Universität Hamburg
Studiengang Mathematik

Bachelorarbeit im Bereich der Optimierung und Approximation
zum erlangen des akademischen Grades
Bachelor of Science

Support Vector Machines

als empirische Risikominimierer

Autor:	Max Lewerenz
Matrikelnummer:	7048126
Erstgutachter:	Dr. Matthias Beckmann
Zweitgutachter:	Prof. Dr. Armin Iske

Inhaltsverzeichnis

1 Einführung

1.1 Voraussetzungen

Diese Bachelorarbeit setzt grundlegende Begriffe und Resultate der Einführungsveranstaltungen Analysis und Lineare Algebra, sowie der Stochastik voraus. Die Kapitel 1 und 2 des Skriptes *Mathematics of Machine Learning* von Martin Lotz decken die geforderten Kenntnisse ab.

1.2 Einführung in die Thematik

Die Support Vector Machine (folgend auch SVM, dt. Stütz Vektor Maschine), ist eine weit verbreitete Methode des maschinellen Lernens. Die Support Vector Machine dient zur Klassifikation von Daten. Die SVM erstellt zu einer Trainingsmenge, eine Zweiklassenpartition von Punkten, ein Modell, welches neue Datenpunkte einer der beiden Klassen zuordnet. Sie tut dies indem sie die Trainingsbeispiele in einen Raum abbildet und in diesem eine möglichst große Lücke zwischen den beiden Klassen findet. Um neue Beispiele zu klassifizieren, werden diese ebenfalls in den Raum abgebildet und dort je nachdem auf welcher Seite der Lücke sie liegen einer der beiden Klassen zugeordnet. Im einfachsten Fall geschieht dies durch eine Ebene im Raum. Der in Kapitel 4 besprochene Kernel Trick lässt jedoch auch nicht lineare Entscheidungsfunktionen zu.
Ein modernes Beispiel für die Anwendung wäre zum Beispiel das Vorhersagen von Hautkrebs anhand von Bildern. Gegeben einer Datenmenge von Bildern von Pigmentzellen von denen man weiß ob diese bösartige Melanome sind oder nicht, kann man eine SVM anlernen. Die SVM ordnet dann neue Bilder einer gutartigen oder bösartigen Klasse zu.
Die Theorie der Support Vector Machine lässt sich in das Feld der statistischen Lerntheorie einordnen. Dieses Gebiet ist Gegenstand aktueller Forschung. Insbesondere der Umgang mit sehr großen Datenmengen in der Praxis und die Untersuchung wie sich das Finden der Entscheidungsfunktion beschleunigen kann beschäftigen die Wissenschaftler.
Das Ziel dieser Arbeit ist eine Einführung in die Thematik der Support Vector Machine zu liefern und diese aus statistischer Sicht genauer zu betrachten.
Um dieses Ziel zu erfüllen wurden Beweise aus der Literatur an einigen Stellen in Eigenleistung ergänzt und ausführlicher gestaltet. Darüber hinaus wurden Risikoschranken und Oracle-Inequalities für spezifische Wahlen von Kernelfunktionen vom Autor untersucht. An Stellen an denen Beweise, Korollare oder Lemma aus der Literatur übernommen wurden, wird an entsprechender Stelle auf die Literatur verwiesen auch wenn der entsprechende Beweis vom Autor ergänzt, kommentiert oder ausführlicher gestaltet wird. Darüber hinaus sind alle graphischen Darstellungen in dieser Arbeit vom Autor erstellt und dienen der Veranschaulichung und dem besseren Verständnis.

1.3 Übersicht

Die Arbeit ist wie folgt gegliedert. Zunächst wird die Support Vector Machine über die geometrische Interpretation eingeführt. Dann wird die äquivalente Formulierung der Support Vector Machine als empirischer Risikominimierer vorgestellt um im weiteren Verlauf die statistische Untersuchung zu ermöglichen. Bis zu diesem Punkt wurden nur lineare Entscheidungsfunktionen untersucht, Kapitel 4 präsentiert den weit verbreiteten Kernel Trick um auch nicht lineare Entscheidungsfunktionen zuzulassen. Abschließend werden eine Risikoabschätzung und eine Oracle Ungleichung vorgestellt und diese für verschiedene Wahlen der Kernel untersucht.

1.4 Literatur

Als Quellen dieser Arbeit dienten hauptsächlich die Skripte

[6] Martin Lotz: *Mathematics of Machine Learning*

[9] Mathias Trabs: *The Mathematics of Machine Learning*

Unterstützend diente das sehr ausführliche Buch über Support Vector Machines

[3] Andreas Christmann, Ingo Steinwart: *Support Vector Machines*

Zusätzlich dienten die folgenden Bücher über maschinelles Lernen im weiteren Rahmen als Quellen

[1] Shai Ben-David, Shai Shalev-Shwartz: *Understanding Machine Learning: From Theory to Algorithms*

[4] Floris Ernst, Achim Schweikard: *Fundamentals of Machine Learning*

[7] Mehryar Mohri, Afshin Rostamizadeh, Ameet Talwalkar: *Foundations of Machine Learning*

Um sich über den aktuellen Stand der Forschung und engere statistische Abschätzungen zu informieren wurden unter anderem die Folgenden Artikel gesichtet, diese dienten jedoch im weiteren Verlauf nicht als Quelle dieser Arbeit.

[2] Gilles Blanchard, Oliver Busquet, Pascal Massart: *Statistical Performance of Support Vector Machines*

[8] Simon Fischer, Ingo Steinwart: *A closer look at covering number bounds for Gaussian kernels*

2 Grundlagen

2.1 Ausgangssituation

Die Situationen in denen Support Vector Machines angewendet werden lassen sich wie folgt modellieren. Die Daten die klassifiziert werden sollen stammen aus einer Menge $\mathcal{X} \subseteq \mathbb{R}^d$, $d \in \mathbb{N}$. Jedes Element aus der Datenmenge ist genau einer von zwei Klassen zugeordnet. Die Menge der Klassen wird im Folgenden mit $\mathcal{Y} = \{-1, 1\}$ bezeichnet. Um später statistische Untersuchungen anstellen zu können wird angenommen, dass die Daten einer Wahrscheinlichkeitsverteilung $\mathcal{D}$ auf $\mathcal{X} \times \mathcal{Y}$ unterliegen. Hierbei ist das Wahrscheinlichkeitsmaß eingeschränkt auf $\mathcal{Y}$ vollständig von $\mathcal{X}$ abhängig. Gleiche Datenpunkte sollen immer der gleichen Klasse zugeordnet werden. Das heißt für zwei Zufallsvariablen $(X, Y) \sim \mathcal{D}$ und $(x, y) \in \mathcal{X} \times \mathcal{Y}$ gilt

$$\mathbb{P}_{\mathcal{D}}(Y = y \mid X = x) = \begin{cases} 1 & \text{falls } y \text{ die Klasse von } x \text{ beschreibt} \\ 0 & \text{sonst.} \end{cases}$$

Stichproben der Größe $m \in \mathbb{N}$ können nun als Zufallsvariablen mit der Verteilung $(X, Y) \sim \mathcal{D}^m$ angesehen werden. Wobei auf $\mathcal{D}^m$ die Produkt-σ-Algebra betrachtet wird.

2.2 Support Vector Machine als Abstandsmaximierer

2.2.1 Hard-margin SVM

Wir starten mit einer Menge von Bezeichnern $\mathcal{Y} = \{-1, 1\}$ und Trainingspunkten $\mathcal{X} = \{x_1, ..., x_m\} \subseteq \mathbb{R}^d$. Jedem Trainingspunkt $x_i \in \mathcal{X}$ ist eine Klasse $y_i \in \mathcal{Y}$ zugeordnet. Die Datenmenge ergibt sich zu $D = \{(x_1, y_1), ..., (x_m, y_m)\} \subseteq \mathbb{R}^d \times \mathcal{Y}$.
Im einfachsten Fall sind die Daten linear trennbar. Das heißt, dass eine affine Hyperebene $h(x) = w^T x + b$ existiert, mit $h(x) = \begin{cases} h(x_i) > 0 & \text{, falls } y_i = 1 \\ h(x_i) < 0 & \text{, falls } y_i = -1 \end{cases}$. Das Problem kann beschrieben werden als Suche nach Vektoren $w \in \mathbb{R}^d, b \in \mathbb{R}$ so, dass

$$\begin{aligned} w^T x_i + b &\geq 1 \quad \text{für } y_i = 1 \\ w^T x_i + b &\leq -1 \quad \text{für } y_i = -1 \end{aligned} \tag{1}$$

für alle $i = 1, ..., m$.

Wir bemerken, dass sich die Hyperebene durch skalieren von w und b nicht ändert, wir können also 1 und -1 durch beliebige positive respektive negative Werte ersetzen.
Außerdem sehen wir, dass wir die Ungleichungen 1 äquivalent als $y_i(w^T x + b) - 1$ beschreiben können. Ziel ist es die Ebene zu finden, die den Abstand zu den Punkten die der Ebene am nächsten liegen zu maximieren.
Im folgenden beschreibt $H = \{x \in \mathbb{R}^d \mid w^T x + b = 0\}, w \in \mathbb{R}^d, b \in \mathbb{R}$ eine Hyperebene im $\mathbb{R}^d$.

Definition 2.2.1 (Margin)**.** Seien δ_+ und δ_- die Abstände der Hyperebene zu den nächsten positiven beziehungsweise negativen Datenpunkten. Dann heißt $\delta := \delta_+ + \delta_-$ **Margin** (dt. **Rand**).

Satz 2.2.2 (Lotz, 2020,[6] Chapter 17)**.** *Die Margin einer Hyperebene H, beschrieben durch $w^T x + b = 0$, erfüllt die Gleichung*

$$\delta = \frac{2}{\|w\|_2}.$$

Beweis. Sei $x \in \mathbb{R}^d$ mit $w^T x + b = 1$. Also ist x ein nächster Punkt an H. Wähle $p = cw \in H \subseteq \mathbb{R}^d$ mit $c \in \mathbb{R}$ ein vielfaches von w, das auf der Hyperebene liegt. Es gilt also $\langle w, cw \rangle + b = 0$. Es folgt

$$\langle w, cw \rangle + b = 0 \;\Leftrightarrow\; c\langle w, w \rangle + b = 0 \;\Leftrightarrow\; c = -\frac{b}{\|w\|_2^2}.$$

Insbesondere $p = -(\frac{b}{\|w\|^2})w$. Der Abstand von x zur Ebene lässt sich nun mithilfe der orthogonalen Projektion berechnen zu

$$\begin{aligned}\delta_+ &= \langle x - p, \frac{w}{\|w\|_2} \rangle = \langle x + \frac{b}{\|w\|_2^2} w, \frac{w}{\|w\|_2} \rangle = \frac{\langle x, w \rangle}{\|w\|_2} + \frac{b}{\|w\|_2^3} \langle w, w \rangle \\ &= \frac{1 - b}{\|w\|_2} + \frac{b}{\|w\|_2} = \frac{1}{\|w\|_2}.\end{aligned}$$

Analog berechnet man δ_- für ein $x \in \mathbb{R}$ mit $w^T x + b = -1$. Also

$$\delta = \delta_+ + \delta_- = \frac{2}{\|w\|_2}.$$

□

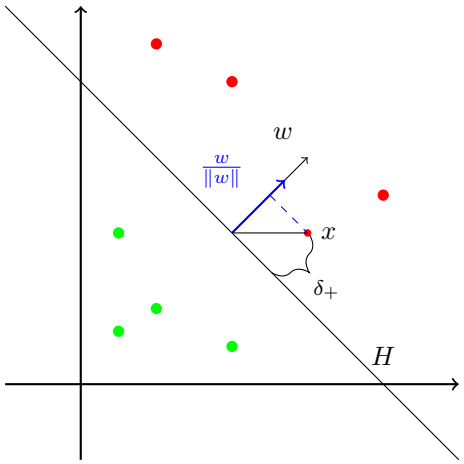

Abbildung 1: Berechnung der Margin

Definition 2.2.3 (Stützvektor, support Vector)**.** Sei H eine Hyperebene. Dann heißt ein Punkt $x \in \mathbb{R}^d$ mit $w^T x + b \in \{-1, 1\}$ **Stützvektor**.

Wir bemerken, dass Stützvektoren genau die Punkte der Datenmenge sind, die am nächsten an der Hyperebene liegen.
Wollen wir nun zu einer Datenmenge $D = \{(x_1, y_1), ..., (x_m, y_m)\}$ die Hyperebene mit dem größten Rand δ finden, so ergibt sich das Optimierungsproblem

$$\begin{aligned}&\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \|w\|_2^2 \\ &\text{so dass } y_i(w^T x_i + b) - 1 \geq 0 \text{ für alle } i \in \{1, ..., m\}.\end{aligned} \tag{2}$$

2.2.2 Soft-margin SVM

Betrachten wir nun den Fall, dass sich die Daten nicht linear durch eine Hyperebene im $\mathbb{R}^d$ trennen lassen. Die Nebenbedingungen in (2) lassen sich in diesem Fall nicht erfüllen. Wir passen das Problem an, indem wir Missklassifikation erlauben. Dies geschieht durch das Einführen sogenannter slack Variables (dt. Schlupfvariablen) $\xi = (\xi_1, ..., \xi_m)^T \in \mathbb{R}^d$. Die Nebenbedinugungen aus (2) ändern sich zu

$$\begin{aligned} w^T x_i + b &\geq 1 - \xi_i \quad \text{, falls } y_i = 1 \\ w^T x_i + b &\leq -1 + \xi_i \text{ , falls } y_i = -1. \end{aligned}$$

Liegt der i-te Datenpunkt auf der richtigen Seite der Hyperebene und jenseits der Margin δ, so ist $\xi_i = 0$. Liegt der i-te Datenpunkt zwischen Margin und Hyperebene, jedoch auf der richtigen Seite der Hyperebene, so gilt $\xi_i \in (0, 1)$. Liegt der i-te Datenpunkt auf der falschen Seite der Hyperebene, so wird x_i falsch klassifiziert und es gilt $\xi_i > 1$. Das Soft-margin SVM Optimierungsproblem ergibt sich zu

$$\begin{aligned} &\min_{w,\xi \in \mathbb{R}^d, b \in \mathbb{R}} \|w\|_2^2 + \frac{C}{m} \sum_{i=1}^{m} \xi_i \\ &\text{so dass } y_i(w^T x_i + b) - 1 + \xi_i \geq 0,\ \xi_i \geq 0 \text{ für alle } i \in \{1, ..., m\} \end{aligned} \tag{3}$$

für einen Regularisierungsparameter $C \in \mathbb{R}$.

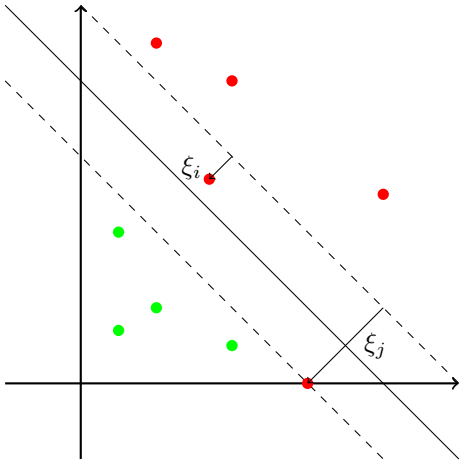

Abbildung 2: Soft-margin SVM erlaubt auch Punkte auf der falschen Seite der Hyperebene

3 Support Vector Machine als empirische Risikominimierer

Wir haben nun die geometrische Interpretation der Support Vector Machine kennengelernt. Es gibt jedoch eine weitere Möglichkeit die SVM zu interpretieren. Der Zugang über die statistische Lerntheorie führt zu einer äquivalenten Formulierung.

3.1 Einführung in die empirische Risikominimierung

Um das Risiko einer Entscheidungsfunktion zu messen benötigen wir eine Verlustfunktion.

Definition 3.1.1 (Verlustfunktion)**.** Eine Verlustfunktion ist eine Abbildung $L : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$, wobei $\mathcal{H}$ die Menge aller Entscheidungsfunktionen ist.

Eine Verlustfunktion soll messen wie richtig oder falsch eine Entscheidungsfunktion $f \in \mathcal{H}$ einen Punkt $x \in \mathcal{X}$ mit zugehöriger Klasse $y \in \mathcal{Y}$ klassifiziert. Typische Beispiele für Verlustfunktionen sind die **0-1-Verlustfunktion** $L(f, x, y) = \mathbb{1}_{\{f(x) \neq y\}}$, welche allen Misklassifikationen den Fehler-Wert 1 zuweist, oder der **quadratische Verlust** $L(f, x, y) = (f(x) - y)^2$, welcher größer ist, je weiter die Entscheidung $f(x) \in \mathcal{Y}$ von der tatsächlichen Klasse $y \in \mathcal{Y}$ abweicht.

Definition 3.1.2 (Risiko)**.** Das Risiko einer Entscheidungsfunktion $f \in \mathcal{H}$ ist definiert als

$$\mathcal{R}_L \coloneqq \mathbb{E}_{(X,Y)\sim\mathcal{D}}[L(f, X, Y)].$$

Hierbei bezeichnen X und Y Zufallsvariablen mit der zugrundeliegenden Wahrscheinlichkeitsverteilung $\mathcal{D}$ und $\mathbb{E}$ den Erwartungswert.

Im folgenden gilt die Konvention, dass falls der Subindex nicht genannt wird, so meint $\mathcal{R} := \mathcal{R}_L$ wobei L in diesem Fall der 0-1-Verlust ist.

Ziel ist es f so zu wählen, dass das Risiko, also der erwartete Verlust möglichst gering wird.

Definition 3.1.3 (Bayes Risiko)**.** Gegeben einer Verteilung $\mathcal{D}$ auf $\mathcal{X} \times \mathcal{Y}$, so ist das **Bayes Risiko** (bezüglich einer Verlustfunktion L) definiert als das Infimum der Risiken aller messbaren Funktionen $h : \mathcal{X} \to \mathcal{Y}$

$$\mathcal{R}_L^* = \inf_{h:\mathcal{X}\to\mathcal{Y}} \mathcal{R}_L(h).$$

Eine Funktion h mit $\mathcal{R}_L(h) = \mathcal{R}_L^*$ heißt **Bayes Klassifikator**.

Nach Funktionen zu suchen, dessen Risiko möglichst gering ist, ist in gewisser Weise natürlich. Die Verteilung welcher die Datenpunkte und ihre Klassen unterliegen, ist jedoch unbekannt. Anstatt des Risikos nutzen wir deshalb das empirische Risiko.

Definition 3.1.4 (Empirisches Risiko)**.** Das empirische Risiko einer Entscheidungsfunktion $f \in \mathcal{H}$ ist definiert als

$$\mathcal{R}_{m,L}(f) = \frac{1}{m} \sum_{i=1}^{m} L(f, x_i, y_i)$$

für eine Menge von Trainingsbeispielen $\{x_1, ..., x_m\} \subseteq \mathcal{X}$ und zugehörigen Klassen $\{y_1, ..., y_m\}$ mit $y_i \in \mathcal{Y}$, für alle $i \in \{1, ..., m\}$.

Nach dem Gesetz der großen Zahlen aus der Stochastik konvergiert das empirische Risiko gegen das tatsächliche Risiko.

Allerdings haben Funktionen mit niedrigem empirischen Risiko nicht automatisch auch ein geringes Risiko. Definieren wir zum Beispiel eine Funktion f mit $D = \{(x_1, y_1), ..., (x_m, y_m)\} \subseteq (\mathcal{X} \times \mathcal{Y})^m$, $m \in \mathbb{N}$, durch

$$f(x) = \begin{cases} y_i & \text{, falls } (x_i, y_i) \in D \\ 0 & \text{, sonst.} \end{cases}$$

So gilt für das empirische Risko $\mathcal{R}_{m,L}(f) = 0$ für die Menge D. f klasifiziert alle Datenpunkte aus D korrekt. Für $(x, y) \in (\mathcal{X} \times \mathcal{Y}) \setminus D$ macht f jedoch Vorhersagen unabhängig von der Verteilung. Warum die Support Vector Machine dennoch mit hoher Wahrscheinlichkeit gute Vorhersagen liefert wird in Kapitel 5 bearbeitet.

3.2 Interpretation als empirische Risikominimierer

Wir wollen nun die Äquivalenz des geometrischen Minimierungsproblems (3) und einer Formulierung als empirische Risikominimierer zeigen. Dazu definieren wir uns eine passende Verlustfunktion.

Definition 3.2.1 (Hinge Loss)**.** Die Hinge Loss Verlustfunktion ist definiert als

$$\begin{aligned} L_{hinge}(f, x, y) : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} &\to \mathbb{R}_+, \\ (f, x, y) &\mapsto (1 - yf(x))_+ := \max\{0, 1 - yf(x)\}. \end{aligned}$$

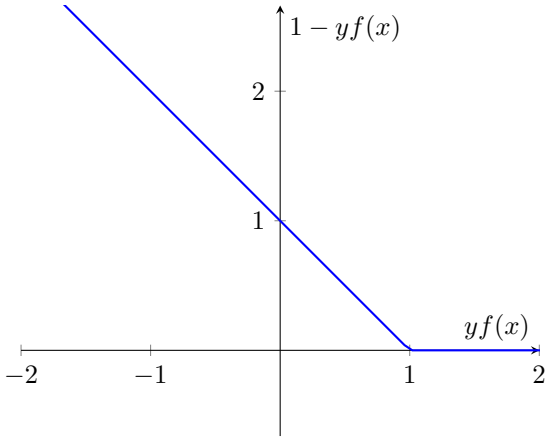

Abbildung 3: Plot von $L_{hinge}(f, x, y)$

Betrachten wir nun erneut das Optimierungsproblem (3).

$$\begin{aligned} &\min_{w,\xi \in \mathbb{R}^d, b \in \mathbb{R}} \|w\|_2^2 + \frac{C}{m} \sum_{i=1}^{m} \xi_i \\ &\text{so dass } y_i(w^T x_i + b) - 1 + \xi_i \geq 0,\ \xi_i \geq 0 \text{ für alle } i \in \{1, ..., m\}. \end{aligned}$$

Wir lösen die Nebenbedingungen für alle $i \in \{1, ..., m\}$ nach ξ_i auf und erhalten

$$\begin{aligned} 1 - (y_i(w^T x_i + b)) &\leq \xi_i \\ 0 &\leq \xi_i. \end{aligned}$$

Es folgt also, dass $\sum_{i=1}^{m} \xi_i$ minimal wird, wenn

$$\xi_i = \max\{0, 1 - (y_i(w^T x_i + b))\}.$$

Das Minimierungsproblem (3) lässt sich also äquivalent beschreiben als

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \|w\|_2^2 + \frac{C}{m} \sum_{i=1}^{m} \max\{0, 1 - (y_i(w^T x_i + b))\}.$$

Mit $\lambda := \frac{1}{C}$ schreiben wir das Minimierungsproblem um, zu

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \lambda\|w\|_2^2 + \frac{1}{m} \sum_{i=1}^{m} \max\{0, 1 - (y_i(w^T x_i + b))\}.$$

Definiere $\mathcal{H} = \{h : \mathbb{R}^d \to \mathbb{R} \mid h(x) = w^T x + b, w \in \mathbb{R}^d, b \in \mathbb{R}\}$.

$$\min_{h \in \mathcal{H}} \lambda\|w\|_2^2 + \frac{1}{m} \sum_{i=1}^{m} \max\{0, 1 - (y_i h_{w,b}(x_i))\}.$$

Wir erkennen hier die Hinge Loss Verlustfunktion aus Definition 3.2.1 wieder und erhalten somit das empirische Risiko.

$$\min_{h \in \mathcal{H}} \lambda\|w\|_2^2 + \mathcal{R}_{n, L_{hinge}}(h). \tag{4}$$

Das Optimierungsproblem ist nun eines der empirischen Risikominimierung welches durch die Norm von w regularisiert wird. Die gefundene SVM-Klassifizierungsfunktion ist die gleiche wie die durch (3) gefundene. Die Äquivalenz der Minimierungsprobleme ist gezeigt. Das vorliegende Minimierungsproblem lässt sich durch das Verfahren der Lagrange-Multiplikatoren lösen. Entschcheidend ist hier, dass in der Lösung die Datenpunkte $\{x_1, ..., x_m\} \subseteq \mathcal{X}$ nur in Form von Skalarprodukten $\langle x_i, x_j \rangle_{\mathbb{R}^n}$, $i, j \in \{1, ..., m\}$ auftreten ([1], Kapitel 15.2). Somit können wir im nächsten Kapitel den sogenannten Kernel Trick anwenden um auch nicht lineare Entscheidungsfunktionen zuzulassen.

4 Kernel Trick

Wir haben bisher lineare Entscheidungsfunktionen betrachtet. Nun wollen wir auch nicht lineare Entscheidungsfunktionen zulassen. Wir tun dies indem wir den sogenannten Kernel Trick anwenden. Die Idee ist, dass wir die Daten in einen höher-dimensionalen Raum mit einem Skalarprodukt abbilden und in diesem eine Entscheidungsfunktion ermitteln.

Definition 4.0.1 (Kernel, Feature Map, Feature Space)**.** Sei $\mathcal{X} \neq \varnothing$ eine nicht leere Menge. Eine Funktion $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ heißt **Kernel**, falls ein $\mathbb{R}$-Hilbertraum H und eine Abbildung $\Phi : \mathcal{X} \to H$ so existieren, dass für alle $x, x' \in \mathcal{X}$ gilt

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

Wir nennen Φ **Feature Map** (dt. Merkmalsabbildung) und H **Feature Space** (dt. Merkmalsraum) von k.

Definition 4.0.2 (positive Definitheit, Symmetrie)**.** Eine Funktion $k : \mathcal{X} \times \mathcal{X}$ heißt **positiv definit**, falls für alle $m \in \mathbb{N}$, $\alpha_1, ..., \alpha_m \in \mathbb{R}$ und alle $x_1, ..., x_m \in \mathcal{X}$ gilt

$$\sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j k(x_j, x_i) \geq 0. \tag{5}$$

Darüber hinaus heißt k **strikt positiv definit**, falls für paarweise verschiedene $x_1, ..., x_m \in \mathcal{X}$ Gleichheit in 5 nur für $\alpha_1 = ... = \alpha_m = 0$ gilt.
Abschließend heißt k **symmetrisch**, falls $k(x, x') = k(x', x)$ für alle $x, x' \in \mathcal{X}$.

Für feste $x_1, ..., x_m \in \mathcal{X}$ heißt die $m \times m$ Matrix

$$K := (k(x_j, x_i))_{i,j=1,...,m}.$$

Gram Matrix.

Bemerkung 4.0.3. *Wir bemerken, dass (5) äquivalent dazu ist, dass die Gram Matrix positiv semi-definit ist.*

Definition 4.0.4 (Hilbertraum mit reproduzierendem Kern, Reproducing kernel hilbert space)**.** Sei $\mathcal{X} \neq \varnothing$ und $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ ein Hilbertraum bestehend aus Funktionen die von $\mathcal{X}$ nach $\mathbb{R}$ abbilden. Eine Funktion $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ heißt **reproduzierender Kern (von $\mathcal{H}$)**, falls

i) $k(\cdot, x) \in \mathcal{H}$ für alle $x \in \mathcal{X}$,

ii) für alle $x \in \mathcal{X}$ und $f \in \mathcal{H}$ gilt $f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{H}}$. (Reproduktionseigenschaft)

$\mathcal{H}$ heißt dann **Hilbertraum mit reproduzierendem Kern** (engl.: reproducing kernel hilbert space) im folgenden genannt **RKHS**.

4.1 Charakterisierung von Kerneln

Es stellt sich nun die Frage, welche Funktionen Kernel sind. Darüber gibt der Satz in folgendem Abschnitt Aufschluss.

Lemma 4.1.1 (Cauchy-Schwarz-Ungleichung für positive symmetrische Bilinearformen)**.** *Sei E ein $\mathbb{R}$-Vektorraum und $\langle \cdot, \cdot \rangle : E \to \mathbb{R}$ eine positive, symmetrische Bilinearform, dann gilt*

$$|\langle x, y \rangle|^2 \leq \langle x, x \rangle \langle y, y \rangle.$$

Beweis. Seien $x, y \in E$. Wir bemerken, dass

$$0 \leq \langle x + \alpha y, x + \alpha y \rangle$$

für alle $\alpha \in \mathbb{R}$.
Fall 1: $\langle x, x \rangle = \langle y, y \rangle = 0$
Es ist zu zeigen, dass $|\langle x, y \rangle|^2$
Betrachte $\alpha = 1$:

$$0 \leq \langle x + y, x + y \rangle = \underbrace{\langle x, x \rangle}_{=0} + 2\langle x, y \rangle + \underbrace{\langle y, y \rangle}_{=0} = 2\langle x, y \rangle.$$

Betrachte $\alpha = -1$:

$$0 \leq \langle x - y, x - y \rangle = \underbrace{\langle x, x \rangle}_{=0} - 2\langle x, y \rangle + \underbrace{\langle y, y \rangle}_{=0} = -2\langle x, y \rangle.$$

Insbesondere gilt $0 \leq \langle x, y \rangle$ und $0 \leq -\langle x, y \rangle$ woraus folgt, dass $\langle x, y \rangle = 0$ und insbesondere $|\langle x, y \rangle|^2 = 0$.
Fall 2: $\langle y, y \rangle \neq 0$
Betrachte $\alpha := -\frac{\langle x, y \rangle}{\langle y, y \rangle}$

$$0 \leq \langle x + \alpha y, x + \alpha y \rangle = \langle x - \frac{\langle x, y \rangle}{\langle y, y \rangle} y, x - \frac{\langle x, y \rangle}{\langle y, y \rangle} y \rangle = \langle x, x \rangle - 2\frac{\langle x, y \rangle^2}{\langle y, y \rangle} + \frac{\langle x, y \rangle^2}{\langle y, y \rangle} = \langle x, x \rangle - \frac{\langle x, y \rangle^2}{\langle y, y \rangle}.$$

Also folgt

$$\langle x, y \rangle^2 \leq \langle x, x \rangle \langle y, y \rangle.$$ □

Nun folgt das Hauptresultat dieses Kapitels.

Satz 4.1.2 (Symmetrisch, positiv semi-definite Funktionen sind Kernel (Steinwart und Christmann, 2008, [3] Thm. 4.16))**.** *Eine Funktion $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ ist Kernel genau dann, wenn k symmetrisch und positiv semi-definit ist.*

Beweis. Sei k zunächst ein Kernel mit Feature Map $\Phi : \mathcal{X} \to H$ und Feature Space H. Dann ist k symmetrisch, da das Skalarprodukt in H symmetrisch ist.

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

Außerdem ist k positiv definit, denn für $n \in \mathbb{N}, \alpha_1, ..., \alpha_n \in \mathbb{R}, x_1, ..., x_n \in \mathcal{X}$ gilt

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \langle \Phi(x_i), \Phi(x_j) \rangle = \left\langle \sum_{i=1}^{n} \alpha_i \Phi(x_i), \sum_{j=1}^{n} \alpha_j \Phi(x_j) \right\rangle \geq 0.$$

Sei nun umgekehrt k symmetrisch und positiv definit. Wir definieren zunächst

$$H_{\text{pre}} := \left\{ \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) \;\middle|\; n \in \mathbb{N}, \alpha_1, ..., \alpha_n \in \mathbb{R}, x_1, ..., x_n \in \mathcal{X} \right\}.$$

Dies ist ein Vektorraum. Für $f := \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) \in H_{\text{pre}}$ und $g := \sum_{j=1}^{m} \beta_j k(\cdot, x'_j) \in H_{\text{pre}}$ definieren wir

$$\langle f, g \rangle_{H_{\text{pre}}} := \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(x_i, x'_j).$$

Wir bemerken, dass diese Definition unabhängig von der Darstellung von f ist, denn

$$\langle f, g \rangle_{H_{\text{pre}}} = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(x_i, x'_j) \overset{k \text{ symmetrisch}}{=} \sum_{j=1}^{m} \beta_j \sum_{i=1}^{n} \alpha_i k(x'_j, x_i) = \sum_{j=1}^{m} \beta_j f(x'_j).$$

Außerdem ist $\langle \cdot, \cdot \rangle_{H_{\text{pre}}}$ auch von der Repräsentation des zweiten Arguments unabhängig

$$\langle f, g \rangle_{H_{\text{pre}}} = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(x_i, x'_j) = \sum_{i=1}^{n} \alpha_i \sum_{j=1}^{m} \beta_j k(x_i, x'_j) = \sum_{i=1}^{n} \alpha_i g(x_i).$$

Um zu zeigen, dass $\langle \cdot, \cdot \rangle_{H_{\text{pre}}}$ tatsächlich ein Skalarprodukt definiert müssen wir drei Eigenschaften zeigen.
(i) Symmetrie:

$$\langle f, g \rangle_{H_{\text{pre}}} = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(x_i, x'_j) \overset{k \text{ symmetrisch}}{=} \sum_{j=1}^{m} \sum_{i=1}^{n} \beta_j \alpha_i k(x'_j, x_i) = \langle g, f \rangle_{H_{\text{pre}}}.$$

(ii) Bilinearität: Seien $f_1, f_2 \in H_{\text{pre}}$. Wie bereits gesehen, gilt

$$\langle f, g \rangle_{H_{\text{pre}}} = \sum_{j=1}^{m} \beta_j f(x'_j).$$

Also

$$\begin{aligned} \langle f_1 + f_2, g \rangle_{H_{\text{pre}}} &= \sum_{j=1}^{m} \beta_j (f_1 + f_2)(x'_j) = \sum_{j=1}^{m} \beta_j (f_1(x'_j) + f_2(x'_j)) \\ &= \sum_{j=1}^{m} \beta_j f_1(x'_j) + \sum_{j=1}^{m} \beta_j f_2(x'_j) = \langle f_1, g \rangle_{H_{\text{pre}}} + \langle f_2, g \rangle_{H_{\text{pre}}}. \end{aligned}$$

Ebenso gilt für $\lambda \in \mathbb{R}$, dass

$$\lambda \langle f, g \rangle_{H_{\text{pre}}} = \lambda \sum_{j=1}^{m} \beta_j f(x'_j) = \sum_{j=1}^{m} \beta_j (\lambda f(x'_j)) = \langle \lambda f, g \rangle_{H_{\text{pre}}}.$$

(iii) Positive Definitheit: Aus der positiven Definitheit von k folgt

$$\langle f, f \rangle_{H_{\text{pre}}} = \sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) = \sum_{i,j=1}^{n} \alpha_i \alpha_j (K)_{ij} \geq 0.$$

Also ist $\langle \cdot, \cdot \rangle_{H_{\text{pre}}}$ positiv semidefinit.
Im vorherigen Lemma 4.1.1 haben wir gesehen, dass $\langle \cdot, \cdot \rangle_{H_{\text{pre}}}$ die Cauchy-Schwarz-Ungleichunge erfüllt, das heißt

$$|\langle f, g \rangle_{H_{\text{pre}}}|^2 \leq \langle f, f \rangle_{H_{\text{pre}}} \langle g, g \rangle_{H_{\text{pre}}} \text{ für alle } f, g \in H_{\text{pre}}.$$

Sei nun $f := \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) \in H_{\text{pre}}$ mit $\langle f, f \rangle_{H_{\text{pre}}} = 0$. Dann gilt für alle $x \in \mathcal{X}$

$$|f(x)|^2 = \left| \sum_{i=1}^{n} \alpha_i k(x, x_i) \right|^2 = \left| \langle f, k(\cdot, x) \rangle_{H_{\text{pre}}} \right|^2 \leq \underbrace{\langle f, f \rangle_{H_{\text{pre}}}}_{=0} \langle k(\cdot, x), k(\cdot, x) \rangle_{H_{\text{pre}}} = 0.$$

Also ist $f \equiv 0$ die Nullfunktion, folglich ist $\langle \cdot, \cdot \rangle_{H_{\text{pre}}}$ strikt postiv definit und definiert ein Skalarprodukt auf H_{pre}.
$\langle f, k(\cdot, x) \rangle_{H_{\text{pre}}} = \sum_{i=1}^{n} \alpha_i k(x, x_i) = f(x)$ ist die in Definition 4.0.4 genannte Reproduktionseigenschaft eines Hilbertraums mit reproduzierendem Kern. Es gilt außerdem die Eigenschaft i), dass $k(\cdot, c) \in H_{\text{pre}}$ nach der Definition von H_{pre}.
Sei nun H die Vervollständigung von H_{pre}, d.h.

$$H := \overline{H_{\text{pre}}} = \overline{\left\{ \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) \;\middle|\; n \in \mathbb{N}, \alpha_1, ..., \alpha_n \in \mathbb{R}, x_1, ..., x_n \right\}}$$

sodass H alle Grenzwerte von Folgen enthält die in der Norm $\|f\|_{H_{\text{pre}}} = \langle f, f \rangle_{H_{\text{pre}}}^{1/2}$ konvergieren. Sei $I : H_{\text{pre}} \to H$ die isometrische Einbettung, dann ist H ein Hilbertraum. Darüber hinaus sogar Hilbertraum mit reproduzierendem Kern, wie wir oben bereits bemerkt haben. Es gilt außerdem

$$\langle Ik(\cdot, x), Ik(\cdot, x') \rangle_H = \langle k(\cdot, x), k(\cdot, x') \rangle_{H_{\text{pre}}} = k(x, x')$$

für alle $x, x' \in \mathcal{X}$. Das heißt $x \mapsto Ik(\cdot, x),\ x \in \mathcal{X}$ definiert eine Feature Map von k. □

4.2 Rechenregeln und Beispiele für Kernel

Im folgenden werden einige Rechenregeln für Kernel mithilfe des vorherigen Satzes gezeigt. Diese Rechenregeln liefern konkrete Beispiele für Kernelfunktionen. Die Folgenden Sätze folgen den Beweisen aus *Foundations of Machine Learning* [7]. Insbesondere Lemma 6.9 und Theorem 6.10 werden ausgeführt und erweitert.

Satz 4.2.1 (Kernel sind abgeschlossen unter Summation)**.** *Seien k, k' Kernel, dann ist auch $k + k'$ ein Kernel.*

Beweis. Seien $K, K' \in \mathbb{R}^{m \times m}$ die zu k und k' gehörigen Gram-Matrizen. Nach Bemerkung 4.0.3 sind K und K' positiv definit. Für beliebiges $c \in \mathbb{R}^m$ gilt dann

$$c^T (K + K') c = \underbrace{c^T K c}_{\geq 0} + \underbrace{c^T K' c}_{\geq 0} \geq 0.$$

Außerdem gilt

$$(k + k')(x, x') = k(x, x') + k'(x, x') = k(x', x) + k(x', x) = (k + k')(x', x) \text{ für alle } x, x' \in \mathcal{X}.$$

Also ist $k + k'$ symmetrisch und positiv definit und somit nach Satz 4.1.2 ein Kernel. □

Definition 4.2.2. Für $c > 0$ und $n \in \mathbb{N}$ heißt die Funktion

$$k : \mathcal{X} \times \mathcal{X}, k(x, x') \mapsto (\langle x, x' \rangle_{\mathbb{R}^m} + c)^n$$

Polynomieller Kernel vom Grad n. Mit Satz 4.2.1 ist k ein Kernel.

Satz 4.2.3 (Kernel sind abgeschlossen unter Produktbildung)**.** *Seien k, k' Kernel, dann ist auch das Produkt kk' ein Kernel.*

Beweis. Seien $K, K' \in \mathbb{R}^{m \times m}$ die zu k und k' gehörigen Gram-Matrizen. Nach Bemerkung 4.0.3 sind K und K' positiv definit. Für die symmetrische und positiv definite Matrix existiert nach der Cholesky-Zerlegung eine Matrix $M \in \mathbb{R}^{m \times m}$ mit $K = MM^T$. Die Kernel Matrix assoziiert mit KK' ist $(K_{ij}K'_{ij})$ Für jedes $c \in \mathbb{R}^m$ gilt

$$\begin{aligned}\sum_{i,j=1}^{m} c_i c_j (K_{ij}K'_{ij}) &= \sum_{i,j=1}^{m} c_i c_j ([\sum_{k=1}^{m} M_{ij}M_{jk}]K'_{ij}) \\ &= \sum_{k=1}^{m} [\sum_{i,j=1}^{m} c_i c_j M_{ik} M_{jk} K'_{ij}] \\ &= \sum_{k=1}^{m} z_k^T K' z_k \geq 0\end{aligned}$$

mit $z_k = (c_1 M_{1k}, ..., c_m M_{mk})^T \in \mathbb{R}^m$. Außerdem gilt

$$(kk')(x, x') = k(x, x')k'(x, x') = k(x', x)k(x', x) = (kk')(x', x) \text{ für alle } x, x' \in \mathcal{X}.$$

Also ist kk' symmetrisch und positiv definit und somit nach 4.1.2 ein Kernel. □

Satz 4.2.4. *Sei $(k_n)_{n \in \mathbb{N}}$ eine Folge von Kerneln mit punktweisem Grenzwert k. Dann ist auch k ein Kernel.*

Beweis. Sei K die Gram-Matrix assoziiert mit k und K_n die Gram-Matrix von k_n für alle $n \in \mathbb{N}$. Nach Bemerkung 4.0.3 gilt für alle $n \in \mathbb{N}$ und $c \in \mathbb{R}^m$

$$c^T K_n c \geq 0.$$

Also auch

$$\lim_{n \to \infty} c^T K_n c = c^T K c \geq 0.$$

Außerdem gilt

$$k(x, x') = \lim_{n \to \infty} k_n(x, x') = \lim_{n \to \infty} k_n(x', x) = k(x', x) \text{ für alle } x, x' \in \mathcal{X}.$$

Also ist k symmetrisch und positiv definit und somit nach 4.0.3 ein gültiger Kernel. □

Korollar 4.2.5. *Sei k ein Kernel mit $|k(x,x')| \leq \rho$ für alle $x, x' \in \mathcal{X}$ und $f(x) := \sum_{i=0}^{\infty} a_i x^i$, $a_i \geq 0$ eine Potenzreihe mit Konvergenzradius $\rho > 0$. Dann ist auch die Komposition $f \circ k$ ein Kernel.*

Beweis. Für jedes $n \in \mathbb{N}$ ist nach 4.2.3 auch k^n ein Kernel. Da $a_n \geq 0$ ist auch $a_n k^n$ ein Kernel für jedes $n \in \mathbb{N}$. Mit 4.2.1 ist auch $\sum_{i=0}^{n} a_i k^i$ ein Kernel für alle $n \in \mathbb{N}$. Durch Grenzwertbildung $n \to \infty$ ist nach 4.2.4 auch $\sum_{i=0}^{\infty} a_i k^i$ ein Kernel. □

Korollar 4.2.6. *Für alle $\sigma^2 > 0$ ist $k : (x, x') \mapsto exp(\frac{\langle x, x' \rangle_{\mathbb{R}^m}}{\sigma^2})$ ein Kernel.*

Beweis. Dies folgt mit Korollar 4.2.5 direkt aus der Darstellung der Exponentialfunktion als Potenzreihe mit Konvergenzradius $\rho = \infty$ und der Tatsache, dass $\langle x, x' \rangle_{\mathbb{R}^m}$ ein Kernel mit Feature Map $\phi(x) = x$ ist. □

Lemma 4.2.7 (Cauchy-Schwarz-Ungleichung für Kernel)**.** *Sei k ein Kernel. Dann gilt für alle $x, x' \in \mathcal{X}$*

$$k(x,x')^2 \leq k(x,x)k(x',x').$$

Beweis. Seien $x, x' \in \mathcal{X}$. Betrachte die Gram-Matrix $K = \left(\begin{smallmatrix} k(x,x) & k(x,x') \\ k(x',x) & k(x',x') \end{smallmatrix}\right)$ bezüglich dieser Elemente. Nach Bemerkung 4.0.3 ist K symmetrisch und positiv definit. Insbesondere ist das Produkt der Eigenwerte, also insbesondere $\det(K) \geq 0$. Durch die Symmetrie $k(x,x') = k(x',x)$ gilt

$$\det(K) = k(x,x)k(x',x') - k(x,x')^2 \geq 0$$

□

Lemma 4.2.8. *Sei k ein Kernel mit Feature Map Φ und Feature Space H. Dann ist der **normalisierte Kernel** k' definiert durch*

$$k'(x,x') = \begin{cases} 0 & k(x,x) = 0 \vee k(x',x') = 0 \\ \frac{k(x,x')}{\sqrt{k(x,x)k(x',x')}} & sonst \end{cases}$$

ebenfalls ein Kernel.

Beweis. Seien $\{x_1, ..., x_m\} \subseteq \mathcal{X}$ und $c \in \mathbb{R}^m$ beliebig. Wir zeigen, dass die Summe $\sum_{i,j=1}^{m} c_i c_j k'(x_i, x_j)$ nicht negativ ist. Wegen der Cauchy-Schwarz Ungleichung für Kernel 4.2.7 gilt, falls $k(x_i, x_i) = 0$, so $k(x_i, x_j) = 0$. Also nach Definition auch $k'(x_i, x_j) = 0$ für alle $j = 1, ..., m$. Wir können also $k(x_i, x_i) > 0$ annehmen für alle $i = 1, ..., m$. Die Summe kann also umgeschrieben werden zu

$$\begin{aligned} \sum_{i,j=1}^{m} c_i c_j k'(x_i, x_j) &= \sum_{i,j=1}^{m} \frac{c_i c_j k(x_i, x_j)}{\sqrt{k(x_i, x_i)k(x_j, x_j)}} = \sum_{i,j=1}^{m} \frac{c_i c_j \langle \Phi(x_i), \Phi(x_j) \rangle_H}{\|\Phi(x_i)\|_H \|\Phi(x_j)\|_H} \\ &= \left\langle \sum_{i=1}^{m} \frac{c_i \Phi(x_i)}{\|\Phi(x_i)\|_H}, \sum_{j=1}^{m} \frac{c_j \Phi(x_j)}{\|\Phi(x_j)\|_H} \right\rangle_H = \left\| \sum_{i=1}^{m} \frac{c_i \Phi(x_i)}{\|\Phi(x_i)\|_H} \right\|_H^2. \end{aligned}$$

Darüber hinaus ist k' symmetrisch, da k symmetrisch ist. Also ist auch k' ein Kernel. □

Korollar 4.2.9. *Für $\sigma^2 > 0$ ist der* ***Gauss Kernel*** $k(x, x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right)$ *der normalisierte Kernel von* $k'(x, x') = \exp\left(\frac{\langle x,x'\rangle}{\sigma^2}\right)$.

Beweis. Seien $x, x' \in \mathcal{X}$.

$$\frac{k'(x,x')}{\sqrt{k'(x,x)k'(x',x')}} = \frac{\exp\left(\frac{\langle x,x'\rangle}{\sigma^2}\right)}{\exp\left(\frac{\|x\|^2}{2\sigma^2}\right)\exp\left(\frac{\|x'\|^2}{2\sigma^2}\right)} = \exp\left(\frac{\langle x,x'\rangle}{\sigma^2} - \frac{\|x\|^2}{2\sigma^2} - \frac{\|x'\|^2}{2\sigma^2}\right) = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right).$$

□

4.3 Repräsentations Theorem

Der folgende Abschnitt liefert eine Aussage darüber, dass die von der SVM gefundene Entscheidungsfunktion eine Linearkombination der Kernelfunktionen, ausgewertet in den Datenpunkten x_i ist.

Satz 4.3.1 (Repräsentations Theorem (Trabs, 2019, [9] Thm. 3.27))**.** *Sei W ein RKHS in Bezug auf einen Kern $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Weiterhin seien $\Phi : \mathbb{R} \to \mathbb{R}$ streng monoton steigend und $G : \mathbb{R}^n \to \mathbb{R}$ beliebig. Dann hat für beliebige $\{x_1, ..., x_n\} \in \mathcal{X}$, jede Lösung des Minimierungsproblems*

$$G(f(x_1), ..., f(x_n)) + \Phi(\|f\|_W) \to \min_{f \in W}! \tag{6}$$

die Form

$$f(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x)$$

mit $\alpha_i \in \mathbb{R}$, für $i = \{1, ..., n\}$.

Beweis. Betrachte $V := \text{span}\{k(x_1, \cdot), ..., k(x_n, \cdot)\}$ und das Orthogonale Komplement $V^\perp = \{u \in W \mid \langle u, v\rangle_W = 0, \text{ für alle } v \in V\}$. Für alle $u \in V^\perp$ gilt wegen der Reproduktionseigenschaft, dass

$$u(x_i) = \langle u, \underbrace{k(x_i, \cdot)}_{\in V}\rangle_W = 0.$$

Daher erhalten wir für ein $f = u + v \in W$ mit $u \in V^\perp$ und $v \in V$, dass für alle $i = 1, ..., n$ gilt

$$f(x_i) = v(x_i) + \underbrace{u(x_i)}_{=0} = v(x_i).$$

Außerdem gilt

$$\begin{aligned}\|f\|_W^2 = \langle f, f\rangle_W = \langle u + v, u + v\rangle_W &= \langle u, u\rangle_W + 2\underbrace{\langle u, v\rangle_W}_{=0} + \langle v, v\rangle_W \\ &= \|u\|_W + \|v\|_W.\end{aligned}$$

Daraus folgt direkt, dass $f = u+v \in W$ nur dann Minimierer ist, wenn $\|u\|_W = 0$, also insbesondere $u = 0$.

Also ist $f \in V$ und somit eine Linearkombination

$$f = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot) \in V. \qquad \square$$

Es folgt das Resultat, dass unter bestimmten Bedingungen Eindeutigkeit gilt.

Korollar 4.3.2 (Eindeutigkeit (Trabs, 2019, [9] Thm. 3.27))**.** *Im Setting von Satz 4.3.1 gilt, falls G konvex und nicht negativ ist, existiert für jedes $\lambda > 0$ eine eindeutige Lösung für das Minimierungsproblem*

$$K(f) := G(f(x_1), ..., f(x_n)) + \lambda \|f\|_W^2 \to \min_{f \in w}!$$

Beweis. Sei G konvex und nicht negativ, dann ist auch K konvex und nicht negativ. Für jedes $f \in W$ mit $\lambda\|f\|_W^2 > G(0, ..., 0)$ gilt, dass $K(f) > K(0)$. Denn, sei $\lambda\|f\|_W^2 > G(0, ..., 0)$, dann

$$\begin{aligned} K(f) &= \underbrace{G(f(x_1), ..., f(x_n))}_{\geq 0} + \underbrace{\lambda\|f\|_W^2}_{> G(0,...,0)} \\ &> G(0, ..., 0) \\ &= K(0). \end{aligned}$$

Falls $(f_n)_{n\in\mathbb{N}} \in W$ eine eine Funktionenfolge ist, mit $\lim_{n\to\infty} K(f_n) = \inf_{f\in W} K(f)$ Angenommen es existiert ein $n \in \mathbb{N}$, mit $\|f_n\|_W > (1/\sqrt{\lambda})\sqrt{G(0, ..., 0)}$. Dann gilt mit obiger Überlegung, dass $K(f_n) > K(0)$. Definiere nun die Folge $(\tilde{f}_n)_{n\in\mathbb{N}} \in W$ mit $\tilde{f}_n = \begin{cases} 0 & \text{, falls } \|f_n\|_W > (1/\sqrt{\lambda})\sqrt{G(0, ..., 0)} \\ f_n & \text{, sonst.} \end{cases}$ Nun gilt $K(f_n) \geq K(\tilde{f}_n)$ für alle $n \in \mathbb{N}$. Also auch $\lim_{n\to\infty} K(\tilde{f}_n) = \inf_{f\in W} K(f)$. Also können wir für alle $n \in \mathbb{N}$ annehmen, dass $\|f_n\|_W \leq (1/\sqrt{\lambda})\sqrt{G(0, ..., 0)}$.
Wenden wir nun die Zerlegung $f_n = u_n + v_n$ mit $u_n \in V^\perp$ und $v_n \in V$ aus Satz 4.3.1 an, so zeigt dies $\lim_{n\to\infty} K(v_n) = \min_{f\in W} K(f)$.
Wir bemerken, dass v_n im kompakten endlichdimensionalen Ball $\{v \in V \mid \|v_n\|_W \leq (1/\sqrt{\lambda})\sqrt{(G(0, ..., 0)}\}$ liegt und jeder Häufungspunkt von v_n das Minimierungsproblem löst.
Beweisen wir nun die Eindeutigkeit.
Angenommen es existieren zwei Lösungen $f_1, f_2 \in W$, dann erfüllt $g = \frac{1}{2}(f_1 + f_2)$

$$\begin{aligned} \|g\|_W^2 = \|\frac{1}{2}(f_1 + f_2)\|_W^2 &= \frac{1}{4}\|f_1 + f_2\|_W^2 = \frac{1}{4}\langle f_1 + f_2, f_1 + f_2\rangle_W \\ &= \frac{1}{4}(\langle f_1, f_1\rangle_W + \langle f_2, f_2\rangle_W + 2\langle f_1, f_2\rangle_W) \\ &= \frac{1}{4}(2\langle f_1, f_1\rangle_W + 2\langle f_2, f_2\rangle_W - \langle f_1, f_1\rangle_W - \langle f_2, f_2\rangle_W + \langle f_1, f_2\rangle_W + \langle f_2, f_1\rangle_W) \\ &= \frac{1}{4}(2\langle f_1, f_1\rangle_W + 2\langle f_2, f_2\rangle_W - \langle f_1 - f_2, f_1 - f_2\rangle_W) \\ &= \frac{1}{4}(2\|f_1\|_W^2 + 2\|f_2\|_W^2 - \|f_1 - f_2\|_W^2) \\ &< \frac{1}{2}(\|f_1\|_W^2 + \|f_2\|_W^2). \end{aligned}$$

Nun gilt mit der Konvexität von G und $\|g\|_W^2 < \frac{1}{2}(\|f_1\|_W^2 + \|f_2\|_W^2)$

$$\begin{aligned} K(g) &= G(g(x_1), ..., g(x_n)) + \lambda\|g\|_W^2 \\ &= G(\frac{1}{2}(f_1 + f_2)(x_1), ..., \frac{1}{2}(f_1 + f_2)(x_n)) + \lambda\|g\|_W^2 \\ &< \frac{1}{2}G(f_1(x_1), ..., f_1(x_n)) + \frac{\lambda}{2}\|f_1\|_W^2 + \frac{1}{2}G(f_2(x_1), ..., f_2(x_n)) + \frac{\lambda}{2}\|f_2\|_W^2 \\ &= \frac{1}{2}K(f_1) + \frac{1}{2}K(f_2). \end{aligned}$$

Ein Widerspruch dazu, dass f_1, f_2 minimierer für K sind. Also ist die Lösung des Minimierungsproblems $\min_{f\in W} K(f)$ eindeutig. □

Die im folgenden definierte Support Vector Machine mit Kernelfunktionen erfüllt die Voraussetzungen des Repräsentationstheorem 4.3.1 und des Satzes über die Eindeutigkeit 4.3.2.

4.4 Support Vector Machine mit Kernel

Definition 4.4.1. Für einen reproduzierenden Kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, den korrespondierenden RKHS H und ein $\lambda > 0$ setze

$$\hat{f}_n^{SVM} := \underset{f\in H, \|f\|_H \leq \lambda}{\arg\min} (\frac{1}{m}\sum_{i=1}^{m}(1 - y_i f(x_i))_+).$$

Der resultierende Klassifizierer $\hat{h}_m^{SVM} := h_{\hat{f}_n^{SVM}} = \operatorname{sgn}(\hat{f}_n^{SVM})$ heißt **SVM Klassifizierer** oder **Support Vector Machine**.

Mit dem Hinge Verlust $\varphi(f, x, y) = (1 - yf(x))_+$ ist die Support Vector Machine ein φ-Risiko-Minimierungsproblem für welches die Menge der möglichen Klassifizierungsfunktionen gegeben ist durch den Ball $\{f \in H : \|f\|_H \leq \lambda\}$ in einem RKHS auf $\mathcal{X}$ mit einem Radius $\lambda > 0$. Nach der Theorie der Langrange Multiplikatoren existiert nun ein $\lambda' > 0$, so dass wir die folgende Repräsentation erhalten

$$\hat{f}_m^{SVM} = \underset{f\in H}{\arg\min}(\mathcal{R}_{m,\varphi})(f) + \lambda'\|f\|_H^2).$$

Diese Darstellung entspricht Gleichung (4) wobei wir nun den Funktionen aus einem anderen Raum betrachten. Nach dem Repräsentationstheorem 4.3.1 angewendet auf $G(f(x_1), ..., f(x_m)) = \mathcal{R}_{m,\varphi}(f)$ kann das vorherige Optimierungsproblem als endlichdimensionales Problem geschrieben werden. Die Lösung muss von der Form $\hat{f}_m^{SVM}(x) = \sum_{i=1}^m \hat{\alpha}_i k(x_i, x)$ für passende $\hat{\alpha} = (\hat{\alpha}_1, ..., \hat{\alpha}_m) \in \mathbb{R}^m$ sein. Es gilt wegen der Reproduktionseigenschaft von H

$$\left\|\sum_{i=1}^{m}\alpha_i k(x_i, \cdot)\right\|_H^2 = \left\langle \sum_{i=1}^{m}\alpha_i k(x_i, \cdot), \sum_{j=1}^{m}\alpha_j k(x_j, \cdot)\right\rangle_H = \sum_{i,j=1}^{m}\alpha_i\alpha_j k(x_i, x_j).$$

Die Koeffizienten $\hat{\alpha}$ von $\hat{h}_m^{SVM}$ sind gegeben durch

$$\hat{\alpha} = \underset{\alpha\in\mathbb{R}^m}{\arg\min}(\frac{1}{m}\sum_{i=1}^{m}(1 - y_i\sum_{j=1}^{m}\alpha_j k(x_j, x_i))_+ + \lambda'\sum_{i,j=1}^{m}\alpha_i\alpha_j k(x_i, x_j)).$$

Es ist zu bemerken, dass der RKHS H in dieser Repräsentation nicht mehr auftaucht.

Die Lösung ist nur abhängig vom Kernel k, dem Parameter λ' und den Datenpunkten $\{(x_1, y_1), ..., (x_m, y_m)\} \subseteq (\mathcal{X} \times \mathcal{Y})^m$. Mit

$$I = \{i = 1, ..., m : y_i \hat{f}_m^{SVM}(x_i) \leq 1\}$$

gilt, dass $(1 - \hat{f}_m^{SVM})_+ = 0$ für alle $i \in I$. Also ist $\hat{h}_m^{SVM}$ auch die Lösung des folgenden Minimierungsproblems.

$$\left(\frac{1}{m} \sum_{i \in I} (1 - y_i f(x_i))_+ + \lambda' \|f\|_H^2 \right) \to \min_{f \in H}!$$

Da das Repräsentationsthorem $\hat{f}_m^{SVM}(x) = \sum_{i \in I} \hat{\alpha} k(x_i, x)$ impliziert, schließen wir, dass $\hat{\alpha}_i = 0$ für alle i mit $y_i \hat{f}_m^{SVM}(x_i) > 1$. Das bedeutet, dass der Datenpunkt (x_i, y_i) „signifikant“ richtig klassifiziert wurde. Die Daten $(x_i)_{i \in I}$ sind dann die Stützvektoren wie in Definition 2.2.3.

5 Statistische Lerntheorie

Im folgenden werden zwei Abschätzungen vorgestellt. Um eine davon später zu deuten, schauen wir uns das Risiko einer Entscheidungsfunktion erneut genauer an. In Definition 3.1.3 hatten wir das Bayes Risiko $\mathcal{R}_L^*$ als das Infimum der Risiken aller messbaren $h : \mathcal{X} \to \mathcal{Y}$ definiert. Sei $h^* : \mathcal{X} \to \mathcal{Y}$ ein Bayes Klassifikator, das heißt $\mathcal{R}_L(h^*) = \mathcal{R}_L^*$. Außerdem sei $\hat{h} \in \mathcal{H}$ ein empirischer Risikominimierer, d.h. $\mathcal{R}_{m,L}(\hat{h}) = \min_{h\in\mathcal{H}} \mathcal{R}_{m,L}(h)$. Dann lässt sich die Differenz der Risiken der beiden Funktionen zerlegen in

$$\mathcal{R}_L(\hat{h}) - \mathcal{R}_L(h^*) = \underbrace{\mathcal{R}_L(\hat{h}) - \inf_{h\in\mathcal{H}} \mathcal{R}_L(h)}_{\text{stochastischer Fehler}} + \underbrace{\inf_{h\in\mathcal{H}} \mathcal{R}_L(h) - \mathcal{R}_L(h^*)}_{\text{Approximationsfehler}} \tag{7}$$

für eine Menge von Klassifikatoren $\mathcal{H}$. Der stochastische Fehler vegleicht hier wie weit das Risiko von $\hat{h}$ von dem geringsten Risiko in $\mathcal{H}$ entfernt ist. Der Approximationsfehler gibt an wie flexibel die Menge $\mathcal{H}$ ist.

Bemerkung 5.0.1. *Sei $\bar{h} \in \mathcal{H}$ mit $\mathcal{R}_L(\bar{h}) = \inf_{h\in\mathcal{H}} \mathcal{R}_L(h)$ der Minimierer von $\mathcal{R}_L(h)$ über $\mathcal{H}$ und $\hat{h}$ weiterhin der empirische Risikominimierer aus $\mathcal{H}$. Der stochastische Fehler lässt sich beschränken durch*

$$\begin{aligned}\mathcal{R}_L(\hat{h}) - \mathcal{R}_L(\bar{h}) &= \mathcal{R}_{m,L}(\hat{h}) - \mathcal{R}_{m,L}(\bar{h}) + \mathcal{R}_L(\hat{h}) - \mathcal{R}_{m,L}(\hat{h}) + \mathcal{R}_{m,L}(\bar{h}) - \mathcal{R}_L(\bar{h}) \\ &\leq 2 \sup_{h\mathcal{H}} |\mathcal{R}_L(h) - \mathcal{R}_{m,L}(h)|.\end{aligned}$$

5.1 Oracle Ungleichung

Das in diesem Abschnitt folgende Theorem gibt eine Schranke für den stochastischen Fehler in Abhängigkeit von dem Parameter $\lambda > 0$ und der Wahl der Kernelfunktion $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. In der Praxis sollte die Wahl von $\lambda > 0$ natürlich den stochastischen Fehler und den Approximationsfehler gleichzeitig gering halten. Die Wahl von λ und k geschieht meist durch ein Verfahren namens Kreuzvalidierung.

Lemma 5.1.1 (Trabs, 2019, [9] Lemma 3.24)**.** *Sei $(H, \langle\cdot,\cdot\rangle_H)$ ein RKHS mit reproduzierendem Kern $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, dann gilt für alle $f \in H$*

$$\|f\|_\infty = \sup_{x\in\mathcal{X}} |f(x)| \leq \|f\|_H \sup_{x\in\mathcal{X}} \sqrt{k(x,x)}.$$

Beweis. Zunächst bemerken wir, dass mit $f := k(\cdot, x) \in H$ für festes $x \in \mathcal{X}$, nach der Reproduktionseigenschaft von H, gilt

$$k(x,y) = f(y) = \langle f, k(\cdot,y)\rangle_H = \langle k(\cdot,x), k(\cdot,y)\rangle_H, \quad \text{für alle } y \in \mathcal{X}.$$

Hierdurch, und durch die Cauchy-Schwarz Ungleichung und die Reproduktionseigenschaft von H gilt, dass

$$\begin{aligned}\sup_{x\in\mathcal{X}} f(x)^2 &= \sup_{x\in\mathcal{X}} \langle f, k(\cdot,x)\rangle_H^2 \leq \|f\|_H^2 \sup_{x\in\mathcal{X}} \|k(\cdot,x)\|_H^2 \\ &= \|f\|_H^2 \sup_{x\in\mathcal{X}} \langle k(\cdot,x), k(\cdot,x)\rangle_H = \|f\|_H^2 \sup_{x\in\mathcal{X}} k(x,x).\end{aligned}$$

Das ziehen der Wurzel auf beiden Seiten der Ungleichung liefert die Behauptung. □

Das folgende Lemma wird für den späteren Beweis der Oracle Ungleichung benötigt.

Lemma 5.1.2. *Mit* $\varphi : \mathbb{R} \to \mathbb{R}_+,\ \varphi(x) = (1+x)_+$ *und* $0 < L < \infty$ *ist*

$$\psi : [-1,1] \to \mathbb{R},\ \psi(u) := \frac{\varphi(Lu) - 1}{L}$$

eine Kontraktion auf $[-1,1]$ *und* $\psi(0) = 0$.

Beweis. Es gilt

$$\psi(0) = \frac{\varphi(0) - 1}{L} = \frac{1-1}{L} = 0.$$

Es bleibt zu zeigen, dass ψ eine Kontraktion ist. Seien dazu $u, v \in [-1,1]$.

$$|\psi(u) - \psi(v)| = \left| \frac{\varphi(Lu) - 1}{L} - \frac{\varphi(Lu) - 1}{L} \right| = \frac{1}{L} |\varphi(Lu) - \varphi(Lv)|.$$

Falls $1 + Lu \geq 0$ *und* $1 + Lv \geq 0$:

$$\frac{1}{L}|\varphi(Lu) - \varphi(Lv)| = \frac{1}{L}|(1+Lu) - (1+Lv)| = \frac{1}{L}|Lu - Lv| = |u - v|.$$

Falls $1 + Lu \geq 0$ *und* $1 + Lv < 0$:
Dann gilt $u \geq -\frac{1}{L}$, $-v > \frac{1}{L}$ und insbesondere

$$\frac{1}{L}|\varphi(Lu) - \varphi(Lv)| = \frac{1}{L}|1 + Lu - 0| = \left| \frac{1}{L} + u \right| = \frac{1}{L} + u \leq -v + u = |u - v|.$$

Falls $1 + Lu < 0$ *und* $1 + Lv \geq 0$:
Dann gilt $-u > \frac{1}{L}$, $v \geq -\frac{1}{L}$ und insbesondere

$$\frac{1}{L}|\varphi(Lu) - \varphi(Lv)| = \frac{1}{L}|0 - (1+Lv)| = \left| -\frac{1}{L} - v \right| = -(-\frac{1}{L} - v) = \frac{1}{L} + v \leq -u + v = |u - v|.$$

Insgesamt gilt somit $|\psi(u) - \psi(v)| \leq |u - v|$ und ψ ist eine Kontraktion. □

Satz 5.1.3 (Oracle Ungleichung(Trabs, 2019, [9] Thm. 3.27))**.** *Sei* k *ein Kernel des Hilbertraums mit reproduzierendem Kern* H *mit* $\sup_{x \in \mathcal{X}} k(x,x) < \infty$. *Dann erfüllt die korrespondierende SVM Klassfizierungsfunktion* $\hat{h}_m^{SVM} : \mathcal{X} \to \{-1, 1\}$ *mit Parameter* $\lambda > 0$ *erfüllt*

$$\mathbb{E}[\mathcal{R}(\hat{h}_m^{SVM})] \leq \inf_{\|f\|_H \leq \lambda} \mathcal{R}_\varphi(f) + \frac{8\lambda}{\sqrt{m}} \mathbb{E}[k(X,X)]^{1/2}.$$

Beweis. Schritt 1: Für $\varphi = (1+x)_+$ gilt

$$\begin{aligned}
\mathcal{R}(\hat{h}_m^{\mathrm{SVM}}) &= \mathbb{P}^{(X,Y)}(\hat{h}_m^{\mathrm{SVM}}(X) \neq Y) = \mathbb{P}^{(X,Y)}(\mathrm{sgn}(\hat{f}_m^{\mathrm{SVM}}) \neq Y) \\
&= \mathbb{P}^{(X,Y)}(-Y \hat{f}_m^{\mathrm{SVM}}(X) > 0) = \mathbb{E}^{(X,Y)}[\mathbb{1}_{\{-Y \hat{f}_m^{\mathrm{SVM}}(X) > 0\}}] \\
&\leq \mathbb{E}^{(X,Y)}[(1 - Y \hat{f}_m^{\mathrm{SVM}}(X))_+] = \mathcal{R}_\varphi(\hat{f}_m^{\mathrm{SVM}}).
\end{aligned}$$

Mit Bemerkung 5.0.1 und der vorherigen Überlegung folgt

$$\begin{aligned}\mathcal{R}(\hat{h}_m^{\mathrm{SVM}}) &\leq \mathcal{R}_\varphi(\hat{f}_m^{\mathrm{SVM}}) - \inf_{\|f\|_H \leq \lambda} \mathcal{R}_\varphi(f) + \inf_{\|f\|_H \leq \lambda} \mathcal{R}_\varphi(f) \\ &\leq 2 \sup_{\|f\|_H \leq \lambda} |\mathcal{R}_{m,\varphi}(f) - \mathcal{R}_\varphi(f)| + \inf_{\|f\|_H \leq \lambda} \mathcal{R}_\varphi(f).\end{aligned}$$

Es bleibt zu zeigen, dass

$$\mathbb{E}[\sup_{\|f\|_H \leq \lambda} |\mathcal{R}_{m,\varphi}(f) - \mathcal{R}_\varphi(f)|] \leq 4\lambda\sqrt{\mathbb{E}[k(X,X)/m]}.$$

Schritt 2: Wir nutzen ein symmetrierungs Vorgehen um das Supremum zu beschränken. Dazu sei $(X_i', Y_i')_{i=1,...,m}$ eine identische Kopie von $(X_i, Y_i)_{i=1,...,m}$ definiert auf dem gleichen Wahrscheinlichkeitsraum (ein sogenanntes ghost Sample). Jensens Ungleichung und $\sup_t \mathbb{E}[Z_t] \leq \mathbb{E}[\sup_t Z_t]$ implizieren

$$\begin{aligned}&\mathbb{E}\left[\sup_{\|f\|_H \leq \lambda} |\mathcal{R}_{m,\varphi}(f) - \mathcal{R}_\varphi(f)|\right] \\ &= \mathbb{E}\left[\sup_{\|f\|_H \leq \lambda} \left|\frac{1}{n}\sum_{i=1}^n (\varphi(-Y_i f(X_i)) - \mathbb{E}[\varphi(-Y_i' f(X_i'))])\right|\right] \\ &\overset{1}{=} \mathbb{E}\left[\sup_{\|f\|_H \leq \lambda} \left|\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (\varphi(-Y_i f(X_i)) - \varphi(-Y_i' f(X_i'))) \;\middle|\; X_1, ..., X_m, Y_1, ..., Y_m,\right]\right|\right] \\ &\overset{2}{\leq} \mathbb{E}\left[\sup_{\|f\|_H \leq \lambda} \mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n (\varphi(-Y_i f(X_i)) - \varphi(-Y_i' f(X_i')))\right| \;\middle|\; X_1, ..., X_m, Y_1, ..., Y_m,\right]\right] \\ &\overset{3}{\leq} \mathbb{E}\left[\mathbb{E}\left[\sup_{\|f\|_H \leq \lambda} \left|\frac{1}{n}\sum_{i=1}^n (\varphi(-Y_i f(X_i)) - \varphi(-Y_i' f(X_i')))\right| \;\middle|\; X_1, ..., X_m, Y_1, ..., Y_m,\right]\right] \\ &\overset{4}{=} \mathbb{E}\left[\sup_{\|f\|_H \leq \lambda} \left|\frac{1}{n}\sum_{i=1}^n (\varphi(-Y_i f(X_i)) - \varphi(-Y_i' f(X_i')))\right|\right].\end{aligned}$$

Darüber hinaus sei $(\varepsilon_i)_{i=1,...,m}$ eine Rademacher Folge, d.h. $\mathbb{P}(\varepsilon_i = 1) = \mathbb{P}(\varepsilon_i = -1) = \frac{1}{2}$, die unabhängig von $(X_i, Y_i)_{i=1,...,m}$ und $(X_i', Y_i')_{i=1,...,m}$. Da die Verteilung von

$$Z_i := (\varphi(-Y_i f(X_i)) - \varphi(-Y_i' f(X_i')))$$

symmetrisch ist , d.h. $Z_i \overset{d}{=} -Z_i$, gilt auch $\varepsilon_i Z_i \overset{d}{=} Z_i$ für $i = 1, ..., m$, denn

$$\mathbb{P}(\varepsilon_i Z_i \in A) = \frac{1}{2}\mathbb{P}(Z_i \in A) + \frac{1}{2}\mathbb{P}(-Z_i \in A) = \mathbb{P}(Z_i \in A) \qquad \text{für alle } A \in \mathcal{B}_\mathbb{R}.$$

[1] unabhängige Zufallsvariablen in bedingtem Erwartungswert
[2] Jensen's Ungleichung mit $|\cdot|$
[3] $\sup \mathbb{E}(\cdot) \leq \mathbb{E}(\sup(\cdot))$
[4] Satz vom totalen Erwartungswert

Wir schließen daraus

$$\begin{aligned}
\mathbb{E}\left[\sup_{\|f\|_H\leq\lambda}|\mathcal{R}_{m,\varphi}(f)-\mathcal{R}_\varphi(f)|\right] &\leq \mathbb{E}\left[\sup_{\|f\|_H\leq\lambda}\left|\frac{1}{m}\sum_{i=1}^m(\varphi(-Y_if(X_i))-\varphi(-Y_i'f(X_i')))\right|\right]\\
&= \mathbb{E}\left[\sup_{\|f\|_H\leq\lambda}\left|\frac{1}{m}\sum_{i=1}^m Z_i\right|\right]\\
&= \mathbb{E}\left[\sup_{\|f\|_H\leq\lambda}\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i Z_i\right|\right]\\
&= \mathbb{E}\left[\sup_{\|f\|_H\leq\lambda}\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i(\varphi(-Y_if(X_i))-\varphi(-Y_i'f(X_i')))\right|\right]\\
&= \mathbb{E}\left[\sup_{\|f\|_H\leq\lambda}\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i(\varphi(-Y_if(X_i))-1+1-\varphi(-Y_i'f(X_i')))\right|\right]\\
&= \mathbb{E}\left[\sup_{\|f\|_H\leq\lambda}\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i(\varphi(-Y_if(X_i))-1)+\varepsilon_i(1-\varphi(-Y_i'f(X_i')))\right|\right]\\
&= \mathbb{E}\left[\sup_{\|f\|_H\leq\lambda}\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i(\varphi(-Y_if(X_i))-1)+\varepsilon_i(\varphi(-Y_i'f(X_i'))-1)\right|\right]\\
&= 2\mathbb{E}\left[\sup_{\|f\|_H\leq\lambda}\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i(\varphi(-Y_if(X_i))-1)\right|\right]
\end{aligned}$$

wobei wir im letzten Schritt genutzt haben, dass (X_i, Y_i) und (X_i', Y_i') die gleiche Verteilung besitzen.

Als nächsten nutzen wir ein Kontraktionsargument von Ledoux und Talagrand (2011, Thm. 4.12)[5]. Falls $\psi : [-1,1] \to \mathbb{R}$ eine Kontraktion ist, d.h. $|\psi(x)-\psi(y)| \leq |x-y|$ und $\psi(0)=0$, dann gilt für jede Familie $\mathcal{G} \subseteq \{g : \mathcal{X}\times\{-1,+1\} \to [-1,1] \text{ messbar}\}$

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i\psi(g(X_i,Y_i))\right|\right] \leq 2\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i g(X_i,Y_i)\right|\right]. \tag{8}$$

Durch nutzen von $\|f\|_H \leq \lambda$ und Lemma 5.1.1 erhalten wir $\|f\|_\infty \leq \lambda \sup_{x\in\mathcal{X}}\sqrt{k(x,x)} =: L < \infty$. Definiere die Funktionen $\psi(u) \coloneqq \varphi(Lu) - 1/L$ und $g(x,y) = -yf(x)/L \in [0,1]$. Dann ist ψ nach Lemma 5.1.2 eine Kontraktion und $\psi(0)=0$ und wir können somit Ungleichung (8) anwenden.

$$\mathbb{E}\left[\sup_{\|f\|_H\leq\lambda}\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i\frac{\varphi(-Y_if(X_i))-1}{L}\right|\right] \leq 2\mathbb{E}\left[\sup_{\|f\|_H\leq\lambda}\left|\frac{1}{m}\sum_{i=1}^m \epsilon_i\frac{Y_if(X_i)}{L}\right|\right].$$

Da $\varepsilon_i \overset{d}{=} \varepsilon_i Y_i$, schließen wir zusammen mir der vorherigen Ungleichung

$$\mathbb{E}\left[\sup_{\|f\|_H\leq\lambda}|\mathcal{R}_{m,\varphi}(f)-\mathcal{R}_\varphi(f)|\right] \leq 4\mathbb{E}\left[\sup_{\|f\|_H\leq\lambda}\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i f(X_i)\right|\right].$$

Schritt 3: Durch nutzen der Hilbertraum Struktur des RKHS H, liefert die Cauchy-Schwarz Ungleichung

$$\begin{aligned}\sup_{\|f\|_H\leq\lambda}\left|\frac{1}{m}\sum_{i=1}^{m}\varepsilon_i f(X_i)\right|^2 &= \sup_{\|f\|_H\leq\lambda}\left|\frac{1}{m}\sum_{i=1}^{m}\varepsilon_i\langle f,k(X_i,\cdot)\rangle_H\right|^2\\ &= \sup_{\|f\|_H\leq\lambda}\left|\langle f,\frac{1}{m}\sum_{i=1}^{m}\varepsilon_i k(X_i,\cdot)\rangle_H\right|^2\\ &\leq \sup_{\|f\|_H\leq\lambda}\|f\|_H^2\left\|\frac{1}{m}\sum_{i=1}^{m}\varepsilon_i k(X_i,\cdot)\right\|_H^2\\ &= \lambda^2\frac{1}{m^2}\sum_{i,j=1}^{m}\varepsilon_i\varepsilon_j k(X_i,X_j).\end{aligned}$$

Wegen $\mathbb{E}[\varepsilon_i\varepsilon_j]=\hat{\delta}_{ij}$ erhalten wir schließlich

$$\begin{aligned}\mathbb{E}\left[\sup_{\|f\|_H\leq\lambda}\left|\frac{1}{m}\sum_{i=1}^{m}\varepsilon_i f(X_i)\right|\right] &\leq \lambda\mathbb{E}\left[\frac{1}{m^2}\sum_{i,j=1}^{m}\varepsilon_i\varepsilon_j k(X_i,X_j)\right]^{1/2}\\ &= \frac{\lambda}{m}\left(\sum_{i=1}^{m}\mathbb{E}[k(X_i,X_i)]\right)^{1/2}\\ &= \frac{\lambda}{\sqrt{m}}\sqrt{\mathbb{E}[k(X,X)]}.\end{aligned}$$

Wobei wir in der letzen Gleichung genutzt haben, dass die X_i für alle $i=1,...,m$ die Gleiche Verteilung X besitzen und unabhängig voneinander sind. Gemeinsam mit den vorherigen Schritten ist die Behauptung bewiesen. □

Das vorherige Theorem sagt also aus, dass sich der Erwartungswert des stochastischen Fehlers beschränken lässt.

5.2 Risiko Schranke

Im folgenden Abschnitt lernen wir eine Schranke für das Risiko einer SVM-Entscheidungsfunktion kenen.

Definition 5.2.1 (Empirische Rademacher Komplexität)**.** Sei $\mathcal{G}$ eine Familie von Funktionen die von $\mathcal{Z}=\mathcal{X}\times\mathcal{Y}$ nach $[a,b]\subseteq\mathbb{R}$ abbilden und $S=(z_1,...,z_m)=((x_1,y_1),...,(x_m,y_m))$ eine feste Stichprobe. Dann ist die **empirische Rademacher Komplexität** von $\mathcal{G}$ in Bezug auf S definiert als

$$\hat{R}_S(\mathcal{G})=\mathbb{E}_\sigma\left[\sup_{g\in\mathcal{G}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i g(z_i)\right]$$

wobei $\sigma=(\sigma_1,...,\sigma_m)^T$ mit unabhängigen, identisch verteilten Zufallsvariablen σ_i mit Werten in $\{-1,1\}$ und $\mathbb{P}(\sigma_i=-1)=\mathbb{P}(\sigma_i=1)=1/2$ für alle $i=1,...,m$.

Definition 5.2.2 (Rademacher Komplexität)**.** Sei $\mathcal{D}$ die Verteilung auf $\mathcal{X}\times\mathcal{Y}$ aus welcher die Stichproben gezogen werden. Für alle $m\in\mathbb{N}$ ist die **Rademacher Komplexität** von $\mathcal{G}$ die

Erwartung der empirischen Rademacher Komplexität über alle Stichproben der Größe m, d.h.

$$R_m(\mathcal{G}) = \mathbb{E}_{S\sim\mathcal{D}^m}[\hat{R}_S(\mathcal{G})].$$

Die Rademacher Komplexität ist ein Maß für die Komplexität einer Familie von Funktionen $\mathcal{H}$. Bezogen auf die Situation der Klassifikation von Daten gibt sie an wie flexibel $\mathcal{H}$ ist für verschiedene Label der Datenpunkte.

Definition 5.2.3 (Margin loss function, Rand-Verlustfunktion)**.** Für jedes $\rho > 0$ ist die **ρ-Rand-Verlustfunktion** $L_\rho : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ definiert für alle $f \in \mathcal{H}, x \in \mathcal{X}, y \in \mathcal{Y}$ durch $L_\rho(f, x, y) = \Phi_\rho(yf(x))$. Wobei

$$\Phi_\rho(z) = \min(1, \max(0, 1 - \frac{z}{\rho})) = \begin{cases} 1, & \text{falls } z \leq 0 \\ 1 - \frac{z}{\rho}, & \text{falls } 0 \leq z \leq \rho \\ 0, & \text{falls } \rho \leq z \end{cases}$$

für alle $z \in \mathbb{R}$.

Die eingeführte Verlustfunktion ist Dargestellt in Abbildung 4. Der Parameter $\rho > 0$ kann interpretiert werden als Konfidenzrand der von einer Hypothese $f \in \mathcal{H}$ gefordert ist. Falls $\text{sgn}(f(x)) \neq y$ so klassifiziert f x falsch und $L_\rho(f, x, y) = 1$. Gilt hingegen $\text{sgn}(f(x)) = y$ so klassifiziert f x richtig. Sei $(x, y) \in \mathcal{X} \times \mathcal{Y}$ ein Datenpunkt und $f \in \mathcal{H}$ eine Entscheidungsfunktion. Falls zusätzlich gilt dass $f(x) > \rho$ so bestraft L_ρ diese Vorhersage nicht. Für Vorhersagen in $[0, \rho)$ so wird diese Vorhersage Linear von L_ρ bestraft. Der empirische Verlust wird analog zu anderen Verlustfunktionen definiert.

Definition 5.2.4 (Empirischer Rand-Verlust)**.** Gegeben einer Stichprobe $S = \{(x_1, y_1), ..., (x_m, y_m)\} \subseteq (\mathcal{X} \times \mathcal{Y})^m$ und einer Entscheidungsfunktion $f \in \mathcal{H}$ so ist der **empirische Rand-Verlust** definiert durch

$$\mathcal{R}_{m,\rho}(f) = \frac{1}{m} \sum_{i=1}^{m} L_\rho(f, x_i, y_i).$$

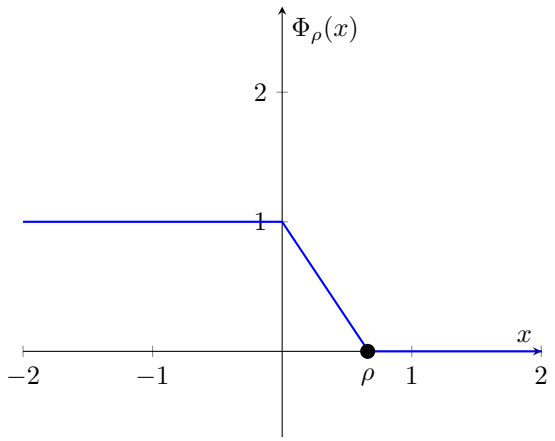

Abbildung 4: Plot von $\Phi_\rho(x)$

Wir bemerken, dass für jedes $i \in \{1, ..., m\}$ gilt

$$L_\rho(f, x_i, y_i) = \Phi_\rho(y_i f(x_i)) \leq \mathbb{1}_{\{y_i f(x_i) \leq \rho\}}.$$

Daher kann auch der empirische Rand-Verlust beschränkt werden durch

$$\mathcal{R}_{m,\rho}(f) \leq \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{\{y_i f(x_i) \leq \rho\}}.$$

Dieser Wert gibt den Anteil der Trainingspunkte vom gesamten Datenset $S \subseteq (\mathcal{X} \times \mathcal{Y})^m$ an die von einer Entscheidungsfunktion $f \in \mathcal{H}$ mit einer Konfidenz von weniger als ρ klassifiziert werden. Wir betrachten nun mit den eingeführten Begriffen eine Abschätzung des Risikos.

Bemerkung 5.2.5. *Die ρ-Rand-Verlustfunktion L_ρ lässt sich für $\rho \in (0, 1]$ beschränken durch die Hinge-Loss-Funktion. Insbesondere gilt dann auch für das empirische Risiko*

$$\mathcal{R}_{m,\rho}(f) \leq \mathcal{R}_{m,L_{hinge}}(f)$$

für alle $f : \mathcal{X} \to \mathcal{Y}$.

Satz 5.2.6 (Mohri, Rostamizadeh und Talwalkar, 2018, Thm. 5.8)**.** *Sei $\mathcal{H}$ eine Menge reellwertiger Funktionen. Fixiere $\rho > 0$, dann gelten die beiden folgenden Abschätzungen für beliebiges $\delta > 0$ mit Wahrscheinlichkeit $1 - \delta$ für alle $h \in \mathcal{H}$.*

$$\mathcal{R}(h) \leq \mathcal{R}_{m,\rho}(h) + \frac{2}{\rho} R_m(\mathcal{H}) + \sqrt{\frac{\log(\frac{1}{\delta})}{2m}}$$

$$\mathcal{R}(h) \leq \mathcal{R}_{m,\rho}(h) + \frac{2}{\rho} \hat{R}_S(\mathcal{H}) + 3\sqrt{\frac{\log(\frac{2}{\delta})}{2m}}.$$

Beweis. Der Beweis befindet sich in *Foundations of Machine learning*(2018, Thm. 5.8)[7] □

Diese Schranken schlagen einen Kompromiss vor. Große Terme ρ verringern den Beitrag der Rademacher Komplexität zur Abschätzung. Gleichzeitig führen größere Werte von ρ dazu, dass der empirische Rand Verlust größer wird da von h eine höhere Konfidenz verlangt wird. Falls $\mathcal{R}_{m,\rho}(h)$ auch für größere Werte von ρ klein bleibt, so zeigt h eine sehr gute Garantie für ein kleines Risiko auf.

Satz 5.2.7 (Rademacher Komplexität Kernel-basierter Hypothesen (Mohri, Rostamizadeh und Talwalkar, 2018, Thm. 6.12))**.** *Sei $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ ein Kernel und sei $\Phi : \mathcal{X} \to H$ eine Merkmalsabbildung assoziiert mit k. Außerdem sei $S \subseteq \{x \in \mathbb{R}^d \mid k(x, x) \leq r^2\}$ eine Stichprobe der Größe m und sei $\mathcal{H} = \{x \mapsto \langle w, \Phi(x) \rangle \mid \|w\|_H \leq \lambda\}$ für ein $\lambda \geq 0$. Dann*

$$\hat{R}_S(\mathcal{H}) \leq \frac{\lambda\sqrt{\mathrm{Tr}(K)}}{m} \leq \sqrt{\frac{(r\lambda)^2}{m}}.$$

Beweis. Es gilt

$$\hat{R}_S(\mathcal{H}) = \frac{1}{m} \mathbb{E}_\sigma \left[\sup_{\|w\| \leq \lambda} \langle w, \sum_{i=1}^{m} \sigma_i \Phi(x_i) \rangle \right].$$

Mit der Cauchy-Schwarz Ungleichung gilt dann

$$\begin{aligned}\hat{R}_S(\mathcal{H}) &\leq \frac{1}{m}\mathbb{E}_\sigma\left[\sup_{\|w\|\leq\lambda}\|w\|_H\left\|\sum_{i=1}^m\sigma_i\Phi(x_i)\right\|_H\right]\\ &\leq \frac{\lambda}{m}\mathbb{E}_\sigma\left[\left\|\sum_{i=1}^m\sigma_i\Phi(x_i)\right\|_H\right].\end{aligned}$$

Mit der Jensen's Ungleichung gilt weiterhin

$$\begin{aligned}\hat{R}_S(\mathcal{H}) &\leq \frac{\lambda}{m}\left(\mathbb{E}_\sigma\left[\left\|\sum_{i=1}^m\sigma_i\Phi(x_i)\right\|_H^2\right]\right)^{1/2}\\ &= \frac{\lambda}{m}\left(\mathbb{E}_\sigma\left[\left\langle\sum_{i=1}^m\sigma_i\Phi(x_i),\sum_{j=1}^m\sigma_j\Phi(x_j)\right\rangle_H\right]\right)^{1/2}\\ &= \frac{\lambda}{m}\sum_{i,j=1}^m\mathbb{E}_\sigma\left[\sigma_i\sigma_j\langle\Phi(x_i),\Phi(x_j)\rangle_H\right]^{1/2}\\ &= \frac{\lambda}{m}\sum_{i=1}^m\mathbb{E}_\sigma\left[\langle\Phi(x_i),\Phi(x_i)\rangle_H\right]^{1/2}\\ &= \frac{\lambda}{m}\mathbb{E}_\sigma\left[\sum_{i=1}^m k(x_i,x_i)\right]\\ &= \frac{\lambda}{m}\sqrt{\mathrm{Tr}(K)} \leq \sqrt{\frac{(r\lambda)^2}{m}}.\end{aligned}$$

Insbesondere wegen $\mathbb{E}_\sigma[\sigma_i\sigma_j] = \hat{\delta}_{ij}$ und im letzten Schritt $\mathrm{Tr}(K) = \sum_{i=1}^m \underbrace{k(x_i,x_i)}_{\leq r^2} \leq mr^2$. □

Bemerkung 5.2.8. *Insbesondere gilt auch für die Rademacher Komplexität*

$$R_m(\mathcal{H}) = \mathbb{E}_{S\sim\mathcal{D}^m}[\hat{R}_S(\mathcal{H})] \leq \mathbb{E}_{S\sim\mathcal{D}^m}\left[\sqrt{\frac{(r\lambda)^2}{m}}\right] = \sqrt{\frac{(r\lambda)^2}{m}}.$$

und außerdem

$$R_m(\mathcal{H}) = \mathbb{E}_{S\sim\mathcal{D}^m}[\hat{R}_S(\mathcal{H})] \leq \mathbb{E}_{S\sim\mathcal{D}^m}\left[\frac{\lambda\,\mathrm{Tr}(K)}{m}\right] = \frac{\lambda}{m}\mathbb{E}_{S\sim\mathcal{D}^m}\left[\sum_{i=1}^m k(x_i,x_i)\right] = \lambda\mathbb{E}_{X\sim\mathcal{D}}[k(X,X)].$$

Wobei wir im letzten Schritt verwendet haben, dass die Datenpunkte identisch Verteilt sind.

Diese Schranken werden im folgenden Abschnitt für verschiedene Wahlen der Kernel-Funktion k genauer untersucht. Insbesondere für translationsinvariante Kernelfunktionen liefert der Abschnitt ein interessantes Resultat.
Die Schranken für die Rademacher Komplexität können in beliebige Risikoschranken eingesetzt werden die auf der Rademacher Komplexität basieren. So auch in die Risikoschranken aus Satz 5.2.9. Somit ergibt sich folgendes Korollar.

Korollar 5.2.9 (Mohri, Rostamizadeh und Talwalkar, 2018, Thm. 6.13)**.** *Sei $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ ein Kernel mit $r^2 = \sup_{x \in \mathcal{X}} k(x,x)$. Sei $\Phi : \mathcal{X} \to H$ eine Merkmalsabbildung assoziiert mit k und sei $\mathcal{H} = \{x \mapsto \langle w, \Phi(x) \rangle \mid \|w\|_H \leq \lambda\}$ für $\lambda \geq 0$. Für festes $\rho > 0$ gelten die folgenden beiden Ungleichung mit Wahrscheinlichkeit $1 - \delta$ für $\delta \in (0,1)$ für alle $h \in \mathcal{H}$*

$$\mathcal{R}(h) \leq \mathcal{R}_{m,\rho}(h) + 2\sqrt{\frac{(r\lambda)^2}{\rho^2 m}} + \sqrt{\frac{\log(\frac{1}{\delta})}{2m}}$$

$$\mathcal{R}(h) \leq \mathcal{R}_{m,\rho}(h) + 2\frac{\sqrt{\mathrm{Tr}(K)\lambda^2}}{\rho m} + 3\sqrt{\frac{\log(\frac{2}{\delta})}{2m}}.$$

Wegen Bemerkung 5.2.5 sagen diese Aussagen aus, dass falls eine Funktion $h : \mathcal{X} \to \mathcal{Y}$ gut im Sinne des empirischen Hinge-Loss-Risikos klassifiziert, so ist auch das wahre Risiko mit hoher Wahrscheinlichkeit gering. Die Abschätzungen liefern somit in gewisser Weise eine Rechtfertigung der Support Vector Machine.

5.3 Untersuchung der Risiko Abschätzungen für verschiedene Kernelfunktionen

Wir haben gesehen, dass der Wert von $k(x,x)$ in beiden eingeführten Risikoschranken eine besondere Rolle spielt. In der Oracle-Ungleichung 5.1.3 tritt der Wert in Form des Erwartungswertes $\mathbb{E}[k(X,X)]$ der Zufallsvariable X auf. In der Risikoschranke 5.2.9 in Form von $r^2 = \sup_{x \in \mathcal{X}} k(x,x)$ beziehungsweise der Spur $Tr(K) = \sum_{i=1}^m k(x_i, x_i)$ der Gram-Matrix K. Wir untersuchen nun dieser Größen für verschiedene Wahlen von Kernelfunktionen k.

5.3.1 Gauß-Kernel

In Korollar 4.2.9 wurde der Gauß-Kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+, (x,x') \mapsto \exp\left(-\|x - x'\|^2/\sigma^2\right)$ eingeführt. Dieser wird auch in der Praxis häufig in Support Vector Machines verwendet und liefert dort gute Ergebnisse.
Sei nun $x \in \mathcal{X}$ und $\sigma^2 > 0$, dann gilt für den Gauß-Kernel

$$k(x,x) = \exp\left(-\frac{\|x - x\|^2}{\sigma^2}\right) = \exp(0) = 1.$$

Daraus folgt insbesondere für die Zufallsvariable X, wobei $(X,Y) \sim \mathcal{D}$ der Verteilung der Daten unterliegen.

$$\mathbb{E}[k(X,X)] = \mathbb{E}[1] = 1.$$

Und für die Spur von der Gram-Matrix K

$$\mathrm{Tr}(K) = \sum_{i=1}^m \underbrace{k(x_i, x_i)}_{=1} = m.$$

Die Feststellung ist, dass die Terme konstant sind. Die Abschätzungen sind unabhängig von der zugrundeliegenden Verteilung $\mathcal{D}$ der Daten. Daher ergibt sich für die Oracle-Ungleichung 5.1.3

$$\mathbb{E}[\mathcal{R}(\hat{h}_n^{\mathrm{SVM}})] \leq \inf_{\|f\|_H \leq \lambda} \mathcal{R}_\varphi(f) + \frac{8\lambda}{\sqrt{m}} \tag{9}$$

und für die Risikoschranken aus 5.2.9

$$\begin{aligned}\mathcal{R}(h) &\leq \mathcal{R}_{m,\rho}(h) + 2\frac{\lambda}{\rho\sqrt{m}} + \sqrt{\frac{\log(\frac{1}{\delta})}{2m}} \\ \mathcal{R}(h) &\leq \mathcal{R}_{m,\rho}(h) + 2\frac{\lambda}{\rho\sqrt{m}} + 3\sqrt{\frac{\log(\frac{2}{\delta})}{2m}}\end{aligned} \tag{10}$$

wobei die erste Abschätzung enger ist, was die zweite in diesem Fall obsolet macht.

Der Paramter σ^2 tritt hier nur in der Berechnung des (empirischen) Risikos auf und beeinflusst über diesen Term die Schranke.
Die Eigenschaft des Gauß-Kernels die für dieses Phänomen verantwortlich ist, ist die Translationsinvarianz dieses Kernels.

Definition 5.3.1 (Tanslationsinvarianz)**.** Eine Funktion $f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ heißt **Translationsinvariant**, falls eine Funktion $f_0 : \mathbb{R}^d \to \mathbb{R}$ existiert mit

$$f(x, y) = f_0(x - y).$$

Bemerkung 5.3.2. *Der Gauß-Kernel, mit $\sigma^2 > 0$,*

$$k(x, x') \mapsto \exp\left(-\frac{\|x - x'\|_{\mathbb{R}^d}}{\sigma^2}\right) \textit{ für alle } x, x' \in \mathbb{R}^d$$

ist Translationsinvariant vermöge der Abblidung

$$k_0(x) \mapsto \exp\left(-\frac{\|x\|_{\mathbb{R}^d}}{\sigma^2}\right) \textit{ für alle } x \in \mathbb{R}^d.$$

Für translationsinvariante Kernel sind die Schranken also im allgemeinen unabhängig von der Verteilung der Daten sondern basieren nur auf der Wahl des translationsinvarianten Kernels. Für Kernel mit dieser Eigenschaft sind die Werte $\mathbb{E}[k(X, X)]$, $\mathrm{Tr}(K)$ und $r^2 = \sup_{x\in\mathcal{X}} k(x, x)$ nämlich konstant und lassen sich durch das Kennen von $k_0(0)$ berechnen. Meist findet sich die Konvention, dass translationsinvariante Kernel durch $k_0(0) = 1$ normiert sind. Es ergeben sich damit die gleichen Abschätzungen wie in (9) und (10) für allgemeine translationsinvariante Kernelfunktionen.

5.3.2 Polynomieller Kernel vom Grad 1 / linearer Kernel

Der polynomielle Kernel vom Grad 1 stellt in gewisser Weise den einfachsten Kernel dar.
Der polynomielle Kernel vom Grad 1 war definiert als

$$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+, k(x, y) \mapsto \langle x, y\rangle_{\mathbb{R}^d} + c$$

für ein $c \geq 0$. Mit dieser Wahl des Kernels gelten für $x, x_1, ..., x_m \in \mathcal{X}$ und für eine Zufallsvariable X, wobei $(X, Y) \sim \mathcal{D}$ der Verteilung der Daten unterliegen, dass

$$k(x, x) = \langle x, x\rangle_{\mathbb{R}^d} = \|x\|^2_{\mathbb{R}^d}.$$

Insbesondere

$$\mathbb{E}[k(X, X)] = \mathbb{E}[\|X\|^2_{\mathbb{R}^d}]$$

und

$$\mathrm{Tr}(K) = \sum_{i=1}^{m} k(x_i, x_i) = \sum_{i=1}^{m} \|x_i\|_{\mathbb{R}^d}^2$$

sowie

$$r^2 = \sup_{x \in \mathcal{X}} k(x, x) = \sup_{x \in \mathcal{X}} \|x\|_{\mathbb{R}^d}^2.$$

Diese Werte können bereits in die Schranken aus 5.1.3 und 5.2.9 eingesetzt werden.

5.3.3 Polynomielle Kernel von höherem Grad

Nach Definition 4.2.2 ist der Polynomielle Kernel vom Grad $n \in \mathbb{N}$ gegeben durch $k(x, y) = (\langle x, y \rangle_{\mathbb{R}^d} + c)^d$ für alle $x, y \in \mathcal{X}$ und $c > 0$. Durch anwenden des binomischen Lehrsatzes t ergibt sich

$$k(x, y) = (\langle x, y \rangle_{\mathbb{R}^d} + c)^d = \sum_{j=1}^{n} \binom{n}{j} \langle x, y \rangle_{\mathbb{R}^d}^{n-j} c^j$$

Für $x \in \mathcal{X}$ heißt das

$$k(x, x) = \sum_{j=1}^{n} \binom{n}{j} \langle x, x \rangle_{\mathbb{R}}^{n-j} c^j = \sum_{j=1}^{n} \binom{n}{j} \|x\|_{\mathbb{R}^d}^{2(n-j)} c^j$$

Für den Erwartungswert der Zufallsvariable X, wobei $(X, Y) \sim \mathcal{D}$ der Verteilung der Daten unterliegen gilt dann

$$\mathbb{E}[k(X, X)] = \sum_{j=1}^{n} \binom{n}{j} \|X\|_{\mathbb{R}^d}^{2(n-j)} c^j$$

sowie für die Spur der Kernel-Matrix K und das Supremum über alle $x \in \mathcal{X}$

$$\mathrm{Tr}(K) = \sum_{i=1}^{m} k(x_i, x_i) = \sum_{i=1}^{m} \sum_{j=1}^{n} \binom{n}{j} \|x_i\|_{\mathbb{R}^d}^{2(n-j)} c^j$$

$$r^2 = \sup_{x \in \mathcal{X}} k(x, x) = \sup_{x \in \mathcal{X}} \sum_{j=1}^{n} \binom{n}{j} \|x\|_{\mathbb{R}^d}^{2(n-j)} c^j.$$

Es lässt sich also feststellen, dass eine große Norm der Datenpunkte die Risikoabschätzung gröber macht. Leider beobachtet man, dass sich die Norm der Datenpunkte durch einfaches skalieren veringern kann und somit die Abschätzung beeinflussen kann. Eine Aussage über die Güte verschiedener polynomieller Kernel erschwert sich hierdurch. Beziehungsweise wird sie nicht von dieser Ungleichung abgedeckt.

6 Rückblick und Ausblick

Ziel dieser Bachelorarbeit war es eine Einführung in die Support Vector Machine zu geben. Dafür wurden die geometrische Interpretation als Abstandsmaximierer und die Interpretation als empirische Risikominimierer vorgestellt. Daraufhin wurde der weit verbreitete Kernel Trick vorgestellt welcher die SVMs in der Praxis sehr wettbewerbsfähig machen. Abschließend wurden eine Risikoschranke und eine Oracle-Inequality vorgestellt die eine theoretische Erklärung dafür geben warum die Support Vector Machines in der Praxis gute Ergebnisse liefern. Die Untersuchung für verschiedene Wahlen von Kerneln hat einige Einblicke gegeben, lässt jedoch einige Fragen offen. Als Auslblick bleibt die weitere Untersuchung der Schranken. Eventuell könnten engere Schranken gefunden werden oder die Schranken könnten mehr Informationen über die Wahl der Kernel beinhalten. Außerdem könnte weiterhin untersucht werden ob SVMs mit gleichen Risikoabschätzungen bei unterschiedlichen Kernel Wahlen ähnliche Ergebnisse erzielen.
An den Leser, der das hier erworbene Wissen vertiefen möchte verweise ich auf das Buch *Support Vector Machines*[3] von Steinwart und Christmann dieses liefert eine sehr aufschlussreiche und tiefgehende Untersuchung der Support Vector Machine und beide Autoren veröffentlichen weiterhin Forschungen auf diesem Gebiet.

Literatur

[1] S. Ben-David und S. Shalev-Shwarz. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014. ISBN: 978-1-107-05713-5.

[2] Gilles Blanchard, Olivier Bousquet und Pascal Massart. "Statistical performance of support vector machines". In: *arXiv:0804.0551 [math, stat]* (Apr. 2008). arXiv: 0804.0551. DOI: 10.1214/009053607000000839. URL: http://arxiv.org/abs/0804.0551 (besucht am 16. 03. 2021).

[3] A. Christmann und I. Steinwart. *Support Vector Machines*. Information Science and Statistics. Springer, 2008. ISBN: 978-0-387-77241-7.

[4] F. Ernst und A. Schweikard. *Fundamentals of Machine Learning. Support Vector Machines Made Easy*. UVK Verlag, 2020. ISBN: 978-3-8385-5251-4.

[5] M. Ledoux und M. Talagrand. *Probability in Banach spaces: isoperimetry and processes*. Classics in mathematics. OCLC: ocn751525992. Berlin ; London: Springer, 2011. ISBN: 9783642202117 9783642202124.

[6] M. Lotz. *(Vorlesungsskript): Mathematics of Machine Learning*. http://homepages.warwick.ac.uk/staff/Martin.Lotz/files/learning/lectnotes-all.pdf. Besucht: 28.02.2021. März 2020.

[7] M. Mohri, A. Rostamizadeh und A. Talwalkar. *Foundations of Machine Learning*. Adaptive computation and machine learning series. The MIT Press, 2018. ISBN: 9780262039406.

[8] Ingo Steinwart und Simon Fischer. "A closer look at covering number bounds for Gaussian kernels". In: *Journal of Complexity* 62 (2021), S. 101513. ISSN: 0885-064X. DOI: https://doi.org/10.1016/j.jco.2020.101513. URL: https://www.sciencedirect.com/science/article/pii/S0885064X2030056X.

[9] M. Trabs. *(Vorlesungsskript): The Mathematics of Machine Learning*. https://www.math.uni-hamburg.de/home/trabs/Lehre/MathMachLearnSS19.pdf. Besucht: 28.02.2021. 2019.

Die vorliegende Arbeit habe ich selbständig verfasst und keine anderen als die angegebenen Hilfsmittel – insbesondere keine im Quellenverzeichnis nicht benannten Internet-Quellen – benutzt. Die Arbeit habe ich vorher nicht in einem anderen Prüfungsverfahren eingereicht. Die eingereichte schriftliche Fassung entspricht genau der auf dem elektronischen Speichermedium.

Unterschrift: