Statistische Lerntheorie

Vorlesungsskript

Thorsten Dickhaus
Universität Bremen
Wintersemester 2017 / 2018
Version: 24. Januar 2018

Vorbemerkungen

Das Material zu diesem Skript habe ich im Wesentlichen aus den Büchern von Vapnik (2000, 1998) entnommen. Sollten sich in den übernommenen Teilen Fehler finden, so bin dafür natürlich ich verantwortlich. Lob und positive Kritik gebührt indes den Original-Autoren.

Für die Manuskripterstellung danke ich Nico Steffen.

Übungsaufgaben zu diesem Kurs stelle ich auf Anfrage gerne zur Verfügung. Einige Referenzen dazu finden sich im Text an den zugehörigen Stellen.

Inhaltsverzeichnis

Kapitel 1

Problemstellung und Beispiele

Ein-/Ausgabebeziehungen der Form

$$\boldsymbol{x} \longrightarrow \boxed{\text{Natur}} \longrightarrow y$$

sind allgegenwärtig in vielen wissenschaftlichen Bereichen.

Beispiel 1.1

a) *Landwirtschaft:*
$y \mathrel{\widehat{=}}$ *Ernteertrag,*
$\boldsymbol{x} \mathrel{\widehat{=}}$ *(Feldgröße, Düngemittelmenge, Niederschlag, Temperatur, Schädlingsbefall)*$^\top$.

b) *Gesundheitswissenschaften/Epidemiologie:*
$y \mathrel{\widehat{=}}$ *Typ II-Diabetes (ja/nein),*
$\boldsymbol{x} \mathrel{\widehat{=}}$ *(Alter, Geschlecht, Ernährung, Lebensstil)*$^\top$.

c) *Physik (Gasgesetz):*
$y \mathrel{\widehat{=}}$ *Gasdruck,*
$\boldsymbol{x} \mathrel{\widehat{=}}$ *(Volumen, Masse, Temperatur, spezifische Gaskonstante)*$^\top$.

Häufig stellen sich uns diese Ein-/Ausgabebeziehungen als nicht-deterministisch (stochastisch) dar. Mögliche Gründe dafür sind:

1) Nicht-Vorhersehbarkeit (z.B. Wetter, Schädlingsbefall in Beispiel 1.1.a)),

2) Nicht-Erhebung mancher relevanter Einflussgrößen (z.B. genetisches Profil in Beispiel 1.1.b)),

3) Nicht perfektes Messinstrumentarium (z.B. Thermometer in Beispiel 1.1.c)).

Dies führt zu einer statistischen Modellierung zur Analyse interessierender Ein-/Ausgabebeziehungen, da typischerweise Unsicherheit über das zu Grunde liegende Zufallsgesetz herrscht.

In einem viel beachteten Aufsatz unterscheidet Breiman (2001) dabei zwei unterschiedliche „Kulturen "der statistischen Modellierung.

Schema 1.2

(a) Daten-Modellierung:

$$\boldsymbol{x} \longrightarrow \boxed{\begin{array}{c}\textit{lineares Regressionsmodell}\\ \textit{logistisches Regressionsmodell}\\ \textit{Cox' proportional hazards-Modell}\end{array}} \longrightarrow y$$

$y = f(\boldsymbol{x}, \text{Parameter(-vektor), Fehlerterme})$. Eine Schätzung $\hat{f}$ erfolgt vermittels der Schätzung der Parameter.

(b) Algorithmische Modellierung:

$$\boldsymbol{x} \longrightarrow \boxed{\textit{„Black Box "}} \longrightarrow y$$

Man beobachtet Beispiele $(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_n, y_n)$ und versucht, daraus (irgend)einen Algorithmus (eine Abbildung) $\hat{f}$ zu konstruieren, so dass $\hat{f}(\boldsymbol{x}_{neu})$ für einen bislang ungesehenen Eingabe-Datenpunkt $\boldsymbol{x}_{neu}$ eine „möglichst gute" Vorhersage der zugehörigen Ausgabe y_{neu} ist. Hierbei wird (im allgemeinsten Falle) keinerlei Vorannahme bzgl. der konkreten Gestalt von $\hat{f}$ gemacht.
Die Konstruktion von $\hat{f}$ auf der Basis von $(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_n, y_n)$ bezeichnet man als statistisches Lernen. Schematisch:

$$(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_n, y_n) \longrightarrow \boxed{\begin{array}{c}\textit{Lernverfahren}\\ \textit{Lern-Algorithmus}\\ \textit{Lern-Maschine}\end{array}} \longrightarrow \hat{f}$$

Man spricht auch von maschinellem Lernen, da $\hat{f}$ statt durch Modellierung und Parameterschätzung, durch einen (Lern-)Algorithmus bestimmt wird.
Im statistischen Jargon könnte man indes auch von nichtparametrischer Funktionenschätzung sprechen.

Definition 1.3 (Komponenten eines statistischen Lernproblems)
Die drei Komponenten eines statistischen Lernproblems sind

(i) ein Generator (G). Dieser erzeugt Eingabe-Zufallsvektoren $\boldsymbol{x}_i \in D \subseteq \mathbb{R}^d$ gemäß einer Wahrscheinlichkeitsverteilung $\mathbb{P}^{\mathbf{X}}$.

(ii) ein Überwacher (englisch Supervisor, S), der für jedes $\boldsymbol{x} \in D$ ein $y \in W$ zurückgibt, gemäß einer bedingten Wahrscheinlichkeitsverteilung $\mathbb{P}^{Y|\mathbf{X}}$. Die gemeinsame Verteilung von $(\mathbf{X}, Y)$ ist demnach gegeben durch $\mathbb{P} := \mathbb{P}^{(\mathbf{X},Y)} = \mathbb{P}^{\mathbf{X}} \otimes \mathbb{P}^{Y|\mathbf{X}}$, wobei wir annehmen, dass $\mathbf{X}$ und Y auf dem selben Wahrscheinlichkeitsraum definiert sind. Wir beachten, dass hiermit auch der (deterministische) Spezialfall $y_i \equiv f(\boldsymbol{x}_i)$ für eine feste Funktion f abgedeckt ist.

(iii) Eine Lern-Maschine (LM), die Funktionen $f \in \mathcal{M}$ implementieren kann. Häufig schreiben wir

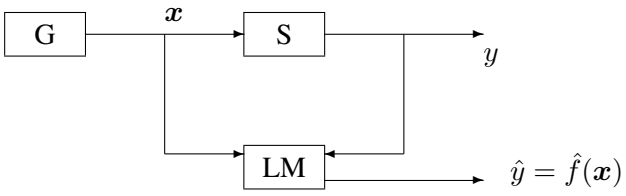
$$\mathcal{M} = \{f(\cdot,\cdot) : D \times \Theta \to W$$
$$(\boldsymbol{x}, \theta) \mapsto f(\boldsymbol{x}, \theta)\},$$

wobei indes typischerweise dim$(\Theta) = \infty$ gilt, d.h., Θ ein Funktionenraum ist.
(Dennoch wird θ häufig als „Parameter“ bezeichnet.)

Das Lernproblem besteht also darin, dasjenige $\hat{f} \in \mathcal{M}$ zu finden, dass die Antwort des Supervisors am besten (in einem gegeben stochastischen Sinne) approximiert.
Dazu dienen Trainingsbeispiele $(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_n, y_n)$. Als Zufallsvariablen aufgefasst, nehmen wir an, dass für den Trainingsdatensatz gilt:

$$(\mathbf{X}_1, Y_1), ..., (\mathbf{X}_n, Y_n) \sim \mathbb{P}^{\otimes n}$$

(stochastisch unabhängige und identisch verteilte (i.i.d.) Beobachtungseinheiten mit $(\mathbf{X}_1, Y_1) \overset{\mathcal{D}}{=} (\mathbf{X}, Y)$).
Schema:



Definition 1.4 (Verlustfunktion, Risiko)
Sei

$$L : W \times W \to \mathbb{R} \tag{1.1}$$
$$(y, \hat{y}) \mapsto L(y, \hat{y}) \in \mathbb{R}$$

eine vorgegebene Verlustfunktion (Diskrepanz). Die Funktion L quantifiziert, wie schlecht die Vorhersage $\hat{y}$ von y ist (schlechte Vorhersage $\Rightarrow$ großer Verlust).
Dann heißt R, gegeben durch

$$R(f) = \mathbb{E}[L(Y, f(\mathbf{X}))],\ (\mathbf{X}, Y) \sim \mathbb{P} \tag{1.2}$$

für $f \in \mathcal{M}$, das zu L gehörige Risikofunktional.

Bezeichnet $F(\cdot,\cdot)$ die gemeinsame Verteilungsfunktion von $(\mathbf{X}, Y)$ und schreiben wir $f \equiv f(\cdot,\theta)$ für $\theta \in \Theta$, so gilt äquivalenterweise

$$R(\theta) = \int L(y, f(\boldsymbol{x},\theta))\, dF(\boldsymbol{x}, y),\ \theta \in \Theta.$$

Ziel: Finde

$$\theta^* = \arg\min_{\theta\in\Theta} R(\theta).$$

Problem: $F(\cdot,\cdot)$ ist unbekannt und es steht nur die Information zur Verfügung, die uns der Trainingsdatensatz liefert! Insofern wird das Ziel in der Praxis nur approximativ oder asymptotisch (für $n \to \infty$) zu erreichen sein, falls überhaupt.

Beispiel 1.5

(a) Klassifikation (Mustererkennung):

Wir betrachten (der Einfachheit halber) $W = \{0, 1\}$ (binäre Klassifikation, Mehrklassen-Klassifikation kann analog behandelt werden).

Konsequenterweise wird hier $\mathcal{M}$ als eine Menge von Indikatorfunktionen gewählt, so dass $f(\boldsymbol{x}) \in \{0, 1\} = W$ für alle $f \in \mathcal{M}$ und alle $\boldsymbol{x} \in D$ gilt.

Eine sinnvolle Verlustfunktion ist gegeben durch

$$L(y, f(\boldsymbol{x})) = \left\{ \begin{array}{ll} 0, & \textit{falls } y = f(\boldsymbol{x}) \\ 1, & \textit{falls } y \neq f(\boldsymbol{x}) \end{array} \right\}.$$

Damit ist

$$\begin{aligned} R(f) &= \mathbb{P}(f(\boldsymbol{X}) \neq Y) \\ &= \mathbb{P}(f(\boldsymbol{X}) = 0, Y = 1) + \mathbb{P}(f(\boldsymbol{X}) = 1, Y = 0) \end{aligned}$$

(Summe aus Fehlerwahrscheinlichkeiten 1. und 2. Art).

(b) (Mittelwert-)Regression:

Sei $W = \mathbb{R}$ und sei Θ so, dass $\mathcal{M}$ die wahre Regressionsfunktion enthält, d.h.,

$$\exists \theta^* \in \Theta : \forall \boldsymbol{x} \in D : f(\boldsymbol{x}, \theta^*) = \int y\, dF(y|\boldsymbol{x}),$$

wobei $F(\cdot|\boldsymbol{x})$ die bedingte Verteilungsfunktion von Y gegeben $\mathbf{X} = \boldsymbol{x}$ bezeichnet.

Es ist bekannt (L_2-Projektionseigenschaft des (bedingten) Erwartungswertes), dass $f(\cdot, \theta^*)$ das Risikofunktional zur quadratischen Verlustfunktion

$$L(y, f(\boldsymbol{x},\theta)) = (y - f(\boldsymbol{x},\theta))^2 \tag{1.3}$$

minimiert. Im Lernkontext ist indes $F(\cdot,\cdot)$ und auch $F(\cdot|\boldsymbol{x}), \boldsymbol{x} \in D$, unbekannt und nur die Information vorhanden, die der Trainingsdatensatz liefert. Dennoch ist L eine sinnvolle Verlustfunktion.

(c) Dichteschätzung:

Nehmen wir an, $\mathbb{P}^{\mathbf{X}}$ besitzt eine (Lebesgue)-Dichte und wir möchten diese auf der Basis der Trainingsdaten schätzen. Offenbar benötigen wir dazu nur $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ und nicht $y_1, ..., y_n$. Man spricht in einem solchen Fall von einem unüberwachten (unsupervised) Lernproblem. Im Gegensatz dazu sind Klassifikation und Regression überwachte Lernprobleme.
Sei also $\mathcal{M}$ eine Menge von (Lebesgue-)Dichten $p = p(\cdot, \theta), \theta \in \Theta$. Eine sinnvolle Verlustfunktion in diesem Kontext ist gegeben durch

$$L(p(\boldsymbol{x}, \theta)) = -\log p(\boldsymbol{x}, \theta).$$

Die wahre Dichte von $\mathbf{X}$ minimiert das zugehörige Risikofunktional.
Dies sieht man wie folgt. Es gilt

$$R(\theta) = -\int \log p(\boldsymbol{x}, \theta)\, p^*(\boldsymbol{x}) dx,$$

wobei p^* die wahre Dichte von $\mathbf{X}$ bezeichnet.
Addieren wir nun zu $R(\theta), \theta \in \Theta$, die Konstante $c := \int \log p^*(\boldsymbol{x})\, p^*(\boldsymbol{x}) dx$, so erhalten wir

$$\begin{aligned} R(\theta) + c &= -\int \log p(\boldsymbol{x}, \theta)\, p^*(\boldsymbol{x}) dx + \int \log p^*(\boldsymbol{x})\, p^*(\boldsymbol{x}) dx \\ &= -\int \log \left\{ \frac{p(\boldsymbol{x}, \theta)}{p^*(\boldsymbol{x})} \right\} p^*(\boldsymbol{x}) dx. \end{aligned} \tag{1.4}$$

Die rechte Seite von (1.4) ist die Kullback-Leibler-Divergenz von $p(\cdot, \theta)$ bezüglich p^*. Diese ist stets nicht-negativ und gleich Null genau dann, wenn $p(\cdot, \theta) = p^*(\cdot)$ $\mathbb{P}$-fast sicher gilt.

Bemerkung 1.6

Möchten wir überwachte und unüberwachte statistische Lernprobleme in einem allgemeinen formalen Rahmen zusammenfassen, so können wir dies wie folgt erreichen.
Sei $(\mathcal{Z}, \mathcal{F}_{\mathcal{Z}})$ ein messbarer Raum und P ein Wahrscheinlichkeitsmaß auf $(\mathcal{Z}, \mathcal{F}_{\mathcal{Z}})$. Betrachte eine Funktionenmenge

$$\{Q(\cdot, \cdot) : \mathcal{Z} \times \Theta \to \mathbb{R}$$
$$(\boldsymbol{z}, \theta) \mapsto Q(\boldsymbol{z}, \theta) \in \mathbb{R}\}$$

und minimiere das Risikofunktional R, gegeben durch

$$R(\theta) = \int Q(\boldsymbol{z}, \theta)\, P(d\boldsymbol{z}) \tag{1.5}$$

über $\Theta \ni \theta$.
Hierbei ist P unbekannt, aber Information über P in Form einer Trainingsstichprobe $\boldsymbol{z}_1, ...\boldsymbol{z}_n$ mit $\mathbf{Z}_1, ..., \mathbf{Z}_n$ i.i.d., $\mathbf{Z}_1 \sim P$, gegeben.

Definition 1.7 (Prinzip der empirischen Risikominimierung (ERM))
Da unter den Bezeichnungen von Bemerkung 1.6 die Verteilung P unbekannt ist, liegt es nahe, P in (1.5) *durch das empirische Maß $\hat{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\boldsymbol{Z}_i}$ zu ersetzen (Plug-in-Methode, Substitutionsprinzip). Das empirische Analogon zu $R(\theta)$ in* (1.5) *ist somit gegeben durch*

$$R_{emp}(\theta) = \frac{1}{n}\sum_{i=1}^{n} Q(\boldsymbol{Z}_i, \theta). \tag{1.6}$$

Das Prinzip der empirischen Risikominimierung (ERM) ersetzt nun die Minimierungsaufgabe bezüglich R durch die Minimierungsaufgabe bezüglich R_{emp} (für gegebene Realisierungen $\boldsymbol{Z}_1 = \boldsymbol{z}_1, \ldots, \boldsymbol{Z}_n = \boldsymbol{z}_n$).

Beispiel 1.8
Klassische statistische Inferenzmethoden lassen sich als Spezialfälle des ERM-Prinzip auffassen.

(a) Kleinste Quadrate-Methode in der Regression:

$$R_{emp}(\theta) = n^{-1}\sum_{i=1}^{n}(y_i - f(\boldsymbol{x}_i, \theta))^2;$$

$\boldsymbol{z}_i \widehat{=} (\boldsymbol{x}_i, y_i)$, $Q(\boldsymbol{z}_i, \theta) = Q(\boldsymbol{x}_i, y_i, \theta) = (y_i - f(\boldsymbol{x}_i, \theta))^2$, *vgl. Beispiel 1.5.(b).*

(b) Maximum-Likelihood-Dichteschätzung:

$$R_{emp}(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \ln p(\boldsymbol{x}_i, \theta);$$

$\boldsymbol{z}_i \widehat{=} \boldsymbol{x}_i$, $Q(\boldsymbol{x}_i, \theta) = -\ln p(\boldsymbol{x}_i, \theta)$, *vgl. Beispiel 1.5.(c).*

Schema 1.9 (Überblick über die restlichen Kapitel)
In den weiteren Kapiteln werden wir die folgenden Fragen untersuchen:

(i) Was sind notwendige und hinreichende Bedingungen für die Konsistenz eines auf ERM basierenden Lernverfahrens? (→ Kapitel 2)

(ii) Wie schnell ist die Konvergenz des Lernverfahrens? (→ Kapitel 3)

(iii) Wie lässt sich die Konvergenzrate (die Generalisierungsfähigkeit) eines Lernverfahrens kontrollieren? (→ Kapitel 4)

(iv) Wie konstruiert man „gute“ statistische Lernverfahren? (→ ab Kapitel 5)

Kapitel 2

Konsistenz von statistischen Lernverfahren

Unter den Voraussetzungen von Definition 1.7 (ERM-Prinzip) sei $\hat{\theta}(n)$ so, dass $Q(\cdot, \hat{\theta}(n))$ das empirische Risikofunktional R_{emp} minimiert, d.h.,

$$\hat{\theta}(n) = \underset{\theta \in \Theta}{\operatorname{argmin}} \left\{ n^{-1} \sum_{i=1}^{n} Q(\boldsymbol{z}_i, \theta) \right\}$$

für beobachtete Werte $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$ mit $\mathbf{Z}_1, ..., \mathbf{Z}_n$ i.i.d., $\mathbf{Z}_1 \sim P$.

Untersuchungsgegenstand: Asymptotisches Verhalten $(n \to \infty)$ von $\hat{\theta}(n)$ bzw. von $R(\hat{\theta}(n))$ und $R_{emp}(\hat{\theta}(n))$, wobei wir $\hat{\theta}(n)$ als Zufallsvariable bzw. „Schätzvorschrift“ auffassen.

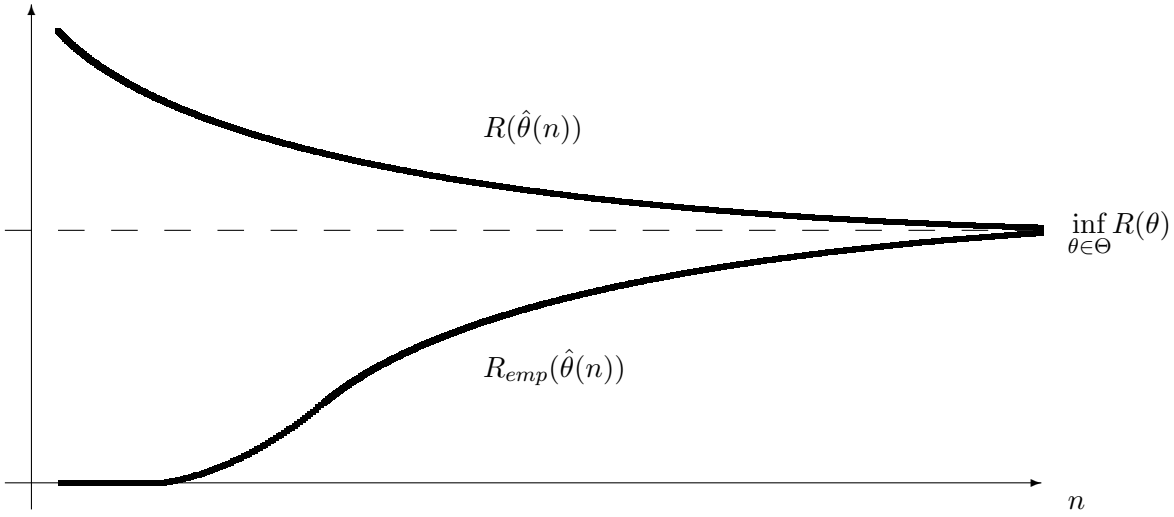
Definition 2.1 (Konsistenz von ERM)
Wir sagen, dass das ERM-Prinzip konsistent für das durch (1.5) *gegebene statistische Lernproblem ist, falls für $n \to \infty$ gilt:*

$$R(\hat{\theta}(n)) \xrightarrow{P} \inf_{\theta \in \Theta} R(\theta), \textit{ und} \tag{2.1}$$

$$R_{emp}(\hat{\theta}(n)) \xrightarrow{P} \inf_{\theta \in \Theta} R(\theta) \tag{2.2}$$

Mit anderen Worten heißt das ERM-Prinzip konsistent für das Lernproblem (1.5), falls es eine Funktionenfolge $(Q(\cdot, \hat{\theta}(n)))_{n \geq 1}$ liefert, für die sowohl das theoretische (erwartete) Risiko als auch das empirische Risiko stochastisch gegen das optimale Risiko über $\theta \in \Theta$ konvergiert.

Schema 2.2



Bemerkung 2.3

(i) In der Praxis ist der Stichprobenumfang n typischerweise fest vorgegeben, oder strebt zumindest nicht gegen unendlich, und man ist daran interessiert, auf der Basis einer limitierten Anzahl an Trainingsbeispielen $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$ eine „gute"Funktion $\hat{f}$ zu konstruieren. Dennoch sind Konsistenzuntersuchungen wichtig, denn sie sichern die konzeptionelle Validität des ERM-Ansatzes.

(ii) Die Funktionenmenge $\{Q(\cdot, \theta) : \theta \in \Theta\}$ ist eine Wahl des/der Datenanalysten/-in. Damit können Fälle auftreten, in denen Konsistenz trivialerweise erfüllt ist. Sei nämlich angenommen, die ERM-Methode ist nicht konsistent für (1.5)*, falls $\mathcal{M} = \{Q(\cdot, \theta) : \theta \in \Theta\}$ gewählt wird. Sei ferner angenommen, es lässt sich eine weitere Funktion $q : \mathcal{Z} \to \mathbb{R}$ finden (die nicht von θ abhängt), so dass*

$$\inf_{\theta\in\Theta} Q(\boldsymbol{z}, \theta) > q(\boldsymbol{z})$$

für alle $\boldsymbol{z} \in \mathcal{Z}$ gilt.
Dann ist die ERM-Methode trivialerweise konsistent, wenn die erweiterte Menge $\mathcal{M}_{extended} = \mathcal{M} \cup \{q\}$ bzw. die entsprechende Menge $\Theta_{extended}$ betrachtet wird, denn (2.1) *und* (2.2) *sind offenbar über $\Theta_{extended}$ für q erfüllt (unabhängig von P !). Um solche Trivialfälle auszuschließen, muss Definition 2.1 verfeinert werden.*

Definition 2.4 (Nicht-triviale Konsistenz von ERM)
Seien die Voraussetzungen von Bemerkung 1.6 erfüllt.
Sei für $c \in \mathbb{R}$ die Teilmenge $\Theta(c)$ gegeben durch

$$\Theta(c) = \{\theta \in \Theta : R(\theta) > c\}.$$

Dann sagen wir, dass das ERM-Prinzip nicht-trivial konsistent für das durch (1.5) *gegebene statistische Lernproblem ist, falls gilt:*

$$\forall c \text{ mit } \Theta(c) \neq \emptyset : \inf_{\theta \in \Theta(c)} R_{emp}(\theta) \xrightarrow{P} \inf_{\theta \in \Theta(c)} R(\theta) \tag{2.3}$$

für $n \to \infty$.

Mit anderen Worten ist ERM dann nicht-trivial konsistent, falls Konvergenz im Sinne von (2.3) auch dann noch stattfindet, wenn die Funktionen mit kleinem Risiko aus $\mathcal{M}$ entfernt werden.

Bemerkung 2.5
Es lässt sich zeigen, dass (2.3) *automatisch* (2.1) *impliziert. (→ Übungsaufgabe)*

Satz 2.6 (Charakterisierung der Konsistenz von ERM, Vapnik and Chervonenkis (1991))
Sei Θ *so, dass reelle Konstanten* a *und* A *existieren mit*

$$\forall P \in \mathcal{P} : \forall \theta \in \Theta : a \leq \int Q(\boldsymbol{z}, \theta) P(d\boldsymbol{z}) = R(\theta) \leq A,$$

wobei $\mathcal{P}$ *eine Menge von Wahrscheinlichkeitsmaßen bezeichnet, die das Modell für* $\mathbf{Z}_1$ *beschreibt. Dann ist ERM genau dann nicht-trivial konsistent, wenn gilt:*

$$\forall \varepsilon > 0 : \lim_{n \to \infty} P(\sup_{\theta \in \Theta} \{R(\theta) - R_{emp}(\theta)\} > \varepsilon) = 0. \tag{2.4}$$

Ein Konvergenzverhalten der Form (2.4) wird gleichmäßige einseitige (stochastische) Konvergenz genannt, wobei hier indes Gleichmäßigkeit über einen ganzen Funktionenraum gefordert wird, während z.B. Sätze vom Glivenko-Cantelli-Typ lediglich Gleichmäßigkeit über die reelle Achse bzw. über $\mathbb{R}^d$, $d \in \mathbb{N}$, liefern.
Im Weiteren wird es bei der Analyse von Bedingung (2.4) daher entscheidend darauf ankommen, die Komplexität von Θ geeignet zu formalisieren (und zu beschränken).

Beweis: von Satz 2.6
Unter den Bezeichnungen von Definition 2.4 sei $c \in \mathbb{R}$ beliebig so, dass $\Theta(c) \neq \emptyset$ ist. Gemäß der definierenden Eigenschaft (2.3) ist ERM nicht-trivial konsistent, falls gilt:

$$\inf_{\theta \in \Theta(c)} n^{-1} \sum_{i=1}^{n} Q(\mathbf{Z}_i, \theta) \xrightarrow{P} \inf_{\theta \in \Theta(c)} \int Q(\boldsymbol{z}, \theta) P(d\boldsymbol{z}) \tag{2.5}$$

Wir folgern nun zunächst, dass (2.5) die gleichmäßige einseitige Konvergenz (2.4) impliziert.
Wir wählen dazu eine endliche Folge $\{a_k\}_{1 \leq k \leq K}$ derart, dass $a_1 = a, a_K = A$ und für alle $1 \leq k \leq K-1$: $|a_{k+1} - a_k| < \frac{\varepsilon}{2}$ ist. Sei für $1 \leq k \leq K$ das Ereignis T_k gegeben durch

$$T_k = \left\{ \inf_{\theta \in \Theta(a_k)} n^{-1} \sum_{i=1}^{n} Q(\mathbf{Z}_i, \theta) < \inf_{\theta \in \Theta(a_k)} \int Q(\boldsymbol{z}, \theta) P(d\boldsymbol{z}) - \frac{\varepsilon}{2} \right\}.$$

Wegen (2.5) gilt $P(T_k) \to 0, n \to \infty$, für alle $1 \leq k \leq K$. Sei nun $T = \bigcup_{k=1}^{K} T_k$.
Da K endlich ist, gilt

$$\lim_{n\to\infty} P(T) = 0. \qquad (\star)$$

Definiere

$$E := \left\{ \sup_{\theta\in\Theta} \left[\int Q(\boldsymbol{z},\theta)P(d\boldsymbol{z}) - n^{-1}\sum_{i=1}^{n} Q(\mathbf{Z}_i,\theta) \right] > \varepsilon \right\}.$$

Angenommen, E tritt ein. Dann gibt es ein $\theta^* \in \Theta$ mit

$$\int Q(\boldsymbol{z},\theta^*)P(d\boldsymbol{z}) - \varepsilon > n^{-1}\sum_{i=1}^{n} Q(\boldsymbol{z}_i,\theta^*).$$

Zu diesem θ^* lässt sich ein $k \in \{1,...,K\}$ finden, so dass $\theta^* \in \Theta(a_k)$ und

$$\int Q(\boldsymbol{z},\theta^*)P(d\boldsymbol{z}) - a_k < \frac{\varepsilon}{2}$$

ist.
Für die so ausgewählte Teilmenge $\Theta(a_k)$ gilt dann die Ungleichung

$$\int Q(\boldsymbol{z},\theta^*)P(d\boldsymbol{z}) - \inf_{\theta\in\Theta(a_k)} \int Q(\boldsymbol{z},\theta)P(d\boldsymbol{z}) < \frac{\varepsilon}{2}.$$

Damit ist insgesamt (nach Dreiecksungleichung)

$$\begin{aligned} \inf_{\theta\in\Theta(a_k)} \int Q(\boldsymbol{z},\theta)P(d\boldsymbol{z}) - \frac{\varepsilon}{2} &> \int Q(\boldsymbol{z},\theta^*)P(d\boldsymbol{z}) - \varepsilon \\ &> n^{-1}\sum_{i=1}^{n} Q(\boldsymbol{z}_i,\theta^*) \\ &\geq \inf_{\theta\in\Theta(a_k)} n^{-1}\sum_{i=1}^{n} Q(\boldsymbol{z}_i,\theta), \end{aligned}$$

d.h., das Ereignis T_k tritt ein.
Damit tritt dann auch (nach Konstruktion von T_k und T) das Ereignis T ein. Insgesamt ist also $E \subseteq T$ und damit $P(E) \leq P(T)$. Aus $(\star)$ folgern wir $\lim_{n\to\infty} P(E) = 0$. Dies ist aber gerade äquivalent zu (2.4), womit eine Richtung der in Satz 2.6 behaupteten Äquivalenz gezeigt ist.

Zum Nachweis der Rückrichtung dürfen wir voraussetzen, dass für alle $\varepsilon > 0$ gilt:

$$P\left(\sup_{\theta\in\Theta} \left\{ \int Q(\boldsymbol{z},\theta)P(d\boldsymbol{z}) - n^{-1}\sum_{i=1}^{n} Q(\mathbf{Z}_i,\theta) \right\} > \varepsilon \right) \to 0,\ n \to \infty. \qquad (\star\star)$$

Wir müssen zeigen, dass aus $(\star\star)$ folgt:

$$\forall \varepsilon > 0 : \forall c \in \mathbb{R} \text{ mit } \Theta(c) \neq \emptyset : \lim_{n\to\infty} P(\tilde{E}) = 0 \text{ für}$$

$$\tilde{E} = \left\{ \left| \inf_{\theta\in\Theta(c)} \int Q(\boldsymbol{z},\theta)P(d\boldsymbol{z}) - \inf_{\theta\in\Theta(c)} n^{-1}\sum_{i=1}^{n} Q(\mathbf{Z}_i,\theta) \right| > \varepsilon \right\},$$

wobei wir zur Vereinfachung der Notation die Abhängigkeit des Ereignisses $\tilde{E}$ von ε und c notationell unterdrücken.

Wir schreiben $\tilde{E} = \tilde{E}_1 \cup \tilde{E}_2$ mit

$$\tilde{E}_1 = \left\{ \inf_{\theta\in\Theta(c)} \int Q(\boldsymbol{z},\theta)P(d\boldsymbol{z}) + \varepsilon < \inf_{\theta\in\Theta(c)} n^{-1}\sum_{i=1}^{n} Q(\mathbf{Z}_i,\theta) \right\},$$

$$\tilde{E}_2 = \left\{ \inf_{\theta\in\Theta(c)} \int Q(\boldsymbol{z},\theta)P(d\boldsymbol{z}) - \varepsilon > \inf_{\theta\in\Theta(c)} n^{-1}\sum_{i=1}^{n} Q(\mathbf{Z}_i,\theta) \right\}.$$

Wir schätzen $P(\tilde{E}_1)$ und $P(\tilde{E}_2)$ separat ab und beobachten, dass $P(\tilde{E}) \leq P(\tilde{E}_1) + P(\tilde{E}_2)$ ist.

Abschätzung von $P(\tilde{E}_1)$:

Wähle θ^* so, dass

$$\int Q(\boldsymbol{z},\theta^*)P(d\boldsymbol{z}) < \inf_{\theta\in\Theta(c)} \int Q(\boldsymbol{z},\theta)P(d\boldsymbol{z}) + \frac{\varepsilon}{2}$$

ist. Tritt $\tilde{E}_1$ ein, so ist

$$n^{-1}\sum_{i=1}^{n} Q(\boldsymbol{z}_i,\theta^*) > \int Q(\boldsymbol{z},\theta^*)P(d\boldsymbol{z}) + \frac{\varepsilon}{2}.$$

Also ist

$$P(\tilde{E}_1) \leq P\left(n^{-1}\sum_{i=1}^{n} Q(\mathbf{Z}_i,\theta^*) - \int Q(\boldsymbol{z},\theta^*)P(d\boldsymbol{z}) > \frac{\varepsilon}{2} \right) =: P(\tilde{\tilde{E}}_1).$$

Nach dem Gesetz der großen Zahlen ist

$$\lim_{n\to\infty} P(\tilde{\tilde{E}}_1) = 0 \Rightarrow \lim_{n\to\infty} P(\tilde{E}_1) = 0.$$

Abschätzung von $P(\tilde{E}_2)$:

Falls $\tilde{E}_2$ eintritt, dann $\exists\theta^{**} \in \Theta(c)$, so dass

$$n^{-1}\sum_{i=1}^{n} Q(\boldsymbol{z}_i,\theta^{**}) + \frac{\varepsilon}{2} < \inf_{\theta\in\Theta(c)} \int Q(\boldsymbol{z},\theta)P(d\boldsymbol{z}) < \int Q(\boldsymbol{z},\theta^{**})P(d\boldsymbol{z}).$$

Also ist

$$\begin{aligned} P(\tilde{E}_2) &\leq P\left(\int Q(\boldsymbol{z},\theta^{**})P(d\boldsymbol{z}) - n^{-1}\sum_{i=1}^{n} Q(\mathbf{Z}_i,\theta^{**}) > \frac{\varepsilon}{2} \right) \\ &\leq P\left(\sup_{\theta\in\Theta} \left[\int Q(\boldsymbol{z},\theta)P(d\boldsymbol{z}) - n^{-1}\sum_{i=1}^{n} Q(\mathbf{Z}_i,\theta) \right] > \frac{\varepsilon}{2} \right) =: P(\tilde{\tilde{E}}_2). \end{aligned}$$

Wegen ($\star\star$) ist

$$\lim_{n\to\infty} P(\tilde{\tilde{E}}_2) = 0 \Rightarrow \lim_{n\to\infty} P(\tilde{E}_2) = 0.$$

Insgesamt erhalten wir somit schließlich $\lim_{n\to\infty} P(\tilde{E}) = 0$ für jede beliebige Wahl von $\varepsilon > 0$. ■

Definition 2.7 (Empirische Prozesse)
Unter den Voraussetzungen von Satz 2.6 setzen wir

$$\forall 1 \leq i \leq n : \forall \theta \in \Theta : \xi_i^{(\theta)} := Q(\mathbf{Z}_i, \theta) \text{ mit Werten in } \mathbb{R}.$$

Das Objekt

$$\left(n^{-1} \sum_{i=1}^{n} \xi_i^{(\theta)} - E[\xi_1^{(\theta)}] \right)_{\theta \in \Theta}$$

heißt empirischer Prozess, indiziert in der (Funktionen)-Klasse Θ.

Die Beurteilung der (nicht-trivialen) Konsistenz des ERM-Verfahrens beruht also auf der Theorie der gleichmäßigen (über $\theta \in \Theta$) Konvergenz empirischer Prozesse. Das Kriterium (2.4) lässt sich äquivalent formulieren als

$$\sup_{\theta \in \Theta} \left\{ E[\xi_1^{(\theta)}] - n^{-1} \sum_{i=1}^{n} \xi_i^{(\theta)} \right\} \xrightarrow{P} 0. \tag{2.6}$$

Beispiel 2.8

(a) Angenommen, $|\Theta| = 1, \Theta = \{\theta^*\}$. Wir schreiben vereinfachend ξ_i statt $\xi_i^{(\theta^*)}, 1 \leq i \leq n$. Nach dem starken Gesetz der großen Zahlen gilt

$$n^{-1} \sum_{i=1}^{n} \xi_i \to E[\xi_1] \qquad\qquad P\text{-fast sicher für } n \to \infty.$$

Damit ist zweiseitige fast sichere Konvergenz der Form

$$\left| E[\xi_1] - n^{-1} \sum_{i=1}^{n} \xi_i \right| \xrightarrow{P-f.s.} 0, \ n \to \infty,$$

gegeben, was selbstverständlich (2.6) *impliziert.*

(b) Angenommen, $\mathcal{Z} = \Theta = \mathbb{R}$ und $Q(Z_i, \theta) = \xi_i^{(\theta)} = \mathbf{1}_{(-\infty, \theta]}(Z_i)$. Bezeichnet F die zu P gehörige Verteilungsfunktion und $\hat{F}_n$ die zu $\hat{P}_n$ gehörige empirische Verteilungsfunktion, so ist die linke Seite von (2.6) *hier gegeben durch*

$$\sup_{\theta \in \mathbb{R}} \{ F(\theta) - \hat{F}_n(\theta) \}.$$

Der Satz von Glivenko-Cantelli liefert nun

$$\sup_{\theta \in \mathbb{R}} \left| \hat{F}_n(\theta) - F(\theta) \right| \xrightarrow{P-f.s.} 0, \ n \to \infty,$$

somit ist auch hier (2.6) *erfüllt.*

Satz 2.9 (Hoeffding-Ungleichung)
Seien $\xi_1, ..., \xi_n$ reellwertige, stochastisch unabhängige, zentrierte und beschränkte Zufallsvariablen, so dass

$$\forall 1 \leq i \leq n : a_i \leq \xi_i \leq b_i,$$

mit $a_i \neq b_i \in \mathbb{R}$. Dann gilt für jedes $\varepsilon > 0$, dass

$$\mathbb{P}\left(\sum_{i=1}^{n} \xi_i \geq \varepsilon\right) \leq \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^{n} \Delta_i^2}\right), \tag{2.7}$$

wobei $\Delta_i = b_i - a_i$ ist, $1 \leq i \leq n$.

Beweis: Wir folgen der Argumentation in Appendix B von Pollard (1984).
Sei $1 \leq i \leq n$ beliebig. Wegen der Konvexität von $\exp(\cdot)$ ist für $t \in \mathbb{R}$

$$e^{t\xi_i} \leq \frac{e^{ta_i}(b_i - \xi_i)}{\Delta_i} + \frac{e^{tb_i}(\xi_i - a_i)}{\Delta_i}.$$

$$\Rightarrow \mathbb{E}\left[e^{t\xi_i}\right] \leq \frac{e^{ta_i} b_i}{\Delta_i} - \frac{e^{tb_i} a_i}{\Delta_i},$$

da ξ_i zentriert ist. Setze

$$\alpha_i := -\frac{a_i}{\Delta_i}, \;\; \beta_i := 1 - \alpha_i = \frac{b_i}{\Delta_i}, \;\; u_i := t\Delta_i$$

und beachte

$$\begin{aligned} \alpha_i + \beta_i &= 1, \\ \alpha_i u_i &= -ta_i, \\ \beta_i u_i &= tb_i, \\ \alpha_i &> 0, \text{ da } a_i < 0 < b_i. \end{aligned}$$

Damit ist

$$\begin{aligned} \log \mathbb{E}\left[e^{t\xi_i}\right] &\leq \log\left(\beta_i e^{-\alpha_i u_i} + \alpha_i e^{\beta_i u_i}\right) \\ &= \log\left(e^{-\alpha_i u_i}\left[\beta_i + \alpha_i e^{(\alpha_i + \beta_i) u_i}\right]\right) \\ &= -\alpha_i u_i + \log\left(\beta_i + \alpha_i e^{u_i}\right) \\ &=: L(u_i). \end{aligned}$$

Es ist

$$\frac{d}{du_i} L(u_i) = -\alpha_i + \frac{\alpha_i e^{u_i}}{\beta_i + \alpha_i e^{u_i}} = -\alpha_i + \frac{\alpha_i}{\alpha_i + \beta_i e^{-u_i}},$$

$$\begin{aligned}\frac{d^2}{du_i^2}L(u_i) &= \frac{\alpha_i\beta_i e^{-u_i}}{[\alpha_i+\beta_i e^{-u_i}]^2}\\ &= \left[\frac{\alpha_i}{\alpha_i+\beta_i e^{-u_i}}\right]\left[\frac{\beta_i e^{-u_i}}{\alpha_i+\beta_i e^{-u_i}}\right] \leq \frac{1}{4},\end{aligned}$$

denn $x(1-x) \leq \frac{1}{4}$ für $0 \leq x \leq 1$.

Taylor-Entwicklung von L um 0 ergibt

$$\begin{aligned}L(u_i) &= L(0) + u_i L'(0) + \frac{1}{2}u_i^2 L''(u^*)\\ &\leq 0 + 0 + \frac{1}{2}u_i^2\frac{1}{4}\\ &= \frac{1}{8}t^2\Delta_i^2.\end{aligned}$$

Also ist

$$\forall 1 \leq i \leq n : \log \mathbb{E}\left[e^{t\xi_i}\right] \leq \frac{1}{8}t^2\Delta_i^2,\ t \in \mathbb{R}.$$

Nach der exponentiellen Markov-Ungleichung gilt mit $S_n := \sum_{i=1}^n \xi_i$ für alle $t \geq 0$:

$$\begin{aligned}\mathbb{P}(S_n \geq \varepsilon) &\leq \exp(-\varepsilon t)\mathbb{E}\left[e^{tS_n}\right]\\ &= \exp(-\varepsilon t)\prod_{i=1}^n \mathbb{E}\left[e^{t\xi_i}\right]\\ &\leq \exp\left(-\varepsilon t + \frac{1}{8}t^2\sum_{i=1}^n \Delta_i^2\right). \qquad (2.8)\end{aligned}$$

Setze nun speziell $t = \frac{4\varepsilon}{\sum_{i=1}^n \Delta_i^2}$ und erhalte schließlich

$$\begin{aligned}\mathbb{P}(S_n \geq \varepsilon) &\leq \exp\left(-\frac{4\varepsilon^2}{\sum_{i=1}^n \Delta_i^2} + \frac{2\varepsilon^2}{\sum_{i=1}^n \Delta_i^2}\right)\\ &= \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n \Delta_i^2}\right)\end{aligned}$$

wie gewünscht. ■

Bemerkung 2.10

Die Wahl $t = \frac{4\varepsilon}{\sum_{i=1}^n \Delta_i^2}$ im Beweis von Satz 2.9 ist optimal in dem Sinne, dass sie zur schärfsten Abschätzung in (2.8) *(über alle $t \geq 0$) führt, siehe Übungsaufgabe.*

Korollar 2.11

Wendet man die Hoeffding-Ungleichung (2.7) *auf $(\xi_i)_{1\leq i\leq n}$ und $(-\xi_i)_{1\leq i\leq n}$ (jeweils) an und verwendet die Bonferroni-Ungleichung, so erhält man unter den Voraussetzungen von Satz 2.9, dass $\forall \varepsilon > 0$ gilt:*

$$\mathbb{P}\left(\left|\sum_{i=1}^n \xi_i\right| \geq \varepsilon\right) \leq 2\exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n \Delta_i^2}\right). \qquad (2.9)$$

Korollar 2.12
Unter den Voraussetzungen von Definition 2.7 sei $|\Theta| = K \in \mathbb{N}, \Theta = \{\theta_1, ..., \theta_K\}$.

Wir rechnen:

$$P\left(\max_{1\le k\le K}\left|n^{-1}\sum_{i=1}^{n}\xi_i^{(\theta_k)} - E\left[\xi_1^{(\theta_k)}\right]\right| > \varepsilon\right) \le \sum_{k=1}^{K} P\left(\left|n^{-1}\sum_{i=1}^{n}\xi_i^{(\theta_k)} - E\left[\xi_1^{(\theta_k)}\right]\right| > \varepsilon\right)$$
$$\le 2K\exp(-2\varepsilon^2 n),$$

nach Korollar 2.11, angewendet auf $\left(\xi_i^{(\theta_k)} - E\left[\xi_1^{(\theta_k)}\right]\right)_{1\le i\le n}$, wobei wir der Einfachheit halber (und ohne Beschränkung der Allgemeinheit im Falle beschränkter Verlustfunktionen) $\Delta_i \equiv 1$ annehmen.
Da $2K\exp(-2\varepsilon^2 n) = 2\exp\left(\left[\frac{\ln K}{n} - 2\varepsilon^2\right] n\right)$ ist und

$$\lim_{n\to\infty}\frac{\ln K}{n} = 0 \tag{2.10}$$

gilt, erhalten wir die Gültigkeit von (2.6).
Es stellt sich heraus, dass Bedingungen der Form (2.10) auch im Falle nicht-endlicher Parameterräume Θ von entscheidender Bedeutung sind, wobei K durch ein geeignetes Komplexitätsmaß zu ersetzen ist.

Definition 2.13 (Entropie einer Menge von Indikatorfunktionen)
Sei $\{Q(\cdot,\theta) : \theta \in \Theta\}$ *eine Menge von Indikatorfunktionen, d.h.,*

$$\forall \boldsymbol{z} \in \mathcal{Z} : \forall \theta \in \Theta : Q(\boldsymbol{z},\theta) \in \{0,1\}.$$

Seien Punkte $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$ *gegeben mit* $\boldsymbol{z}_i \in \mathcal{Z}$ *für alle* $1 \le i \le n$.
Sei die Zahl $N^{\Theta}(\boldsymbol{z}_1, ..., \boldsymbol{z}_n)$ *die Anzahl unterschiedlicher Möglichkeiten, die Punkte* $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$ *mit Hilfe der Indikatorfunktionen* $Q(\cdot,\theta), \theta \in \Theta$, *in zwei Klassen aufzuteilen*

$$(\textit{1. Klasse: solche } \boldsymbol{z}_i \textit{ mit } Q(\boldsymbol{z}_i,\theta) = 0,$$

$$\textit{2. Klasse: solche } \boldsymbol{z}_i \textit{ mit } Q(\boldsymbol{z}_i,\theta) = 1).$$

Dies kann auch wie folgt formalisiert werden. Für jedes feste $\theta \in \Theta$ *kann der Binärvektor* $(Q(\boldsymbol{z}_1,\theta), ..., Q(\boldsymbol{z}_n,\theta))^{\top} \in \{0,1\}^n$ *mit einer Ecke des* n*-dimensionalen Einheitswürfel identifiziert werden. Damit ist* $N^{\Theta}(\boldsymbol{z}_1, ..., \boldsymbol{z}_n)$ *die Anzahl unterschiedlicher Eckpunkte, die man mit den* $\theta \in \Theta$ *auf der Basis der gegebenen Werte* $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$ *erreichen kann.*

$$\textit{Offenbar gilt stets}: \ 1 \le N^{\Theta}(\boldsymbol{z}_1, ..., \boldsymbol{z}_n) \le 2^n.$$

Nehmen wir nun an, dass $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$ *Realisierungen von Zufallsvariablen* $\mathbf{Z}_1 = \boldsymbol{z}_1, ..., \mathbf{Z}_n = \boldsymbol{z}_n$ *sind, wobei* $\mathbf{Z}_1, ..., \mathbf{Z}_n$ *i.i.d. mit* $\mathbf{Z}_1 \sim P$, P *ein Wahrscheinlichkeitsmaß auf* $(\mathcal{Z}, \mathcal{F}_{\mathcal{Z}})$, *und dass*

die Abbildung $N^{\Theta}(\mathbf{Z}_1, ..., \mathbf{Z}_n)$ messbar ist.
Dann nennen wir

$$H^{\Theta}(\mathbf{Z}_1, ..., \mathbf{Z}_n) = \ln N^{\Theta}(\mathbf{Z}_1, ..., \mathbf{Z}_n)$$

die (zufällige) Entropie von $\{Q(\cdot, \theta) : \theta \in \Theta\}$ bezüglich P und

$$H^{\Theta}(n) := \mathbb{E}_{P^{\otimes n}}[H^{\Theta}(\mathbf{Z}_1, ..., \mathbf{Z}_n)]$$

die (erwartete) Entropie von $\{Q(\cdot, \theta) : \theta \in \Theta\}$ bezüglich P.

Satz 2.14 (Theorem 3.3 in Vapnik (1998))
Unter den Voraussetzungen von Definition 2.13 gilt

$$\forall \varepsilon > 0 : P\left(\sup_{\theta \in \Theta} \left| \int Q(\boldsymbol{z}, \theta) P(d\boldsymbol{z}) - n^{-1} \sum_{i=1}^{n} Q(\mathbf{Z}_i, \theta) \right| > \varepsilon \right) \to 0 \text{ für } n \to \infty$$

genau dann, wenn

$$\lim_{n \to \infty} \frac{H^{\Theta}(n)}{n} = 0. \tag{2.11}$$

Bemerkung 2.15

(a) Man beachte die strukturelle Analogie von (2.10) *und* (2.11)*. Somit ist die (erwartete) Entropie hier das geeignete Komplexitätsmaß für Θ (unter P), mit dem z.B. Konsistenz von ERM im Kontext der binären Klassifikation beurteilt werden kann.*

(b) Gilt $N^{\Theta}(\mathbf{Z}_1, ..., \mathbf{Z}_n) = 2^n$ P-fast sicher für alle $n \in \mathbb{N}$, so ist (2.11) *verletzt. Dann ist Θ so „reichhaltig", dass man mit den entsprechenden Indikatorfunktionen (fast) jeden Datensatz (der gemäß P zustande kommt) „perfekt erklären"kann. Dies führt zu Überanpassung und Inkonsistenz von ERM.*

Definition 2.16 (ε-Netz)
Sei (M, ρ) ein metrischer Raum und G eine Teilmenge von M. Dann heißt eine Teilmenge B_ε von M ein ε-Netz von G, falls

$$\forall g \in G \; \exists b \in B_\varepsilon : \rho(b, g) < \varepsilon, \; \varepsilon > 0.$$

Ferner sagen wir, dass G eine Überdeckung durch endliche ε-Netze besitzt, falls für jedes $\varepsilon > 0$ ein ε-Netz B_ε von G existiert, das aus endlich vielen Elementen besteht. Im letzteren Fall nennen wir das ε-Netz B_ε^* von G minimal, falls es die minimal mögliche Anzahl an Elementen enthält.

Definition 2.17 (Entropie einer Menge beschränkter reellwertiger Funktionen)
Sei $\{Q(\cdot, \theta) : \theta \in \Theta\}$ eine Menge beschränkter reellwertiger Verlustfunktionen, so dass eine reelle Konstante A existiert mit

$$\forall \theta \in \Theta : \; \forall \boldsymbol{z} \in \mathcal{Z} : \; |Q(\boldsymbol{z}, \theta)| \leq A.$$

Seien zudem $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$ gegebene Punkte mit $\boldsymbol{z}_i \in \mathcal{Z}, 1 \leq i \leq n$.
Für alle $\theta \in \Theta$ sei der n-dimensionale Vektor $q^*(\theta)$ gegeben durch

$$q^*(\theta) = (Q(\mathbf{z}_1, \theta), ..., Q(\mathbf{z}_n, \theta))^\top \in [-A, A]^n.$$

Die Menge $\{q^*(\theta) : \theta \in \Theta\}$ ist eine Teilmenge des n-dimensionalen Würfels mit Kantenlänge $2A$.
Wir betrachten nun auf $\mathbb{R}^n$ die Chebyshev-Metrik ρ_C, gegeben durch

$$\rho_C(\mathbf{x}, \mathbf{y}) = \max_{1 \leq i \leq n} |x_i - y_i|,\ \mathbf{x} = (x_1, ..., x_n)^\top \in \mathbb{R}^n,\ \mathbf{y} = (y_1, ..., y_n)^\top \in \mathbb{R}^n.$$

Sei $N^\Theta(\varepsilon; \boldsymbol{z}_1, ..., \boldsymbol{z}_n)$ die Anzahl der Elemente eines minimalen ε-Netzes von $\{q^*(\theta) : \theta \in \Theta\}$ bezüglich der Metrik $\rho_C, \varepsilon > 0$. Wie in Definition 2.13 nehmen wir nun an, dass die Abbildung $N^\Theta(\varepsilon; \mathbf{Z}_1, ..., \mathbf{Z}_n)$ messbar ist, wobei $\mathbf{Z}_1, ..., \mathbf{Z}_n$ i.i.d. sind mit $\mathbf{Z}_1 \sim P$.
Dann nennen wir

$$H^\Theta(\varepsilon; \mathbf{Z}_1, ..., \mathbf{Z}_n) := \ln N^\Theta(\varepsilon; \mathbf{Z}_1, ..., \mathbf{Z}_n)$$

die zufällige ε-Entropie von $\{Q(\cdot, \theta) : \theta \in \Theta\}$ bezüglich P, und $H^\Theta(\varepsilon; n) = \mathbb{E}_{P^{\otimes n}}[H^\Theta(\varepsilon; \mathbf{Z}_1, ..., \mathbf{Z}_n)]$ die (erwartete) ε-Entropie von $\{Q(\cdot, \theta) : \theta \in \Theta\}$ bezüglich P.

Bemerkung 2.18
Da $[-A, A]^n$ eine kompakte Teilmenge des $\mathbb{R}^n$ ist, ist die Existenz eines minimalen ε-Netzes von $\{q^*(\theta) : \theta \in \Theta\}$ sichergestellt.

Satz 2.19 (Theorem 3.4 in Vapnik (1998))
Unter den Voraussetzungen von Definition 2.17 gilt $\forall \varepsilon > 0$:

$$P\left(\sup_{\theta \in \Theta} \left| \int Q(\boldsymbol{z}, \theta) P(d\boldsymbol{z}) - n^{-1} \sum_{i=1}^{n} Q(\mathbf{Z}_i, \theta) \right| > \varepsilon \right) \to 0 \text{ für } n \to \infty$$

genau dann, wenn

$$\forall \varepsilon > 0 : \lim_{n \to \infty} \frac{H^\Theta(\varepsilon; n)}{n} = 0. \tag{2.12}$$

Erneut ist also die (erwartete) Entropie das geeignete Komplexitätsmaß für Θ.

Bemerkung 2.20

(a) *Satz 2.14 und Satz 2.19 beschäftigen sich mit gleichmäßiger zweiseitiger (stochastischer) Konvergenz. Für die Konsistenz von ERM ist indes gemäß Satz 2.6 die gleichmäßige einseitige Konvergenz bereits hinreichend.*
Die Beschränkung der (erwarteten) Entropie ist indes auch in diesem Fall essentiell, vgl. Abschnitt 2.4 in Vapnik (2000).

(b) Verallgemeinerungen auf unbeschränkte Verlustfunktionen (wie z.B. den quadratischen Verlust bei Regressionsproblemen) finden sich in Abschnitt 3.9 von Vapnik (1998). Im Wesentlichen wird dabei die Entropie-Bedingung (2.12) *für jede Funktionenmenge* $\{Q_A(\cdot, \theta) : \theta \in \Theta\}$ *mit* $A > 0$ *gefordert, wobei*

$$Q_A(\boldsymbol{z}, \theta) = \begin{cases} A, & Q(\boldsymbol{z}, \theta) > A, \\ Q(\boldsymbol{z}, \theta), & |Q(\boldsymbol{z}, \theta)| \leq A, \\ -A, & Q(\boldsymbol{z}, \theta) < -A. \end{cases}$$

Ferner muss eine (bezüglich P*) integrierbare Funktion* K *existieren, mit*

$$\sup_{\theta \in \Theta} |Q(\boldsymbol{z}, \theta)| \leq K(\boldsymbol{z}),$$

für alle $\boldsymbol{z} \in \mathcal{Z}$.

Kapitel 3

Konvergenzgeschwindigkeit statistischer Lernverfahren

In Kapitel 2 haben wir notwendige und hinreichende Bedingungen für die Konsistenz von ERM bezüglich einer (festen) Verteilung P von $\mathbf{Z}_1$ kennengelernt.
Defizite dabei:

1) Konsistenz ist ein rein qualitatives (konzeptionelles) Kriterium, das nichts darüber aussagt, wie schnell $R_{emp}(\hat{\theta}(n))$ sich dem Wert $\inf_{\theta\in\Theta} R(\theta)$ (stochastisch) nähert. Insbesondere kann man mit diesem Konzept in der Praxis nicht abschätzen, wie groß der Stichprobenumfang n gewählt werden sollte, um eine hinreichend präzise Funktionenschätzung zu erhalten.

2) Die Entropie-Untersuchungen in Satz 2.14 und Satz 2.19 sind jeweils an ein festgelegtes Wahrscheinlichkeitsmaß P gebunden, während in der Praxis typischerweise Unsicherheit über den Daten-generierenden probabilistischen Prozess herrscht.

Beide Aspekte werden in diesem Kapitel 3 behandelt.

Definition 3.1 (Schnelle Konvergenz)

(a) Wir sagen, dass ERM unter P schnell konvergiert, falls es zwei positive reelle Konstanten b und c gibt, so dass für alle $n > n_0 = n_0(\varepsilon, \Theta, P)$ die Ungleichung

$$P\left(\sup_{\theta\in\Theta}\left|\int Q(\boldsymbol{z},\theta)P(d\boldsymbol{z}) - n^{-1}\sum_{i=1}^{n} Q(\mathbf{Z}_i,\theta)\right| > \varepsilon\right) < b\exp(-c\,\varepsilon^2 n) \tag{3.1}$$

gilt.

(b) Wir sagen, dass ERM stets schnell konvergiert, falls es zwei positive reelle Konstanten b und

c gibt, so dass für alle $n > n_0 = n_0(\varepsilon, \Theta)$ die Ungleichung

$$\sup_P P\left(\sup_{\theta\in\Theta}\left|\int Q(\boldsymbol{z},\theta)P(d\boldsymbol{z}) - n^{-1}\sum_{i=1}^{n} Q(\mathbf{Z}_i,\theta)\right| > \varepsilon\right) < b\exp(-c\,\varepsilon^2 n) \tag{3.2}$$

gilt, wobei das $\sup_P$ in (3.2) über alle Wahrscheinlichkeitsverteilungen auf $(\mathcal{Z}, \mathcal{F}_{\mathcal{Z}})$ gebildet wird.

Es zeigt sich, dass zur Analyse der Gültigkeit von (3.1)und (3.2) weitere Entropie- bzw. Komplexitätsbegriffe für Θ gebraucht werden.

Definition 3.2 (Entropiebegriffe für Familien von Indikatorfunktionen)
Unter den Voraussetzungen von Definition 2.13 heißt

$$H_{ann}^{\Theta}(n) := \ln\left(\mathbb{E}_{P^{\otimes n}}\left[N^{\Theta}(\mathbf{Z}_1, ..., \mathbf{Z}_n)\right]\right) \tag{3.3}$$

die verschärfte (englisch: annealed) Entropie von $\{Q(\cdot,\theta) : \theta \in \Theta\}$ bezüglich P und

$$G^{\Theta}(n) := \ln \sup_{\mathbf{z}_1,...,\mathbf{z}_n} N^{\Theta}(\mathbf{z}_1, ..., \mathbf{z}_n) \tag{3.4}$$

die Wachstumsfunktion von $\{Q(\cdot,\theta) : \theta \in \Theta\}$. Wegen der Jensen'schen Ungleichung gilt

$$H^{\Theta}(n) \le H_{ann}^{\Theta}(n) \le G^{\Theta}(n) \le n\ln(2).$$

Definition 3.3 (Entropiebegriffe für Familien von beschränkten, reellwertigen Funktionen)
Unter den Voraussetzungen von Definition 2.17 heißt

$$H_{ann}^{\Theta}(\varepsilon; n) := \ln\left(\mathbb{E}_{P^{\otimes n}}\left[N^{\Theta}(\varepsilon; \mathbf{Z}_1, ..., \mathbf{Z}_n)\right]\right) \tag{3.5}$$

die verschärfte (annealed) ε-Entropie von $\{Q(\cdot,\theta) : \theta \in \Theta\}$ bezüglich P und

$$G^{\Theta}(\varepsilon; n) := \ln \sup_{\mathbf{z}_1,...,\mathbf{z}_n} N^{\Theta}(\varepsilon; \mathbf{z}_1, \ldots, \mathbf{z}_n) \tag{3.6}$$

die ε-Wachstumsfunktion von $\{Q(\cdot,\theta) : \theta \in \Theta\}$.
Auch hier gilt

$$H^{\Theta}(\varepsilon; n) \le H_{ann}^{\Theta}(\varepsilon; n) \le G^{\Theta}(\varepsilon; n).$$

Lemma 3.4
Sei $\{Q(\cdot,\theta) : \theta \in \Theta\}$ eine Menge von Indikatorfunktionen. Angenommen, $\theta^* = \operatorname{argmin}_{\theta\in\Theta} R(\theta)$ existiert. Dann gilt für alle $\eta \in (0,1)$, dass

$$P\left(R(\theta^*) > R_{emp}(\theta^*) - \sqrt{\frac{-\ln(\eta)}{2n}}\right) \ge 1 - \eta. \tag{3.7}$$

Beweis: Wir wenden Satz 2.9 (Hoeffding-Ungleichung) an auf $\xi_i := \left[Q(\mathbf{Z}_i, \theta^*) - \int Q(\boldsymbol{z}, \theta^*) P(d\boldsymbol{z})\right]$, $1 \leq i \leq n$, und beachten, dass dann $\Delta_i \equiv 1$ für alle $1 \leq i \leq n$ ist. Also ist für alle $\varepsilon > 0$:

$$P\left(\sum_{i=1}^{n} \xi_i \geq \varepsilon\right) \leq \exp\left(\frac{-2\varepsilon^2}{n}\right)$$

$$\Leftrightarrow \quad P\left(\frac{1}{n}\sum_{i=1}^{n} \xi_i \geq \frac{\varepsilon}{n}\right) \leq \exp\left(\frac{-2\varepsilon^2}{n}\right).$$

Wir setzen $\varepsilon := n\delta$ für $\delta > 0$ beliebig und erhalten, dass

$$P\left(\frac{1}{n}\sum_{i=1}^{n} \xi_i \geq \delta\right) \leq \exp\left(-2n\delta^2\right)$$

$$\Leftrightarrow \quad P(R_{emp}(\theta^*) - R(\theta^*) \geq \delta) \leq \exp\left(-2n\delta^2\right)$$

$$\Leftrightarrow \quad P(R_{emp}(\theta^*) - R(\theta^*) < \delta) \geq 1 - \exp\left(-2n\delta^2\right)$$

$$\Leftrightarrow \quad P(R(\theta^*) > R_{emp}(\theta^*) - \delta) \geq 1 - \exp\left(-2n\delta^2\right).$$

Setzen wir nun speziell $\delta = \sqrt{\frac{-\ln(\eta)}{2n}}$, so ergibt sich schließlich

$$\begin{aligned} P\left(R(\theta^*) > R_{emp}(\theta^*) - \sqrt{\frac{-\ln(\eta)}{2n}}\right) &\geq 1 - \exp\left(-2n\left[\frac{-\ln(\eta)}{2n}\right]\right) \\ &= 1 - \eta, \end{aligned}$$

wie gewünscht. ■

Satz 3.5 (Theorem 4.1 in (Vapnik, 1998))

Unter den Voraussetzungen von Lemma 3.4 gilt für jedes $\varepsilon > 0$:

$$P\left(\sup_{\theta \in \Theta} |R(\theta) - R_{emp}(\theta)| > \varepsilon\right) < 4 \exp\left(\left[\frac{H_{ann}^{\Theta}(2n)}{n} - \left(\varepsilon - \frac{1}{n}\right)^2\right] n\right) \tag{3.8}$$

Korollar 3.6

Unter den Voraussetzungen von Satz 3.5 ist die Bedingung

$$\lim_{n \to \infty} \frac{H_{ann}^{\Theta}(n)}{n} = 0 \tag{3.9}$$

hinreichend dafür, dass ERM unter P schnell konvergiert.

Korollar 3.7

Unter den Voraussetzungen von Satz 3.5 gilt

$$P\left(R(\hat{\theta}(n)) - R(\theta^*) \leq \sqrt{\frac{H_{ann}^{\Theta}(2n) - \ln\left(\frac{\eta}{4}\right)}{n}} + \sqrt{\frac{-\ln(\eta)}{2n}} + \frac{1}{n}\right) \quad \geq \quad 1 - 2\eta.$$

Also konvergiert $R(\hat{\theta}(n))$ exponentiell schnell stochastisch gegen $R(\theta^*)$.

Beweis: Wegen (3.8) ist für jedes feste $n \in \mathbb{N}$

$$P\left(R(\hat{\theta}(n)) < R_{emp}(\hat{\theta}(n)) + \varepsilon\right) > 1 - 4\exp\left(\left[\frac{H_{\text{ann}}^{\Theta}(2n)}{n} - \left(\varepsilon - \frac{1}{n}\right)^2\right] n\right).$$

Wir setzen nun speziell

$$\varepsilon := \sqrt{\frac{H_{\text{ann}}^{\Theta}(2n) - \ln\left(\frac{\eta}{4}\right)}{n}} + \frac{1}{n}.$$

Damit ist

$$4\exp\left(\left[\frac{H_{\text{ann}}^{\Theta}(2n)}{n} - \left(\varepsilon - \frac{1}{n}\right)^2\right] n\right) = 4\exp\left(\ln\left(\frac{\eta}{4}\right)\right) = \eta$$

und somit gilt mit Wahrscheinlichkeit mindestens $1 - \eta$, dass

$$R(\hat{\theta}(n)) < R_{emp}(\hat{\theta}(n)) + \sqrt{\frac{H_{\text{ann}}^{\Theta}(2n) - \ln\left(\frac{\eta}{4}\right)}{n}} + \frac{1}{n}. \qquad (\star)$$

Andererseits liefert Lemma 3.4, dass mit Wahrscheinlichkeit mindestens $1 - \eta$ gilt:

$$R(\theta^*) > R_{emp}(\theta^*) - \sqrt{\frac{-\ln(\eta)}{2n}} \qquad (\star\star)$$

Nehmen wir $(\star)$ und $(\star\star)$ zusammen und beachten, dass $P(A \cup B) \leq P(A) + P(B)$ ist, so ergibt sich, dass mit Wahrscheinlichkeit mindestens $1 - 2\eta$ gilt:

$$\begin{aligned} R(\hat{\theta}(n)) - R(\theta^*) &< R_{emp}(\hat{\theta}(n)) - R_{emp}(\theta^*) + \sqrt{\frac{H_{\text{ann}}^{\Theta}(2n) - \ln\left(\frac{\eta}{4}\right)}{n}} + \frac{1}{n} + \sqrt{\frac{-\ln(\eta)}{2n}} \\ &\leq \sqrt{\frac{H_{\text{ann}}^{\Theta}(2n) - \ln\left(\frac{\eta}{4}\right)}{n}} + \sqrt{\frac{-\ln(\eta)}{2n}} + \frac{1}{n}, \end{aligned}$$

da $R_{emp}(\hat{\theta}(n)) - R_{emp}(\theta^*) \leq 0$ ist. ■

Definition 3.8 (Vapnik-Chervonenkis (VC)-Dimension)
Sei $\mathcal{M} = \{Q(\cdot, \theta) : \theta \in \Theta\}$ eine Menge von Indikatorfunktionen auf $\mathcal{Z}$.
Wir definieren

$$h(\Theta) := \max\left\{\begin{array}{c} j \in \mathbb{N} : \textit{Jede der } 2^j \textit{ verschiedenen Möglichkeiten, } j \textit{ Elemente von } \mathcal{Z} \\ \textit{in zwei disjunkte Klassen aufzuteilen, lässt sich mit Funktionen} \\ \textit{aus } \mathcal{M} \textit{ verwirklichen (durch geschickte Wahl der } j \textit{ Elemente)} \end{array}\right\}.$$

Falls sich für jedes $n \in \mathbb{N}$ Elemente $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$ von $\mathcal{Z}$ finden lassen, die auf alle 2^n verschiedenen Möglichkeiten mit Funktionen aus $\mathcal{M}$ in zwei disjunkte Klassen aufgeteilt werden können, so setzen wir $h(\Theta) = \infty$.
Wir nennen $h(\Theta)$ die Vapnik-Chervonenkis (VC)-Dimension von $\mathcal{M}$. Ferner nennen wir $\mathcal{M}$ eine Vapnik-Chervonenkis (VC)-Klasse, falls $h(\Theta) < \infty$ ist.

Lemma 3.9

Sei $n \in \mathbb{N}$ und $h \leq n$. Dann gilt:

$$\sum_{j=0}^{h} \binom{n}{j} \leq \left(\frac{n\,e}{h}\right)^h. \tag{3.10}$$

Beweis: Für jedes $0 \leq j \leq h$ ist

$$\binom{n}{j} = \frac{n(n-1)\cdots(n-j+1)}{j!} \leq \frac{n^j}{j!}.$$

Damit ist

$$\begin{aligned}
\sum_{j=0}^{h} \binom{n}{j} \leq \sum_{j=0}^{h} \frac{n^j}{j!} &= \sum_{j=0}^{h} \frac{h^j}{j!} \left(\frac{n}{h}\right)^j \\
&\leq \left(\frac{n}{h}\right)^h \sum_{j=0}^{h} \frac{h^j}{j!} \\
&\leq \left(\frac{n}{h}\right)^h e^h = \left(\frac{n\,e}{h}\right)^h.
\end{aligned}$$

■

Korollar 3.10

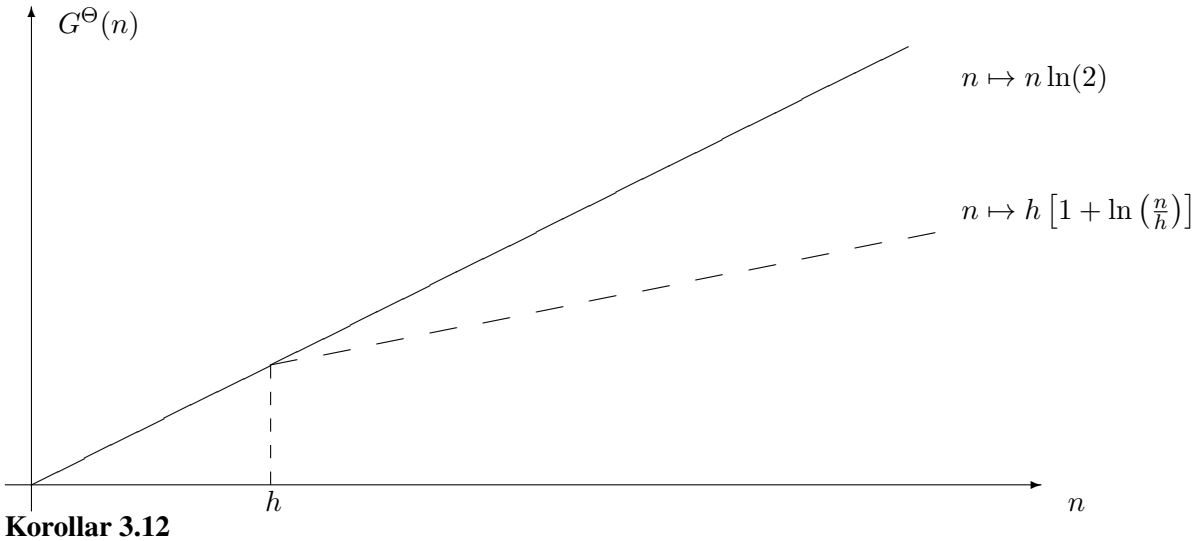
Unter den Voraussetzungen von Definition 3.8 gilt für die Wachstumsfunktion $G^{\Theta}(\cdot)$ von $\{Q(\cdot,\theta) : \theta \in \Theta\}$:

(a) Falls $h = h(\Theta) = \infty$ ist, so ist $G^{\Theta}(n) = n\ln(2), n \in \mathbb{N}$.

(b) Falls $h = h(\Theta) < \infty$ ist, so ist für alle $n \in \mathbb{N}$

$$G^{\Theta}(n) \begin{cases} = n\ln(2), & n \leq h \equiv h(\Theta). \\ \leq \ln\left(\sum_{j=0}^{h} \binom{n}{j}\right) \leq \ln\left(\left(\frac{n\,e}{h}\right)^h\right) & \\ \qquad = h\ln\left(\frac{n\,e}{h}\right) & n > h \equiv h(\Theta). \\ \qquad = h\left[1 + \ln\left(\frac{n}{h}\right)\right], & \end{cases}$$

Schema 3.11



Korollar 3.12

Unter den Voraussetzungen von Definition 3.8 gilt

$$\lim_{n\to\infty} \frac{G^{\Theta}(n)}{n} = 0$$

genau dann, wenn $\mathcal{M}$ eine VC-Klasse ist, d.h., wenn $h(\Theta) < \infty$ ist.

Korollar 3.13 (zu Satz 3.5)

Sei $\mathcal{M} = \{Q(\cdot, \theta) : \theta \in \Theta\}$ eine Menge von Indikatorfunktionen mit endlicher VC-Dimension $h \equiv h(\Theta)$. Dann gilt für jedes Wahrscheinlichkeitsmaß P auf $(\mathcal{Z}, \mathcal{F}_{\mathcal{Z}})$ und jedes $n > h$, dass

$$P\left(\sup_{\theta\in\Theta} |R(\theta) - R_{emp}(\theta)| > \varepsilon\right) < 4 \exp\left(\left[\frac{h\left[1 + \ln\left(\frac{2n}{h}\right)\right]}{n} - \left(\varepsilon - \frac{1}{n}\right)^2\right] n\right). \qquad (3.11)$$

Beweis: Es gilt stets (für jedes P), dass $H^{\Theta}_{\text{ann}}(2n) \leq G^{\Theta}(2n) \leq h\left[1 + \ln\left(\frac{2n}{h}\right)\right]$, $n > h$, gemäß Korollar 3.10. Damit folgt (3.11) sofort aus (3.8). ■

Bemerkung 3.14

In Analogie zur Argumentation in Korollar 3.7 kann (3.11) *auch in einen „Konfidenzbereich“ für das „Exzess-Risiko“ $R(\hat{\theta}(n)) - R(\theta^*)$ umgerechnet werden.*

Korollar 3.15

Unter den Voraussetzungen von Korollar 3.13 gilt:
Falls $h \equiv h(\Theta) < \infty$ ist, d.h., falls $\lim_{n\to\infty} G^{\Theta}(n)/n = 0$ gilt, so konvergiert ERM stets schnell.

Satz 3.16

Gleichmäßige zweiseitige (stochastische) Konvergenz von $n^{-1} \sum_{i=1}^{n} Q(\mathbf{Z}_i, \theta)$ gegen $\int Q(\boldsymbol{z}, \theta) P(d\boldsymbol{z})$ für jedes Wahrscheinlichkeitsmaß P auf $(\mathcal{Z}, \mathcal{F}_{\mathcal{Z}})$ gilt unter den Voraussetzungen von Korollar 3.13 genau dann, wenn $h \equiv h(\Theta) < \infty$ ist. Man sagt auch, dass die Eigenschaft „VC-Klasse“ äquivalent zur Eigenschaft „Glivenko-Cantelli-Klasse“ ist.

Beweis: Es bleibt wegen Korollar 3.15 nur noch, die Notwendigkeit von $h < \infty$ für die behauptete gleichmäßige zweiseitige (stochastische) Konvergenz zu zeigen.
Nehmen wir dazu also an, die Menge $\mathcal{M} = \{Q(\cdot, \theta) : \theta \in \Theta\}$ ist keine VC-Klasse. Dann gilt für jedes $n \in \mathbb{N}$ die Gleichheit

$$\sup_{\boldsymbol{z}_1,...,\boldsymbol{z}_n} N^{\Theta}(\boldsymbol{z}_1, ..., \boldsymbol{z}_n) = 2^n. \qquad (\star)$$

Wir müssen zeigen, dass unter ($\star$) für jedes $n \in \mathbb{N}$ und jedes $\varepsilon > 0$ ein Wahrscheinlichkeitsmaß P auf $(\mathcal{Z}, \mathcal{F}_{\mathcal{Z}})$ konstruiert werden kann, so dass mit Wahrscheinlichkeit 1 gilt:

$$\sup_{\theta\in\Theta} \left| \int Q(\boldsymbol{z}, \theta) P(d\boldsymbol{z}) - n^{-1} \sum_{i=1}^{n} Q(\mathbf{Z}_i, \theta) \right| > 1 - \varepsilon, \qquad (\star\star)$$

wobei $\mathbf{Z}_1, ..., \mathbf{Z}_n$ i.i.d. mit $\mathbf{Z}_1 \sim P$ sind.
Sei dazu $K \in \mathbb{N}$ so gewählt, dass $K > n/\varepsilon$ ist. Dann ist es wegen ($\star$) (angewendet auf K statt n) möglich, K Elemente $\boldsymbol{z}_1, ..., \boldsymbol{z}_K$ von $\mathcal{Z}$ so auszuwählen, dass diese Elemente von Funktionen aus $\mathcal{M}$ auf alle 2^K verschiedenen Möglichkeiten in die Klassen „0“ und „1“ eingeteilt werden können. Sei P nun die diskrete Gleichverteilung auf $\{\boldsymbol{z}_1, ..., \boldsymbol{z}_K\}$.
Ist dann $\mathbf{Z}_1, ..., \mathbf{Z}_n$ eine i.i.d.-Stichprobe mit $\mathbf{Z}_1 \sim P$, so bezeichne

$$\mathcal{Z}^* = \{\boldsymbol{z} \in \{\boldsymbol{z}_1, ..., \boldsymbol{z}_K\} : \nexists 1 \le i \le n \text{ mit } \mathbf{Z}_i = \boldsymbol{z}\}.$$

Es ist evident, dass $|\mathcal{Z}^*| \ge K - n$ ist. Da $N^{\Theta}(\boldsymbol{z}_1, ..., \boldsymbol{z}_K) = 2^K$ ist, existiert ein $\theta^* \in \Theta$, so dass

$$\begin{aligned} \forall \boldsymbol{z} \in \mathcal{Z}^* : Q(\boldsymbol{z}, \theta^*) &= 1, \\ \forall 1 \le i \le n : Q(\mathbf{Z}_i, \theta^*) &= 0 \text{ (mit Wahrscheinlichkeit } 1). \end{aligned}$$

Demnach ist $n^{-1} \sum_{i=1}^{n} Q(\mathbf{Z}_i, \theta^*) = 0$ mit Wahrscheinlichkeit 1, aber

$$\int Q(\boldsymbol{z}, \theta^*) P(d\boldsymbol{z}) \ge \frac{K - n}{K} = 1 - \frac{n}{K} > 1 - \varepsilon,$$

wegen der Konstruktion (Wahl) von K. Somit folgt ($\star\star$). ■

Wenden wir uns nun allgemein Klassen von reellwertigen (beschränkten) Verlustfunktionen zu.

Satz 3.17 (Theorem 15.2 in Vapnik (1998))
Sei $\{Q(\cdot, \theta) : \theta \in \Theta\}$ eine Menge von beschränkten, reellwertigen Verlustfunktionen mit

$$\forall \theta \in \Theta : \forall \boldsymbol{z} \in \mathcal{Z} : -\infty < A \le Q(\boldsymbol{z}, \theta) \le B < \infty.$$

Sei P ein gegebenes Wahrscheinlichkeitsmaß auf $(\mathcal{Z}, \mathcal{F}_{\mathcal{Z}})$ und seien $\mathbf{Z}_1, ..., \mathbf{Z}_n$ i.i.d.mit $\mathbf{Z}_1 \sim P$. Dann gibt es für alle genügend große $n \in \mathbb{N}$ eine Konstante c, so dass gilt:

$$\begin{aligned} &P\left(\sup_{\theta\in\Theta} \left| \int Q(\boldsymbol{z}, \theta) P(d\boldsymbol{z}) - n^{-1} \sum_{i=1}^{n} Q(\mathbf{Z}_i, \theta) \right| > \varepsilon \right) \\ &\le \exp\left(\left[\frac{H^{\Theta}_{ann}(\varepsilon/[6(B-A)]; n)}{n} - \frac{\varepsilon^2}{36(B-A)^2} + \frac{c + \ln(n)}{n} \right] n \right). \end{aligned}$$

Also ist die Bedingung

$$\forall \varepsilon > 0 : \lim_{n\to\infty} \frac{H^{\Theta}_{ann}(\varepsilon; n)}{n} = 0 \tag{3.12}$$

hinreichend dafür, dass ERM unter P schnell konvergiert.

Definition 3.18 (VC-Dimension von Klassen reellwertiger Funktionen)
Sei $\mathcal{M} = \{Q(\cdot, \theta) : \theta \in \Theta\}$ eine Menge reellwertiger Verlustfunktionen.
Definiere

$$\begin{aligned} A &:= \inf_{\boldsymbol{z},\theta} Q(\boldsymbol{z}, \theta) \in \mathbb{R} \cup \{-\infty\}, \\ B &:= \sup_{\boldsymbol{z},\theta} Q(\boldsymbol{z}, \theta) \in \mathbb{R} \cup \{+\infty\}. \end{aligned}$$

Dann ist die VC-Dimension $h \equiv h(\Theta)$ von $\mathcal{M}$ definiert als die VC-Dimension der Menge $\{I(\cdot, \theta, \gamma) : \theta \in \Theta, \gamma \in (A, B)\}$ von Indikatorfunktionen, wobei

$$I(\boldsymbol{z}, \theta, \gamma) = \mathbf{1}\{Q(\boldsymbol{z}, \theta) \geq \gamma\},\ \boldsymbol{z} \in \mathcal{Z}, \theta \in \Theta, \gamma \in (A, B).$$

Satz 3.19 (vgl. Abschnitt 3.7 in Vapnik (2000))
Angenommen, unter den Voraussetzungen von Definition 3.18 sind $A, B \in \mathbb{R}$ und $h \equiv h(\Theta) < \infty$.
Definiere

$$\varepsilon \equiv \varepsilon(\eta; n, h) := 4\frac{h\left[\ln\left(\frac{2n}{h}\right) + 1\right] - \ln\left(\frac{\eta}{4}\right)}{n}.$$

Dann gelten die folgenden Aussagen für hinreichend großes $n \in \mathbb{N}$:

(a) Mit Wahrscheinlichkeit mindestens $1 - \eta$ simultan über alle $\theta \in \Theta$ ist

$$|R(\theta) - R_{emp}(\theta)| \leq \frac{B - A}{2}\sqrt{\varepsilon(\eta; n, h)}. \tag{3.13}$$

(b) Mit Wahrscheinlichkeit mindestens $1 - 2\eta$ ist

$$R(\hat{\theta}(n)) - \inf_{\theta\in\Theta} R(\theta) \leq (B - A)\sqrt{\frac{-\ln(\eta)}{2n}} + \frac{B - A}{2}\sqrt{\varepsilon(\eta; n, h)}. \tag{3.14}$$

Bemerkung 3.20

(a) Die Abschätzung (3.14) *folgt sofort aus* (3.13) *zusammen mit Lemma 3.4, vgl. Korollar 3.7.*

(b) Falls $|\Theta| = K \in \mathbb{N}$ ist, so kann in (3.13) *und* (3.14) *statt $\varepsilon(\eta; n, h)$ die Größe*

$$\varepsilon(\eta; n, K) = 2\frac{\ln(K) - \ln(\eta)}{n}$$

verwendet werden, vgl. Korollar 2.12.

(c) Es existieren ebenfalls Abschätzungen im Falle von $A = 0$ und $B = +\infty$, d.h., im Falle von nicht-negativen, unbeschränkten Verlustfunktionen.

Kapitel 4

Strukturelle Risikominimierung

Lemma 4.1 (siehe Section 4.2.1 in Cherkassky and Mulier (2007))

(a) *VC-Dimension bei binärer Klassifikation mit 0-1 Verlust*
Sei $\{f(\cdot,\theta) : \theta \in \Theta\}$ *eine Menge von Indikatorfunktionen, wobei*

$$\forall \theta \in \Theta : f(\cdot,\theta) : D \to \{0,1\} = W$$
$$x \mapsto f(x,\theta) = \hat{y} \in \{0,1\} = W$$

eine binäre Klassifikationsfunktion ist. Bezeichne die VC-Dimension von $\{f(\cdot,\theta) : \theta \in \Theta\}$ *mit* h_f.
Sein nun für $\boldsymbol{z} = (\boldsymbol{x}, y)$ *mit* $\boldsymbol{x} \in D$ *und* $y \in W = \{0,1\}$ *die Verlustfunktion* $Q(\cdot,\cdot)$ *gegeben durch*

$$Q(\boldsymbol{z},\theta) = |y - f(\boldsymbol{x},\theta)| \in \{0,1\}.$$

Diese Verlustfunktion entspricht offenbar der Verlustfunktion $L(\cdot,\cdot)$ *aus Beispiel 1.5.(a). Dann ist die VC-Dimension* h *von* $\{Q(\cdot,\theta) : \theta \in \Theta\}$ *gleich* h_f.

(b) *VC-Dimension bei (Mittelwert-) Regression mit quadratischem Verlust*
Sei $\{f(\cdot,\theta) : \theta \in \Theta\}$ *eine Menge von reellwertigen Funktionen, wobei*

$$\forall \theta \in \Theta : f(\cdot,\theta) : D \to \mathbb{R} = W$$
$$x \mapsto f(x,\theta) = \hat{y} \in \mathbb{R} = W$$

eine Regressionsfunktion ist. Bezeichne wieder h_f *die VC-Dimension von* $\{f(\cdot,\theta) : \theta \in \Theta\}$.
Sei für $\boldsymbol{z} = (\boldsymbol{x}, y)$ *mit* $\boldsymbol{x} \in D$ *und* $y \in W = \mathbb{R}$ *die Verlustfunktion* $Q(\cdot,\cdot)$ *gegeben durch*

$$Q(\boldsymbol{z},\theta) = (y - f(\boldsymbol{x},\theta))^2.$$

Dann gilt für die VC-Dimension h von $\{Q(\cdot, \theta) : \theta \in \Theta\}$, dass

$$h_f \leq h \leq c \cdot h_f, \tag{4.1}$$

wobei c eine universelle Konstante ist.

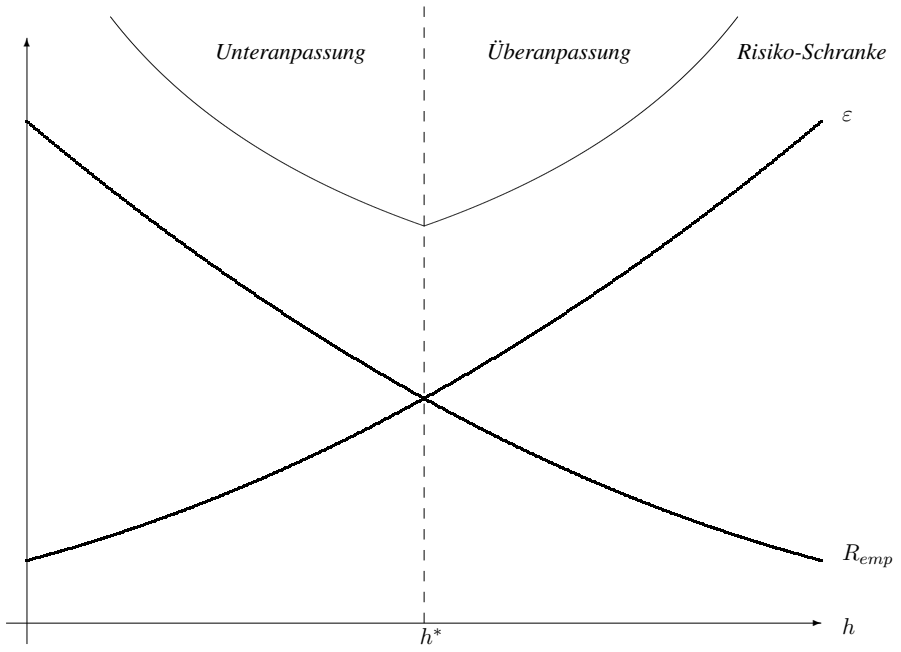
Basierend auf Lemma 4.1 werden wir in der Folge nicht mehr streng zwischen den VC-Dimensionen h und h_f unterscheiden, denn wegen (4.1) bleiben die Schranken in (3.13) und (3.14) auch dann noch gültig, wenn in der Definition von $\varepsilon \equiv \varepsilon(\eta; n, h)$ die VC-Dimension h_f statt h verwendet wird.

Schema 4.2

Die Abschätzung (3.13) *[und die analoge Abschätzung für gegebenenfalls nicht beschränkte Verlustfunktionen] lässt sich wie folgt paraphrasieren:*

$$\textit{Theoretisches Risiko} \leq \textit{empirisches Risiko + Komplexität von } \Theta, \tag{4.2}$$

wobei die Aussage nur mit einer gewissen „Konfidenzwahrscheinlichkeit “und nur für hinreichend große n gilt, so dass h/n nicht zu groß ist. Das empirische Risiko kann typischerweise dadurch verringert oder sogar auf Null gebracht werden („Überanpassung“), dass die Komplexität von Θ gesteigert wird.



Die Idee des Prinzips der strukturellen Risikominimierung (SRM) ist es daher, die Komplexität von Θ (gemessen in ihrer VC-Dimension h) mit in das Optimierungsproblem bezüglich $\hat{\theta}(n)$ aufzunehmen.

Definition 4.3 (Struktur auf $\mathcal{M}$)
Sei $\mathcal{M} = \{Q(\cdot, \theta) : \theta \in \Theta\}$ eine Menge von (nicht-negativen) Verlustfunktionen. Dann nennen wir eine aufsteigende Folge

$$\mathcal{M}_1 \subseteq \mathcal{M}_2 \subseteq ... \subseteq \mathcal{M}_d \subseteq ...$$

von Teilmengen der Form $\mathcal{M}_k = \{Q(\cdot, \theta) : \theta \in \Theta_k\}$ eine (zulässige) Struktur auf $\mathcal{M}$, falls gilt:

(a) Die VC-Dimension h_k von $\mathcal{M}_k$ ist endlich für alle $k \geq 1$. Selbstverständlich gilt

$$h_1 \leq h_2 \leq ... \leq h_d \leq ...$$

(b) Für alle $k \geq 1$ gilt

(i) Es existiert ein $B_k \in \mathbb{R}$ mit $\forall \theta \in \Theta_k : \forall \boldsymbol{z} \in \mathcal{Z} : 0 \leq Q(\boldsymbol{z}, \theta) \leq B_k$

oder

(ii) Es exisitieren $p > 2$ und $\tau_k \in \mathbb{R}$ mit

$$\sup_{\theta \in \Theta_k} \frac{\left[\int_{\mathcal{Z}} Q^p(\boldsymbol{z}, \theta) P(d\boldsymbol{z})\right]^{\frac{1}{p}}}{\int_{\mathcal{Z}} Q(\boldsymbol{z}, \theta) P(d\boldsymbol{z})} \leq \tau_k.$$

Nach Konstruktion gilt

$$B_1 \leq B_2 \leq ... \leq B_d \leq ...$$

beziehungsweise

$$\tau_1 \leq \tau_2 \leq ... \leq \tau_d \leq ...$$

Analog verfahren wir für eine Menge $\mathcal{M}_k = \{f(\cdot, \theta) : \theta \in \Theta_k\}$ von Klassifikations- oder Regressionsfunktionen, vgl. Lemma 4.1.

Definition 4.4 (Strukturelle Risikominimierung (SRM))
Sei eine Struktur $\{M_k : k \geq 1\}$ auf $\mathcal{M}$ gegeben. Das SRM-Prinzip zur Schätzung einer Funktion f besteht aus zwei Schritten.

1) Modellauswahl: Wähle k^* gemäß der Balancierung von R_{emp} in Schema 4.2.

2) Schätzung: Minimiere R_{emp} über Θ_{k^*}.

Bemerkung 4.5 (Regularisierung)
In vielen Anwendungsfällen (insbesondere bei Regressionsproblemen) kann der Modellauswahlschritt auch durch die Einführung eines Straf- bzw. Penalisierungsterms realisiert werden.
Das zu lösende Minimierungsproblem ist dann von der Form

$$f^* = \underset{f \in \mathcal{M}}{\operatorname{argmin}} \{R_{emp}(f) + pen_n(f)\} \tag{4.3}$$

mit einem Penalisierungsterm $pen_n(f)$, der die Komplexität von f „bestraft“.
Beispielsweise könnte im Fall der polynomiellen Regression (vgl. Aufgabe 2) der Strafterm als der (Höchst-)Grad von f gewählt werden. Die zugehörige Struktur auf

$$\mathcal{M} = \{f : \mathbb{R} \to \mathbb{R}\}$$

wäre dann gegeben durch

$$\mathcal{M}_k = \{f : f \text{ ist Polynom von Höchstgrad } k\}, k \geq 1.$$

Beispiel 4.6 (Basis-Entwicklung)
Sei $\{g(\cdot, \gamma) : \gamma \in \Gamma\}$ eine Funktionen-Basis und definiere für $k \geq 1$ die Funktion f_k durch

$$f_k(\boldsymbol{x}, \theta) = \sum_{j=1}^{k} w_j \cdot g(\boldsymbol{x}, \gamma_j),\ w_j \in \mathbb{R}, \tag{4.4}$$

$\theta = (w_1, ..., w_k, \gamma_1, ..., \gamma_k)^\top$.
Häufig wird $g(\boldsymbol{x}, \gamma_1) \equiv 1$ gesetzt, so dass w_1 als der „Offset“ des Modells interpretiert werden kann. Setzen wir $g(\boldsymbol{x}, \gamma_j) \equiv \phi_j(\boldsymbol{x})$, so sind nur $w_1, ..., w_k$ freie Modellparameter, und die VC-Dimension von $\mathcal{M}_k$ ist gleich k (für den Fall $\boldsymbol{x} \in D = \mathbb{R}$ siehe Übungsaufgabe 13.(a)). Somit bildet dann $\{\mathcal{M}_k : k \geq 1\}$ eine Struktur auf $\mathcal{M} = \{f : D \to \mathbb{R}\}$, die die Eigenschaft (a) aus Definition 4.3 besitzt.

Beispiel 4.7 (Merkmalsauswahl, „feature selection“)
Sei f_k wie in (4.4)*, wobei die k Basisfunktionen aus eine Menge von $K >> k$ Basisfunktionen ausgewählt werden. Man spricht hier auch von Merkmalsauswahl (englisch: feature selection), da $g(\boldsymbol{x}, \gamma_j)$ als ein Merkmal von $\boldsymbol{x}$ interpretiert werden kann.*
Im Falle von $D = \mathbb{R}$ könnte zum Beispiel

$$f_k(x, w_1, ..., w_k, \gamma_1, ..., \gamma_k) = \sum_{j=1}^{k} w_j x^{\gamma_j},\ \gamma_j \in \mathbb{N}_0,$$

gewählt werden. Es sind dann also Monome $\{x^{\gamma_j} : 1 \leq j \leq k\}$ auszuwählen, die eine optimale Datenanpassung ergeben.

Beispiel 4.8 (Datenglättung /-vorverarbeitung)
Insbesondere in der Bildverarbeitung werden die Orginal-(Bild-)Daten typischerweise in einem Vorverarbeitungsschritt geglättet, bevor die eigentliche Datenanalyse erfolgt. Für Originaldaten $\boldsymbol{x}$ sei dazu $\tilde{\boldsymbol{x}} := K(\boldsymbol{x}, \beta)$ definiert, wobei K eine Glättungsfunktion (Kern) und β die zugehörige Bandweite (Glättungsintensität) bezeichnet.
Betrachten wir ein Gitter $c_1 > c_2 > ...$ von möglichen Werten für β, so induziert dies eine Struktur $\{\mathcal{M}_k : k \geq 1\}$ vermittels

$$\mathcal{M}_k = \{f(K(\boldsymbol{x}, \beta), \theta) : \beta \geq c_k\}.$$

Hier ist also der „Parameterraum“ Θ für alle $k \geq 1$ identisch, und die Struktur bezieht sich auf die „Nutzdaten“ $\tilde{\boldsymbol{x}}$, die aus den „Rohdaten“ $\boldsymbol{x}$ gewonnen und dann für die eigentliche Datenanalyse eingesetzt werden.

Beispiel 4.9 (Informationskriterien in der Regression)
Betrachten wir, um es möglichst konkret zu machen, ein multiples lineares Regressionsmodell der Form

$$\mathbf{Y} = \mathbf{X}\theta + \varepsilon, \text{ wobei} \tag{4.5}$$

$$\mathbf{Y} = (Y_1, ..., Y_n)^\top \in \mathbb{R}^n \text{ der Responsevektor},$$

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & \dots & x_{1,p} \\ \\ \vdots & & \vdots \\ x_{n,1} & \dots & x_{n,p} \end{pmatrix} \in \mathbb{R}^{n \times p} \text{ die Design-Matrix},$$

und $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)^\top \in \mathbb{R}^n$ der Vektor der Fehlerterme ist. Nehmen wir zusätzlich der Einfachheit halber an, dass $\varepsilon_1, ..., \varepsilon_n$ stochastisch unabhängig und identisch verteilt sind, mit $\mathbb{E}[\varepsilon_1] = 0$ und $Var(\varepsilon_1) = \sigma^2 \in (0, \infty)$. Ziel der statistischen Inferenz ist $\theta = (\theta_1, ..., \theta_p)^\top$. Die Modellgleichung (4.5) *macht (entgegen des generellen Setups der algorithmischen Modellierung aus Kapitel 1) eine qualitative Annahme über den Daten-generierenden Prozess (additives Rauschen). Bezeichnen wir mit $\Gamma = 2^{\{1,...,p\}} \setminus \emptyset$ die Menge aller nicht-leeren Teilmengen von $\{1, ..., p\}$, so lässt sich das Problem der Modellauswahl in diesem Fall dadurch formalisieren, dass ein $\gamma \in \Gamma$ gewählt wird und nur diejenigen Spalten der Design-Matrix sowie die entsprechenden Koordinaten von θ in die Datenanalyse bzw. -modellierung einbezogen werden, deren Indizes in γ liegen.*
Zur Auswahl von γ wurden von Akaike (1974) bzw. Schwarz (1978) die folgenden Informationskriterien vorgeschlagen.

$$AIC(\gamma) = R_{emp}(\gamma) + \frac{2\,|\gamma|}{n}\hat{\sigma}^2_{voll}, \tag{4.6}$$

$$BIC(\gamma) = R_{emp}(\gamma) + \frac{\log(n)\,|\gamma|}{n}\hat{\sigma}^2_{voll}. \tag{4.7}$$

Dabei ist $\hat{\sigma}^2_{voll}$ eine Schätzung der Fehlervarianz σ^2 im vollen Modell ($\gamma = \{1, ..., p\}$) und

$$R_{emp}(\gamma) = n^{-1} \sum_{i=1}^{n} (y_i - \hat{y}_i(\gamma))^2$$

mit $\hat{y}_i(\gamma) = \left(\mathbf{X}_\gamma \hat{\theta}_\gamma\right)_i$, wobei $\mathbf{X}_\gamma$ die (reduzierte) Design-Matrix bezeichnet, die nur die durch γ vorgegebenen Spalten von $\mathbf{X}$ beinhaltet, und $\hat{\theta}_\gamma$ den Kleinste-Quadrate-Schätzer in diesem reduzierten Modell bezeichnet, d.h.,

$$\hat{\theta}_\gamma = (\mathbf{X}_\gamma^\top \mathbf{X}_\gamma)^{-1} \mathbf{X}_\gamma^\top \mathbf{Y}.$$

Im Hinblick auf Bemerkung 4.5 kann die Minimierung von (4.6) *bzw.* (4.7) *als eine Anwendung der SRM-Prinzips aufgefasst werden.*

Beispiel 4.10 (Regularisierte Regression, Ridge und LASSO)
Nehmen wir unter dem Modell aus (4.5) *an, dass* $\varepsilon_1 \sim \mathcal{N}(0, \sigma^2)$ *für* $\sigma^2 \in (0, \infty)$ *ist, so ergibt sich für die gemeinsame Likelihoodfunktion von* $\mathbf{Y} = (Y_1, ..., Y_n)^\top$*, dass*

$$p(\boldsymbol{y}|\theta) = (2\pi)^{-\frac{n}{2}} \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \mathbf{X}\theta)^\top(\boldsymbol{y} - \mathbf{X}\theta)\right), \ \boldsymbol{y} = (y_1, ..., y_n)^\top \in \mathbb{R}^n.$$

Wir nehmen nun einen Bayesianischen Standpunkt ein und betrachten eine a priori-Dichte $\pi(\cdot)$ *auf* $\Theta = \mathbb{R}^p$*.*
Dann ist die a posteriori-Dichte von $\vartheta = \theta$*, gegeben*$\mathbf{Y} = \boldsymbol{y}$*, proportional zu* $p(\boldsymbol{y}|\theta)\pi(\theta)$*. Der Maximierer* $\hat{\theta}_{MAP}$ *dieses Ausdrucks heißt Maximum a posteriori-Schätzer von* θ*.*
Da die natürliche Logarithmusfunktion strikt isoton ist, gilt äquivalenterweise, dass

$$\begin{aligned}
\hat{\theta}_{MAP} &= \operatorname*{argmax}_{\theta \in \mathbb{R}^p} \{\ln(p(\boldsymbol{y}|\theta)) + \ln(\pi(\theta))\} \\
&= \operatorname*{argmin}_{\theta \in \mathbb{R}^p} \{-\ln(p(\boldsymbol{y}|\theta)) - \ln(\pi(\theta))\} \\
&= \operatorname*{argmin}_{\theta \in \mathbb{R}^p} \{\|\boldsymbol{y} - \mathbf{X}\theta\|_2^2 - \ln(\pi(\theta))\}.
\end{aligned}$$

Dieses Minimierungsproblem ist offenbar von der Form (4.3)*, mit dem Penalisierungsterm* $pen_n(\theta) = -\ln(\pi(\theta))$*.*
Zwei beliebte Wahlen für $\pi(\cdot)$ *führen zur „Ridge Regression“ bzw. zur „LASSO-Regression“.*

Beispiel 4.11 (Ridge Regression)
Wählen wir, unter den Gegebenheiten von Beispiel 4.10, eine a priori $\mathcal{N}_p(\mathbf{0}, \tau^2 I_p)$*-Verteilung, wobei* $\tau^2 \in (0, \infty)$ *ein Hyperparameter ist, und setzen wir* $\lambda := (2\tau^2)^{-1}$*, so erhalten wir*

$$-\ln(\pi(\theta)) \propto \frac{1}{2\tau^2}\theta^\top\theta = \lambda\|\theta\|_2^2.$$

Wir erhalten also eine L_2*-regularisierte Regression, die auch „Ridge Regression“genannt wird.*

Beispiel 4.12 (LASSO-Regression)
Die Doppelexponentialverteilung (auch: Laplace-Verteilung) mit Skalenparameter $\lambda > 0$ *ist eine absolut stetige (bezüglich des Lebesguemaßes) Wahrscheinlichkeitsverteilung auf* $\mathbb{R}$ *mit Lebesguedichte* f_λ*, gegeben durch*

$$f_\lambda(t) = \frac{\lambda}{2}\exp(-\lambda|t|), \ t \in \mathbb{R}.$$

Wählen wir unter den Gegebenheiten von Beispiel 4.10

$$\pi(\theta) = \prod_{j=1}^{p} f_\lambda(\theta_j), \ \theta = (\theta_1, ..., \theta_p)^\top,$$

also a priori stochastisch unabhängige, identisch Laplace (λ)-verteilte Parameter an, so ist

$$-\ln(\pi(\theta)) \propto \lambda \sum_{j=1}^{p} |\theta_j| = \lambda \|\theta\|_1.$$

Wir erhalten also eine L_1-regularisierte Regression, die auch „LASSO-Regression“genannt wird („least absolute shrinkage and selection operator“, nach Tibshirani (1996)).

Schema 4.13 (Ridge Regression und LASSO-Regression als Instanzen des SRM-Prinzips)
Fassen wir die Erkenntnisse aus den Beispielen 4.11 und 4.12 zusammen, so erhalten wir:

$$\hat{\theta}_{Ridge} = \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} \{\|\boldsymbol{y} - \mathbf{X}\theta\|_2^2 + \lambda \|\theta\|_2^2\},$$

$$\hat{\theta}_{LASSO} = \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} \{\|\boldsymbol{y} - \mathbf{X}\theta\|_2^2 + \lambda \|\theta\|_1\}.$$

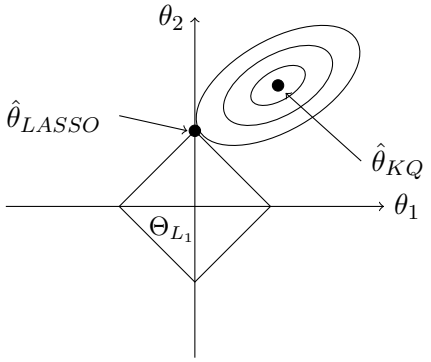
Diese Definitionen lassen sich wie folgt umformulieren.

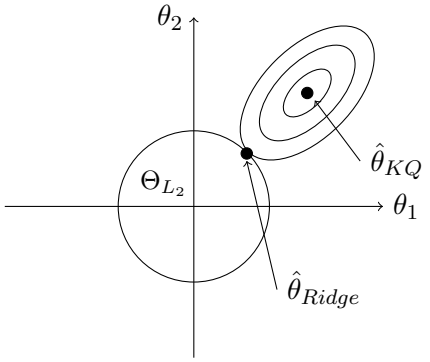
$$\hat{\theta}_{Ridge} = \underset{\theta \in \Theta_{L_2}(\lambda)}{\operatorname{argmin}} \|\boldsymbol{y} - \mathbf{X}\theta\|_2^2,$$

wobei $\Theta_{L_2}(\lambda) = \{\gamma \in \mathbb{R}^p : \|\gamma\|_2^2 \leq C_2(\lambda)\}$ für eine geeignete Konstante $C_2(\lambda)$;

$$\hat{\theta}_{LASSO} = \underset{\theta \in \Theta_{L_1}(\lambda)}{\operatorname{argmin}} \|\boldsymbol{y} - \mathbf{X}\theta\|_2^2,$$

wobei $\Theta_{L_1}(\lambda) = \{\gamma \in \mathbb{R}^p : \|\gamma\|_1 \leq C_1(\lambda)\}$ für eine geeignete Konstante $C_1(\lambda)$.
Offenbar sind $C_1(\lambda)$ und $C_2(\lambda)$ monoton in λ, induzieren also eine Struktur auf $\Theta = \mathbb{R}^p$.
Wir erhalten die folgenden beiden Schaubilder (adaptiert nach Tibshirani (1996)) für $p = 2$.





Aufgrund der geometrischen Struktur von $\Theta_{L_1} \equiv \Theta_{L_1}(\lambda)$ führt die L_1-Regularisierung (im Gegensatz zur L_2-Regularisierung) oft implizit zu einer Merkmalsauswahl, denn typischerweise werden einige θ_j exakt auf Null „geschrumpft“.

Bemerkung 4.14
Der Bayesianische (MAP-) Ansatz lässt sich auch in vielen anderen Modellen als eine Instantiierung des SRM-Prinzips begreifen; vergleiche dazu zum Beispiel Section 4.11 von Vapnik (2000).

Satz 4.15 (Theorem 6.2 in Vapnik (1998))

Unter den Voraussetzungen von Definition 4.3 sei eine Modellauswahlregel gegeben, die für gegebenen Stichprobenumfang n des Trainingsdatensatzes ein $k(n) \in \mathbb{N}$ liefert, so dass das Strukturelement $\mathcal{M}_{k(n)}$im Modellauswahlschritt gemäß Definition 4.4 gewählt wird. Dann ist SRM konsistent, falls gilt:

$$\frac{D_{k(n)}^2 h_{k(n)} \ln(n)}{n} \to 0,\ n \to \infty, \tag{4.8}$$

$$k(n) \to \infty,\ n \to \infty. \tag{4.9}$$

Dabei ist $D_k = B_k$ für beschränkte Verlustfunktionen (siehe Definition 4.3 (b).(i)) beziehungsweise $D_k = \tau_k$ unter den Annahmen von Definition 4.3 (b).(ii).
Ferner existieren explizite Risikoschranken unter (4.8) *und* (4.9)*.*

Kapitel 5

Methoden zur binären Klassifikation

In diesem Kapitel studieren wir spezifische binäre Klassifikationsfunktionen

$$\hat{f} : D \to W = \{-1, +1\}$$

$$\boldsymbol{x} \in D \mapsto \hat{y} = \hat{f}(\boldsymbol{x}) \in W = \{-1, +1\}.$$

Die Funktion $\hat{f}$ wird „gelernt“ auf der Basis eines Trainingsdatensatzes $(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_n, y_n)$ für $n \in \mathbb{N}$. Die Kodierung ± 1 für das „Label“ y wird gewählt, weil (i) hiermit eine Dichotomisierung eines stetigen Merkmals $g(\boldsymbol{x})$ einfach mit Hilfe der Vorzeichenfunktion $\text{sgn}(\cdot)$ formalisiert werden kann und (ii) $\mathbb{E}[Y] = 0$ ist, falls die erwartete relative Häufigkeit der beiden Klassen jeweils gleich $1/2$ ist (balancierte Klassen).

Beispiel 5.1 (Perceptron, Rosenblatt (1958, 1962))
Rosenblatt (1958, 1962) schlug die folgende Familie $\{f(\cdot, \theta) : \theta \in \mathbb{R}^d\}$ zur binären Klassifikation vor:

$$f(\boldsymbol{x}, \theta) = sgn\left(\sum_{i=1}^{d} \theta_j \psi_j(\boldsymbol{x})\right), \tag{5.1}$$

wobei für alle $1 \leq j \leq d$ die Funktion $\psi_j : D \to \mathbb{R}$ beliebig gewählt werden kann.
Die Modellgleichung (5.1) *lässt sich auch in der folgenden (linearisierten) Form schreiben.*

$$f(\boldsymbol{u}, \theta) = sgn(\langle \boldsymbol{u}, \theta \rangle_{\mathbb{R}^d}),\ \boldsymbol{u} = (u_1, ..., u_d)^\top, \tag{5.2}$$

wobei $u_j = \psi_j(\boldsymbol{x})$ gesetzt wird, $1 \leq j \leq d$. Wir nennen $u_j = \psi_j(\boldsymbol{x})$ auch das j-te Merkmal von $\boldsymbol{x}$ und $\mathbb{R}^d \supseteq U \ni \boldsymbol{u}$ den zu D gehörigen Merkmalsraum (englisch: feature space).
Durch Übergang von D zu U kann also ohne Beschränkung der Allgemeinheit auch sofort $D = \mathbb{R}^d$ angenommen werden und

$$f(\boldsymbol{x}, \theta) = sgn(\langle \boldsymbol{x}, \theta \rangle_{\mathbb{R}^d}) \tag{5.3}$$

betrachtet werden. Die Gleichung (5.3) *hat die Interpretation, dass von vorne herein lediglich die* d *interessierenden Merkmale beobachtet werden.*
Zur Kalibrierung von $\hat{f}$ *(„Schätzung" der Koeffizienten* $\theta_1, ..., \theta_d$*) schlug Frank Rosenblatt das folgende iterative Schema vor*

1) Initialisiere $\theta(0) = \mathbf{0} \in \mathbb{R}^d$.

2) a) For i *from* 1 *to* n *do:*

$$\theta(i) = \begin{cases} \theta(i-1), & \text{falls } y_i \langle \theta(i-1), \boldsymbol{x}_i \rangle_{\mathbb{R}^d} > 0 \\ \theta(i-1) + y_i \boldsymbol{x}_i, & \text{falls } y_i \langle \theta(i-1), \boldsymbol{x}_i \rangle_{\mathbb{R}^d} \leq 0 \end{cases} \tag{5.4}$$

End for

b) Setze $\hat{\theta} = \theta(n)$ *und* $\hat{f} = f(\cdot, \hat{\theta})$.

3) Wiederhole Schritt 2) mit $\theta(0) := \hat{\theta}$ *so lange, bis dass* $R_{emp}(\hat{f}) \leq$ *TOL ist, wobei TOL eine vordefinierte Toleranzschwelle bezeichnet, oder eine festgelegte Maximalanzahl an Wiederholungen erreicht ist.*

Bemerkung 5.2

(i) Die Bedingung $y_i \langle \theta(i-1), \boldsymbol{x}_i \rangle_{\mathbb{R}^d} > (\leq) 0$ *in* (5.4) *bedeutet, dass das* i*-te Trainingsbeispiel richtig (falsch) klassifiziert wird, wenn* (5.3) *mit* $\theta = \theta(i-1)$ *angewendet wird.*

(ii) Ein „Intercept" $\hat{\theta}_1$ *kann in die Klassifikationsregel aufgenommen werden, indem das erste (Pseudo-)Merkmal konstant auf den Wert* 1 *gesetzt wird.*

Satz 5.3 (Novikoff (1963))
Angenommen, unter den Gegebenheiten von Beispiel 5.1 sind die folgenden drei Voraussetzungen erfüllt.

(i) Die Norm der Merkmalsvektoren $\boldsymbol{x}$ *in* (5.3) *ist beschränkt durch* R.

(ii) Es herrscht lineare Separierbarkeit, das heißt, $\exists \delta > 0$ *mit*

$$\sup_{\theta \in \mathbb{R}^d} \min_{1 \leq i \leq n} y_i \langle \theta(i-1), \boldsymbol{x}_i \rangle_{\mathbb{R}^d} > \delta. \tag{5.5}$$

(iii) Es werden im dritten Schritt des iterativen Kalibrierungsalgorithmus' hinreichend viele Wiederholungen durchgeführt.

Dann liefert der iterative Kalibrierungsalgorithmus nach höchstens $\lfloor \frac{R^2}{\delta^2} \rfloor$ *Korrekturschritten von* $\theta(0) = \mathbf{0} \in \mathbb{R}^d$ *eine Funktion* $\hat{f}$ *mit* $R_{emp}(\hat{f}) = 0$.

Korollar 5.4

Unter den Annahmen von Satz 5.3 gilt mit Wahrscheinlichkeit mindestens $1-\eta$*, dass*

$$R(\hat{f}) \leq \frac{d\left[1+\ln\left(\frac{2n}{d}\right)\right]-\ln\left(\frac{\eta}{4}\right)}{n} \to 0,\ n\to\infty,$$

denn die VC-Dimension von $\{sgn(\langle\cdot,\theta\rangle_{\mathbb{R}^d}) : \theta\in\mathbb{R}^d\}$ *ist gleich d, vergleiche Übungsaufgabe.*

Falls (5.5) verletzt ist, so ist das Problem der empirischen Risikominimierung vermittels des Perceptrons von erheblicher (kombinatorischer) Komplexität. Aus diesem Grunde werden in solchen Fällen oftmals Relaxationstechniken zum Einsatz gebracht.

Definition 5.5 (Geglättete Vorzeichenfunktion)

Eine Funktion $S:\mathbb{R}\to\mathbb{R}$ *heißt eine geglättete Vorzeichenfunktion, falls* S *monoton wachsend und hinreichend oft stetig differenzierbar ist mit*

$$S(0)=0,\ \lim_{z\to-\infty}S(z)=-1,\ \lim_{z\to+\infty}S(z)=+1.$$

Beispielsweise ist $\tanh_\gamma$, gegeben durch

$$\tanh_\gamma(z)=\frac{\exp(\gamma z)-\exp(-\gamma z)}{\exp(\gamma z)+\exp(-\gamma z)} \text{ für gegebenes } \gamma>0,$$

eine geglättete Vorzeichenfunktion.

Beispiel 5.6 (Verfahren des steilsten Abstiegs)

Sei $\{f(\cdot,\theta):\theta\in\Theta\}$ *mit* $f(\boldsymbol{x},\theta)=sgn(\langle\boldsymbol{x},\theta\rangle)$ *eine Familie von binären Klassifikationsfunktionen, vgl.* (5.3)*. Wir betrachten das empirische Risikofunktional*

$$R_{emp}(\theta)=n^{-1}\sum_{i=1}^{n}(y_i-f(\boldsymbol{x}_i,\theta))^2 \tag{5.6}$$

für Trainingsdatenpunkte $(\boldsymbol{x}_1,y_1),...,(\boldsymbol{x}_n,y_n)$*, wobei* $\forall 1\leq i\leq n:\boldsymbol{x}_i\in D\subseteq\mathbb{R}^d$ *und* $y_i\in W=\{-1,+1\}, d\in\mathbb{N}$*.*

Das Verfahren des steilsten Abstiegs zur (approximativen) Minimierung von (5.6) *ersetzt* $f(\boldsymbol{x}_i,\theta)$ *in* (5.6) *durch* $S(\langle\boldsymbol{x}_i,\theta\rangle)$ *für eine geglättete Vorzeichenfunktion* $S, 1\leq i\leq n$*. Dies führt zu einer (glatten) Approximation*

$$\tilde{R}_{emp}(\theta)=n^{-1}\sum_{i=1}^{n}[y_i-S(\langle\boldsymbol{x}_i,\theta\rangle)]^2$$

von $R_{emp}(\theta)$*. Es ist*

$$\nabla_\theta\tilde{R}_{emp}(\theta)=-\frac{2}{n}\sum_{i=1}^{n}\left\{[y_i-S(\langle\boldsymbol{x}_i,\theta\rangle)]S'(\langle\boldsymbol{x}_i,\theta\rangle)\boldsymbol{x}_i^\top\right\}.$$

Damit kann ein iteratives Verfahren zur Optimierung bezüglich θ *verwendet werden.*

Ausgehend von einem Startwert $\theta(0)$ *wird im* k*-ten Iterationsschritt*

$$\theta(k)=\theta(k-1)-\gamma(k)\{\nabla_\theta\tilde{R}_{emp}(\theta(k-1))\}^\top$$

gesetzt, $1 \leq k \leq K$. *Hierbei wird die Schrittweite* $\gamma(k) \geq 0$ *so gewählt, dass*

$$\sum_{k=1}^{\infty} \gamma(k) = \infty \text{ und } \sum_{k=1}^{\infty} \gamma^2(k) < \infty$$

gelten.

Bemerkung 5.7 (Neuronale Netze)
Die Technik aus Beispiel 5.6 kann zum Verfahren der sogenannten Neuronalen Netze erweitert werden. Hierbei werden die Merkmalsvektoren des Trainingsdatensatzes mit Hilfe eines Systems geglätteter Indikatorfunktionen, die in mehreren Schichten angeordnet sind, verarbeitet beziehungsweise zusammengeschaltet (vgl. Abbildung 5.1 in (Vapnik, 2000)).
Zur Optimierung der Koeffizienten in den Schichten kommt dabei jeweils ein Abstiegsverfahren wie in Beispiel 5.6 zum Einsatz. Durch die Kombinationen (Überlagerung) von geglätteten Indikatorfunktionen mit gegebenenfalls verschiedenen Glättungsparametern ergibt sich eine größere modelliererische Flexibilität als im Falle des Rosenblatt'schen Perceptrons aus Beispiel 5.1.

Bemerkung 5.8 (Probleme der Neuronalen Netze)

(i) Das (geglättete) empirische Risikofunktional kann mehrere lokale Minima besitzen. Deswegen kann das Verfahren sensitiv gegenüber der Wahl der Startwerte sein.

(ii) Die Konvergenz des Verfahrens des steilsten Abstieges kann langsam sein. Es werden also gegebenenfalls viele Iterationsschritte benötigt, bis ein Abbruchkriterium erfüllt ist.

(iii) Das Verfahren benötigt die Festlegung diverser Tuningparameter, zum Beispiel der Skalierungsparameter des tanh.

Definition 5.9 (optimale trennende Hyperebene)
Wir betrachten Trainingsdaten $(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_n, y_n)$ *mit* $\forall 1 \leq i \leq n : \boldsymbol{x}_i \in D \subseteq \mathbb{R}^d, y_i \in W = \{+1, -1\}$. *Angenommen, es existiert ein* $\theta \in \mathbb{R}^d$ *mit* $\|\theta\|_2 = 1$, *so dass die Hyperebene im* $\mathbb{R}^d$, *die durch die Gleichung*

$$\langle \boldsymbol{x}, \theta \rangle_{\mathbb{R}^d} = c \tag{5.7}$$

gegeben ist, die die Daten $(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_n, y_n)$ *ohne Fehler trennt, wobei ohne Beschränkung der Allgemeinheit gelten soll:*

$$y_i = +1 \Rightarrow \langle \boldsymbol{x}_i, \theta \rangle_{\mathbb{R}^d} > c, \tag{5.8}$$

$$y_i = -1 \Rightarrow \langle \boldsymbol{x}_i, \theta \rangle_{\mathbb{R}^d} < c. \tag{5.9}$$

Seien für einen Einheitsvektor $u \in \mathbb{R}^d$ Konstanten $c_1(u)$ und $c_2(u)$ gegeben durch

$$\begin{aligned} c_1(u) &= \min_{i:y_i=+1} \langle \boldsymbol{x}_i, u \rangle_{\mathbb{R}^d}, \\ c_2(u) &= \max_{i:y_i=-1} \langle \boldsymbol{x}_i, u \rangle_{\mathbb{R}^d}. \end{aligned}$$

Dann nennen wir $\theta^* \in \mathbb{R}^d$ mit $\|\theta^*\|_2 = 1$ optimal, falls θ^* die Funktion ρ, gegeben durch

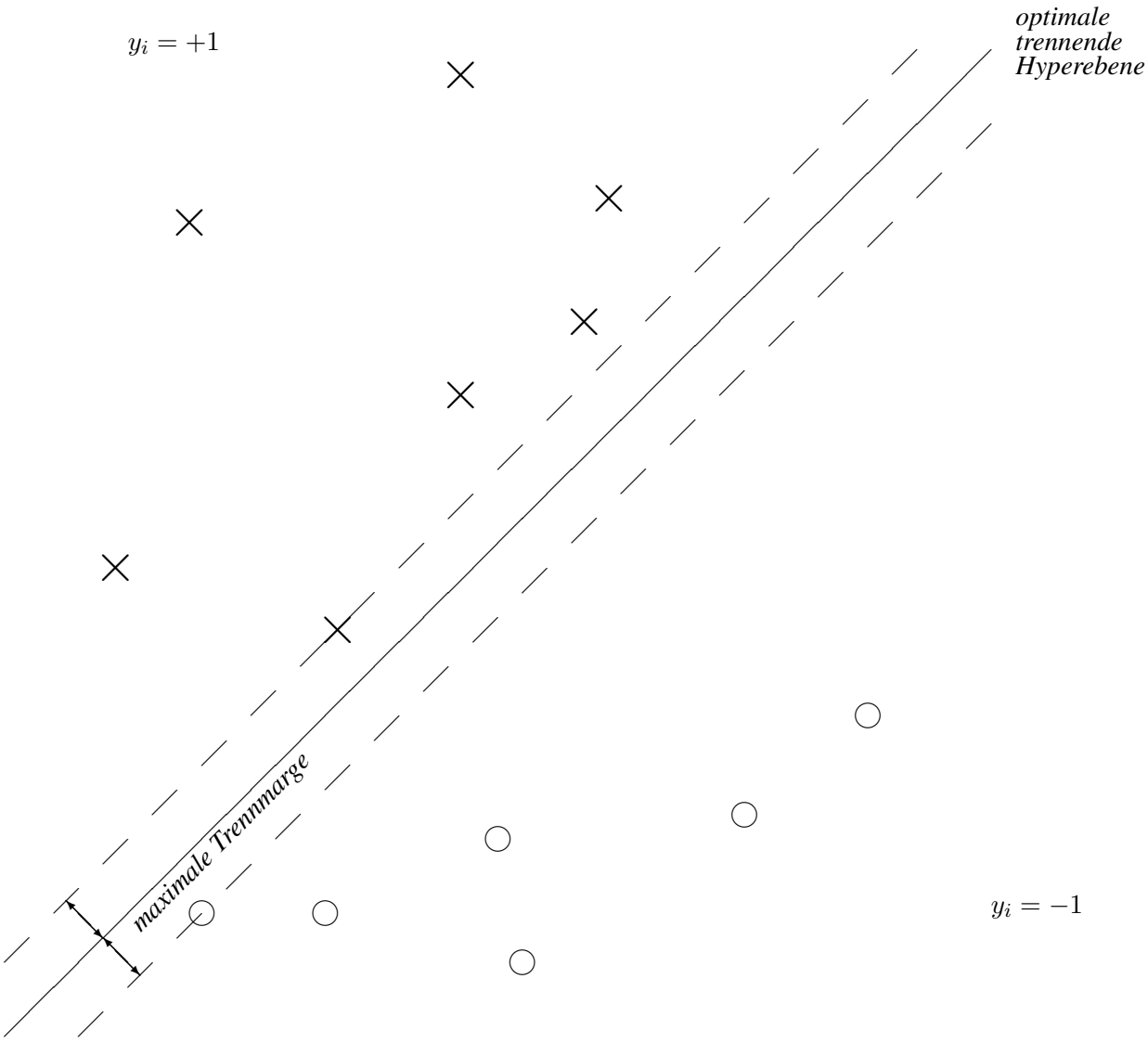
$$\rho(\theta) = \frac{c_1(\theta) - c_2(\theta)}{2}, \tag{5.10}$$

unter den Nebenbedingungen (5.8) *und* (5.9) *maximiert.*
Ferner heißt die durch θ^* und

$$c^* = \frac{c_1(\theta^*) + c_2(\theta^*)}{2} \tag{5.11}$$

gegebene trennende Hyperebene optimal.
Die Zahl $\rho(\theta^*)$ heißt maximale Trennmarge (englisch: maximal margin).

Schema 5.10



Satz 5.11

Unter den Voraussetzungen von Definition 5.9 ist die optimale trennende Hyperebene eindeutig.

Beweis: Die Funktion ρ ist stetig auf $\mathbb{R}^d$. Damit nimmt ρ auf dem beschränkten Bereich $\{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq 1\}$ ihr Maximum an. Ferner muss die Maximalstelle auf dem Rand liegen, das heißt, der Maximierer θ^* erfüllt $\|\theta^*\|_2 = 1$. Wäre dies nämlich nicht der Fall, so könnte $\theta^{**} := \frac{\theta^*}{\|\theta^*\|_2}$

gesetzt werden, mit

$$\rho(\theta^{**}) = \frac{\rho(\theta^*)}{\|\theta^*\|_2} > \rho(\theta^*).$$

Schließlich ist θ^* eindeutig festgelegt, denn falls ein θ^{**} mit $\rho(\theta^{**}) = \rho(\theta^*)$ existieren würde (notwendigerweise mit $\|\theta^{**}\|_2 = 1$), so würde das Maximum wegen der Konkavität von ρ (siehe Übungsaufgabe) auch auf der Verbindungslinie von θ^* und θ^{**} angenommen. Die Punkte echt zwischen θ^* und θ^{**} hätten dann indes eine Länge echt kleiner als 1, was einen Widerspruch zur obigen Argumentation ergäbe. ■

Satz 5.12

Unter den Voraussetzungen von Definition 5.9 kann die optimale trennende Hyperebene äquivalenterweise durch die folgende Optimierungsaufgabe charakterisiert werden:
Minimiere $\|\theta\|_2^2$ unter den Nebenbedingungen

$$\forall 1 \leq i \leq n \, \textit{mit} \, y_i = +1 : \langle \boldsymbol{x}_i, \theta \rangle_{\mathbb{R}^d} + b \geq 1 \tag{5.12}$$

$$\forall 1 \leq i \leq n \, \textit{mit} \, y_i = -1 : \langle \boldsymbol{x}_i, \theta \rangle_{\mathbb{R}^d} + b \leq -1 \tag{5.13}$$

für eine Konstante b. Dabei können (5.12) *und* (5.13) *zusammengefasst werden zu*

$$\forall 1 \leq i \leq n : y_i \left[\langle \boldsymbol{x}_i, \theta \rangle_{\mathbb{R}^d} + b \right] \geq 1. \tag{5.14}$$

Genauer gilt:
Sei θ^* der Minimierer von $\theta \mapsto \|\theta\|_2^2$ unter den Nebenbedingungen (5.12) *und* (5.13) *und sei ξ^* der Maximierer von $\theta \mapsto \rho(\theta)$ aus* (5.10) *unter den Nebenbedingungen* (5.8) *und* (5.9)*. Dann ist*

$$\xi^* = \frac{\theta^*}{\|\theta^*\|_2} \, \textit{und} \, \rho(\xi^*) = \frac{1}{\|\theta^*\|_2}.$$

Beweis: Zunächst ist θ^* eindeutig bestimmt, da es sich bei $\theta \mapsto \|\theta\|_2^2$ um eine strikt konvexe Zielfunktion handelt, die unter den linearen Nebenbedingungen (5.14) minimiert werden soll.
Definiere nun $\xi^* = \frac{\theta^*}{\|\theta^*\|_2}$. Offenbar ist dann $\|\xi^*\|_2 = 1$.
Wegen der Nebenbedingungen aus (5.14) ist

$$\begin{aligned} \rho(\xi^*) = \rho\left(\frac{\theta^*}{\|\theta^*\|_2}\right) &= \frac{1}{2}\left[c_1\left(\frac{\theta^*}{\|\theta^*\|_2}\right) - c_2\left(\frac{\theta^*}{\|\theta^*\|_2}\right)\right] \\ &\geq \frac{1}{2\|\theta^*\|_2}[1 - (-1)] \\ &= \frac{1}{\|\theta^*\|_2}. \end{aligned}$$

Genügt also zu zeigen:
Der Fall

$$\rho\left(\frac{\theta}{\|\theta\|_2}\right) > \frac{1}{\|\theta^*\|_2} \tag{$\star$}$$

kann niemals eintreten (für keinen Vektor $\theta \in \mathbb{R}^d$).
Wir führen einen Widerspruchsbeweis und nehmen an, dass ($\star$) für ein $\theta \in \mathbb{R}^d$ gilt.
Definiere dann

$$\gamma = \frac{\theta}{\|\theta\|_2}.$$

Es gilt dann

$$\rho(\gamma) > \frac{1}{\|\theta^*\|_2}.$$

Sei nun der Vektor η gegeben durch

$$\eta = \frac{\gamma}{\rho(\gamma)}.$$

Für diesen Vektor η ist

$$\|\eta\|_2 < \|\theta^*\|_2.$$

Ferner erfüllt η die Nebenbedingungen aus (5.14) mit

$$b = -\frac{c_1(\eta) + c_2(\eta)}{2}.$$

Dies widerspricht aber der Definition von θ^*. ■

Bemerkung 5.13

(a) Die Lösung des in Satz 5.12 aufgeworfenen restringierten Optimierungsproblems kann (numerisch) mit Hilfe der Methode der Lagrange-Multiplikatoren bestimmt werden.

(b) Die Vektoren $\boldsymbol{x}_i$ mit Abstand zu optimalen trennenden Hyperebene exakt gleich der maximalen Trennmarge heißen Support-Vektoren.

Definition 5.14 (Verallgemeinerte optimale Hyperebene)
Angenommen, die Trainingsdaten $(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_n, y_n)$ mit $\boldsymbol{x}_i \in \mathbb{R}^d$ und $y_i \in \{-1, +1\}, 1 \leq i \leq n$, sind nicht linear separierbar.
Dann lässt sich die Optimierungsaufgabe bezüglich $\theta \mapsto \|\theta\|_2^2$ also unter den Nebenbedingungen aus (5.14) *nicht lösen. Wir führen daher nicht-negative Schlupfvariablen $\xi_1, ..., \xi_n$ ein.*
Zwei relaxierte Optimierungsprobleme zur Bestimmung einer verallgemeinerten optimalen Hyperebene sind dann gegeben durch
(a) Harte Trennmargen-Verallgemeinerung:
Minimiere $\Phi(\theta, b) = \sum_{i=1}^n \xi_i$ unter den Nebenbedingungen

$$\forall 1 \leq i \leq n : y_i(\langle \boldsymbol{x}_i, \theta \rangle + b) \geq 1 - \xi_i,\ \xi_i \geq 0 \text{ und} \tag{5.15}$$

$$\|\theta\|_2^2 \leq A^2 \text{ für gegebenes } A^2 > 0. \tag{5.16}$$

(b) Weiche Trennmargen-Verallgemeinerung: Minimiere

$$\Phi(\theta, b) = \frac{1}{2}\|\theta\|_2^2 + C\sum_{i=1}^{n}\xi_i \tag{5.17}$$

für einen gegebenen Wert C unter den Nebenbedingungen aus (5.15)*.*
Die verallgemeinerte optimale Hyperebene ist dann jeweils gegeben durch die Gleichung

$$\langle\theta^*, \boldsymbol{x}\rangle_{\mathbb{R}^d} + b^* = 0,\ \boldsymbol{x} \in \mathbb{R}^d,$$

wobei (θ^*, b^*) die Lösung des jeweiligen Optimierungsproblems bezeichnet.

Lemma 5.15 (siehe Abschnitt 10.2.2 in Vapnik (1998))
Die Lösung des durch (5.17) *und* (5.15) *gegebenen Optimierungsproblems lässt sich wie folgt charakterisieren.*

$$\theta^* = \sum_{i=1}^{n}\lambda_i y_i \boldsymbol{x}_i \tag{5.18}$$

für Lagrange-Multiplikatoren $\lambda_1, ..., \lambda_n$. Diese sind gegeben als Lösung des (dualen) Optimierungsproblems

$$\textit{Maximiere } \sum_{i=1}^{n}\lambda_i - \frac{1}{2}\sum_{i,j=1}^{n} y_i y_j \lambda_i \lambda_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle_{\mathbb{R}^d} \tag{5.19}$$

unter den Nebenbedingungen

$$0 \le \lambda_i \le C, 1 \le i \le n, \textit{ und } \sum_{i=1}^{n}\lambda_i y_i = 0.$$

Die Lösung ist darüber hinaus derart, dass nur die zu den Support-Vektoren (den am nächsten zur verallgemeinerten optimalen Hyperebene gelegenen Trainings-Merkmals-Vektoren) gehörigen λ_i von Null verschieden sind. Der resultierende Klassifikator ist demnach von der folgenden Form:

$$\hat{f}(\boldsymbol{x}) = sgn\left(\sum_{i:\boldsymbol{x}_i \textit{ist Support-Vektor}} y_i\lambda_i\langle\boldsymbol{x}_i, \boldsymbol{x}\rangle_{\mathbb{R}^d} - b^*\right), \tag{5.20}$$

wobei

$$b^* = \frac{1}{2}\left[\langle\theta^*, \boldsymbol{x}_{supp}^{+1}\rangle + \langle\theta^*, \boldsymbol{x}_{supp}^{-1}\rangle\right]$$

für beliebige Support-Vektoren $\boldsymbol{x}_{supp}^{+1}, \boldsymbol{x}_{supp}^{-1}$ aus den beiden Klassen ist.

Bemerkung 5.16

(a) Wegen der Darstellung (5.20) *heißt der Klassifikationsalgorithmus auch „Support-Vektor-Maschine“(SVM).*

(b) Wegen (5.19)*und* (5.20) *genügt es, Skalarprodukte im Merkmalsraum auswerten zu können.*

Lemma 5.17 (Satz von Mercer)
Wir gehen zurück zum Original-Eingaberaum $D \ni \boldsymbol{x}$ (statt des Merkmalsraums $U \subseteq \mathbb{R}^d$). Eine symmetrische, quadrat-integrierbare Funktion $K : D \times D \to \mathbb{R}$ besitzt genau dann eine Darstellung der Form

$$K(\boldsymbol{x}^1, \boldsymbol{x}^2) = \sum_{k=1}^{\infty} a_k \psi_k(\boldsymbol{x}^1)\psi_k(\boldsymbol{x}^2) \tag{5.21}$$

mit positiven Koeffizienten $(a_k)_{k\geq 1}$, wenn die Mercer-Bedingung

$$\int\int K(\boldsymbol{x}^1, \boldsymbol{x}^2)g(\boldsymbol{x}^1)g(\boldsymbol{x}^2)d\boldsymbol{x}^1 d\boldsymbol{x}^2 > 0 \tag{5.22}$$

für alle Funktionen $g \neq 0$ mit

$$\int g^2(\boldsymbol{x})d\boldsymbol{x} < \infty$$

erfüllt ist.
Die Darstellung (5.21) *besagt, dass $K(\cdot,\cdot)$ ein Skalarprodukt in einem (nicht explizit angegebenen) Merkmalsraum beschreibt. Die Funktion K heißt auch ein Kern.*
Umgekehrt existiert zu jeder Kernfunktion K auf $D\times D$ ein Merkmalsraum mit Merkmalen $(\psi_k)_{k\geq 1}$, in dem Skalarprodukte durch (5.21) *gegeben sind. Deswegen können SVMs auch ohne den „Umweg“über die explizite Definition der Merkmale direkt vermittels einer Kernfunktion auf $D \times D$ definiert werden. Dies nennt man auch den „Kern-Trick“.*
Mit Hilfe eines (beliebig gewählten) Kerns K ergibt sich also die Darstellung

$$\hat{f}(\boldsymbol{x}) = sgn\left(\sum_{i:\boldsymbol{x}_i \text{ist Support-Vektor}} y_i\lambda_i K(\boldsymbol{x}_i, \boldsymbol{x}) - b^*\right) \tag{5.23}$$

einer SVM.

Beispiel 5.18 (SVMs mit Kernfunktionen)
(a) Polynomieller Kern:
Sei $D \subseteq \mathbb{R}^d$ und

$$K(\boldsymbol{x}, \boldsymbol{x}_i) = [\langle \boldsymbol{x}, \boldsymbol{x}_i\rangle_{\mathbb{R}^d} + 1]^m,\ m \in \mathbb{N}.$$

Durch Übergang in den höher-dimensionalen Merkmalsraum können nicht-lineare Separierungen von Trainingsdaten im Original-Datenraum D erreicht werden.
(b) Radial Basis Funktion (RBF)-Kern:
Sei $K(\boldsymbol{x}, \boldsymbol{x}_i) = K_\gamma(\|\boldsymbol{x} - \boldsymbol{x}_i\|)$, wobei $K_\gamma : \mathbb{R} \to \mathbb{R}$ positiv definit und monoton ist mit

$$\lim_{\|z\|\to\infty} K_\gamma(\|z\|) \to 0$$

für alle $\gamma > 0$. Besonders beliebt ist der Gauß-Kern, gegeben durch

$$K_\gamma(\|\boldsymbol{x} - \boldsymbol{x}_i\|) = \exp(-\gamma\|\boldsymbol{x} - \boldsymbol{x}_i\|_2^2).$$

(c) Neuronales Netz mit zwei Schichten:
Sei $K(\boldsymbol{x}, \boldsymbol{x}_i) = S(\langle \boldsymbol{x}, \boldsymbol{x}_i \rangle_{\mathbb{R}^d})$ für $\boldsymbol{x}, \boldsymbol{x}_i \in D \subseteq \mathbb{R}^d$ und eine geglättete Vorzeichen- (sigmoidale) Funktion S. Hier ist die SVM also gegeben durch eine Überlagerung (Zusammenschaltung) von sigmoidalen Funktionen, was einem neuronalen Netz mit zwei Schichten entspricht; vergleiche Bemerkung 5.7.

Bemerkung 5.19 (Mehrklassen-Klassifikation)
Klassifikationsprobleme mit $K > 2$ Klassen (das heißt, $W = \{1, 2, ..., K\} \ni y$) können mit einem zweistufigen Verfahren behandelt werden:

1) *Konstruiere K binäre Klassifikatoren $\hat{f}_k, 1 \leq k \leq K$, so dass $\hat{f}_k(\boldsymbol{x}) = +1$ Zugehörigkeit zu Klasse k bedeutet und $\hat{f}_k(\boldsymbol{x}) = -1$ Nicht-Zugehörigkeit zu Klasse k.*

2) *Für $1 \leq i \leq n$, klassifiziere*

$$\hat{y}_i = \underset{1 \leq k \leq K}{\operatorname{argmax}} \{\hat{f}_1(\boldsymbol{x}_i), ..., \hat{f}_K(\boldsymbol{x}_i)\}.$$

Es existieren aber auch SVM-Implementierungen, die in einem Schritt direkt $\hat{y}_i$ liefern; vergleiche Abschnitt 10.10 in Vapnik (1998).

Kapitel 6

Methoden zur Funktionenschätzung

Wir kehren zurück zum statistischen Lernproblem der (Mittelwert-) Regression, das in Beispiel 1.5.(b) aufgeworfen wurde. Hierbei ist $W = \mathbb{R}$.
In Beispiel 1.5.(b) hatten wir die quadratische Verlustfunktion L, gegeben durch

$$L(y, f(\boldsymbol{x}, \theta)) = (y - f(\boldsymbol{x}, \theta))^2 \tag{6.1}$$

betrachtet, die zur kleinsten Quadrate-Methode führt; vergleiche Beispiel 1.8.(a). Ein Nachteil der durch (6.1) gegebenen Verlustfunktion ist, das sie nicht robust gegenüber Ausreißern ist.

Definition 6.1 (Verlustfunktionen für Regressionsprobleme)
Für eine gegebene reelle Zahl $\varepsilon > 0$ sei

$$|y - f(\boldsymbol{x}, \theta)|_\varepsilon = \begin{cases} 0, & \textit{falls } |y - f(\boldsymbol{x}, \theta)| \leq \varepsilon, \\ |y - f(\boldsymbol{x}, \theta)| - \varepsilon, & \textit{sonst}. \end{cases}$$

(a) Lineare ε-insensitive Verlustfunktion:
Wir nennen die durch

$$L(y, f(\boldsymbol{x}, \theta)) = |y - f(\boldsymbol{x}, \theta)|_\varepsilon \tag{6.2}$$

gegebene Verlustfunktion L eine lineare ε-insensitive Verlustfunktion, wobei $\varepsilon > 0$ vorgegeben ist.

(b) Quadratische ε-insensitive Verlustfunktion:
Wir nennen die durch

$$L(y, f(\boldsymbol{x}, \theta)) = \{|y - f(\boldsymbol{x}, \theta)|_\varepsilon\}^2 \tag{6.3}$$

gegebene Verlustfunktion L eine quadratische ε-insensitive Verlustfunktion, wobei $\varepsilon > 0$ vorgegeben ist.

(c) Huber'sche Verlustfunktion:
Für vorgegebenes $c > 0$ heißt die durch

$$L(y, f(\boldsymbol{x}, \theta)) = \begin{cases} c|y - f(\boldsymbol{x}, \theta)| - \frac{c^2}{2}, & |y - f(\boldsymbol{x}, \theta)| > c, \\ \frac{1}{2}|y - f(\boldsymbol{x}, \theta)|^2, & \textit{sonst}, \end{cases} \tag{6.4}$$

gegebene Verlustfunktion Huber'sche Verlustfunktion mit Parameter c, nach Huber (1964).

Die in Definition 6.1 eingeführten Verlustfunktionen legen weniger Gewicht auf große Werte von $|y - f(\boldsymbol{x}, \theta)|$ als die quadratische Verlustfunktion aus (6.1).

Definition 6.2 (Support-Vektor-Regression)
Der Support-Vektor-Maschinen-Ansatz zur Lösung von Regressionsproblemen ist durch die folgenden drei Eigenschaften gekennzeichnet.

(i) Es wird die Funktionenmenge

$$\mathcal{M} = \{f(\cdot, \cdot) : D \times \Theta \to W \\ (\boldsymbol{x}, \theta) \mapsto f(\boldsymbol{x}, \theta)\}$$

betrachtet, wobei

$$f(\boldsymbol{x}, \theta) = \langle \boldsymbol{w}, \boldsymbol{u} \rangle_{\mathbb{R}^d} + b, \quad \theta = (\boldsymbol{w}, b), \tag{6.5}$$

gilt, mit einem zu $\boldsymbol{x}$ gehörigen Merkmalsvektor $\boldsymbol{u} \equiv \boldsymbol{u}(\boldsymbol{x}) \in U \subseteq \mathbb{R}^d$.

(ii) Es wird eine der in Definition 6.1 eingeführten Verlustfunktionen verwendet.

(iii) Es wird das SRM-Prinzip verfolgt, wobei das Strukturelement $\mathcal{M}_k$ gegeben ist durch die Bedingung

$$\|\boldsymbol{w}\|_2^2 \leq c_k \tag{6.6}$$

für eine wachsende Folge $(c_k)_{k \geq 1}$ nicht-negativer reeller Zahlen.

Lemma 6.3
Für ein gegebenes Strukturelement $\mathcal{M}_k$ ist der optimale (bezüglich der Minimierung des empirischen Risikos) Richtungsvektor $\hat{\boldsymbol{w}}$ in (6.5) *gegeben als eine Linearkombination der Merkmalsvektoren $\boldsymbol{u}_1, ..., \boldsymbol{u}_n$. Das heißt, es gilt:*

$$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{n} \beta_i \langle \boldsymbol{u}(\boldsymbol{x}), \boldsymbol{u}(\boldsymbol{x}_i) \rangle_{\mathbb{R}^d} + b$$

für Koeffizienten $\beta_1, ..., \beta_n$ und mit $\boldsymbol{u}_i := \boldsymbol{u}(\boldsymbol{x}_i), 1 \leq i \leq n$.

Beweis: Wir beweisen die Aussage nur für die lineare ε-insensitive Verlustfunktion aus (6.2); der Beweis für die anderen Verlustfunktionen wird analog geführt.
Wir betrachten das (empirische) Risikofunktional

$$R_{emp}(\boldsymbol{w}, b) = n^{-1} \sum_{i=1}^{n} |y_i - \langle \boldsymbol{w}, \boldsymbol{u}_i \rangle_{\mathbb{R}^d} - b|_{\varepsilon},$$

wobei $\boldsymbol{u}_1, ..., \boldsymbol{u}_n$ die Merkmalsvektoren aus dem Trainingsdatensatz sind. Es gilt, $R_{emp}(\boldsymbol{w}, b)$ unter der Nebenbedingung (6.6) zu minimieren. Dieses Optimierungsproblem kann äquivalenterweise wie folgt charakterisiert werden.
Seien $\xi_1, ..., \xi_n, \xi_1^*, ..., \xi_n^*$ nicht-negative Schlupfvariablen. Minimiere die Funktion F, gegeben durch

$$F(\xi, \xi^*) = \sum_{i=1}^{n} \xi_i^* + \sum_{i=1}^{n} \xi_i$$

(mit $\xi = \xi_1, ..., \xi_n)^\top$ und $\xi^* = (\xi_1^*, ..., \xi_n^*)^\top$) unter den Nebenbedingungen

$$y_i - \langle \boldsymbol{w}, \boldsymbol{u}_i \rangle_{\mathbb{R}^d} - b \leq \varepsilon + \xi_i^*, \; 1 \leq i \leq n, \tag{$\star$}$$

$$\langle \boldsymbol{w}, \boldsymbol{u}_i \rangle_{\mathbb{R}^d} + b - y_i \leq \varepsilon + \xi_i, \; 1 \leq i \leq n, \tag{$\star\star$}$$

$$\xi_i, \xi_i^* \geq 0, \; 1 \leq i \leq n, \tag{$\star\star\star$}$$

$$\|\boldsymbol{w}\|_2^2 \leq c_k. \tag{$\star\star\star\star$}$$

Wir betrachten dazu eine Lagrange-Funktion $\mathcal{L}$, gegeben durch

$$\begin{aligned}
\mathcal{L}(\theta, \xi^*, \xi; \alpha^*, \alpha, C^*, \gamma, \gamma^*) \;=\; & \sum_{i=1}^{n} (\xi_i^* + \xi_i) \\
& - \sum_{i=1}^{n} \alpha_i [y_i - \langle \boldsymbol{w}, \boldsymbol{u}_i \rangle_{\mathbb{R}^d} - b + \varepsilon + \xi_i] \\
& - \sum_{i=1}^{n} \alpha_i^* [\langle \boldsymbol{w}, \boldsymbol{u}_i \rangle_{\mathbb{R}^d} + b - y_i + \varepsilon + \xi_i^*] \\
& - \sum_{i=1}^{n} (\gamma_i^* \xi_i^* + \gamma_i \xi_i) \\
& - \frac{C^*}{2} (c_k - \langle \boldsymbol{w}, \boldsymbol{w} \rangle_{\mathbb{R}^d}).
\end{aligned}$$

Diese Lagrange-Funktion muss bezüglich $\boldsymbol{w}, b, (\xi_i)_{1 \leq i \leq n}$ und $(\xi_i^*)_{1 \leq i \leq n}$ minimiert werden und bezüglich $C^* \geq 0, \alpha_i^* \geq 0, \alpha_i \geq 0, \gamma_i \geq 0$, sowie $\gamma_i^* \geq 0$ $(1 \leq i \leq n)$ maximiert werden.
Für die partielle Ableitung von $\mathcal{L}$ nach $\boldsymbol{w}$ erhalten wir

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}} = \sum_{i=1}^{n} \alpha_i \boldsymbol{u}_i - \sum_{i=1}^{n} \alpha_i^* \boldsymbol{u}_i + C^* \boldsymbol{w}.$$

Somit ist

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}} = \mathbf{0} \Leftrightarrow \boldsymbol{w} = \sum_{i=1}^{n} \frac{(\alpha_i^* - \alpha_i)}{C^*} \boldsymbol{u}_i,$$

was die Aussage impliziert. ■

Bemerkung 6.4

Diejenigen $\boldsymbol{x}_i$, für die die $\beta_i = \frac{\alpha_i^* - \alpha_i}{C^*}$ von Null verschieden ist, werden die Support-Vektoren des Regressionsproblems genannt.

Korollar 6.5

Wegen Lemma 6.3 in Verbindung mit dem Kern-Trick gilt für die Lösung des SVM-Regressionsproblems, dass die optimale Funktion $\hat{f}$ in der Form

$$\hat{f}(\boldsymbol{x}) = \sum_{i:\boldsymbol{x}_i \text{ ist Support-Vektor}} \beta_i K(\boldsymbol{x}, \boldsymbol{x}_i) + b$$

geschrieben werden kann, wobei $K : D \times D \to \mathbb{R}$ eine Kern-Funktion ist, die die Mercer-Bedingung (5.22) *erfüllt.*

Beispiel 6.6 (Polynomielle Approximation)

Sei $D = W = \mathbb{R}$ und betrachte ein System $(P_\ell)_{\ell \geq 1}$ von orthonormalen Polynomen. Angenommen, wir möchten eine Kern-Funktion $K(\cdot, \cdot)$ verwenden, die einer Entwicklung der Funktion f in die durch $(P_\ell)_{\ell \geq 1}$ gegebene Polynombasis entspricht. Dann sind die folgenden Christoffel-Darboux-Formeln hilfreich.

$$\begin{aligned} \sum_{\ell=1}^{L} P_\ell(x_1) P_\ell(x_2) &= a_L \frac{P_{L+1}(x_1) P_L(x_2) - P_L(x_1) P_{L+1}(x_2)}{x_1 - x_2}, \; x_1 \neq x_2, \\ \sum_{\ell=1}^{L} P_\ell^2(x) &= a_L [P'_{L+1}(x) P_L(x) - P'_L(x) P_{L+1}(x)] \end{aligned}$$

für eine Konstante a_L, die von der Wahl des Polynomsystems abhängt.
Eine regularisierter Kern ist gegeben durch

$$K(x_1, x_2) = \sum_{\ell=1}^{L} r_\ell P_\ell(x_1) P_\ell(x_2),$$

wobei $(r_\ell)_{\ell \geq 1}$ eine Folge positive reeller Zahlen ist mit $\lim_{\ell \to \infty} r_\ell = 0$.
Zum Beispiel kann $r_\ell = q^\ell$ für $0 < q < 1$ gewählt werden. Für manche Polynomsysteme (z.B., Hermite-Polynome) existieren sogar geschlossene Ausdrücke für $\sum_{\ell=1}^{\infty} q^\ell P_\ell(x_1) P_\ell(x_2)$; vergleiche, zum Beispiel, Theorem 53 in Titchmarsh (1948) bzw. Watson (1933).

Bemerkung 6.7

Sei $D = \mathbb{R}^d, d \in \mathbb{N}$, und seien univariate Kernfunktionen K_k, $1 \leq k \leq d$ gegeben, wobei $\forall 1 \leq k \leq d : K_k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. Dann ist eine Kernfunktion $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ gegeben durch

$$K(\boldsymbol{x}, \boldsymbol{z}) = \prod_{k=1}^{d} K_k(x_k, z_k),$$

$\boldsymbol{x} = (x_1, ..., x_d)^\top \in \mathbb{R}^d, \boldsymbol{z} = (z_1, ..., z_d)^\top \in \mathbb{R}^d$.

Beispiel 6.8 (Spline-Approximation)

Sei $D = [0, a]$ für gegebenes $a > 0$. Wir betrachten einen Spline (stückweises Polynom) der Ordnung $d \geq 0$ mit m äquidistanten Stützstellen der Form

$$t_k = \frac{ka}{m},\ 1 \leq k \leq m.$$

Dieser lässt sich wie folgt darstellen:

$$f(x) = \sum_{r=0}^{d} a_r x^r + \sum_{k=1}^{m} a_{d+k}(x - t_k)_+^d,\ x \in D = [0, a],$$

für $d + m + 1$ freie Parameter $a_0, ..., a_{d+m}$. Dabei ist

$$(x - t_k)_+^d = \begin{cases} 0, & \text{falls } x \leq t_k, \\ (x - t_k)^d, & \text{falls } x > t_k. \end{cases}$$

Betrachte nun die Abbildung von $D = [0, a]$ nach $\mathbb{R}^{d+m+1}$, die gegeben ist durch

$$x \in D \mapsto \boldsymbol{u} \equiv \boldsymbol{u}(x) := (1, x, ..., x^d, (x - t_1)_+^d, ..., (x - t_m)_+^d)^\top \in \mathbb{R}^{d+m+1}.$$

Dann ist , mit $\boldsymbol{a} = (a_0, ..., a_{d+m})^\top$,

$$f(x) = \langle \boldsymbol{a}, \boldsymbol{u}(x) \rangle_{\mathbb{R}^{d+m+1}}.$$

Eine Spline-generierende Kernfunktion $K : D \times D \to \mathbb{R}$ ist daher gegeben durch

$$\begin{aligned} K(x, z) &= \langle \boldsymbol{u}(x), \boldsymbol{u}(z) \rangle_{\mathbb{R}^{d+m+1}} \\ &= \sum_{r=0}^{d} x^r z^r + \sum_{k=1}^{m} (x - t_k)_+^d\, (z - t_k)_+^d. \end{aligned}$$

Beispiel 6.9 (Fourier-Approximation)

Sei $x \in D = \mathbb{R}$ und betrachte eine Fourier-Entwicklung von $f(x)$ der Ordnung $L \in \mathbb{N}$.
Sei dazu $\boldsymbol{u} \equiv \boldsymbol{u}(x)$ gegeben durch

$$\boldsymbol{u}(x) = \left(\frac{1}{\sqrt{2}}, \sin(x), ..., \sin(Lx), cos(x), ..., \cos(Lx) \right)^\top \in \mathbb{R}^{2L+1}.$$

Damit ist, für Fourier-Koeffizienten $\boldsymbol{a} = (a_0, ..., a_L, b_1, ..., b_L)^\top$,

$$\langle \boldsymbol{a}, \boldsymbol{u}(x) \rangle_{\mathbb{R}^{2L+1}} = \frac{a_0}{\sqrt{2}} + \sum_{\ell=1}^{L} \{a_\ell \sin(\ell x) + b_\ell \cos(\ell x)\}.$$

Die zugehörige Kernfunktion $K : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ *ist gegeben durch*

$$\begin{aligned}
K(x,z) &= \langle \boldsymbol{u}(x), \boldsymbol{u}(z) \rangle_{\mathbb{R}^{2L+1}} \\
&= \frac{1}{2} + \sum_{\ell=1}^{L} \{\sin(\ell x)\sin(\ell z) + \cos(\ell x)\cos(\ell z)\} \\
&= \frac{1}{2} + \sum_{\ell=1}^{L} \left\{ \frac{1}{2}\cos(\ell(x-z)) - \frac{1}{2}\cos(\ell(x+z)) + \frac{1}{2}\cos(\ell(x-z)) + \frac{1}{2}\cos(\ell(x+z)) \right\} \\
&= \frac{1}{2} + \sum_{\ell=1}^{L} \cos(\ell(x-z)) \\
&= \frac{1}{2} \left[1 + 2\sum_{\ell=1}^{L} \cos(\ell(x-z)) \right] \\
&= \frac{1}{2} \left[\frac{\sin\left((L+1/2)(x-z)\right)}{\sin\left((x-z)/2\right)} \right];
\end{aligned}$$

siehe zum Beispiel Abschnitt 1.1 in Zygmund (2002).

Bemerkung 6.10

Der SVM-Ansatz zur Funktionenschätzung kann auch zur Approximation von (bedingten) Lebesguedichten eingesetzt werden; vergleiche Abschnitt 11.10 und 11.11 in (Vapnik, 1998).

Literaturverzeichnis

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Autom. Control 19*, 716–723.

Breiman, L. (2001). Statistical modeling: the two cultures. *Statist. Sci. 16*(3), 199–231. With comments and a rejoinder by the author.

Cherkassky, V. S. and F. M. Mulier (2007). *Learning from data: concepts, theory, and methods. 2nd ed.* Hoboken, NJ: John Wiley & Sons.

Huber, P. (1964). Robust estimation of a location parameter. *Ann. Math. Stat. 35*, 73–101.

Novikoff, A. (1963). On convergence proofs for perceptrons. Proc. Sympos. math. Theor. Automata, New York, April 24-26, 1962, 615-622 (1963).

Pollard, D. (1984). *Convergence of stochastic processes.* Springer Series in Statistics. New York etc.: Springer-Verlag.

Rosenblatt, F. (1958, Nov). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol Rev 65*(6), 386–408.

Rosenblatt, F. (1962). *Principles of neurodynamics. Perceptrons and the theory of brain mechanisms*. Spartan Books, Washington, D.C.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat. 6*, 461–464.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc., Ser. B 58*(1), 267–288.

Titchmarsh, E. C. (1948). *Introduction to the theory of Fourier integrals. Second Edition*. Oxford University Press.

Vapnik, V. (2000). *The nature of statistical learning theory. 2nd ed.* New York, NY: Springer.

Vapnik, V. N. (1998). *Statistical learning theory.* Chichester: Wiley.

Vapnik, V. N. and A. Y. Chervonenkis (1991). The necessary and sufficient conditions for consistency of the method of empirical risk minimization. *Pattern Recogn. Image Anal. 1*, 284–305.

Watson, G. N. (1933). Notes on generating functions of polynomials. (2) Hermite polynomials. *J. Lond. Math. Soc. 8*, 194–199.

Zygmund, A. (2002). *Trigonometric series. Volumes I and II combined. With a foreword by Robert Fefferman. 3rd Edition.* Cambridge: Cambridge University Press.