $n \geq 1$. Wähle nun N so groß, dass $\|f_n - f_N\|_{H_0} \leq \epsilon/M$ für alle $n \geq N$. Da $f_N \in H_0$, gibt es $m \in \mathbb{N}, x_i \in S, a_i \in \mathbb{R}, i = 1, \ldots, m$, mit $f_N = \sum_{i=1}^m k(x_i, \cdot)$. Nun gilt

$$\begin{aligned}\|f_n\|_{H_0}^2 = \langle f_n, f_n\rangle_{H_0} &= \langle f_n - f_N, f_n\rangle_{H_0} + \langle f_N, f_n\rangle_{H_0} \\ &\overset{(3.7)}{=} \langle f_n - f_N, f_n\rangle_{H_0} + \sum_{i=1}^m a_i f_n(x_i) \\ &\leq \epsilon + \sum_{i=1}^m a_i f_n(x_i).\end{aligned}$$

Daher gilt $\limsup_{n\to\infty} \|f_n\|_{H_0}^2 \leq \epsilon$. Da $\epsilon > 0$ beliebig war, folgt die Hinrichtung. Für die Rückrichtung bemerken wir zunächst, dass jede Nullfolge auch eine Cauchyfolge ist. Des Weiteren gilt wegen (3.8)

$$|f(x)| \leq \sqrt{k(x,x)}\|f_n\|_{H_0} \xrightarrow{n\to\infty} 0.$$

Also impliziert die Konvergenz in H_0 Norm die punktweise Konvergenz. □

Sei nun H_k der Raum aller Funktionen $f : \mathcal{X} \to \mathbb{R}$ die punktweise Limiten von Cauchyfolgen (f_n) aus H_0 sind. Für $f, g \in H_k$ setze

$$\langle f, g\rangle_{H_k} = \lim_{n\to\infty} \langle f_n, g_n\rangle_{H_0},$$

wobei (f_n) und (g_n) Cauchyfolgen sind die punktweise gegen f und g konvergieren. Verwendet man die Lemmata 3.21 und 3.22, so sieht man, dass diese Konstruktion gerade mit dem Raum aller Cauchyfolgen modulo Nullfolgen übereinstimmt.

3.23 Lemma. *$(H_k, \langle\cdot,\cdot\rangle_{H_k})$ ist ein Hilbertraum.*

Beweis. Sind (f_n) und (g_n) zwei Cauchyfolgen in H_0, so ist $(\langle f_n, g_n\rangle_{H_0})$ eine Cauchyfolge in $\mathbb{R}$. Diese konvergiert da $\mathbb{R}$ vollständig ist. Seien nun (f_n') und (g_n') zwei weitere Cauchyfolgen welche punktweise gegen f und g konvergieren. Dann sind $(f_n - f_n')$ und $(g_n - g_n')$ Cauchyfolgen die punktweise gegen 0 konvergieren und Lemma 3.22 impliziert $\|f_n - f_n'\|_{H_0}, \|f_n - f_n'\|_{H_0} \to 0$ für $n \to 0$. Es folgt, dass

$$|\langle f_n, g_n\rangle_{H_0} - \langle f_n', g_n'\rangle_{H_0}| \leq \|f_n - f_n'\|_{H_0}\|g_n\|_{H_0} + \|f_n'\|_{H_0}\|g_n - g_n'\|_{H_0} \to 0$$

für $n \to 0$. Daher ist $\langle\cdot,\cdot\rangle_{H_k}$ wohldefiniert. Es ist nun einfach zu sehen, dass $\langle\cdot,\cdot\rangle_{H_k}$ ein Skalarprodukt auf H_k ist welches $\langle\cdot,\cdot\rangle_{H_0}$ erweitert. Gilt z.B. $\|f\|_{H_k} = 0$, d.h. $\lim_{n\to 0} \|f_n\|_{H_0} = 0$ mit (f_n) Cauchyfolge in H_0 die

punktweise gegen f konvergiert, so folgt $f = 0$ aus Lemma 3.22. Ist nun $f \in H_k$ und (f_n) Cauchyfolge in H_0 die punktweise gegen f konvergiert, so folgt aus der Konstruktion des Skalarproduktes, dass $\lim_{n\to\infty} \|f - f_n\|_{H_k} = 0$. Somit liegt H_0 dicht in H_k. Wir zeigen nun, dass H_k vollständig ist. Sei hierfür (f_n) eine Cauchyfolge in H_k. Aus der Dichtheit von H_0 folgt, dass es für alle $n \geq 1$ Funktionen $f_n' \in H_0$ gibt mit $\|f_n - f_n'\| < 1/n$. Dann ist (f_n') eine Cauchyfolge und es folgt aus Lemma 3.21, dass diese punktweise gegen ein $f \in H_k$ konvergiert. Es folgt $\lim_{n\to 0} \|f - f_n'\|_{H_k} = 0$ und somit $\lim_{n\to 0} \|f - f_n\|_{H_k} = 0$. □

Wir zeigen nun, dass der so konstruierte Hilbertraum H_k die Eigenschaften (a) und (b) erfüllt. Eigenschaft (a) ist klar. Um (b) zu zeigen seien $f \in H_k$, $x \in \mathcal{X}$ und (f_n) eine Cauchyfolge in H_0 die punktweise gegen f konvergiert. Dann konvergiert f_n auch in Norm gegen f und es folgt aus der Stetigkeit des Skalarprodukts, dass

$$f(x) = \lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} \langle f_n, k(x,\cdot)\rangle = \langle f, k(x,\cdot)\rangle.$$

Es bleibt die Eindeutigkeit zu zeigen. Sei also H_k' ein weiterer Hilbertraum von Funktionen $f : \mathcal{X} \to \mathbb{R}$ welcher (a) und (b) erfüllt. Dann ist $(H_0, \langle\cdot,\cdot\rangle_{H_0})$ ein Unterraum von H_k' und es folgt aus der Vollständigkeit, dass auch $(H_k, \langle\cdot,\cdot\rangle_{H_k})$ ein Unterraum von H_k' ist. Ist nun $f \in H_K'$ mit $f \perp H_k$, so gilt $f(x) = \langle f, k(x,\cdot)\rangle = 0$ für alle $x \in \mathcal{X}$ und wir erhalten $f = 0$. Also gilt $H_k' = H_k$. □

3.24 Bemerkung. Der in Satz 3.16 konstruierte RKHS hat folgende zusätzliche Eigenschaften.

(a) Ist $k(x,\cdot)$ messbar für alle $x \in \mathcal{X}$, so ist H_k ein Raum messbarer Funktionen. (Beweis hierfür: Nach Konstruktion besteht H_k aus Linearkombinationen von $k(x,\cdot)$ und (gewissen) punktweise Limiten von diesen.) Hieraus kann man auch schließen, dass die kanonische Featureabbildung $\Phi : \mathcal{X} \to H_k$ messbar ist.

(b) Ist $(\mathcal{X}, d)$ ein metrischer Raum, $k(x,\cdot)$ stetig für alle $x \in \mathcal{X}$ und k beschränkt, so ist H_k ein Raum stetiger Funktionen. (Beweis hierfür: Sei $f \in H_k$, $x \in \mathcal{X}$ und $\epsilon > 0$. Sei (f_n) eine Cauchyfolge in H_0 die punktweise gegen f konvergiert. Dann gibt es ein $m \geq 1$ mit $\|f - f_m\|_{H_k} < \epsilon/(3M)$ (siehe Beweis von Lemma 3.23) und $|f(x) - f_m(x)| < \epsilon/3$, wobei $M = \sup_{x\in\mathcal{X}} \sqrt{k(x,x)} < \infty$. Da f_m stetig ist, existiert ein $\delta > 0$, so dass $|f_m(x) - f_m(y)| < \epsilon/3$ für alle y mit $d(x,y) < \delta$. Für alle y mit $d(x,y) < \epsilon$ folgt somit

$$\begin{aligned} |f(x) - f(y)| &\leq |f(x) - f_m(x)| + |f_m(x) - f_m(y)| + |f_m(y) - f(y)| \\ &< \epsilon/3 + \epsilon/3 + \sqrt{k(x,x)}\|f - f_m\|_{H_k} < \epsilon \end{aligned}$$

und die Behauptung folgt.)

Ist $(\mathcal{X}, d)$ zusätzlich separabel, so ist H_k separabel. (Beweis hierfür: Sei $\mathcal{X}_0 \subseteq S$ dicht und abzählbar. Ist $f \perp \overline{\text{span}}(k(x,\cdot) : x \in \mathcal{X}_0)$, so gilt $f(x) = \langle f, k(x,\cdot)\rangle = 0$ für alle $x \in \mathcal{X}_0$. Da f stetig ist folgt $f(x) = 0$ für alle $x \in \mathcal{X}$.)

RKHS einer Featureabbildung

3.25 Beispiel. Sei $\mathcal{X} = \mathbb{R}^d$ und $k(x,y) = \langle x, y\rangle^2$. Setze

$$\Phi : \mathbb{R}^d \to \mathbb{R}^{\frac{(d+1)d}{2}}, \quad \Phi(x) = \begin{pmatrix} x_j^2 & 1 \le j \le d \\ \sqrt{2}x_i x_j & 1 \le i < j \le d \end{pmatrix}.$$

Dann gilt $k(x,y) = \langle \Phi(x), \Phi(y)\rangle$ für alle $x, y \in \mathcal{X}$. Somit ist Φ eine zugehörige Featureabbildung. Wir fragen uns wie der zu k gehörige RKHS aussieht. Man sieht zunächst leicht, dass wegen (a) aus Satz 3.16 alle Koordinatenabbildungen von Φ (d.h. die Funktionen $x \mapsto x_j$, $x \mapsto \sqrt{2}x_i x_j$) in H_k liegen müssen. Verwendet man nun (b), so sieht man, dass die Koordinatenabbildungen außerdem orthonormal sein müssen. Wir erhalten also, dass

$$H_k = \left\{ f_w = \langle w, \Phi(\cdot)\rangle : w \in \mathbb{R}^{\frac{(d+1)d}{2}} \right\}$$

versehen mit dem Skalarprodukt $\langle f_w, f_v\rangle_{H_k} = \langle w, v\rangle$, $v, w \in \mathbb{R}^{\frac{(d+1)d}{2}}$, der zu k gehörige RKHS ist. In der Tat ist $(H_k, \langle\cdot,\cdot\rangle_{H_k})$ ein Hilbertraum mit

$$k(x,\cdot) = \langle \Phi(x), \Phi(\cdot)\rangle = f_{\Phi(x)} \in H_k, \quad \forall x \in \mathbb{R}^d$$

und

$$\langle f_w, k(x,\cdot)\rangle_{H_k} = \langle f_w, f_{\Phi(x)}\rangle_{H_k} = \langle w, \Phi(x)\rangle = f_w(x), \quad \forall x \in \mathbb{R}^d, w \in \mathbb{R}^{\frac{(d+1)d}{2}}.$$

Allgemeiner kann man zeigen:

3.26 Satz. *Sei $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ ein Kern und $\Phi : \mathcal{X} \to H$ eine zugehörige Featureabbildung mit $k(x,y) = \langle \Phi(x), \Phi(y)\rangle$ für alle $x, y \in \mathcal{X}$. Dann ist*

$$H_k = \{f : \mathcal{X} \to \mathbb{R} : \exists w \in H \text{ mit } f(x) = \langle w, \Phi(x)\rangle \text{ für alle } x \in \mathcal{X}\}$$

versehen mit der Norm $\|f\|_{H_k} = \min\{\|w\| : w \in H \text{ mit } f(\cdot) = \langle w, \Phi(\cdot)\rangle\}$ der zu k gehörige RKHS.

Mit Hilfe von Satz 3.26 kann man die folgende Aufgabe lösen:

3.27 Aufgabe. *Sind H_1 und H_2 zwei RKHS mit reproduzierenden Kernen $k_1, k_2 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, so ist $H_1 + H_2$ ein RKHS mit reproduzierendem Kern $k_1 + k_2$. Bestimme den RKHS von $k_1 + k_2$ in dem Fall, dass $\mathcal{X} = [0,1]$, $k_1(x,y) = \min(x,y)$ und $k_2(x,y) = 1$.*

Man betrachtet hierfür die Featureabbildung $\Phi : h \to H_1 \oplus H_2, x \mapsto (k_1(x,\cdot), k_2(x,\cdot))$ und verwendet $\langle (f_1, f_2), \Phi(x)\rangle = f_1(x) + f_2(x)$.

Beweis von Satz 3.26. Betrachte die lineare Abbildung

$$L : H \to H_k, w \mapsto \langle w, \Phi(\cdot)\rangle.$$

Des Weiteren sei $\ker L = \{w \in H : Lw = 0\}$ der Kern von L. Man sieht zunächst leicht, dass $\ker L$ eine abgeschlossener Unterraum von H ist: für eine Folge (w_n) in $\ker L$ mit $w_n \to w \in H$ gilt $\langle w, \Phi(x)\rangle = \lim_{n\to\infty}\langle w_n, \Phi(x)\rangle$ für alle $x \in \mathcal{X}$, d.h. $w \in \ker L$. Sei nun $(\ker L)^\perp$ das orthogonale Komplement von $\ker L$ und

$$L_\perp : (\ker L)^\perp \to H_k, w \mapsto \langle w, \Phi(\cdot)\rangle,$$

die Einschränkung von L auf $(\ker L)^\perp$. Wir zeigen zunächst, dass $L_\perp$ surjektiv ist. Ist $f \in H_k$, so existiert nach Konstruktion ein $w \in H$ mit $f(\cdot) = \langle w, \Phi(\cdot)\rangle$. Aus Satz A.17 folgt, dass $w = w_0 + w_\perp$ mit $w_0 \in \ker L$ und $w_\perp \in (\ker L)^\perp$. Hieraus erhalten wir, dass $Lw = Lw_0 + Lw_\perp = Lw_\perp$ und somit die Surjektivität. Da $L_\perp$ außerdem injektiv und linear ist und $(\ker L)^\perp$ als abgeschlossener Unterraum von H selbst ein Hilbertraum ist (siehe Satz A.17), können wir H_k mit einer Hilbertraumstruktur versehen, indem wir

$$\langle f, g\rangle_{H_k} = \langle L_\perp^{-1} f, L_\perp^{-1} g\rangle, \qquad f, g \in H_k$$

setzen. Es gilt nun $\Phi(x) \in (\ker L)^\perp$ für alle $x \in \mathcal{X}$ und somit $k(x,\cdot) = \langle \Phi(x), \Phi(\cdot)\rangle \in H_k$ für alle $x \in \mathcal{X}$ und

$$\langle f, k(x,\cdot)\rangle_{H_k} = \langle L_\perp^{-1} f, \Phi(x)\rangle = f(x), \qquad \forall x \in \mathcal{X}, f \in H_k.$$

Es bleibt noch die Behauptung über die Norm zu zeigen, d.h. $\|f\|_{H_k} = \|L_\perp^{-1} f\|$ stimmt mit der Definition aus dem Satz überein. Sei hierfür $w \in H$ mit $f = \langle w, \Phi(\cdot)\rangle$. Dann gilt mit $w_\perp = L_\perp^{-1} f$, dass $w = w_\perp + w - w_\perp$ mit $w_\perp \in (\ker L)^\perp$ und $w - w_\perp \in \ker L$. Es folgt $\|w\|^2 = \|w_\perp\|^2 + \|w - w_\perp\|^2 = \|f\|_{H_k}^2 + \|w - w_\perp\|^2$. Also gilt $\|w\| \geq \|f\|_{H_k}$ und Gleichheit gilt für $w = w_\perp$. □

Der Generalisierungsfehler von kernel SVM

Sind nun $(X_1, Y_1), \cdots, (X_n, Y_n) \in \mathcal{X} \times \{\pm 1\}$ Daten, $\Phi : \mathcal{X} \to H$ eine Featureabbildung mit Kern $k(\cdot,\cdot) = \langle \Phi(\cdot), \Phi(\cdot)\rangle$ und H_k der zu k gehörige RKHS, so ist das Minimierungsproblem

$$\text{minimiere} \quad \frac{1}{n}\sum_{i=1}^{n}(1 - Y_i\langle w, \Phi(X_i)\rangle)_+ + \lambda\|w\|^2 \quad \text{über } w \in H$$

äquivalent zu

$$\text{minimiere} \quad \frac{1}{n}\sum_{i=1}^{n}(1 - Y_i f(X_i))_+ + \lambda\|f\|_{H_k}^2 \quad \text{über } f \in H_k \tag{3.9}$$

Wir sehen also, dass der RKHS der gewissermaßen kleinste Featureraum von k ist und somit eine kanonische Wahl darstellt. Betrachtet man anstelle von (3.9) für $\lambda' > 0$

$$\hat{h}_n^{k\text{-SVM}} = \operatorname{sign}(\hat{f}_n^{k\text{-SVM}})$$

mit

$$\hat{f}_n^{k\text{-SVM}} \in \operatorname*{argmin}_{\|f\|_{H_k} \le \lambda'} \frac{1}{n}\sum_{i=1}^{n}(1 - Y_i f(X_i))_+,$$

so erhält man für den Generalisierungsfehler von kernel-SVM unter der Annahme, dass $(X_1, Y_1), \ldots, (X_n, Y_n), (X, Y)$ i.i.d.

3.28 Satz. *Es gelte $k(x, x) \le M^2$ für alle $x \in \mathcal{X}$. Dann gilt*

$$\mathbb{E}R(\hat{h}_n^{k\text{-SVM}}) \le \min_{\|f\|_{H_k} \le \lambda'} R^{hinge}(f) + \frac{2\lambda' M}{\sqrt{n}}.$$

Beweis. Setzt man $f(x) = \langle f, k(x, \cdot)\rangle_{H_k}$, so ist der Beweis völlig analog zum Beweis von Satz 3.28, in dem man $(\mathbb{R}^p, \langle\cdot,\cdot\rangle)$ durch $(H_k, \langle\cdot,\cdot\rangle_{H_k})$ und X_i durch $k(X_i, \cdot)$ ersetzt, und am Schluss noch $\|k(X_i, \cdot)\|_{H_k} = \sqrt{k(X_i, X_i)} \le M$ verwendet. □

4 Hochdimensionale Statistik

4.1 Das dünn besetzte lineare Modell in hoher Dimension

Ziel des Kapitels ist ein kurzer Einblick in das dünn besetzte lineare Modell in hoher Dimension und dessen Bedeutung in der Statistik und im Compressed Sensing. Für $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$ und $\beta^* \in \mathbb{R}^p$ betrachten wir das lineare Gleichungssystem (LGS)

$$Y = X\beta^*$$

oder mit $\epsilon \in \mathbb{R}^n$ das lineare Modell

$$Y = X\beta^* + \epsilon.$$

Dabei sind Y, X bekannt und β^* unbekannt. Wir sind an dem Fall $p \ge n$ bzw. p sehr viel größer als n, d.h. $p \gg n$ interessiert. Wir wollen also im ersten Fall ein (hochgradig) unterbestimmtes LGS lösen. Um dies zu ermöglichen nehmen wir an, dass β^* dünn besetzt, d.h. sparse ist:

$$\|\beta^*\|_0 = |\{j : \beta_j^* \neq 0\}| \le s \qquad \text{mit} \qquad s \ll n.$$

Compressed Sensing

In diesem Fall heißt β^* oft auch Signal (zum Beispiel lassen sich Fotos mittels Waveletsystem oft sparse darstellen), X Messmatrix und $Y = X\beta^* = (\langle X_{i\cdot}, \beta^* \rangle)_{i=1}^n$ erfasstes Signal. Das Ziel ist es eine Messmatrix X mit möglichst kleinem n zu finden, so dass sich sparse Vektoren β^* aus Y rekonstruieren lassen.

4.1 Beispiel. Betrachte $f(t) = \sum_{j=-M}^{M} \beta_k^* e^{2\pi i k t}$, $t \in [0,1]$. Wir tasten das Signal f zu Zeitpunkten $t_1, \dots, t_n$ ab:

$$Y = \begin{pmatrix} f(t_1) \\ \vdots \\ f(t_n) \end{pmatrix} = \begin{pmatrix} e^{-2\pi i M t_1} & \dots & e^{2\pi i M t_1} \\ \vdots & & \vdots \\ e^{-2\pi i M t_n} & \dots & e^{2\pi i M t_n} \end{pmatrix} \beta^* = X\beta^*$$

mit $p = 2M+1$. Man kann leicht sehen, dass die Matrix X im Fall $n = 2M+1$ und $t_j = j/(2M+1)$, $j = 0, \dots, 2M$ invertierbar ist. Diese Tatsache ist mit Shannons Abtasttheorem verbunden, welches (unter anderem) besagt, dass

$$f(t) = \frac{1}{2M+1} \sum_{j=0}^{2M+1} f\Big(\frac{j}{2M+1}\Big) D_M\Big(t - \frac{j}{2M+1}\Big) \qquad t \in [0,1]$$

mit Dirichlet-Kern $D_m(t) = \sum_{k=-M}^{M} e^{2\pi i k t}$. Es reicht also f an $2M+1$ Punkten abzutasten um f exakt zu rekonstruieren. Im Compressed Sensing zeigt man hingegen, dass man im Fall $t_1, \dots, t_n \overset{i.i.d.}{\sim} \mathrm{Unif}[0,1]$ alle s-sparsen Signale β^* exakt rekonstruieren kann (mit hoher Wahrscheinlichkeit), falls $n \geq Cs\log(p)$.

Motivation aus der Statistik

In diesem Fall heißt Y_i Antwortvariable und $X_{i\cdot}$ zugehöriger Kovariatenvektor. In vielen aktuellen Anwendungen ist $X_{i\cdot}$ ein hoch-dimensionaler Vektor mit $p \gg n$, von dem die meisten Einträge keinen oder nur einen sehr geringen Einfluss auf Y_i haben. Zum Beispiel kann Y_i ein Merkmal der i-ten Pflanze sein (z.B. deren Größe) und $X_{i\cdot}$ verschiedene Genexpressionen der i-ten Pflanze.

Notationen

Wir verwenden im Folgenden für $\beta \in \mathbb{R}^p$,

$$\|\beta\|_0 = |\{j : \beta_j \neq 0\}|, \quad \|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|, \quad \|\beta\|_2 = \Big(\sum_{j=1}^{p} \beta_j^2\Big)^{1/2}$$

und

$$\|\beta\|_\infty = \max_{1 \leq j \leq p} |\beta_j|.$$

Die Menge $\{j : \beta_j \neq 0\}$ heißt der Träger von β^*. Ist $S \subseteq \{1, \dots, p\}$ und $\beta \in \mathbb{R}^p$ so definieren wir $\beta_S \in \mathbb{R}^p$ durch

$$(\beta_S)_j = \begin{cases} \beta_j, & j \in S, \\ 0, & \text{sonst.} \end{cases}$$

4.2 Rekonstruktion mittels ℓ^1-Minimierung

In diesem Kapitel betrachten wir zunächst das LGS $Y = X\beta^*$. Eine erste Lösungsstrategie besteht in

$$\text{minimiere} \quad \|\beta\|_0 \qquad \text{unter der NB} \quad Y = X\beta.$$

Dies ist allerdings im Allgemeinen ein NP-schweres Problem (löse $Y = X_S\beta$ mit $\beta \in \mathbb{R}^{|S|}$ für $|S| = 1, 2, \dots, s$). Ein Ausweg besteht darin $\|\cdot\|_0$ durch $\|\cdot\|_1$ als naheste konvexe Norm zu ersetzen:

$$\text{minimiere} \quad \|\beta\|_1 \qquad \text{unter der NB} \quad Y = X\beta. \tag{4.1}$$

4.2 Definition. Die Matrix X besitzt die Nullraumeigenschaft (null space property) bezüglich $S \subseteq \{1, \dots, p\}$ falls

$$\{\Delta \in \mathbb{R}^p : \|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1\} \cap \ker X = \{0\},$$

wobei $\ker X = \{\Delta \in \mathbb{R}^p : X\Delta = 0\}$.

4.3 Satz. *Für $S \subseteq \{1, \dots, p\}$ sind die folgenden Eigenschaften äquivalent:*

(a) Für alle $\beta^* \in \mathbb{R}^p$ mit Träger enthalten in S hat die ℓ^1-Minimierung in (4.1) *angewendet auf $Y = X\beta^*$ die eindeutige Lösung $\hat{\beta} = \beta^*$.*

(b) X erfüllt die Nullraumeigenschaft.

Beweis. $(b) \Rightarrow (a)$: Die Vektoren $\hat{\beta}$ und β^* erfüllen $Y = X\hat{\beta} = X\beta^*$. Also folgt $\|\hat{\beta}\|_1 \leq \|\beta^*\|_1$ und somit für $\hat{\Delta} = \hat{\beta} - \beta^*$

$$\begin{aligned} \|\beta^*_S\|_1 = \|\beta^*\|_1 \geq \|\hat{\beta}\|_1 = \|\beta^* + \hat{\Delta}\|_1 &= \|\beta^*_S + \hat{\Delta}_S\|_1 + \|\hat{\Delta}_{S^c}\|_1 \\ &\geq \|\beta^*_S\|_1 - \|\hat{\Delta}_S\|_1 + \|\hat{\Delta}_{S^c}\|_1. \end{aligned}$$

Es gilt also $\hat{\Delta} \in \ker X$ und $\|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1$. Die Nullraumeigenschaft liefert $\hat{\Delta} = 0$, d.h. $\hat{\beta} = \beta^*$.

$(a) \Rightarrow (b)$: Für $\Delta \in \ker X$ wende (4.1) auf $Y = X\Delta_S$ an. Dann gilt wegen $X\Delta = X\Delta_S + X\Delta_{S^c} = 0$, dass $Y = X\Delta_S = X(-\Delta_{S^c})$ und es folgt $\|\Delta_S\|_1 < \|\Delta_{S^c}\|_1$ aus der Eindeutigkeit in (a). □

4.4 Definition. Die Matrix X besitzt die Restricted Isometry Property (RIP) der Ordnung s, falls es ein $\delta_s \in (0, 1)$ gibt, so dass

$$(1 - \delta_s)\|\beta\|_2^2 \leq \|X\beta\|_2^2 \leq (1 + \delta_s)\|\beta\|_2^2 \qquad \forall \beta \in \mathbb{R}^p, \|\beta\|_0 \leq s.$$

4.5 Bemerkung. Sind $\alpha, \beta \in \mathbb{R}^p$ s-sparse mit disjunktem Träger, so gilt unter RIP auch

$$|\langle X\alpha, X\beta\rangle| \leq \delta_{2s}\|\alpha\|_2\|\beta\|_2.$$

Um dies zu sehen, reicht es den Fall $\|\alpha\|_2 = 1$ und $\|\beta\|_2 = 1$ zu betrachten. Dann gilt

$$\begin{aligned} 4|\langle X\alpha, X\beta\rangle| &= |\|X(\alpha+\beta)\|_2^2 - \|X(\alpha-\beta)\|_2^2| \\ &= |\|X(\alpha+\beta)\|_2^2 - 2 + 2 - \|X(\alpha-\beta)\|_2^2| \leq 2\delta_{2s} + 2\delta_{2s}. \end{aligned}$$

4.6 Proposition. *Erfüllt X die RIP der Ordnung $2s$ mit $\delta_s < 1/3$, so erfüllt X die Nullraumeigenschaft für alle $S \subseteq \{1, \dots, p\}$ mit $|S| \leq s$.*

Insbesondere können in diesem Fall alle s-sparsen Vektoren $\beta^* \in \mathbb{R}^p$ mittels ℓ^1-Minimierung (4.1) aus $Y = X\beta^*$ rekonstruiert werden.

Beweis. Sei $\Delta \in \ker X$ und S_0 die Menge der Indizes der s größten Einträge von Δ. Es reicht zu zeigen, dass

$$\|\Delta_{S_0}\|_1 < \|\Delta_{S_0^c}\|_1.$$

Konstruiere nun weiter $S_0, S_1, S_2, \dots$ mit $|S_j| = |s|$ ($\leq s$ für die letzte Menge) und $|\Delta_b| \leq |\Delta_a|$ für alle $a \in S_j$ und $b \in S_{j+1}$. Es gilt also $\Delta = \Delta_{S_0} + \sum_{j\geq 1} \Delta_{S_j}$. Verwenden wir sukzessive die RIP, die Identität $X\Delta_{S_0} = -\sum_{j\geq 1} X\Delta_{S_j}$, die Dreicksungleichung und Bemerkung 4.5, so folgt

$$\begin{aligned} \|\Delta_{S_0}\|_2^2 \leq \frac{1}{1-\delta_{2s}}\|X\Delta_{S_0}\|_2^2 &= \frac{1}{1-\delta_{2s}}|\langle X\Delta_{S_0}, \sum_{j\geq 1} X\Delta_{S_j}\rangle| \\ &\leq \frac{1}{1-\delta_{2s}} \sum_{j\geq 1} |\langle X\Delta_{S_0}, X\Delta_{S_j}\rangle| \\ &\leq \frac{\delta_{2s}}{1-\delta_{2s}} \sum_{j\geq 1} \|\Delta_{S_0}\|_2\|\Delta_{S_j}\|_2 \end{aligned}$$

und somit

$$\|\Delta_{S_0}\|_2 \leq \frac{\delta_{2s}}{1-\delta_{2s}} \sum_{j\geq 1} \|\Delta_{S_j}\|_2.$$

Aus der Konstruktion der S_j folgt, dass $\|\Delta_{S_{j+1}}\|_\infty \leq s^{-1}\|\Delta_{S_j}\|_1$ und daher $\|\Delta_{S_{j+1}}\|_2 \leq s^{-1/2}\|\Delta_{S_j}\|_1$. Wir schließen

$$\|\Delta_{S_0}\|_1 \leq \sqrt{s}\|\Delta_{S_0}\|_2 \leq \frac{\delta_{2s}}{1-\delta_{2s}}\sqrt{s}\sum_{j\geq 1}\|\Delta_{S_j}\|_2 \leq \frac{\delta_{2s}}{1-\delta_{2s}}(\|\Delta_{S_0}\|_1 + \sum_{j\geq 1}\|\Delta_{S_j}\|_1).$$

und somit

$$\|\Delta_{S_0}\|_1 \leq \frac{\delta_{2s}}{1-2\delta_{2s}}\|\Delta_{S_0^c}\|_1.$$

Die Behauptung folgt aus der Tatsache, dass $\delta_{2s}/(1-2\delta_{2s}) < 1$ genau dann, wenn $\delta_{2s} < 1/3$. □

4.3 RIP für Zufallsmatrizen

4.7 Satz. *Betrachte* $X = (n^{-1/2} g_{ij}) \in \mathbb{R}^{p \times n}$ *mit* $g_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. *Sei* $\delta \in (0,1)$. *Dann existieren Konstanten* $c_1, c_2 > 0$ *die nur von* δ *abhängen, so dass* X *die RIP der Ordnung* s *mit Konstante* δ *mit Wahrscheinlichkeit mindestens* $1 - e^{-c_2 n}$ *erfüllt falls* $s \log(ep) \leq c_1 n$.

4.8 Bemerkungen.

1) Analoge Resultate gelten für andere Zufallsvariablen, z.B. $g_{ij} \overset{i.i.d.}{\sim}$ Rademacher.

2) Warum Zufallsmatrizen? Für X aus Satz 4.7 gilt zum Beispiel

$$\mathbb{E}\|X\beta\|_2^2 = \mathbb{E}\frac{1}{n}\sum_{i=1}^{n}\Big(\sum_{j=1}^{p} g_{ij}\beta_j\Big)^2 = \|\beta\|^2.$$

   Die RIP kann also als Konzentrationsresultat interpretiert werden.

3) Was das Lösen eines sparse LGS betrifft, so ist die Bedingung $s \leq n$ offensichtlich notwendig für die Existenz einer eindeutigen Lösung. Der zusätzliche $\log p$ Term ist der Preis den wir zahlen dafür dass wir die positiven Koordinaten nicht kennen.

4.9 Lemma. *Für* $X = (n^{-1/2} g_{ij}) \in \mathbb{R}^{p \times n}$ *mit* $g_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ *und* $\delta \in (0,1)$ *gilt*

$$\mathbb{P}\big((1-\delta)\|\beta\|_2^2 \leq \|X\beta\|_2^2 \leq (1+\delta)\|\beta\|_2^2\big) \geq 1 - 2e^{-n\delta^2/8}$$

für alle $\beta \in \mathbb{R}^p$.

Beweis. Sei ohne Einschränkung $\|\beta\|_2 = 1$. Dann gilt

$$\|X\beta\|_2^2 = \frac{1}{n}\sum_{i=1}^{n}\Big(\sum_{j=1}^{p} g_{ij}\beta_j\Big)^2 \sim \frac{1}{n}\chi^2(n)$$

Daher folgt die Behauptung aus standard Konzentrationsungleichungen für χ^2 Verteilungen wie sie im ersten Teil der Vorlesung hergeleitet wurden. □

4.10 Lemma (Überdeckungsargument)**.** *Sei* $B = \{\beta \in \mathbb{R}^s : \|\beta\|_2 \leq 1\}$ *die abgeschlossene Einheitskugel im* $\mathbb{R}^s$ *und sei* $\delta \in (0,1)$. *Dann existiert* $Q \subseteq B$, $|Q| \leq (3/\delta)^s$ *mit der Eigenschaft*

$$\forall \beta \in B \quad \exists q \in Q \quad \textit{mit} \quad \|\beta - q\|_2 \leq \delta.$$

Beweis. Der Beweis beruht auf einem Volumenvergleichsargument. Seien $q_1, \dots, q_M \in B$ mit $\|q_j - q_k\|_2 > \delta$ für alle $j \neq k$ und M maximal. Dann gilt existiert für jedes $\beta \in \mathbb{R}^s$ ein $j \leq M$ mit $\|\beta - q_j\|_2 \leq \delta$. Wir müssen also noch $M \leq (3/\delta)^s$ zeigen. Aus der Konstruktion folgt, dass die Kugeln $B(q_j, \delta) = \{\beta \in \mathbb{R}^s : \|\beta - q_j\|_2 \leq \delta\}$ disjunkt und es gilt $\bigcup_{j=1}^M B(q_j, \delta/2) \subseteq B(0, 1 + \delta/2)$. Wir erhalten also, dass

$$M(\delta/2)^s \operatorname{vol}(B) \leq (1 + \delta/2)^s \operatorname{vol}(B)$$

d.h.

$$M \leq \left(\frac{1 + \delta/2}{\delta/2}\right)^s = \left(\frac{2 + \delta}{\delta}\right)^s \leq \left(\frac{3}{\delta}\right)^s,$$

wobei wir $\delta \leq 1$ in der letzten Ungleichung verwendet haben. □

4.11 Lemma. *Für $S \subseteq \{1, \dots, p\}$ mit $|S| \leq s$ und $\delta \in (0, 1)$ gilt*

$$\begin{aligned}&\mathbb{P}\big((1 - 2\delta)\|\beta\|_2^2 \leq \|X\beta\|_2^2 \leq (1 + 3\delta)\|\beta\|_2^2, \forall \beta \text{ mit Träger in } S\big)\\ &\geq 1 - 2\Big(\frac{12}{\delta}\Big)^s e^{-n\delta^2/32}\end{aligned}$$

Beweis. Sei $X_S \in \mathbb{R}^{n \times s}$ die Matrix der Spalten von X Index aus S. Dann reicht es zu zeigen, dass mit Wahrscheinlichkeit mindestens $1 - 2(12/\delta)^s e^{-n\delta^2/32}$,

$$1 - 2\delta \leq \|X\beta\|_2^2 \leq 1 + 3\delta \qquad \forall \beta \in \mathbb{R}^s, \|\beta\|_2 = 1.$$

Wähle nun Q aus Lemma 4.10 mit $\delta/4$. Dann folgt aus Lemma 4.9 und der Bonferroni-Ungleichung, dass mit Wahrscheinlichkeit mindestens $1 - 2(12/\delta)^s e^{-n\delta^2/32}$,

$$(1 - \delta/2)\|q\|_2^2 \leq \|Xq\|_2^2 \leq (1 + \delta/2)\|q\|_2^2 \qquad \forall q \in Q. \tag{4.2}$$

Wir nehmen nun an, dass (4.2) gilt und setzen $A > 0$ als die kleinste Zahl mit

$$\|X\beta\|_2 \leq 1 + A \qquad \forall \beta \in \mathbb{R}^s, \|\beta\|_2 = 1.$$

Wir behaupten, dass $A \leq \delta$. Um dies zu sehen betrachte $\beta \in \mathbb{R}^s$, $\|\beta\|_2 = 1$, und $q \in Q$ mit $\|\beta - q\|_2 \leq \delta/4$. Dann gilt

$$\|X\beta\|_2 \leq \|Xq\|_2 + \|X(\beta - q)\|_2 \leq 1 + \delta/2 + \delta/4 + A\delta/2,$$

wobei wir die Dreiecksungleichung, (4.2), $(1 + \delta/2)^{1/2}\|q\|_2 \leq 1 + \delta/2$ und die Definition von A verwendet haben. Es folgt $A \leq \delta/2 + \delta/4 + A\delta/2$ und somit

$$A \leq \frac{\delta/2 + \delta/4}{1 - \delta/4} = \frac{3\delta}{4 - \delta} \leq \delta.$$

Analog gilt $\|X\beta\|_2 \geq 1-\delta$. Wir schließen, dass mit Wahrscheinlichkeit mindestens $1 - 2(12/\delta)^s e^{-n\delta^2/32}$,

$$1 - \delta \leq \|X\beta\|_2 \leq 1 + \delta \qquad \forall \beta \in \mathbb{R}^s, \|\beta\|_2 = 1$$

und die Behauptung folgt durch Quadrieren. □

Ende des Beweis des Satzes 4.7. Aus Lemma 4.11 und der Bonferroni-Ungleichung folgt, dass

$$\begin{aligned}&\mathbb{P}((1-2\delta)\|\beta\|_2^2 \leq \|X\beta\|_2^2 \leq (1+3\delta)\|\beta\|_2^2, \forall \beta \in \mathbb{R}^p, \|\beta\|_0 \leq s)\\ &\geq 1 - 2\binom{p}{s}\Big(\frac{12}{\delta}\Big)^s e^{-n\delta^2/32}.\end{aligned}$$

Es gilt und

$$2\binom{p}{s}\Big(\frac{12}{\delta}\Big)^s e^{-n\delta^2/32} \leq \exp\Big(s\log(p) + \log(2) + s\log\Big(\frac{12}{\delta}\Big) - \frac{n\delta^2}{64} - \frac{n\delta^2}{64}\Big)$$

und die Behauptung folgt indem man

$$s\log(p) + \log(2) + s\log\Big(\frac{12}{\delta}\Big) - \frac{n\delta^2}{64} \leq 0$$

fordert. □

4.4 Lasso

Im Fall des linearen Modells $Y = X\beta^* + \epsilon$ kann die ℓ^1-Minimierung auf verschiedene Art und Weise erweitert werden:

$$\text{minimiere} \quad \|\beta\|_0 \qquad \text{unter der NB} \quad \|Y - X\beta\|_2 \leq b.$$

Dabei ist b eine obere Schranke für das Fehlerniveau. Man kann alternativ für $R > 0$ auch folgendes betrachten:

$$\text{minimiere} \quad \|Y - X\beta\|_2 \qquad \text{unter der NB} \quad \|\beta\|_0 \leq R. \tag{4.3}$$

Man spricht auch vom constrained Lasso. Der Lasso hingegen ist die Lösung der folgenden äquivalenten Problems

$$\text{minimiere} \quad \|Y - X\beta\|_2 + \lambda\|\beta\|_0 \qquad \text{über} \quad \beta \in \mathbb{R}^p,$$

wobei λ eine Konstante ist.

4.12 Definition. Die Matrix X erfüllt die Restricted Eigenvalue Condition bezüglich $S \subseteq \{1, \ldots, p\}$ mit Parameter $\kappa > 0$, falls

$$\|X\Delta\|_2^2 \geq \kappa\|\Delta\|_2^2 \qquad \text{für alle } \Delta \text{ mit } \|\Delta_{S^c}\|_1 < \|\Delta_S\|_1.$$

Offensichtlich ist die Restricted Eigenvalue Condition eine Verstärkung der Nullraumeigenschaft.

4.13 Satz. *Es gelte die Restricted Eigenvalue Condition bezüglich $S \subseteq \{1,\dots,p\}$ mit $\kappa > 0$. Des Weiteren sei der Träger von β^* in S enthalten. Dann erfüllt jede Lösung $\hat\beta$ von (4.3) mit $R = \|\beta^*\|_1$ die folgenden Ungleichungen:*

(a) $\|\hat\beta - \beta^*\|_2 \le \frac{4}{\kappa}\sqrt{s}\|X^T\epsilon\|_\infty$

(b) $\|X(\hat\beta - \beta^*)\|_2 \le \frac{4}{\sqrt{\kappa}}\sqrt{s}\|X^T\epsilon\|_\infty$

Sind insbesondere $\epsilon_1,\dots,\epsilon_n \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma^2)$ und gilt $\max_j \|X_{\cdot j}\|_2 \le 1$, so erhalten wir mit Wahrscheinlichkeit mindestens $1-e^{-u}$, $u>0$,

$$\|\hat\beta - \beta^*\|_2 \le \frac{4\sigma}{\kappa}\sqrt{s}\sqrt{2\log(2s)+2u}.$$

Beweis. Wegen $R = \|\beta^*\|_1$ erfüllen sowohl β^* als auch $\hat\beta$ die Nebenbedingung in (4.3). Es folgt also, dass $\|Y - X\hat\beta\|_2^2 \le \|Y - X\beta^*\|_2 = \|\epsilon\|_2^2$. Mit $\hat\Delta = \hat\beta - \beta^*$ gilt also $\|\epsilon - X\hat\Delta\|_1^2 \le \|\epsilon\|_2^2$, oder äquivalent

$$\|X\hat\Delta\|_2^2 \le 2\langle\epsilon, X\hat\Delta\rangle.$$

Also

$$\|X\hat\Delta\|_2 \le 2|\langle\epsilon, X\hat\Delta\rangle| \le 2|\langle X^T\epsilon, \hat\Delta\rangle| \le 2\|X^T\epsilon\|_\infty\|\hat\Delta\|_1.$$

Wegen $\|\hat\beta\|_1 \le \|\beta^*\|_1$ folgt wie im Beweis von Satz 4.3, dass $\|\hat\Delta_{S^c}\|_1 \le \|\hat\Delta_S\|_1$. Wir können also die Restricted Eigenvalue Condition anwenden und erhalten

$$\begin{aligned}\kappa\|\hat\Delta\|_2^2 &\le \|X\hat\Delta\|_2^2 \le 2\|X^T\epsilon\|_\infty\|\hat\Delta\|_1 \le 2\|X^T\epsilon\|_\infty(\|\hat\Delta_{S^c}\|_1 + \|\hat\Delta_S\|_1)\\ &\le 4\|X^T\epsilon\|_\infty\|\hat\Delta_S\|_1 \le 4\sqrt{s}\|X^T\epsilon\|_\infty\|\hat\Delta_S\|_2 \le 4\sqrt{s}\|X^T\epsilon\|_\infty\|\hat\Delta\|_2.\end{aligned}$$

Teilen durch $\|\hat\Delta\|_2$ liefert (a). Die Ungleichung in (b) folgt analog in dem wir in der obigen Kette an Ungleichungen die RIP am Ende anstatt am Anfang anwenden. Um den Zusatz zu sehen, setze $\sigma_j^2 = \sigma^2\|X_{\cdot j}\|_2^2 \le \sigma^2$. Dann gilt unter Verwendung der Bonferroni-Ungleichung

$$\begin{aligned}\mathbb{P}(\|X^T\epsilon\|_\infty > t) &= \mathbb{P}(\max_{j=1,\dots,p}|\langle X_{\cdot j},\epsilon\rangle| > t)\\ &\le \sum_{j=1}^p \mathbb{P}(|\langle X_{\cdot j},\epsilon\rangle| > t) = \sum_{j=1}^p \mathbb{P}(|\sigma_j Z| > t) = 2\sum_{j=1}^p \mathbb{P}(\sigma_j Z > t)\end{aligned}$$

mit $Z \sim \mathcal{N}(0,1)$. Setzen wir nun

$$\mathbb{P}(\sigma_j Z > t) \le \mathbb{E}e^{\sigma_j\lambda Z - \lambda t} = e^{\sigma_j^2\lambda^2/2 - \lambda t} \le e^{\sigma^2\lambda^2/2-\lambda t} \overset{\lambda = t/\sigma^2}{=} e^{-t^2/2\sigma^2}$$

ein, so folgt die Behauptung aus der Wahl $t = \sigma\sqrt{2\log(2p)+2u}$. □

A Anhang

A.1 Rademacher-Komplexitäten

A.1 Lemma (Symmetrisierungstrick)**.** *Seien $Z_1, \dots, Z_n$ i.i.d. Zufallsvariablen mit Werten in $\mathcal{Z}$ und $\mathcal{F}$ eine (abzählbare) Menge von gleichmäßig beschränkten Funktionen $f : \mathcal{Z} \to \mathbb{R}$. Des Weiteren seien $\epsilon_1, \dots, \epsilon_n$ unabhängige Rademacher Zufallsvariablen (d.h. $\mathbb{P}(\epsilon_i = +1) = \mathbb{P}(\epsilon_i = -1) = 1/2$) unabhängig von $Z_1, \dots, Z_n$. Dann gilt*

$$\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} f(Z_i) - \mathbb{E} f(Z_i) \leq 2\, \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(Z_i).$$

Beweis. Sei $(Z_1', \dots, Z_n')$ eine unabhängige Kopie von $(Z_1, \dots, Z_n)$ (d.h. $Z_1, \dots, Z_n, Z_1', \dots, Z_n'$ i.i.d., man spricht von einer Phantom-Stichprobe (ghost sample)). Dann sind die Zufallsvariablen $f(Z_i) - f(Z_i')$ unabhängig, symmetrisch und haben daher die gleiche Verteilung wie $\epsilon_i(f(Z_i) - f(Z_i'))$. Es folgt

$$\begin{aligned}
\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} f(Z_i) - \mathbb{E} f(Z_i) &= \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} f(Z_i) - \mathbb{E}\, f(Z_i') \\
&\leq \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} f(Z_i) - f(Z_i') \\
&= \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i (f(Z_i) - f(Z_i')) \\
&\leq \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(Z_i) + \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} -\epsilon_i f(Z_i') \\
&= 2\, \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(Z_i),
\end{aligned}$$

wobei die erste Ungleichung mit Hilfe der Jensenschen Ungleichung folgt. Alternativ kann man auch verwenden, dass für $a \in \mathbb{R}$ und $U_f = (1/n) \sum_{i=1}^{n} f(Z_i')$ folgendes gilt: $\sup_f (a - \mathbb{E}\, U_f) \leq \sup_f \mathbb{E}(a - U_f) \leq \mathbb{E} \sup_f (a - U_f)$. □

A.2 Definition. Seien $\epsilon_1, \dots, \epsilon_n$ unabhängige Rademacher Zufallsvariablen, $z_1, \dots, z_n \in \mathcal{Z}$ feste Elemente und $\mathcal{F}$ eine (abzählbare) Menge von gleichmäßig beschränkten Funktionen $f : \mathcal{Z} \to \mathbb{R}$. Dann heißt

$$R_n(\mathcal{F}) = R_{n,z}(\mathcal{F}) = \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(z_i)$$

Rademacher-Komplexität von $\mathcal{F}$. Ist allgemeiner $T \subseteq \mathbb{R}^n$, so heißt

$$R_n(T) = \mathbb{E} \sup_{t \in T} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i t_i$$

Rademacher-Komplexität von T.

A.3 Lemma (Kontraktions-Prinzip)**.** *Sei $T \subseteq \mathbb{R}^n$ beschränkt und $\varphi : \mathbb{R} \to \mathbb{R}$ L-Lipschitz-stetig. Weiter seien $\epsilon_1, \dots, \epsilon_n$ unabhängige Rademacher Zufallsvariablen. Dann gilt $R_n(\varphi(T)) \leq R_n(T)$, d.h.*

$$\mathbb{E} \sup_{t \in T} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \varphi(t_i) \leq L \mathbb{E} \sup_{t \in T} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i t_i.$$

Beweis. Wir nehmen o.B.d.A. an, dass $L = 1$. Setzen wir $T_i = \{(t_1, \dots, t_{i-1}, \varphi(t_i), t_{i+1}, \dots, t_n) : t \in T\}$, so reicht es die Aussage für alle T, T_i zu zeigen, d.h.

$$R_n(T_i) \leq R_n(T).$$

Betrachte hierfür $i = 1$, die anderen Fälle folgen analog. Es gilt nun

$$\begin{aligned}
nR_n(T_i) &= \mathbb{E} \sup_{t \in T} \Big\{ \epsilon_1 \varphi(t_1) + \sum_{i=2}^{n} \epsilon_i t_i \Big\} \\
&= \mathbb{E}_{\epsilon_n} \dots \mathbb{E}_{\epsilon_1} \sup_{t \in T} \Big\{ \epsilon_1 \varphi(t_1) + \sum_{i=2}^{n} \epsilon_i t_i \Big\} \\
&= \frac{1}{2} \mathbb{E}_{\epsilon_n} \dots \mathbb{E}_{\epsilon_2} \sup_{t \in T} \Big\{ \varphi(t_1) + \sum_{i=2}^{n} \epsilon_i t_i \Big\} + \sup_{t \in T} \Big\{ -\varphi(t_1) + \sum_{i=2}^{n} \epsilon_i t_i \Big\} \\
&= \frac{1}{2} \mathbb{E}_{\epsilon_n} \dots \mathbb{E}_{\epsilon_2} \sup_{t, t' \in T} \Big\{ \varphi(t_1) - \varphi(t_1') + \sum_{i=2}^{n} \epsilon_i t_i + \sum_{i=2}^{n} \epsilon_i t_i' \Big\} \\
&\leq \frac{1}{2} \mathbb{E}_{\epsilon_n} \dots \mathbb{E}_{\epsilon_2} \sup_{t, t' \in T} \Big\{ |t_1 - t_1'| + \sum_{i=2}^{n} \epsilon_i t_i + \sum_{i=2}^{n} \epsilon_i t_i' \Big\} \\
&= \frac{1}{2} \mathbb{E}_{\epsilon_n} \dots \mathbb{E}_{\epsilon_2} \sup_{t, t' \in T} \Big\{ t_1 - t_1' + \sum_{i=2}^{n} \epsilon_i t_i + \sum_{i=2}^{n} \epsilon_i t_i' \Big\}
\end{aligned}$$

Folgen wir der gleichen Kette an Argumenten rückwärts, so erhalten wir, dass der letzte Ausdruck gleich ist mit

$$\mathbb{E} \sup_{t \in T} \Big\{ \epsilon_1 t_1 + \sum_{i=2}^{n} \epsilon_i t_i \Big\} = R_n(T)$$

und die Behauptung folgt. □

A.4 Aufgabe. *Für $T = \{t_1, \dots, t_M\}$ gilt*

$$R_n(T) \leq \max_{i=1,\dots,n} \|t_i\| \frac{\sqrt{2 \log M}}{n}.$$

A.2 Das Subdifferential einer konvexen Funktion

Wollen wir das Minimum einer stetig differenzierbaren Funktion $f : \mathbb{R}^n \to \mathbb{R}$ finden, so besagt ein Standardresultat aus der Analysis, dass ein solches (sofern es existiert) in der Menge $\{x : \nabla f(x) = 0\}$ zu finden ist. Ein ähnliches Kritertium exisitiert im Fall konvexer (nicht notwendigerweise differenzierbarer) Funktionen unter Verwendung des Subdifferentials. In diesem Anhang führen wir das Subdifferential ein und geben einige Rechenregeln an mit dem sich das Subdifferential in vielen Fällen effektiv berechnen lässt. Insbesondere können diese Rechenregeln im Fall des LASSO-Schätzers und des SVM-Klassifizierers (vgl. Aufgabe 10.3) einfach angewendet werden. Interessierte Leser können weitere Informationen in [6] oder [4] finden.

Konvexe Funktionen

A.5 Definition. Eine Menge $C \subseteq \mathbb{R}^n$ heißt konvex, falls $\lambda x + (1-\lambda)y \in C$ für alle $x, y \in C$ und alle $\lambda \in [0,1]$.

A.6 Definition. Eine Funktion $f : \mathbb{R}^n \to \mathbb{R}$ heißt konvex, falls

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

für alle $x, y \in \mathbb{R}^n$ und alle $\lambda \in [0,1]$.

Es folgt direkt aus den Definitionen, dass eine Funktion $f : \mathbb{R}^n \to \mathbb{R}$ genau dann konvex ist wenn der Epigraph

$$\operatorname{epi} f = \{(x, r) \in \mathbb{R}^n \times \mathbb{R} : r \geq f(x)\}$$

eine konvexe Menge ist. Was Optimierung betrifft, besitzen konvexe Funktionen gute Eigenschaften. Zum einen ist eine lokale Minimalstelle einer konvexen Funktion $f : \mathbb{R}^n \to \mathbb{R}$ immer auch ein globale Minimalstelle (siehe Proposition 3.1 in [6]), zum anderen ist eine konvexe Funktion $f : \mathbb{R}^n \to \mathbb{R}$ immer stetig (siehe Proposition 3.3. in [6]).

Das Subdifferential einer konvexen Funktion

A.7 Definition. Für eine konvexe Funktion $f : \mathbb{R}^n \to \mathbb{R}$ ist das Subdifferential von f and der Stelle x definiert durch

$$\partial f(x) = \{w \in \mathbb{R}^n : f(y) \geq f(x) + \langle w, y - x\rangle \text{ für alle } y \in \mathbb{R}^n\}.$$

Die Wichtigkeit der Definition zeigt sich in der folgenden Charakterisierung von Minimalstellen.

A.8 Lemma. *Für jede konvexe Funktion $f : \mathbb{R}^n \to \mathbb{R}$ gilt*

$$x^* \in \operatorname*{argmin}_{x \in \mathbb{R}^n} f(x) \quad \Longleftrightarrow \quad 0 \in \partial f(x^*).$$

Beweis. Beide Bedingungen sind äquivalent zu $f(x) \geq f(x^*) + \langle 0, x - x^* \rangle$ für alle $x \in \mathbb{R}^n$. □

Das folgende Lemma besagt, dass das Subdifferential einer differenzierbaren konvexen Funktion gerade mit dem Gradienten übereinstimmt und dass das Subdifferential im allgemeinen Fall eine nicht leere (konvexe) Menge ist.

A.9 Lemma. *Ist $f : \mathbb{R}^n \to \mathbb{R}$ eine konvexe Funktion, so ist $\partial f(x)$ nicht leer für alle $x \in \mathbb{R}^n$. Ist f zusätzlich differenzierbar in x, so gilt $\partial f(x) = \{\nabla f(x)\}$.*

Beweis. Die erste Aussage folgt aus dem Trennungssatz für konvexe Mengen. Es folgt zum Beispiel aus Lemma 4.2.1 in [4], dass ein $(w, b) \in \mathbb{R}^n \times \mathbb{R}$ existiert mit

$$\langle w, y \rangle + br \geq \langle w, x \rangle + bf(x) \qquad \forall (y, r) \in \operatorname{epi} f.$$

Mit $r \to +\infty$ folgt $b \geq 0$, mit $y = x - w$ folgt $b \neq 0$. Wir können also o.B.d.A. annehmen, dass $b = 1$ und erhalten mit der Wahl $(y, r) = (y, f(y)) \in \operatorname{epi} f$,

$$f(y) \geq f(x) + \langle w, x - y \rangle$$

und somit $-w \in \partial f(x)$. Es folgt, dass $\partial f(x)$ nicht leer ist.

Sei nun f zusätzlich differenzierbar in x. Ist $n = 1$, so gilt (Bild!)

$$\frac{f(y) - f(x)}{y - x} \leq f'(x) \leq \frac{f(z) - f(x)}{z - x} \qquad \forall y < x < z,$$

was $f'(x) \in \partial f(x)$ impliziert. Für allgemeines f setze $g(t) = f(x + t(y - x))$, $t \in \mathbb{R}$. Dann ist g konvex, differenzierbar in 0 mit $f'(0) = \langle \nabla f(x), y - x \rangle$, und es folgt

$$f(y) = g(1) \geq g(0) + g'(0) = f(x) + \langle \nabla f(x), y - x \rangle.$$

Wir haben also $\nabla f(x) \in \partial f(x)$ gezeigt. Ist nun $w \in \partial f(x)$, so folgt aus der Taylorschen Formel, dass

$$\pm t \langle w, h \rangle \leq f(x \pm th) - f(x) = \pm t \langle \nabla f(x), h \rangle + o(t).$$

Mit $t \to 0$ erhalten wir $\langle \nabla f(x), h \rangle = \langle w, h \rangle$ für alle $h \in \mathbb{R}^n$, und somit $\nabla f(x) = w$. □

Zwei einfache Beispiele

A.10 Beispiel (Betragsfunktion)**.** Sei $f : \mathbb{R} \to \mathbb{R}, x \mapsto |x|$. Dann ist f mit der Ausnahme von $x = 0$ in allen Punkten differenzierbar und es gilt

$$\partial f(x) = \begin{cases} +1, & \text{falls } x > 0, \\ -1, & \text{falls } x < 0, \\ [-1, 1], & \text{falls } x = 0. \end{cases}$$

A.11 Beispiel (Hinge loss)**.** Sei $f : \mathbb{R} \to \mathbb{R}, x \mapsto \max(1 - x, 0)$. Dann ist f mit der Ausnahme von $x = 1$ in allen Punkten differenzierbar und es gilt

$$\partial f(x) = \begin{cases} 0, & \text{falls } x > 1, \\ -1, & \text{falls } x < 1, \\ [-1, 0], & \text{falls } x = 1. \end{cases}$$

Rechenregeln für das Subdifferential

Um Lemma A.8 auch für kompliziertere konvexe Funktionen als in den obigen zwei Beispielen behandeln zu können, benötigen wir noch einige Rechenregeln für das Subdifferential.

Skalare

Ist $f : \mathbb{R}^n \to \mathbb{R}$ eine konvexe Funktion und $a \geq 0$, so gilt

$$\partial(af)(x) = a\partial f(x) \qquad \forall x \in \mathbb{R}^n.$$

Summenbildung

Sind $g, f : \mathbb{R}^n \to \mathbb{R}$ zwei konvexe Funktionen, so gilt

$$\partial(g + f)(x) = \partial g(x) + \partial f(x) \qquad \forall x \in \mathbb{R}^n. \tag{A.1}$$

Die Inklusion $\partial g(x) + \partial f(x) \subseteq \partial(g + f)(x)$ folgt direkt aus der Definition, die umgekehrte Inklusion kann wieder mit dem Trennungssatz für konvexe Mengen gezeigt werden, siehe zum Beispiel Theorem 3.39 in [6] für die Details (oder Theorem 4.1.1 in [4] für einen weiteren Beweis).

Kettenregel unter affine Transformationen

Ist $h : \mathbb{R}^m \to \mathbb{R}$ eine konvexe Funktion, $A \in \mathbb{R}^{m \times n}$ und $b \in \mathbb{R}^m$, so gilt für die (konvexe) Funktion $f : \mathbb{R}^n \to \mathbb{R}, x \mapsto h(Ax + b)$, dass

$$\partial f(x) = A^T \partial h(Ax + b) \qquad \forall x \in \mathbb{R}^n. \tag{A.2}$$

Die Inklusion $A^T \partial h(Ax + b) \subseteq \partial f(x)$ folgt wieder direkt aus der Definition, die umgekehrte Inklusion zeigt man am einfachsten iterativ mit Hilfe der Singulärwertzerlegung von A, siehe Theorem 3.40 in [6] für eine allgemeinere Aussage (oder Theorem 4.2.1 in [4] für einen weiteren Beweis).

Aufgaben

A.12 Aufgabe (LASSO-Schätzer)**.** *Für $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$ und $\lambda > 0$, betrachte*

$$\hat{w} \in \operatorname*{argmin}_{w\in\mathbb{R}^p} \mathscr{L}(w), \qquad \mathscr{L}(w) = \|Y - Xw\|^2 + \lambda\|w\|_1.$$

Dann gilt

$$X^T X\hat{w} = X^T Y - \frac{\lambda}{2}\hat{\alpha}$$

wobei $\hat{\alpha} \in \mathbb{R}^p$ mit $\hat{\alpha}_j = \operatorname{sign}(\hat{w}_j)$ falls $\hat{w}_j \neq 0$ und $\hat{\alpha}_j \in [-1, 1]$ falls $\hat{w}_j = 0$.

Beweis. Unsere Aufgabe ist es das Subdifferential von $\mathscr{L}$ zu bestimmen. Setzen wir $f : \mathbb{R} \to \mathbb{R}, x \mapsto |x|$, so gilt

$$\mathscr{L}(w) = \|Y - Xw\|^2 + \lambda \sum_{j=1}^{p} f(e_j^T w)$$

wobei e_j der j-te Standardvektor ist. Mit (A.1) und (A.2) folgt

$$\partial\mathscr{L}(w) = -2X^T(Y - Xw) + \lambda \sum_{j=1}^{p} e_j \partial f(e_j^T w).$$

Verwenden wir außerdem, dass $0 \in \partial\mathscr{L}(\hat{w})$ aus Lemma A.8, so erhalten wir

$$0 \in -2X^T(Y - X\hat{w}) + \lambda \sum_{j=1}^{p} e_j \partial f(\hat{w}_j)$$

und die Behauptung folgt. □

A.13 Aufgabe (SVM-Klassifizierer)**.** *Für gegebene $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^p \times \{-1, +1\}$ und $\lambda > 0$, betrachte*

$$\hat{w} \in \operatorname*{argmin}_{w\in\mathbb{R}^n} \mathscr{L}(w), \qquad \mathscr{L}(w) = \frac{1}{n}\sum_{i=1}^{n} \max\left(1 - y_i\langle w, x_i\rangle, 0\right) + \lambda\|w\|^2.$$

Zeige, dass $\hat{w}$ von der Form $\hat{w} = \sum_{i=1}^n \hat{\alpha}_i y_i x_i$ ist, wobei

$$\begin{cases} \hat{\alpha}_i = 0, & \text{falls } y_i\langle \hat{w}, x_i\rangle > 1, \\ \hat{\alpha}_i = 1/(2\lambda n), & \text{falls } y_i\langle \hat{w}, x_i\rangle < 1, \\ \hat{\alpha}_i \in [0, 1/(2\lambda n)], & \text{falls } y_i\langle \hat{w}, x_i\rangle = 1. \end{cases}$$

A.3 Hilberträume

A.14 Definition. Eine Abbildung $\langle \cdot, \cdot \rangle : H \times H \to \mathbb{R}$ auf einem rellen Vektorraum H heißt Skalarprodukt, falls

(a) $\langle f + tg, h \rangle = \langle f, h \rangle + t\langle g, h \rangle$ für alle $f, g, h \in H$ und $t \in \mathbb{R}$;

(b) $\langle f, g \rangle = \langle g, f \rangle$ für alle $f, g \in H$;

(c) $\langle f, f \rangle \geq 0$ für alle $f \neq 0$.

(d) $\langle f, f \rangle = 0$ impliziert $f = 0$.

Ein reeller Vektorraum H vershen mit einem Skalarprodukt $\langle \cdot, \cdot \rangle$ heißt auch Innenproduktraum oder Prähilbertraum.

A.15 Satz (Cauchy-Schwarz-Ungleichung)**.** *Unter (a)-(c) aus Defnition A.14 gilt* $\langle f, g \rangle^2 \leq \langle f, f \rangle \langle g, g \rangle$ *für alle* $f, g \in H$.

Beweis. Nach Vorausssetzung gilt für alle $t \in \mathbb{R}$,

$$0 \leq \langle f + tg, f + tg \rangle = \langle f, f \rangle + 2t\langle f, g \rangle + t^2 \langle g, g \rangle.$$

Ist $\langle g, g \rangle = 0$ und $\langle f, f \rangle = 0$, so folgt $\langle f, g \rangle = 0$ in dem wir $t = -\langle f, g \rangle$ setzen. Ansonsten gilt $\langle g, g \rangle \neq 0$ oder $\langle f, f \rangle \neq 0$ und wir können ohne Einschränkung $\langle g, g \rangle \neq 0$ annehmen. In diesem Fall folgt die Behauptung aus der Wahl $t = -\langle f, g \rangle / \langle g, g \rangle$. □

Mit Hilfe der Cauchy-Schwarz-Ungleichung erhält man, dass durch

$$\|f\| := \sqrt{\langle f, f \rangle}$$

eine Norm auf H definiert ist, d.h. es gilt

(a) $\|f + g\| \leq \|f\| + \|g\|$ für alle $f, g \in H$;

(b) $\|tf\| = |t| \|f\|$ für alle $f \in H$ und $t \in \mathbb{R}$;

(c) $\|f\| > 0$ für alle $f \neq 0$.

Eine Folge von Elementen $(f_n)_{n \geq 1}$ heißt Cauchyfolge, falls für alle $\epsilon > 0$ eine natürliche Zahl N exisitiert, so dass $\|f_n - f_m\| < \epsilon$ für all $n, n \geq N$. Konvergiert jede Cauchyfolge $(f_n)_{n \geq 1}$ gegen ein Element $f \in H$, so heißt H (bzw. die Norm $\| \cdot \|$) vollständig.

A.16 Definition. Ein Hilbertraum $(H, \langle \cdot, \cdot \rangle)$ ist ein vollständiger Innenproduktraum.

Für eine Teilmenge $A \subseteq H$ heißt $A^\perp = \{f \in H : \langle f, g \rangle = 0$ für alle $g \in A\}$ orthogonales Komplement von A.

A.17 Satz. *Ist H_1 eine abgeschlossener Unterraum eines Hilbertraumes H, so ist auch $H_1^\perp$ ein abgeschlossener Unterraum von H_1. Des Weiteren besitzt jedes Element eine eindeutige Zerlegung $f = g + h$ mit $g \in H_1$ und $h \in H_1^\perp$.*

Beweis. Offensichtlich ist $H_1^\perp$ ein abgeschlossener Unterraum von H_1. Setze $c = \inf\{\|f - g\| : g \in H_1\}$. Sei (g_n) eine Folge aus H_1 mit $\|f - g_n\| \searrow c$. Dann gilt

$$\begin{aligned}\|g_n - g_m\|^2 &= 2\|g_n - f\|^2 + 2\|g_m - f\|^2 - \|g_n - f + g_m - f\|^2 \\ &= 2\|g_n - f\|^2 + 2\|g_m - f\|^2 - 4\left\|\frac{g_n + g_m}{2} - f\right\|^2.\end{aligned}$$

Da $(g_n + g_m)/2 \in H_1$ folgt, dass der letzte Term größer gleich $4c$ ist. Wir erhalten also, dass (g_n) eine Cauchyfolge ist. Nach Voraussetzung existiert also ein $g \in H_1$ mit $\|g - f\| = c$. Wir setzen $h = f - g$ und behaupten, dass $h \in H_1^\perp$. Dies kann man wie folgt sehen. Für alle $g' \in H_1$ und $t \in \mathbb{R}$ folgt aus der Definition von f, dass

$$\|h\|^2 \leq \|h + tg'\|^2 = \|h\|^2 + 2t\langle h, g'\rangle + t^2\|g'\|^2.$$

Wäre $\langle h, g'\rangle \neq 0$, so würde die Wahl $t = \epsilon\langle h, g'\rangle$ mit $\epsilon > 0$ genügend klein zu einem Widerspruch führen.

Es bleibt nch die Eindeutigkeit der Darstellung zu zeigen. Sei also $f = g + h = g' + h'$ mit $g, g' \in H_1$ und $h, h' \in H_1^\perp$, so folgt $g - g' = h - h'$ und somit $\langle g - g', g - g'\rangle = \langle g - g', h - h'\rangle = 0$. Also gilt $g = g'$ und analog $h = h'$. □

A.18 Satz (Rieszscher Darstellungssatz)**.** *Sei L ein lineares Funktional (d.h. eine lineare Abbildung $L : H \to \mathbb{R}$) welches zusätzlich beschränkt ist, d.h. es existiert $M > 0$ mit $|L(f)| \leq M\|f\|$ für alle $f \in H$. Dann existiert genau ein $g \in H$ mit $L(f) = \langle f, g\rangle$ für alle $f \in H$.*

Beweis. Ist L die Nullabbildung, so setze $g = 0$. Ist $L \neq 0$ so ist wegen Satz A.17 $(\ker L)^\perp$ ein eindimensionaler Unterraum von H. Sei nun $f \in (\ker L)^\perp$ mit $Lf = 1$. Dann ist die Abbildung definiert durch $L - \langle f, \cdot\rangle/\langle f, f\rangle$ sowohl auf $\ker L$ als auch auf $(\ker L)^\perp$ die Nullabbildung. Es folgt $Lh - \langle f, h\rangle/\langle f, f\rangle = 0$ für alle $h \in H$. Also gilt die Behauptung mit $g = f/\langle f, f\rangle$. □

Literatur

[1] Mikhail Belkin, Daniel J Hsu, and Partha Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 2300–2311. Curran Associates, Inc., 2018.

[2] Christophe Giraud. *Introduction to high-dimensional statistics*, volume 139 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL, 2015.

[3] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer, New York, second edition, 2009. Data mining, inference, and prediction.

[4] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Grundlehren Text Editions. Springer-Verlag, Berlin, 2001. Abridged version of ıt Convex analysis and minimization algorithms. I [Springer, Berlin, 1993; MR1261420 (95m:90001)] and ıt II [ibid.; MR1295240 (95m:90002)].

[5] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 103 of *Springer Texts in Statistics*. Springer, New York, 2013. With applications in R.

[6] Jean-Paul Penot. *Calculus without derivatives*, volume 266 of *Graduate Texts in Mathematics*. Springer, New York, 2013.

[7] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, USA, 2014.

[8] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, USA, 2004.

[9] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Information Science and Statistics. Springer, New York, 2008.

[10] M. Trabs, M. Jirak, K. Krenz, and M. Reiß. Methoden der Statistik und des maschinellen Lernens: Eine mathematische Einführung. Buchprojekt, verfügbar unter https://www.math.uni-hamburg.de/home/trabs/Lehre/BuchEntwurf.pdf, 2018.

[11] Martin J. Wainwright. *High-dimensional statistics*, volume 48 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2019. A non-asymptotic viewpoint.