Theorie des maschinellen Lernens

Hans U. Simon

22. Juni 2016

15 Support-Vector Maschinen (SVM)

SVM sind ein wichtiges Werkzeug zum Lernen linearer Voraussagefunktionen in Räumen einer hohen (evtl. sogar unendlich großen) Dimension. Dies wirft die Frage nach der Beherrschbarkeit der Informations- und der Berechnungskomplexität auf. In diesem Kapitel wird aufgezeigt, wie sich die Informationskomplexität mit Hilfe von „Large Margin"-Separatoren kontrollieren lässt. Grob gesagt ist ein „Large Margin"-Separator eine affine Hyperebene, welche die Datenpunkte nicht nur auf der richtigen Seite (Stichwort: positiver versus negativer Halbraum) liegen hat sondern zusätzlich zu diesen einen Sicherheitsabstand, genannt „Margin", einhält. Im folgenden Kapitel besprechen wir den sogenannten „Kernel-Trick", mit welchem sich die Berechnungskomplexität kontrollieren lässt.

Die SVM-Theorie basiert auf der von Vapnik und Chervonenkis in den 1970er Jahren entworfenen Theorie der empirischen und strukturellen Risikominimierung. Sie wurde in den 1990er Jahren (vor allem) von Vapnik perfektioniert. SVM sind eines der derzeit erfolgreichsten Werkzeuge zum maschinellen Lernen.

15.1 Die harte SVM-Lernregel

Es sei $S = [(x_1, y_1), \ldots, (x_m, y_m)] \in (\mathbb{R}^d \times \{-1, 1\})^m$ eine Trainingsmenge. S heißt *linear separierbar*, wenn eine separierende Hyperebene für S existiert, d.h., wenn $w \in \mathbb{R}^d$ und $b \in \mathbb{R}$ existieren, so dass

$$\forall i \in [m] : y_i(\langle w, x_i \rangle + b) > 0 \ . \tag{1}$$

Wir setzen im gesamten Abschnitt 15.1 voraus, dass S linear separierbar ist. Betrachten wir zwei separierende Hyperebenen, von denen eine dicht an einigen der Datenpunkte vorbeischrammt, während die andere zu den Datenpunkten einen „ordentlichen" Sicherheitsabstand einhält. Die Intuition scheint einem zu sagen, dass die zweite separierende Hyperebene der ersten vorzuziehen ist.

In diesem Abschnitt führen wir den Begriff des geometrischen Randabstandes (= Margin) einer Hyperebene zu den Datenpunkten in S ein und beschäftigen uns dann mit „Large Margin"-Separatoren, welche versuchen den Margin zu maximieren. Zunächst aber benötigen wir eine algebraische Formel, welche den Abstand eines Datenpunktes von einer Hyperebene angibt:

Lemma 15.1 *Sei $w \in \mathbb{R}^d$ ein Einheitsvektor (d.h. $\|w\| = 1$) und $b \in \mathbb{R}$. Dann gilt: die Distanz zwischen einem Punkt $x \in \mathbb{R}^d$ und der Hyperebene $H_{w,b}$ beträgt $|\langle w, x\rangle + b|$.*

Beweis Von x aus erreichen wir die Hyperebene $H_{w,b}$ auf dem kürzesten Weg, wenn wir im rechten Winkel auf sie zusteuern. Das wäre für Punkte im positiven (bzw. negativen) Halbraum die Richtung $-w$ (bzw. w). Die Distanz zwischen x und $H_{w,b}$ ist demnach der Absolutbetrag des Skalars λ mit $x - \lambda w \in H_{w,b}$. Die Bedingung $x - \lambda w \in H_{w,b}$ ist äquivalent zu $\langle w, x - \lambda w\rangle + b = 0$. Die folgende Rechnung zeigt, dass $\lambda = \langle w, x\rangle + b$ der gesuchte Skalar ist:

$$\langle w, x - (\langle w, x\rangle + b)w\rangle + b = \langle w, x\rangle - (\langle w, x\rangle + b)\langle w, w\rangle + b = 0 \ .$$

Die letzte Gleichung nutzt aus, dass $1 = \|w\| = \|w\|^2 = \langle w, w\rangle$. **qed.**

Es sei $w \in \mathbb{R}$ ein Einheitsvektor. Die Distanz zwischen $S_{\mathcal{X}} = \{x_1, \ldots, x_m\}$ und $H_{w,b}$ wird auch der *geometrische Randabstand (Margin)* von $H_{w,b}$ zu S genannt. Gemäß Lemma 15.1 ist der Margin gegeben durch

$$\gamma_{w,b}(S) = \min_{i \in [m]} |\langle w, x_i\rangle + b| \ . \tag{2}$$

Wenn w nicht notwendig normiert ist, dann heißt die durch (2) gegebene Größe $\gamma_{w,b}(S)$ der *funktionale Randabstand (Functional Margin)* von $H_{w,b}$ zu S.

Die *harte SVM-Lernregel* besagt: wähle w^*, b^* so, dass $H_{w,b}$ eine separierende Hyperebene für S mit dem größtmöglichen geometrischen Randabstand zu S ist. Formal:

$$(w^*, b^*) = \text{argmax}_{w,b}\gamma_{w,b}(S) \textbf{ s.t. } (\|w\| = 1) \wedge (1) \ .$$

Der maximale Wert bei diesem Optimierungsproblem heißt der *größtmögliche geometrische Randabstand (Maximum Margin)* für S. Er wird im Folgenden mit $\gamma_{max}(S)$ notiert.

Bemerkung 15.2 *Eine Hyperebene $H_{w,b}$ hat genau dann einen funktionalen Randabstand von mindestens 1 zu S, wenn mit dem normierten Vektor $w/\|w\|$ anstelle von w und $b/\|w\|$ anstelle von b ein geometrischer Randabstand von mindestens $1/\|w\|$ erzielt wird. Die für S separierende Hyperebene mit dem größtmöglichen Randabstand zu S ist also identisch mit der für S separierenden Hyperebene, die mit einem Vektor möglichst kleiner Länge einen funktionalen Randabstand von mindestens 1 zu S erzielt.*

Aus Bemerkung 15.2 ergibt sich folgendes Resultat:

Lemma 15.3 *Die harte SVM-Lernregel lässt sich auf äquivalente Weise umformulieren wie folgt: wähle (w^*, b^*) so, dass $H_{w,b}$ eine separierende Hyperebene für S mit einem funktionalen Randabstand von mindestens 1 zu S ist und $\|w\|$ unter dieser Randbedingung minimiert wird. Formal:*

$$(w^*, b^*) = \text{argmin}_{w,b}\|w\|^2 \textbf{ s.t. } \forall i \in [m] : y_i(\langle w, x_i\rangle + b) \geq 1 \ . \tag{3}$$

Es gilt dann:

$$\gamma_{max}(S) = \frac{1}{\|w^*\|}$$

und dieser Margin wird von $(w^*/\|w^*\|, b^*/\|w^*\|)$ erzielt.

Beachte, dass das Minimieren der Kostenfunktion $\|w\|^2$ äquivalent zum Minimieren der Kostenfunktion $\|w\|$ ist. Die Wahl von $\|w\|^2$ statt $\|w\|$ in (3) hat den Vorteil, dass ein *konvexes* quadratisches Optimierungsproblem entsteht, das mit Standardwerkzeugen effizient gelöst werden kann.

Wir merken kurz an, dass die harte SVM-Lernregel im homogenen Fall (also mit der Festlegung $b = 0$) die folgende Form hat:

$$w^* = \text{argmin}_w \|w\|^2 \textbf{ s.t. } \forall i \in [m] : y_i \langle w, x_i \rangle \geq 1 \ . \tag{4}$$

Reduktion auf den homogenen Fall. Wie wir wissen ist $S = [(x_1, y_1), \ldots, (x_m, y_m)] \in (\mathbb{R}^d \times \{-1, 1\})^m$ genau dann durch eine affine Hyperebene im $\mathbb{R}^d$ separierbar, wenn $S' = [(x'_1, y_1), \ldots, (x'_m, y_m)]$ mit $x'_i = (1, x_i)$ durch eine homogene Hyperebene (mit Bias $b = 0$) im $\mathbb{R}^{d+1}$ separierbar ist, sagen wir durch $H_{w',0}$ mit $w' = (w_0, w)$ und $w = (w_1, \ldots, w_d)$. Die entsprechende affine Hyperebene im $\mathbb{R}^d$ wäre dann H_{w,w_0}, d.h., w_0 spielt die Rolle des Bias. Allerdings liefert die harte SVM-Lernregel für das Szenario im $\mathbb{R}^d$ i.A. *nicht* das selbe Resultat wie für das entsprechende Szenario im $\mathbb{R}^{d+1}$. Der Grund ist, dass die Einbettung in den $\mathbb{R}^{d+1}$ die geometrischen Randabstände verändert: die Distanz zwischen H_{w,w_0} und x_i beträgt

$$\frac{1}{\|w\|} |\langle w, x_i \rangle + w_0| \ ,$$

wohingegen die Distanz zwischen $H_{w',0}$ und x'_i gegeben ist durch

$$\frac{1}{\|w'\|} |\langle w', x'_i \rangle| = \frac{1}{\|w'\|} |\langle w, x_i \rangle + w_0| \ .$$

Beim Optimierungsproblem (4) (mit w' in der Rolle von w) macht sich der Unterschied dadurch bemerkbar, dass wir die Kostenfunktion $\|w'\|^2$ verwenden und nicht $\|w\|^2$. In der Sprechweise der Lerntheorie: das Bias $b = w_0$ wird als Bestandteil des Gewichtsvektors w' im homogenen Fall „mitregularisiert“. Die Erfahrung in der Praxis scheint aber dahin zu gehen, dass es keinen dramatischen Unterschied macht, ob wir die Reduktion auf den homogenen Fall vornehmen oder nicht.

15.2 Weiche SVM-Lernregeln

In Anwendungen sind Trainingsmengen selten perfekt linear separierbar. Es ist daher ratsam, die harte SVM-Lernregel etwas aufzuweichen, so dass funktionale Randabstände unterhalb von 1 grundsätzlich zugelassen sind. Das kann zum Beispiel umgesetzt werden, indem für jeden Datenpunkt (x_i, y_i), $i = 1, \ldots, m$, eine nicht-negative Schlupfvariable ξ_i eingeführt und die Bedingung $y_i(\langle w, x_i \rangle + b) \geq 1$ in (3) durch $y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i$ ersetzt wird. Sogar $\xi_i > 1$, d.h. x_i ist auf der falschen Seite der Hyperebene, ist nicht kategorisch ausgeschlossen. Um den exzessiven Gebrauch der ξ_i zu drosseln, muss im Gegenzug ein Strafterm in die Kostenfunktion eingeführt werden, zum Beispiel der Term $\sum_{i=1}^m \xi_i/m$. Das Verhältnis dieses Strafterms zum Regularisierungsterm $\|w\|^2$ könnte über einen Skalar $\lambda > 0$ geregelt werden.

Diese Ideen führen zu der folgenden *weichen SVM-Lernregel*: wähle (w^*, b^*) gemäß

$$(w^*, b^*) = \text{argmin}_{w,b,\xi} \left(\lambda \|w\|^2 + \frac{1}{m} \sum_{i=1}^{m} \xi_i \right) \textbf{ s.t. } \forall i \in [m] : (y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i) \wedge (\xi_i \geq 0) \ . \tag{5}$$

Die *Hinge-Verlustfunktion (hinge loss)* ist gegeben durch

$$\ell^{hinge}((w,b),(x,y)) = \max\{0, 1 - y(\langle w, x \rangle + b)\} \ .$$

Beachte, dass $\ell^{hinge}((w,b),(x,y)) \geq 1$ gilt, falls x auf der falschen Seite der Hyperebene $H_{w,b}$ liegt (also falls $y(\langle w, x \rangle + b) \leq 0$). Daher gilt stets $\ell^{0-1}((w,b),(x,y)) \leq \ell^{hinge}((w,b),(x,y))$. Der mittlere Hinge-Verlust von (w,b) auf S wird dann als $L_S^{hinge}(w,b)$ notiert, d.h.,

$$L_S^{hinge}(w,b) = \frac{1}{m} \sum_{i=1}^{m} \ell^{hinge}((w,b),(x_i,y_i)) \ .$$

Es ist offensichtlich, dass $\ell^{hinge}((w,b),(x_i,y_i))$ exakt dem nicht-negativen Strafterm ξ_i in (5) entspricht. Daher ist die weiche SVM-Lernregel (5) äquivalent zu

$$(w^*, b^*) = \text{argmin}_{w,b} \left(\lambda \|w\|^2 + L_S^{hinge}(w,b) \right) \ .$$

Es ist somit die Summe aus einem Regularisierungsterm, $\lambda \|w\|^2$, und einem empirischen Verlust, $L_S^{hinge}(w,b)$, zu minimieren. Optimierungsprobleme dieser Form werden in der Lerntheorie auch als *regularisierte Verlustminimierungsprobleme* bezeichnet.

Wir merken kurz an, dass die weiche SVM-Lernregel im homogenen Fall (also mit der Festlegung $b = 0$) die folgende Form hat:

$$w^* = \text{argmin}_{w} \left(\lambda \|w\|^2 + L_S^{hinge}(w) \right) \ , \tag{6}$$

wobei $L_S^{hinge}(w) = L_S^{hinge}(w,0)$ gesetzt wird.

15.3 Optimalitätskriterien und Support-Vektoren

Betrachten wir nochmals das Optimierungsproblem (4). Es sei w^* eine optimale Lösung. Ein Vektor x_i aus der Trainingsmenge S heißt *Support-Vektor* für w^*, wenn $y_i \langle w^*, x_i \rangle = 1$. Dies bedeutet, dass x_i einer der Punkte aus S mit der geringsten Distanz zu der homogenen Hyperebene $H_w = H_{w,0}$ ist. Der Name „Support-Vektor“ ist motiviert durch das folgende Resultat:

Theorem 15.4 *Eine optimale Lösung des Problems (4) liegt im Vektorraum, der von ihren Support-Vektoren aufgespannt wird.*

Dieser Satz ist eine einfache Folgerung aus der folgenden allgemeinen „Fritz John Optimalitätsbedingung“:

Lemma 15.5 *Es seien $f, g_1, \ldots, g_m$ auf $\mathbb{R}^d$ definierte differenzierbare Funktionen. Weiter sei*

$$w^* = \text{argmin}_w f(w) \textbf{ s.t. } \forall i \in [m] : g_i(w) \leq 0 \ .$$

Dann existiert $\alpha \in \mathbb{R}^m$ mit

$$\nabla f(w^*) + \sum_{i \in [m]: g_i(w^*)=0} \alpha_i \nabla g_i(w^*) = \vec{0} \ .$$

Eine analoge Aussage gilt auch im inhomogenen Fall. Es wäre eine gute Übung, Theorem 15.4 mit Hilfe von Lemma 15.5 zu beweisen. (Den Beweis des Lemmas lassen wir aus.)

Ein alternativer Beweis von Theorem 15.4 könnte unter Ausnutzen von Dualität und Anwendung der sogenannten „complementary slackness"-Bedingungen geführt werden. Auf Details können wir im Rahmen dieser Vorlesung nicht eingehen.

15.4 Kontrolle der Informationskomplexität

Definition 15.6 *Es sei D eine Verteilung über $\mathbb{R}^d \times \{-1, 1\}$ und es seien $\gamma, \rho > 0$. Wir sagen D ist* separierbar mit (γ, ρ)-Margin, *falls ein Paar $(w, b) \in \mathbb{R}^d \times \mathbb{R}$ mit $\|w\| = 1$ existiert, so dass mit Wahrscheinlichkeit 1 für $(x, y) \sim D$ folgendes gilt: $\|x\| \leq \rho$ und $y(\langle w, x\rangle + b) \geq \gamma$. In Worten: ein mit D zufällig gezogenes Trainigsbeispiel (x, y) liegt fast sicher mit einem Margin von mindestens γ auf der richtigen Seite der Hyperebene $H_{w,b}$ und die (Euklidische) Länge von x ist fast sicher durch ρ nach oben beschränkt.*

Das folgende Resultat liefert eine obere Schranke für die Fehlerrate von Hypothesen, die mit der harten SVM-Lernregel ermittelt werden:

Theorem 15.7 *Es sei D eine mit (γ, ρ)-Margin separierbare Verteilung über $\mathbb{R}^d \times \{-1, 1\}$. Weiter sei $(w^*, b^*) \in \mathbb{R}^d \times \mathbb{R}$ die Ausgabe der harten SVM-Lernregel zu einer gegebenen Trainingsmenge der Größe m. Dann gilt mit einer Wahrscheinlichkeit von mindestens $1 - \delta$ (und den üblichen Notationen):*

$$L_D^{0-1}(w^*, b^*) \leq \frac{1}{\sqrt{m}} \left(\frac{2\rho}{\gamma} + \sqrt{2 \log \frac{2}{\delta}} \right) \ .$$

Den Beweis, der auf theoretischen Resultaten zur sogenannten Rademacher-Komplexität beruht, lassen wir erst einmal aus (und holen ihn evtl. zu einem späteren Zeitpunkt der Vorlesung nach).

Das entsprechende Resultat für die weiche SVM-Lernregel (dessen Beweis wir im Rahmen dieser Vorlesung nicht führen können) liest sich wie folgt:

Theorem 15.8 *Es sei $\rho > 0$, $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\| \leq \rho\}$ und D eine Verteilung über $\mathcal{X} \times \{-1, 1\}$. Weiter sei w^* die Ausgabe der weichen SVM-Lernregel (im homogenen Fall) zu einer gegebenen Trainingsmenge der Größe m. Dann gilt:*

$$\mathbb{E}_{S \sim D^m}[L_D^{hinge}(w^*)] \leq \min_{w \in \mathbb{R}^d} \left\{ L_D^{hinge}(w) + \lambda \|w\|^2 \right\} + \frac{2\rho^2}{\lambda m} \ .$$

Darüberhinaus gilt für jedes $B > 0$ und der Parmeterwahl $\lambda = \frac{\rho}{B}\sqrt{\frac{2}{m}}$ folgendes:

$$\mathbb{E}_{S\sim D^m}[L_D^{hinge}(w^*)] \leq \min_{w\in\mathbb{R}^d:\|w\|\leq B} L_D^{hinge}(w) + 2\rho B\sqrt{\frac{2}{m}} \ .$$

Da die Hinge-Verlustfunktion eine obere Schranke für die 0-1-Verlustfunktion darstellt, gelten die oberen Schranken aus Theorem 15.8 auch für L_D^{0-1} anstelle von L_D^{hinge}.

Das Bemerkenswerte an den Fehlerschranken in den Theoremen 15.7 und 15.8 ist, dass sie gänzlich unabhängig von der Dimension d des zugrunde liegenden Euklidischen Raumes sind. Das bedeutet, dass die Informationskomplexität Margin-basiert auch in extrem hochdimensionalen Räumen beherrschbar ist.

Die Beschränkung auf Vektoren x mit $\|x\| \leq \rho$ ist notwendig: Margins ließen sich sonst durch Skalierung der Punkte x beliebig vergrößern, hätten dann aber keine Aussagekraft mehr.

16 Kernel-Methoden

Dieses Kapitel ist dem sogenannten „Kernel-Trick“ gewidmet, einer Methode,

- welche es erlaubt lineare Voraussagefunktionen mit Erfolg auf hochgradig nicht-lineare Daten anzuwenden
- und welche eine Kontrolle über die Berechnungskomplexität herstellt selbst bei (impliziter) Verwendung extrem hochdimensionaler Räume.

16.1 Daten- und Merkmalsraum

Anwendungsdaten weisen in der Regel Nicht-Linearitäten auf. Eine unmittelbare Anwendung von linearen Voraussagefunktionen ist daher nur selten erfolgreich. Stattdessen ist es besser, die Daten aus $\mathcal{X}$ in einen Merkmalsraum (feature space) $\mathcal{F}$ so abzubilden, dass die Merkmalsvektoren sich (zumindest approximativ) durch lineare Voraussagefunktionen beschreiben lassen. Die in den Daten vorhandenen Nicht-Linearitäten werden dabei, so gut es geht, in die Abbildung von $\mathcal{X}$ in $\mathcal{F}$ eingekapselt. Wir betrachten ein künstliches, aber illustratives,

Beispiel 16.1 *Eine ganze Zahl x erhalte das Label „1“, falls $|x| > 2$ und andernfalls das Label „-1“. Wir betten $\mathcal{X} = \mathbb{Z}$ durch die Abbildung $x \mapsto \psi(x) = (x, x^2)$ in $\mathcal{F} = \mathbb{Z} \times \mathbb{Z}$ ein. Über $\mathcal{F}$ gibt es eine lineare Voraussagefunktion, welche alle Daten perfekt beschreibt. Zum Beispiel gilt für $(w, b) = ((0, 1), -6)$:*

$$\text{sign}(\langle w, \psi(x)\rangle + b) = \text{sign}(x^2 - 6)$$

und diese Funktion ordnet allen ganzen Zahlen das korrekte Label zu.

Die allgemeine Vorgehensweise ist wie folgt:

1. Wähle eine Abbildung $\psi : \mathcal{X} \to \mathcal{F}$, die jedem Datenpunkt $x \in \mathcal{X}$ einen sogenannten *Merkmalsvektor* $\psi(x) \in \mathcal{F}$ zuordnet. Wir werden häufig $\mathcal{F} = \mathbb{R}^d$ wählen, aber wir dürfen an Stelle von $\mathcal{F}$ einen beliebigen *Hilbert-Raum*[1] (mit evtl. unendlicher Dimension) verwenden.

2. Gegeben eine Trainingssequenz $S = [(x_1, y_1), \ldots, (x_m, y_m)]$, bestimme die zugehörige Bildsequenz $S = [(x'_1, y_1), \ldots, (x'_m, y_m)]$ mit $x'_i = \psi(x_i)$ für $i = 1, \ldots, m$.

3. Ermittle eine lineare Voraussagefunktion $h : \mathcal{F} \to \{-1, 1\}$.

4. Sage das Label eines Testpunktes $x \in \mathcal{X}$ mit $h(\psi(x))$ voraus.

Falls D die zugrunde liegende Verteilung auf $\mathcal{X}$ bezeichnet, so dass $S \sim D^m$, dann ist die Bildverteilung D^ψ auf $\mathcal{F}$ gegeben durch

$$D^\psi(A) = D(\psi^{-1}(A))$$

und diese Gleichung gilt für alle „messbaren“ Mengen $A \subseteq \mathcal{F}$. Die Abbildung ψ muss selber „messbar“ sein, d.h., aus der Messbarkeit von $A \subseteq \mathcal{F}$ muss die Messbarkeit von $\psi^{-1}(A) \subseteq \mathcal{X}$ folgen. Bei den Abbildungen, die wir betrachten werden, werden diese „Messbarkeitsanforderungen“ immer erfüllt sein (ohne dass wir explizit darauf eingehen werden).
Es folgt eine naheliegende Verallgemeinerung von Beispiel 16.1:

Beispiel 16.2 *Für alle $x \in \mathcal{X} = \mathbb{R}$ setze $\psi(x) = (1, x, x^2, \ldots, x^k) \in \mathcal{F} = \mathbb{R}^{k+1}$. Wenn die Daten aus $\mathbb{R} \times \{-1, 1\}$ mit einem Polynom vom Grad $\leq k$ separierbar sind, dann sind die zugehörigen markierten Merkmalsvektoren in $\mathbb{R}^{k+1} \times \{-1, 1\}$ linear separierbar.*
Allgemeiner sei nun $\mathcal{X} = \mathbb{R}^n$ und wir nehmen an, dass die Daten aus $\mathcal{X} \times \{-1, 1\}$ durch ein multivariates Polynom (über $x_1, \ldots, x_n$) vom Grad $\leq k$ separierbar sind. Wir setzen der Einfachheit halber $x_0 = 1$. Dann sind die zugehörigen Merkmalsvektoren linear separierbar, wenn wir $\psi(x)$ so definieren, dass für jede Wahl von $(i_1, \ldots, i_k) \in \{0, 1, \ldots, n\}^k$ eine Komponente

$$\psi_{i_1,\ldots,i_k}(x) = \prod_{j=1}^{k} x_{i_j} \tag{7}$$

von $\psi(x)$ vorgesehen ist (so dass $\psi(x) \in \mathbb{R}^{(n+1)^k}$).

16.2 Mathematische Grundlagen

Definition 16.3 *Eine symmetrische Matrix $A \in \mathbb{R}^{m \times m}$ heißt* positiv semidefinit, *falls*

$$\forall u \in \mathbb{R}^m : u^\top A u = \sum_{i,j=1}^{m} A_{i,j} u_i u_j \geq 0 \ .$$

[1] Zur Definition von Hilbert-Räumen s. Abschnitt 16.2.

Falls dabei die Gleichheit $u^\top Au = 0$ *nur für den Fall* $u = \vec{0}$ *eintritt, so heißt* A positiv definit.

Offensichtlich sind positive Linearkombinationen positiv semidefiniter Matrizen ebenfalls positiv semidefinit, da für $\lambda_k > 0$ und positiv semidefinite Matrizen A_k folgendes gilt:

$$u^\top \left(\sum_k \lambda_k A_k \right) u = \sum_k \lambda_k (u^\top A_k u) \geq 0 \ .$$

Beispiel 16.4 *Es sei* $z \in \mathbb{R}^m$ *und* $A = zz^\top$, *d.h.,* $A_{i,j} = z_i z_j$ *für alle* $1 \leq i, j \leq m$. A *ist offensichtlich symmetrisch. Weiterhin gilt für jeden Vektor* $u \in \mathbb{R}^m$*:*

$$u^\top (zz^\top) u = (u^\top z)(z^\top u) = (u^\top z)^2 \geq 0 \ .$$

Daher ist die Matrix $A = zz^\top$ *positiv semidefinit.*

Beispiel 16.5 *Es seien* $z_1, \ldots, z_N \in \mathbb{R}^m$ *und* $M = [z_1 \ldots z_N] \in \mathbb{R}^{m \times N}$. *Dann ist* $M^\top M \in \mathbb{R}^{N \times N}$ *die Matrix mit* $\langle z_i, z_j \rangle$ *in Zeile* i *und Spalte* j. *Die Matrix* $M^\top M$ *ist offensichtlich symmetrisch und wegen*

$$u^\top (M^\top M) u = (u^\top M^\top)(Mu) = (Mu)^\top (Mu) = \|Mu\|^2 \geq 0 \ .$$

außerdem positiv semidefinit.

Definition 16.6 *Ein Vektorraum* $\mathcal{F}$ *heißt* Semi-Innerer-Produkt-Raum, *falls er mit einer bilinearen Abbildung* $\langle \cdot, \cdot \rangle$ *versehen ist, die zudem für alle Wahlen von* $z, z' \in \mathcal{F}$ *folgende Bedingungen erfüllt:*

$$\langle z, z' \rangle = \langle z', z \rangle \text{ und } \langle z, z \rangle \geq 0 \ .$$

Falls $\langle z, z \rangle = 0$ *nur für den Nullvektor* $z = \vec{0}$ *gilt, dann heißt* $\mathcal{F}$ *ein* Innerer-Produkt-Raum.

Wie oben schon mal erwähnt kann die Rolle des Merkmalsraumes $\mathcal{F}$ von einem beliebigen „Hilbert-Raum“ gespielt werden. Letztere sind definiert wie folgt:

Definition 16.7 *Ein* Hilbert-Raum $\mathcal{F}$ *ist ein Semi-Innerer-Produkt-Raum, der (bezüglich der von* $\langle \cdot, \cdot \rangle$ *induzierten Pseudo-Metrik) topologisch vollständig ist. Letzteres bedeutet, dass jede Cauchy-Sequenz in* $\mathcal{F}$ *auch einen Grenzwert in* $\mathcal{F}$ *besitzt.*

Beispiel 16.8 *1. Für jede Wahl von* $d \geq 1$ *ist* $\mathbb{R}^d$ *ein Hilbert-Raum.*

2. Es sei ℓ_2 *die Menge aller Folgen* $(z_n)_{n \geq 1}$ *über* $\mathbb{R}$ *so dass*

$$\|z\|^2 = \sum_{n=1}^{\infty} z_n^2 < \infty \ .$$

Wir definieren ein inneres Produkt wie folgt:

$$\langle z, z' \rangle = \sum_{n=1}^{\infty} z_n z'_n \ .$$

Es ist nicht schwer zu zeigen, dass $\langle \cdot, \cdot \rangle$ ein inneres Produkt und dass ℓ_2, mit diesem inneren Produkt versehen, ein (unendlich-dimensionaler) Hilbert-Raum ist.

Wir werden hauptsächlich ausnutzen, dass „Projektionen“ in Hilbert-Räumen wohldefiniert sind, d.h., wenn $\mathcal{F}'$ ein Unterraum von $\mathcal{F}$ ist, dann gibt es zu jedem Vektor $z \in \mathcal{F}$ eine eindeutige Zerlegung $z = u + v$ mit $u \in \mathcal{F}'$ und für alle $u' \in \mathcal{F}'$ gilt $\langle u', v \rangle = 0$. Der Vektor u ist dann die Projektion von z auf $\mathcal{F}'$ (also der Vektor in $\mathcal{F}'$ mit der kürzesten Distanz zu z).

16.3 Der Kernel-Trick

Wie wir in Kapitel 15 gesehen haben, ist die Verwendung von hochdimensionalen Räumen akzeptabel, sofern Margin-basierte Schranken für den Generalisierungsfehler zum Einsatz kommen. In diesem Abschnitt beschäftigen wir uns mit der Berechnungskomplexität. Wenn die Dimension d sehr groß ist, könnte allein das Aufschreiben eines Vektors mit d Komponenten zuviel Zeit erfordern (ganz zu schweigen von der Zeit, die wir benötigen, um die betreffenden konvexen Optimierungsprobleme im Raum $\mathbb{R}^d$ zu lösen).

An dieser Stelle kommt der sogenannte „Kernel-Trick“ zum Einsatz. Eine Funktion $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ heißt *Kernel* für den Datenraum $\mathcal{X}$, falls ein Hilbert-Raum $\mathcal{F}$ und eine Abbildung $\psi : \mathcal{X} \to \mathcal{F}$ existieren, so dass

$$\forall x, x' \in \mathcal{X} : K(x, x') = \langle \psi(x), \psi(x') \rangle \ . \tag{8}$$

Die Abbildung ψ wird auch *Mermalsabbildung (feature map)* genannt. Ein Algorithmus heißt *kernbasiert*, wenn er über ein „Orakel“ verfügt, welches auf Anfrage (x, x') in einem Schritt den Wert $K(x, x')$ zurückliefert. Ein effizienter kernbasierter Algorithmus wird zu einem „richtigen“ effizienten Algorithmus, wenn wir das Orakel durch eine effiziente Prozedur zur Berechnung von $K(x, x')$ ersetzen. Es wird sich zeigen, dass $K(x, x')$ häufig unmittelbar auf dem Datenraum (also ohne den Umweg über die Gleichung (8)) effizient berechnet werden kann. Weder muss dabei ein Vektor $\psi(x)$ jemals explizit konstruiert werden, noch ist es notwendig die Abbildung ψ auch nur zu kennen! Der Rest dieses Abschnittes ist den kernbasierten Lernalgorithmen gewidmet.

Wenn wir uns die Merkmalsvektoren $\psi(x_i)$ an die Stelle der Datenpunkte x_i denken, dann haben die in (4) und (6) beschriebenen Optimierungsprobleme die folgende allgemeine Form:[2]

$$w^* = \operatorname{argmin}_w \left(R(\|w\|) + f(\langle w, \psi(x_1) \rangle, \ldots, \langle w, \psi(x_m) \rangle) \right) \tag{9}$$

[2]Eine analoge Aussage gilt für die entsprechenden Optimierungsprobleme im inhomogenen Fall.

Dabei ist $f : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ eine beliebige und $R : \mathbb{R}_+ \to \mathbb{R}$ eine monoton wachsende Funktion. Zum Beispiel erhalten wir (4), indem wir $R(a) = a^2$ und

$$f(a_1, \ldots, a_m) = \begin{cases} 0 & \text{falls } y_i a_i \geq 1 \text{ für alle } i \in [m] \\ \infty & \text{sonst} \end{cases}$$

setzen. Wir erhalten (6), indem wir $R(a) = \lambda a^2$ und

$$f(a_1, \ldots, a_m) = \frac{1}{m} \sum_{i=1}^{m} \max\{0, 1 - y_i a_i\}$$

setzen.

Der folgende Satz besagt, dass eine optimale Lösung für (9) existiert, die von den Merkmalsvektoren $\psi(x_1), \ldots, \psi(x_m)$ aufgespannt wird:

Theorem 16.9 (Representer Theorem) *Es sei $\psi : \mathcal{X} \to \mathcal{F}$ für einen Hilbert-Raum $\mathcal{F}$. Dann existiert ein $\alpha \in \mathbb{R}^m$, so dass*

$$w = \sum_{i=1}^{m} \alpha_i \psi(x_i)$$

das Problem (9) optimal löst.

Beweis Es sei $\mathcal{F}' = \text{span}(\psi(x_1), \ldots, \psi(x_m))$. Es sei $w^* \in \mathcal{F}$ eine optimale Lösung von (9). Da $\mathcal{F}$ ein Hilbert-Raum ist können wir w^* zerlegen gemäß $w^* = w + v$, wobei $w = \sum_{i=1}^{m} \alpha_i \psi(x_i)$ die Projektion von w^* auf $\mathcal{F}'$ ist und es gilt $\langle w', v \rangle = 0$ für alle $w' \in \mathcal{F}'$. Insbesondere gilt $\langle w, v \rangle = 0 = \langle \psi(x_i), v \rangle$. Es genügt zu zeigen, dass auch w eine optimale Lösung von (9) ist. Wegen $\langle w, v \rangle = 0$ gilt $\|w^*\|^2 = \|w\|^2 + \|v\|^2$ und somit $\|w\| \leq \|w^*\|$ sowie auch $R(\|w\|) \leq R(\|w^*\|)$. Weiter gilt

$$\langle w^*, \psi(x_i) \rangle = \langle w, \psi(x_i) \rangle + \langle v, \psi(x_i) \rangle = \langle w, \psi(x_i) \rangle$$

für alle $i \in [m]$ und somit auch

$$f(\langle w^*, \psi(x_1) \rangle, \ldots, \langle w^*, \psi(x_m) \rangle) = f(\langle w, \psi(x_1) \rangle, \ldots, \langle w, \psi(x_m) \rangle) \ .$$

Zusammen mit $R(\|w\|) \leq R(\|w^*\|)$ erhalten wir, dass die Kosten der Lösung w nicht höher sind als die der Lösung w^*. Daher ist auch w eine optimale Lösung von (9). **qed.**

Wir bezeichnen die Matrix $K \in \mathbb{R}^{m \times m}$ mit dem Eintrag $K(x_i, x_j) = \langle \psi(x_i), \psi(x_j) \rangle$ in Zeile i und Spalte j als die *Kernel-Matrix*[3] zur Sequenz $x_1, \ldots, x_m$. Auf Grundlage des Representer-Theorems können wir nach einer optimalen Wahl w^* für w suchen, indem wir eine optimale Wahl α^* für α treffen und

$$w^* = \sum_{i=1}^{m} \alpha_i^* \psi(x_i) \tag{10}$$

setzen. Dabei ergibt sich folgendes Optimierungsproblem:

[3]mitunter auch *Gram-Matrix* genannt

Theorem 16.10 *Wir erhalten eine optimale Lösung von (9) gemäß (10), wobei*

$$\alpha^* = \text{argmin}_\alpha \left(R(\sqrt{\alpha^T K \alpha}) + f(K\alpha) \right) \quad . \tag{11}$$

Beweis Ein Vergleich von (11) und (9) zeigt, dass wir folgendes verifizieren müssen:

$$\|w\|^2 = \alpha^\top K \alpha \text{ und } \langle w, \psi(x_i) \rangle = (K\alpha)_i \text{ für } i = 1, \ldots, m \quad .$$

Die erste Gleichung ergibt sich aus

$$\|w\|^2 = \left\langle \sum_{i=1}^m \alpha_i \psi(x_i), \sum_{i=j}^m \alpha_j \psi(x_j) \right\rangle = \sum_{i,j=1}^m \alpha_i \alpha_j \langle \psi(x_i), \psi(x_j) \rangle = \sum_{i,j=1}^m \alpha_i \alpha_j K(x_i, x_j) = \alpha^\top K \alpha \quad .$$

Die zweite gewünschte Gleichung ergibt sich aus

$$\langle w, \psi(x_i) \rangle = \left\langle \sum_{j=1}^m \alpha_j \psi(x_j), \psi(x_i) \right\rangle = \sum_{j=1}^m \alpha_j \langle \psi(x_i), \psi(x_j) \rangle = \sum_{j=1}^m \alpha_j K(x_i, x_j) = (K\alpha)_i \quad .$$

qed.

Mit Hilfe sturen Einsetzens ergibt sich aus Theorem 16.10 der folgende Spezialfall:

Folgerung 16.11 *Wir erhalten eine optimale Lösung von (6) gemäß (10), wobei*

$$\alpha^* = \text{argmin}_\alpha \left(\lambda \alpha^\top K \alpha + \frac{1}{m} \sum_{i=1}^m \max\{0, 1 - y_i (K\alpha)_i\} \right) \quad . \tag{12}$$

(12) ist ein konvexes quadratisches Optimierungsproblem, zu dessen Lösung effiziente (und kernbasierte!) Standardwerkzeuge existieren.

Wir merken kurz an, dass die Vorhersage $\text{sign}(\langle w, \psi(x) \rangle)$, die eine Hypothese H_w auf einer Instanz x tätigt, auch kernbasiert und einfach zu berechnen ist, da

$$\langle w, \psi(x) \rangle = \sum_{j=1}^m \alpha_j \langle x_j, x \rangle = \sum_{j=1}^m \alpha_j K(x_j, x) \quad .$$

Die analoge Bemerkung gilt für die Vorhersage $\text{sign}(\langle w, \psi(x) \rangle + b)$ einer Hypothese $H_{w,b}$ im inhomogenen Fall.

16.4 Effizient auswertbare Kernels

Wir stellen in diesem Abschnitt einige der populärsten (und bei Anwendungen erfolgreichsten) Kernels vor und demonstrieren jeweils, dass sie effizient ausgewertet werden können (ohne dass die zugehörige Abbildung in einen Hilbert-Raum bekannt sein muss).

Polynomieller Kernel. Wir kommen zurück auf die $x \in \mathbb{R}^n$ zugeordneten Merkmalsvektoren $\psi(x)$ aus Beispiel 16.2 mit einer Komponente der Form (7) für jede Wahl von $(i_1, \ldots, i_k) \in \{0, 1 \ldots, n\}^k$. Die folgende Rechnung, bei der wir der Einfachheit halber $x_0 = x'_0 = 1$ gesetzt haben, zeigt, dass

$$K(x, x') = (1 + \langle x, x' \rangle)^k$$

der zu ψ passende Kernel mit $K(x, x') = \langle \psi(x), \psi(x') \rangle$ ist:

$$\begin{aligned} K(x, x') &= (1 + \langle x, x' \rangle)^k = \left(\sum_{i=0}^{n} x_i x'_i \right)^k \\ &\overset{DG}{=} \sum_{i_1, \ldots, i_k = 0}^{n} \prod_{j=1}^{k} x_{i_j} x'_{i_j} \\ &= \sum_{i_1, \ldots, i_k = 0}^{n} \left(\prod_{j=1}^{k} x_{i_j} \right) \left(\prod_{j=1}^{k} x'_{i_j} \right) \\ &\overset{(7)}{=} \sum_{i_1, \ldots, i_k = 0}^{n} \psi_{i_1, \ldots, i_k}(x) \psi_{i_1, \ldots, i_k}(x') = \langle \psi(x), \psi(x') \rangle \ . \end{aligned}$$

Bei der mit „DG“ markierten Gleichung wurde das Distributivitätsgesetz angewendet.

Gauss–Kernel. Zu $x \in \mathbb{R}$ definieren wir einen (unendlich-dimensionalen) Merkmalsvektor $\psi(x) = (\psi_n(x))_{n \geq 0} \in \ell_2$, wobei

$$\psi_n(x) = \frac{x^n}{\sqrt{n!}} e^{-x^2/2} \ .$$

Für das innere Produkt zweier solcher Merkmalsvektoren gilt dann:

$$\begin{aligned} \langle \psi(x), \psi(x') \rangle &= \sum_{n=0}^{\infty} \left(\frac{x^n}{\sqrt{n!}} e^{-x^2/2} \frac{(x')^n}{\sqrt{n!}} e^{-(x')^2/2} \right) \\ &= \exp\left(-\frac{x^2 + (x')^2}{2} \right) \sum_{n \geq 0} \frac{(xx')^n}{n!} \\ &= \exp\left(-\frac{1}{2}(x^2 + (x')^2 - 2xx') \right) = \exp\left(-\frac{(x - x')^2}{2} \right) \end{aligned}$$

Der zu ψ passende Kernel ist demnach

$$K(x, x') = \exp\left(-\frac{(x - x')^2}{2} \right)$$

Dies ist der normierte 1-dimensionale Gauss-Kernel. Der allgemeine *d-dimensionale Gauss-Kernel mit Varianz* σ^2, definiert auf $(x, x') \in \mathbb{R}^d \times \mathbb{R}^d$, ist gegeben durch

$$\exp\left(-\frac{\|x - x'\|^2}{2\sigma} \right) \ .$$

Dieser Kernel hat Werte in der Nähe von 0 (bzw. 1), wenn x und x' weit von einander weg (bzw. nah bei einander) liegen. Mit dem Parameter σ können wir genauer skalieren, was wir unter „nah“ und „weit“ verstehen wollen. Es ist nicht schwer zu zeigen, dass der zum d-dimensionalen Gauss-Kernel passende Merkmalsvektor $\psi(x)$ für jedes $n \geq 0$ und für jeden Vektor $(i_1, \ldots, i_n) \in [d]^n$ eine Komponente

$$\psi_{i_1,\ldots,i_n} = \frac{1}{\sqrt{n!}} \exp\left(-\frac{\|x\|^2}{2\sigma}\right) \prod_{j=1}^{n} x_{i_j}$$

vorsieht. Jede polynomielle Voraussagefunktion kann demnach auch mit dem Gauss-Kernel dargestellt werden. Der Gauss-Kernel wird mitunter auch RBF-Kernel genannt. „RBF“ steht für „Radial Basis Funktions“.

Die nächsten Kernels, die wir diskutieren, gehören zur Familie der „String-Kernels“. Wir schaffen aber zunächst ein paar begriffliche Voraussetzungen.

Es A ein Alphabet (= endliche Menge). A^* bezeichne die Menge aller Wörter (strings) über A (inklusive dem leeren Wort ϵ). Das i-te Zeichen eines Wortes $x \in A^*$ notieren wir als x_i. Für ein Wort $x = x_1 \ldots x_n \in A^*$ bezeichne $|x| = n$ seine Länge. Es gilt $|\epsilon| = 0$, d.h., das leere Wort hat die Länge 0. Es sei $x \in A^*$ ein Wort der Länge n und $I \subseteq [n]$. Dann bezeichnet x_I das Wort, das wir erhalten, wenn wir in x die Buchstaben x_i mit $i \in I$ von links nach rechts aneinander reihen. Für $I = \emptyset$ setzen wir $x_I = \epsilon$. Wort $u \in A^*$ heißt *Teilsequenz* von x, falls $u = x_I$ für eine geeignete Wahl von I. Eine Teilmenge $I \subset \mathbb{N}$ der Form $I = [r : s] = \{r, r+1, \ldots, s\}$ mit $s \geq r - 1$ nennen wir *kohärent*. Die Menge $I = [r : r-1]$ identifizieren wir dabei mit $\emptyset$. Ein Wort $u \in A^*$ heißt *Teilwort* von x, falls $u = x_I$ für eine geeignete Wahl einer kohärenten Menge I. Ein Teilwort u von x heißt *Präfix* (bzw. *Suffix*) von x, falls das Wort x mit u anfängt (bzw. endet). Beachte, dass das leere Wort ε sowohl Präfix als auch Suffix (und damit auch Teilwort und erst recht Teilsequenz) eines jeden Wortes x ist.

Teilsequenz-Kernel. Zu $x \in A^*$ definieren wir einen (unendlich-dimensionalen) Merkmalsvektor $\psi(x) = (\psi_u(x))_{u \in A^*} \in \ell_2$, wobei

$$\psi_u(x) = \sum_I \mathbb{1}_{[x_I = u]} \tag{13}$$

und I rangiert über alle Teilmengen von $\mathbb{N}$, für welche x_I definiert ist (also über alle Teilmengen von $[|x|]$). Anders ausgedrückt: $\psi_u(x)$ ist die Anzahl der Vorkommen von u als Teilsequenz von x. Für das innere Produkt zweier solcher Merkmalsvektoren gilt dann:

$$\begin{aligned}
\langle \psi(x), \psi(x') \rangle &= \sum_{u \in A^*} \psi_u(x) \psi_u(x') \\
&= \sum_{u \in A^*} \left(\sum_I \mathbb{1}_{[x_I = u]} \right) \left(\sum_{I'} \mathbb{1}_{[x'_{I'} = u]} \right) \\
&= \sum_{I, I'} \sum_{u \in A^*} \mathbb{1}_{[x_I = u \wedge x'_{I'} = u]} = \sum_{I, I'} \mathbb{1}_{[x_I = x'_{I'}]} \ .
\end{aligned}$$

Der zu ψ passende Kernel ist demnach

$$K(x, x') = \sum_{I,I'} \mathbb{1}_{[x_I = x'_{I'}]} \ . \tag{14}$$

In Worten: $K(x, x')$ ist die Anzahl der Vorkommen identischer Teilsequenzpaare in (x, x'). Die Berechnung von $K(x, x')$ gemäß (14) würde exponentiell mit $|x| + |x'|$ wachsende Rechenzeit erfordern. Um zu einem effizienten Auswertungsverfahren zu gelangen, ist folgende (leicht zu verifizierende) rekursive Formel hilfreich:

$$K(x, \epsilon) = 1 \text{ und } K(x, x'a) = K(x, x') + \sum_{k:x_k = a} K(x_{[1:k-1]}, x') \ . \tag{15}$$

Aus Effizienzgründen darf man aber die Berechnung von $K(x, x')$ *nicht* rekursiv programmieren (weil dann Einträge der K-Tabelle mehrfach berechnet würden), sondern muss die obige Rekursion verwenden, um eine K-Tabelle anzulegen, welche die K-Werte für alle Präfixe von x und x' bereit hält. Kürzere Präfixe kommen dabei vor längeren zum Einsatz. Die Tabelle kann iterativ effizient ausgefüllt werden: wenn $K(u, u'a)$ gemäß (15) berechnet wird), befindet sich der Eintrag $K(u, u')$ bereits in der Tabelle. S. auch das folgende Beispiel 16.12.

Beispiel 16.12 *Wir legen eine K-Tabelle an, um den Teilsequenz-Kernel an den Worten BEERE und BERT auszuwerten:*

K	ϵ	B	BE	BEE	$BEER$	$BEERE$
ϵ	1	1	1	1	1	1
B	1	2	2	2	2	2
BE	1	2	4	6	6	8
BER	1	2	4	6	12	14
$BERT$	1	2	4	6	12	14

Der Eintrag K(BE,BEERE) = K(BEERE,BE), zum Beispiel, wird berechnet gemäß

$$K(BEERE,BE)) = K(BEERE,B) + K(B,B) + K(BE,B) + K(BEER,B) = 2 + 2 + 2 + 2 = 8$$

oder, alternativ, gemäß

$$K(BE,BEERE)) = K(BE,BEER) + K(B,BEER) = 6 + 2 = 8 \ .$$

Die 14 *Vorkommen identischer Teilsequenzpaare in (BERT,BEERE) lassen sich durch folgende Paare (I, I') spezifizieren:*

$$\begin{array}{ccccc} (\emptyset, \emptyset) & (\{1\}, \{1\}) & (\{2\}, \{2\}) & (\{2\}, \{3\}) & (\{2\}, \{5\}) \\ (\{3\}, \{4\}) & (\{1,2\}, \{1,2\}) & (\{1,2\}, \{1,3\}) & (\{1,2\}, \{1,5\}) & (\{1,3\}, \{1,4\}) \\ (\{2,3\}, \{2,4\}) & (\{2,3\}, \{3,4\}) & (\{1,2,3\}, \{1,2,4\}) & (\{1,2,3\}, \{1,3,4\}) & \end{array}$$

Teilwort-Kernel. Zu $x \in A^*$ definieren wir einen (unendlich-dimensionalen) Merkmalsvektor $\psi(x) = (\psi_u(x))_{u \in A^*} \in \ell_2$. Das Merkmal $\psi_u(x)$ ist wieder durch die Gleichung (13) gegeben, aber diesmal rangiert I nur über alle *kohärenten* Teilmengen von $\mathbb{N}$, für welche x_I definiert ist (also über alle kohärenten Teilmengen von $[|x|]$). Anders ausgedrückt: $\psi_u(x)$ ist die Anzahl der Vorkommen von u als Teilwort von x. Der dazu passende Kernel (mit $K(x, x') = \langle \psi(x), \psi(x') \rangle$) ist wieder durch die Gleichung (14) gegeben, aber freilich rangieren I und I' hier nur über *kohärente* Mengen. In Worten: $K(x, x')$ ist die Anzahl der Vorkommen identischer Teilwortpaare in (x, x'). Es bezeichne $K_0(x, x')$ die Anzahl der gemeinsamen Suffixe von x und x'. Um zu einem effizienten Auswertungsverfahren von $K(x, x')$ zu gelangen, werden folgende (leicht zu verifizierenden) rekursiven Formeln für K_0 und K verwendet:

$$K_0(x, \epsilon) = 1 \quad \text{und} \quad K_0(xa, x'b) = \begin{cases} 1 + K_0(x, x') & \text{falls } a = b \\ 1 & \text{falls } a \neq b \end{cases}$$

$$K(x, \epsilon) = 1 \quad \text{und} \quad K(x, x'a) = K(x, x') + \sum_{k : x_k = a} K_0(x_{[1:k-1]}, x') \ .$$

Diese Rekursion muss (ähnlich wie im oben diskutierten Fall von Teilsequenzen) verwendet werden, um eine K_0- und eine K-Tabelle anzulegen, welche die betreffenden Funktionswerte für alle Präfixe von x und x' bereit hält. S. auch das folgende Beispiel 16.13.

Beispiel 16.13 *Wir legen folgende Tabellen an, um den Teilwort-Kernel an den Worten BEERE und BERT auszuwerten:*

K_0	ϵ	B	BE	BEE	$BEER$	$BEERE$
ϵ	1	1	1	1	1	1
B	1	2	1	1	1	1
BE	1	1	3	2	1	2
BER	1	1	1	1	3	1
$BERT$	1	1	1	1	1	1

K	ϵ	B	BE	BEE	$BEER$	$BEERE$
ϵ	1	1	1	1	1	1
B	1	2	2	2	2	2
BE	1	2	4	5	5	6
BER	1	2	4	5	7	8
$BERT$	1	2	4	5	7	8

Der Eintrag $K(BE,BEERE) = K(BEERE,BE)$, zum Beispiel, wird berechnet gemäß

$$K(BEERE,BE)) = K(BEERE,B) + K_0(B,B) + K_0(BE,B) + K_0(BEER,B) = 2+2+1+1 = 6$$

oder, alternativ, gemäß

$$K(BE,BEERE)) = K(BE,BEER) + K_0(B,BEER) = 5 + 1 = 6 \ .$$

Die 8 *Vorkommen identischer Teilwortpaare in (BERT,BEERE) lassen sich durch folgende Paare* (I, I') *spezifizieren:*

$$\begin{array}{cccc} (\emptyset,\emptyset) & (\{1\},\{1\}) & (\{2\},\{2\}) & (\{2\},\{3\}) \\ (\{2\},\{5\}) & (\{3\},\{4\}) & (\{1,2\},\{1,2\}) & (\{2,3\},\{3,4\}) \end{array}$$

Bei einem Teilwort u von x pochen wir darauf, dass die Buchstaben von u in x unmittelbar aufeinander folgen, während bei einer Teilsequenz beliebig große Abstände zwischen dem Vorkommen von u_i und dem Vorkommen von u_{i+1} in x liegen dürfen. Die Intuition sagt einem, dass ein großer Wert von $K(x, x')$ bei dem Teilwort-Kernel aussagekräftiger ist als bei dem Teilsequenz-Kernel. Allerdings ist das Teilwortkriterium tendenziell zu strikt und durchaus vorhandene Ähnlichkeiten könnten mit dem Teilwort-Kernel evtl. übersehen werden. Es ist daher nicht verwunderlich, dass in der Literatur viele Hybride zwischen dem Teilsequenz- und dem Teilwort-Kernel zu finden sind. Diese werden mathematisch hergestellt, indem Teilsequenzen zwar grundsätzlich mitgezählt werden, aber mit einem Gewicht, welches umso höher ist je näher die Buchstaben des Wortes u in x aufeinander folgen. Die interessierte Leserin und der interessierte Leser seien auf Kapitel 11 im Buch „Kernel Methods for Pattern Analysis“ von John Shawe-Taylor und Nello Cristianini verwiesen.

16.5 Charakterisierung von Kernels

Das folgende Resultat erlaubt es, Kernels K zu designen und zu verwenden, ohne sich Gedanken um den Hilbert-Raum zu machen, in welchem die Gleichung $K(x, x') = \langle \psi(x), \psi(x') \rangle$ gilt.

Theorem 16.14 *Es sei* $\mathcal{X} = \{x_1, \ldots, x_N\}$ *eine endliche Menge und* $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *eine Funktion. Dann gilt folgendes:* K *ist ein Kernel genau dann, wenn* K*, aufgefasst als Matrix* $K \in \mathbb{R}^{\mathcal{X}\times\mathcal{X}}$*, symmetrisch und positiv semidefinit ist.*

Beweis Wir beweisen zunächst die Richtung„⇒“ und setzen also voraus, dass K ein Kernel ist. Sei dann $\psi : \mathcal{X} \to \mathcal{F}$ die Abbildung von $\mathcal{X}$ in einen Hilbert-Raum $\mathcal{F}$ mit $K(x_i, x_j) = \langle \psi(x_i), \psi(x_j) \rangle$. Eine Inspektion des Beispiels 16.5 ergibt, dass K dann symmetrisch und positiv semidefinit ist.
Den Beweis in der Richtung „⇐“ skizzieren wir lediglich. Wir definieren $\psi(x_j)$ als die Funktion von $\mathcal{X}$ nach $\mathbb{R}$, welche x_i auf $K(x_i, x_j)$ abbildet was wir in der Form $\psi(x_j) = K(\cdot, x_j)$ notieren. Es sei $\mathcal{F}_0$ der von $K(\cdot, x_1), \ldots, K(\cdot, x_N)$ aufgespannte Untervektorraum von $\mathbb{R}^{\mathcal{X}}$. Wir definieren folgende Abbildung:

$$\left\langle \sum_i \alpha_i K(\cdot, x_i), \sum_j \beta_j K(\cdot, x_j) \right\rangle = \sum_{i,j} \alpha_i \beta_j K(x_i, x_j) \ . \tag{16}$$

Es ist leicht nachzurechnen, dass die so definierte Abbildung $\langle \cdot, \cdot \rangle$ ein semi-inneres Produkt auf $\mathcal{F}_0$ ist. Zum Beispiel ergibt sich die Forderung $\langle z, z \rangle \geq 0$ für alle $z \in \mathcal{F}_0$ aus

$$\left\langle \sum_i \alpha_i K(\cdot, x_i), \sum_j \alpha_j K(\cdot, x_j) \right\rangle = \sum_{i,j} \alpha_i \alpha_j K(x_i, x_j) = \alpha^T K \alpha$$

und der (vorausgesetzten) positiven Semidefinitheit von K. Der Beweis kann dadurch abgeschlossen werden, dass $\mathcal{F}_0$ in einen *topologisch vollständigen* semi-inneren Produkt-Raum (also einen Hilbert-Raum) $\mathcal{F}$ eingebettet wird. Hierzu müssen alle Grenzwerte von Cauchy-Sequenzen in $\mathcal{F}_0$ in $\mathcal{F}$ aufgenommen werden. Wir verzichten auf die formale Durchführung dieser Konstruktion. **qed.**

Es sei $f = \sum_i \alpha_i K(\cdot, x_i)$ ein beliebiges Element des semi-inneren Produkt-Raums $\mathcal{F}_0$ aus dem Beweis von Theorem 16.14. Es folgt, dass $f(x) = \sum_i \alpha_i K(x, x_i)$. Dann gilt für alle $x \in \mathcal{X}$

$$\langle f, K(\cdot, x)\rangle = f(x) \ ,$$

wie sich durch folgende Rechnung bestätigen lässt:

$$\langle f, K(\cdot, x)\rangle = \left\langle \sum_i \alpha_i K(\cdot, x_i), K(\cdot, x) \right\rangle \overset{(16)}{=} \sum_i \alpha_i K(x_i, x) = f(x) \ .$$

Die Gleichung $\langle f, K(\cdot, x)\rangle = f(x)$ gilt sogar für alle f aus der (topologisch vollständigen) Erweiterung $\mathcal{F}$ von $\mathcal{F}_0$. Sie wird auch als die „reproducing property“ (reproduzierende Eigenschaft) des semi-inneren Produktes bezeichnet (da das semi-innere Produkt von f mit $K(x)$ den Funktionswert von f an der Stelle x reproduziert).

Der Fall eines unendlichen Grundbereiches. Eine Abbildung $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (für eine evtl. unendliche Menge $\mathcal{X}$ heiße *positiv semidefinit*, falls folgendes gilt: für jede endliche Teilmenge $\{x_1, \ldots, x_N\}$ von $\mathcal{X}$ ist die entsprechende Einschränkung von K, aufgefasst als $(N \times N)$-Matrix $(K(x_i, x_j))_{1 \le i,j \le N}$, positiv semidefinit. Ein metrischer Raum $\mathcal{X}$ heißt *separabel*, falls eine abzählbare Teilmenge $\mathcal{X}_0$ von $\mathcal{X}$ existiert, die dicht in $\mathcal{X}$ liegt (d.h., dass sich jeder Punkt in $\mathcal{X}$ als Limes einer Cauchy-Sequenz in $\mathcal{X}_0$ darstellen lässt). Wir merken ohne Beweis kurz an, dass die Charakterisierung von Theorem 16.14 unter folgenden Voraussetzungen auch für unendliche Mengen $\mathcal{X}$ gilt:

- $\mathcal{X}$ ist ein separabler metrischer Raum.
- Die Funktion $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ ist stetig.

Die Separabiltät von $\mathcal{X}$ überträgt sich dann sogar auf den im Beweis konstruierten Hilbert-Raum $\mathcal{F}$.

Theorem 16.14 kann als einfacher Spezialfall dieser allgemeineren Aussage gesehen werden: wenn wir eine endliche Menge $\mathcal{X}$ mit der diskreten Metrik

$$d(x, x') = \begin{cases} 1 & \text{falls } x \neq x' \\ 0 & \text{falls } x = x' \end{cases}$$

versehen, dann ist $\mathcal{X}$ trivialerweise separabel und *jede* Abbildung $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ ist stetig.

Der im Beweis von Theorem 16.14 (bzw. von der Verallgemenerung des Theorems auf evtl. unendliche Mengen $\mathcal{X}$) konstruierte Hilbert-Raum $\mathcal{F}$ wird auch „Reproducing Kernel Hilbert Space (RHKS)“ zum Kernel K genannt.

Zur Wahl eines Kernels. Bei einem konkreten Lernproblem sollte die Wahl des Kernels einen formalen, anwendungsunspezifischen und einen intuitiven, anwendungsgspezifischen Aspekt berücksichtigen:

- Auf der formalen Seite ist es wichtig, dass K positiv semidefinit ist, damit es sich (gemäß Theorem 16.14 bzw. dessen Verallgemeinerung) überhaupt um einen Kernel handelt.

- Auf der intuitiven Seite sollte $K(x, x')$ ein Maß für die Ähnlichkeit von x und x' sein, wobei hohe positive (bzw. negative) Werte eine große Ähnlichkeit (bzw. Unähnlichkeit) ausdrücken. Zum Beispiel könnten zwei Texte als ähnlich gelten, wenn sie viele Vorkommen von gemeinsamen Teilwortpaaren haben.