Eine mathematische Einführung zu SVMs

Florian Jarre, Math. Institut, Heinrich Heine Universität Düsseldorf

1. Februar 2026

Zusammenfassung

Es wird eine in sich geschlossene Herleitung von sogenannten "soft-margin Support Vector Machines mit Kernel" aufgezeigt, die ohne Konzepte aus der Funktionalanalysis wie dem – in diesem Zusammenhang häufig zitierten – Satz von Mercer auskommt.

1. Vorbemerkung

Für das Problem, Ja-Nein-Antworten abzuschätzen – basierend auf einer gegebenen Datenmenge mit unbekannter Strukturierung – sind sogenannte "Support Vector Machines" (SVM) ein Ansatz der unter schwachen Voraussetzungen angewendet werden kann.

Etwas genauer formuliert sind SVM Verfahren zur automatisierten Einteilung neuer Daten in zwei Klassen anhand einer Reihe von alten Daten mit zugehöriger Klassifizierung. Die Daten können dabei z.B. digitalisierte Bilder handschriftlicher Ziffern sein und die Klassifizierung besteht darin, zu entscheiden ob die Pixel der Bilder eine gegebene Ziffer darstellen oder nicht. Die alten Daten werden dabei Trainingsdaten genannt, deren Klassifzierung irgendwie vorab vorgenommen wurde, z.B. vom Menschen, der die Bilder erkennt und die zugehörige Ziffer von Hand eingibt. Sind jetzt viele verschiedene Bilder handschriftlicher Ziffern eingescannt und klassifiziert worden, so soll die SVM neue Bilder handschriftlicher Ziffern automatisch erkennen, ohne dass ein Programmierer erst Spezifizierungen der Art "eine Drei hat folgende Merkmale" eingeben muss.

SVM schätzen ja/nein-Entscheidungen ab; für kompliziertere Antworten können ggf. mehrere unterschiedliche SVM kombiniert werden; in der Regel sind dann aber andere Ansätze sinnvoll. Trotz der eingeschränkten Form der Antwort können SVM bei komplizierten Entscheidungsproblemen ein durchaus hilfreicher Ansatz sein.

1.1 Einschränkungen zu SVM

Anwendungen von SVM sind z.B. das automatische Klassifizieren von Bildern oder die Aufgabe, aus einer Datenbank mit Patientendaten abzulesen, ob ein neuer Patient in eine Risikogruppe für eine spezielle Krankheit fällt – oder anhand einer anderen Datenbak zu entscheiden, ob einem Kunden ein Kredit gewährt werden soll. Diese beiden Beispiele zeigen ein

Problem auf, das in Anwendungen von Verfahren der KI wiederholt aufgetreten ist und z.B. in [7] beschrieben ist. Wenn die SVM genutzt wird um bei Patienten ein Risiko zu erkennen, das sonst leicht übersehen wird, so dient die SVM dem Wohl des Menschen, nicht aber, wenn sie genutzt wird um eine Krankenversicherung oder einen Kredit ohne weitere Untersuchung abzulehnen.

Ein anderes Problem tritt auf, wenn Trainingsdaten selber automatisch klassifiziert wurden. Dabei unterlaufene Klassifizierungsfehler setzen sich dann in der Regel in der damit entwickelten SVM fort. Letzteres gilt natürlich auch, wenn die Trainigsdaten vom Menschen falsch klassifiziert wurden.

Auch wenn die gegebenen Daten korrekt klassifiziert wurden, so sind sie häufig nicht ausreichend um daraus eine eindeutige Klassifizierung abzuleiten. Trotzdem wird letzteres angestrebt.

Ein weiteres Problem tritt auf, wenn die Trainingsdaten nicht gleichverteilt erhoben wurden. Wenn die handgeschriebenen Ziffern aus obigem Beispiel in den USA gesammelt wurden (wo die Ziffern 1 und 7 anders geschrieben werden als in Deutschland) so ist bei Anwendung einer dazu in den USA entwickelten SVM in Deutschland eine erhöhte Fehlerquote zu erwarten. (Dies ist ein eher harmloses Beispiel!)

Auf derart Modellierungs- und Interpretationsfehler soll im Folgenden nicht weiter eingegangen werden sondern der Mechanismus der SVM erklärt werden.

1.2 Ausblick

Die Grundidee ist, dass ähnliche Daten derselben Klassifizierung zugeordnet werden. Allerdings ist der Begriff "ähnliche Daten" sehr unpräzise, so können die Pixel zu zwei scans von handschriftlichen Ziffern z.B. komplett verschieden sein auch wenn dieselbe Ziffer dargestellt ist. Ein Problem, das SVM idealerweise automatisiert lösen, ist es, ein geeignetes Kriterium für "Ähnlichkeit" aus den Daten abzuleiten und zu nutzen.

Die Grundlagen zu SVM sind gut erforscht und verstanden, siehe z.B. [9, 6, 2, 10] und die Referenzen dort. Nachfolgend soll eine einführende mathematische Zusammenfassung angegeben werden. Die hier gewählte Darstellung ist "minimalisitisch" in dem Sinn, dass auf einige Konzepte, die häufig bei der Betrachtung von Kernel-Funktionen genutzt werden, verzichtet wird, und auch die Ergebnisse aus der Statistical Learning Theory nicht aufgegriffen werden. Auch werden die Vollständigkeit des "Feature"-Raums und damit verbundene Sätze wie der Darstellungssatz von Riesz oder der Satz von Mercer nicht benutzt. Trotzdem kann eine zentrale Eigenschaft, die Stetigkeit der Kernel-Funktionen, begründet und analysiert werden.

1.3 Notation

Anstelle des aus dem Englischen übernommen Wortes Kernel-Funktion wird gelegentlich auch kurz das Wort "Kern" benutzt. $A \succeq 0$ bezeichnet dass A eine symmetrische positiv semidefinite Matrix ist. Seien zwei Matrizen $A, B \in \mathbb{R}^{m \times n}$ gegeben. Dann ist deren Hada-

mardprodukt $A \circ B$ durch komponentenweises Ausmultiplizieren definiert,

$$(A \circ B)_{ij} = A_{ij} B_{ij} \qquad \text{für } 1 \le i \le m, 1 \le j \le n.$$

Zu einer Matrix A bezeichnet $\|A\|_2$ die von der 2-Norm induzierte Matrix-Norm. Der Vektor e ist stets der Vektor mit lauter Einsen, $e = (1, \ldots, 1)^T$. Das Landau-Symbol $f = o(g)$ besagt, dass $\lim_{g \to 0} |f|/g = 0$. Die Ableitung $Df(x)$ einer differenzierbaren Funktion $f : \mathbb{R}^n \to \mathbb{R}^m$ wird als $m \times n$-Matrix dargestellt (die Jacobi-Matrix), und im Fall $m = 1$ ist der Gradient $\nabla f(x) = Df(x)^T$ ein Spaltenvektor.

2. Grundform der Support Vector Machines

Die Ausgangssituation ist wie folgt. Gegeben sind Trainingsdaten durch Punkte $x^{(i)}$ aus einer kompakten konvexen Menge $\Omega \subset \mathbb{R}^n$ für $1 \le i \le m$ und zugehörige Klassifizierungen $\zeta_i \in \{-1, 1\}$.

Der einfachste Fall ist dabei der, dass es eine Hyperebene $\{x \in \mathbb{R}^n \mid a^T x = \beta\}$ mit einem festen Vektor $a \in \mathbb{R}^n \backslash \{0\}$ und einer Konstanten $\beta \in \mathbb{R}$ gibt, sodass

$$\begin{aligned} a^T x^{(i)} &> \beta \quad \forall\, i \text{ mit } \zeta_i = 1 \text{ und} \\ a^T x^{(i)} &< \beta \quad \forall\, i \text{ mit } \zeta_i = -1. \end{aligned} \tag{1}$$

In diesem Fall nennen wir die Hyperebene $\{x \mid a^T x = \beta\}$ auch klassifizierend, da sie alle Datenpunkte "korrekt trennt". Das in der Vorbemerkung angesprochene Kriterium der "Ähnlichkeit" von Daten reduziert sich hier also nur auf die Frage ob $a^T x > \beta$ gilt oder nicht.

Bei der späteren Untersuchung von Stetigkeitseigenschaften spielt auch die Skalierung der Daten eine Rolle. So ist offensichtlich, dass für $\lambda > 0$ beim Übergang von

$$x^{(i)} \text{ und } \beta \quad \text{zu} \quad \lambda x^{(i)} \text{ und } \lambda\beta \quad \text{für} \quad 1 \le i \le m \tag{2}$$

die Klassifizierung (1) unverändert bleibt. Es wird daher später vorausgesetzt, dass die Daten so skaliert sind, dass $\|x\|_2$ für $x \in \Omega$ durch eine moderate Konstante beschränkt sei. Zunächst ist diese Annahme aber unerheblich.

Um die "Aussicht" zu maximieren, dass die gewählte Hyperebene neue Punkte korrekt klassifiziert, soll die Hyperebene nun so gewählt werden, dass sie die gegebenen Punkte einerseits korrekt trennt, und dass sie andererseits von allen gegebenen Punkten möglichst weit entfernt liegt – d.h. dass möglichst keiner der Punkte ein Grenzfall ist, der bei geringer Störung in die andere Klasse wechseln würde. Die in diesem Sinne beste Hyperebene zu finden soll nachfolgend automatisiert erfolgen.

Zunächst schreiben wir dazu die Bedingungen (1) äquivalent in der kompakten Form

$$\delta_i := \zeta_i (a^T x^{(i)} - \beta) > 0 \quad \forall\, i. \tag{3}$$

Falls nun ein Punkt $x^{(i)}$ die Bedingung (3) erfüllt und eine Zahl $\lambda \geq 1/\delta_i$ gewählt ist, so erfüllt $x^{(i)}$ auch die Bedingung

$$\zeta_i(\lambda a^T x^{(i)} - \lambda\beta) \geq 1.$$

Durch Übergang von a zu λa und β zu $\lambda\beta$ können also die scharfen Ungleichungen in (3) durch die schwachen Ungleichungen

$$\zeta_i(a^T x^{(i)} - \beta) \geq 1 \quad \forall\ i \tag{4}$$

mit rechter Seite 1 ersetzt werden. Schließlich sei daran erinnert, dass der Abstand eines Punktes $\bar{x}$ von der Hyperebene $\{x \mid a^T x = \beta\}$ durch $|a^T\bar{x} - \beta|/\|a\|_2$ gegeben ist. Die Maximierung des Abstands von $\bar{x}$ zur Hyperebene unter der Voraussetzung, dass $|a^T x - \beta| \geq 1$ gilt, ist daher äquivalent zur Minimierung der Norm von a. Somit lässt sich das Problem, den minimalen Abstand aller Punkte zu einer klassifizierenden Hyperebene zu maximieren in der Form

$$\min_{a,\ \beta}\{\ \tfrac{1}{2}\|a\|_2^2 \ \mid\ \zeta_i(a^T x^{(i)} - \beta) \geq 1 \quad \forall\ 1 \leq i \leq m\ \} \tag{5}$$

schreiben. Hier wurde die übliche Schreibweise benutzt, dass der Ausdruck "min" in Problem (5) im Sinn von "minimiere" zu verstehen ist; wenn die Daten so sind, dass es gar keine trennende Hyperebene gibt, so existiert das Minimum nicht, andernfalls ist es aber eindeutig bestimmt. Wenn nun a, β aus (5) optimal bestimmt sind, so kann das Label $\tilde{\zeta}$ zu einem neuem Punkt $\tilde{x}$ durch

$$\tilde{\zeta} := \operatorname{sign}(a^T\tilde{x} - \beta)$$

geschätzt werden, denn dann erfüllt auch $(\tilde{x}, \tilde{\zeta})$ die Beziehung $\tilde{\zeta}(\tilde{a}^T\tilde{x} - \beta) \geq 0$.

Bei der Lösung von (5) erweisen sich in der Regel viele der Nebenbedingungen $\zeta_i(a^T x^{(i)} - \beta) \geq 1$ als überflüssig. Nur die Punkte mit den kleinsten Werten $|a^T x^{(i)} - \beta|$ sind für die Bestimmung der optimalen Hyperebene relevant. Diejenigen Trainingspunkte, die nicht überflüssig sind, d.h. die minimalen Abstand zur Hyperebene haben, werden "support vectors" oder Stützvektoren genannt und erklären den Namen der SVM.

3. Soft Margin SVM

Häufig liegt auch die Situation vor, dass die gegebenen Daten gar nicht exakt durch eine Hyperebene getrennt werden können, weil z.B. nicht alle Trainingsdaten korrekt klassifiziert wurden. In diesem Fall kann man eine sogenannte "soft margin" SVM (SVM mit "weichem Rand") nutzen, bei der die Restriktionen zu $\zeta_i(a^T x^{(i)} - \beta) \geq 1 - s_i$ mit $s_i \geq 0$ abgeschwächt werden und der Ausdruck

$$\tfrac{1}{2}\|a\|_2^2 + C\sum_{i=1}^{m} s_i \tag{6}$$

für ein festes $C > 0$ minimiert wird[1]. Bei großem C werden also vorrangig die $s_i \geq 0$ minimiert, (in der Hoffnung, dass nur die inkompatiblen, falsch klassifizierten Datenpunkte $x^{(i)}$

[1]Eine andere Formulierung der soft margin ist z.B. in [2] besprochen.

positive Werte $s_i > 0$ behalten) und nachrangig die Norm von a, deren Inverses den Abstand der Stützvektoren zur trennenden Hyperebene beschreibt. Setzt man $e := (1, 1, \ldots, 1)^T \in \mathbb{R}^m$, so ist die Summe in (6) gegeben durch $\sum_{i=1}^m s_i = e^T s$ und insgesamt ergibt sich das Problem: Finde μ^* und a, β, s mit

$$\mu^* = \min_{a,\ \beta,\ s} \{\ \tfrac{1}{2}\|a\|_2^2 + Ce^T s\ \mid\ \zeta_i(a^T x^{(i)} - \beta) \geq 1 - s_i,\ \ \forall\ 1 \leq i \leq m,\ \ s \geq 0\}. \tag{7}$$

Die Lösung von (7) lässt sich wie folgt umformen. Sei L die Lagrange-Funktion zu (7), d.h.

$$L((a, \beta, s), (u, v)) := \tfrac{1}{2}\|a\|_2^2 + Ce^T s + \sum_{i=1}^m u_i \underbrace{(1 - s_i + \zeta_i\beta - \zeta_i a^T x^{(i)})}_{\leq 0 \text{ in } (7)} + v^T \underbrace{(-s)}_{\leq 0 \text{ in } (7)}$$

für sogenannte Lagrange-Multiplikatoren $u, v \geq 0$. Damit lässt sich Problem (7) schreiben als

$$\mu^* := \inf_{a,\ \beta,\ s} \left(\sup_{u \geq 0,\ v \geq 0} L((a, \beta, s), (u, v)) \right)$$

denn in der Infimumbildung werden nur solche (a, β, s), ausgewählt, für die das Supremum endlich ist, und das sind genau die, für die die Nebenbedingungen aus (7) erfüllt sind.

(Wäre z.B. $s_i < 0$ für ein i, so hätte das innere Supremum mit $u_i \to \infty$ für dieses i den Wert $+\infty$. Daher werden bei der Infimumbildung nur Vektoren $s \geq 0$ berücksichtigt. Analog werden nur (a, β, s) berücksichtigt, für die $1 - s_i + \zeta_i\beta - \zeta_i a^T x^{(i)} \leq 0$ gilt.)

Problem (7) ist ein konvexes Problem, und da nur lineare Nebenbedingungen vorliegen ist die Slater-Bedingung trivial erfüllt, sodass die Lagrange-Dualität gilt,

$$\inf_{a,\ \beta,\ s} \sup_{u \geq 0,\ v \geq 0} L((a, \beta, s), (u, v)) = \sup_{u \geq 0,\ v \geq 0} \inf_{a,\ \beta,\ s} L((a, \beta, s), (u, v)). \tag{8}$$

(Dass die linke Seite in (8) größer oder gleich der Rechten ist, lässt sich elementar herleiten, dass bei konvexen Optimierungsproblemen mit linearen Restriktionen beide Seiten gleich sind ist ein Standard-Resultat der Optimierung, siehe z.B. den Beweis von Satz 8.3.4 (3) in [4].) Das innere Problem rechts (die Infimumbildung) hat nun keine Nebenbedingungen mehr und kann aufgrund der Konvexität daher explizit gelöst werden, indem die Ableitung gleich Null gesetzt wird (für gegebene $u, v \geq 0$).

Schreibt man die Lagrangefunktion äquivalent als

$$L((a, \beta, s), (u, v)) = \tfrac{1}{2}\|a\|_2^2 - \left(\sum_{i=1}^m u_i \zeta_i x^{(i)} \right)^T a + \left(\sum_{i=1}^m u_i \zeta_i \right) \beta + (Ce - u - v)^T s + e^T u$$

so ergibt sich für die Minimalstelle des Infimum-Problems der rechten Seite in (8) durch Null-Setzen der Ableitungen bezüglich (a, β, s) dass

$$a = \sum_{i=1}^m u_i \zeta_i x^{(i)}, \quad \sum_{i=1}^m u_i \zeta_i = 0, \quad \text{und} \quad Ce - u - v = 0 \tag{9}$$

erfüllt sein müssen. Mit diesen Bedingungen verschwinden die Terme

$$(\sum_{i=1}^{m} u_i\zeta_i)\beta + (Ce - u - v)^T s$$

in der Lagrangefunktion und die erste Gleichung im (9) besagt, dass die ersten beiden Terme in der Lagrangefunktion sich reduzieren zu

$$\tfrac{1}{2}\|a\|_2^2 - (\sum_{i=1}^{m} u_i\zeta_i x^{(i)})^T a = -\tfrac{1}{2}\|\sum_{i=1}^{m} u_i\zeta_i x^{(i)}\|_2^2,$$

wobei die Variable a eliminiert werden konnte.

Die Bedingungen (9) sind gleichwertig zur inneren Infimum-Bildung der rechten Seite von (8) und können daher als Nebenbedingungen an die Supremumbildung formuliert werden,

$$\mu^* = \sup_{u\geq 0,\ v\geq 0} \left\{ -\tfrac{1}{2}\|\sum_{i=1}^{m} u_i\zeta_i x^{(i)}\|_2^2 + e^T u \ \mid\ u^T\zeta = 0,\ Ce - u - v = 0 \right\}. \tag{10}$$

Der "Schlupf-Vektor" $v \geq 0$ besagt nur, dass $Ce - u \geq 0$ gelte. Er kann oben noch eliminiert werden und mit einem Vorzeichenwechsel in der Zielfunktion ergibt sich

$$-\mu^* = \inf_{u\in\mathbb{R}^m} \left\{ \tfrac{1}{2}\|\sum_{i=1}^{m} u_i\zeta_i x^{(i)}\|_2^2 - e^T u \ \mid\ u^T\zeta = 0,\ Ce \geq u \geq 0 \right\}. \tag{11}$$

(Die Nebenbedingung $u \geq 0$ aus dem Term "$\sup_{u\geq 0}$" in (10) ist rechts in (11) wieder explizit aufgeführt.) Problem (11) wird auch *duales Problem* zu (7) bezeichnet.

In aller Regel sind auch in der ursprünglichen Formulierung (7) viele Datenpunkte $x^{(i)}$ überflüssig, d.h. $\zeta^i(a^T x^{(i)} - \beta) > 1$ in jeder Optimallösung a, β, s von (7). Diese $x^{(i)}$ und die zugehörigen s_i können dann in der Lagrangefunktion einfach weggelassen werden, bzw. die zugehörigen Multiplikatoren u_i können auf Null fixiert werden. Und diese $x^{(i)}$ sind dann auch in der äquivalenten Umformung (11) überflüssig d.h. die zugehörigen Multiplikatoren u_i in einer Optimallösung von Problem (11) sind Null. Bezeichnet man mit $\mathcal{B}$ die Indices, für die $u_i > 0$ in der gegebenen Optimallösung gilt, so folgt mit (9), dass

$$a = \sum_{i\in\mathcal{B}} u_i\zeta_i x^{(i)}. \tag{12}$$

Die Stützvektoren sind hier also $\{x^{(i)}\}_{i\in\mathcal{B}}$. Wenn nun a wie oben aus der Lösung von (11) berechnet ist, so lässt sich das zugehörige β anhand von (7) ermitteln: Für gegebenes β und a ist das zugehörige optimale s in (7) durch die Lösung von

$$\min\ Ce^T s \ \mid\ s_i \geq 1 + \zeta_i(\beta - a^T x(i)) \quad \forall\ 1 \leq i \leq m, \quad s \geq 0$$

gegeben. Setzt man $\zeta \in \mathbb{R}^m$ den Vektor mit Komponenten ζ_i und $b \in \mathbb{R}^m$ den Vektor mit Komponenten $b_i := 1 - \zeta_i a^T x^{(i)}$ so lassen sich die obigen Ungleichungen an die Variablen

s_i schreiben als $s_i \geq 1 - \zeta_i a^T x^{(i)} + \beta\zeta_i = b_i + \beta\zeta_i$. Daher ist zu gegebenem β (und a) das optimale $s \in \mathbb{R}^m$ in (7) explizit darstellbar als

$$s(\beta) = \max\{b + \beta\zeta,\ 0\},$$

wobei das Maximum komponentenweise angewendet wird. Die Abbildung

$$\beta \mapsto e^T s(\beta) = e^T(\max\{b + \beta\zeta,\ 0\}) =: \sigma(\beta)$$

ist stückweise linear, als Maximum linearer Funktionen auch konvex, und der Wert von β, der $e^T s(\beta)$ minimiert löst somit auch (7),

$$\beta = \text{argmin}_{\hat{\beta}\in\mathbb{R}}\ \{e^T \max\{b + \hat{\beta}\zeta,\ 0\}\}.$$

Die Minimierung der konvexen stückweise linearen Funktion $\sigma : \mathbb{R} \to \mathbb{R}$ ist mit minimalem Rechenaufwand möglich.

Die Klassifizierung eines neuen Datenpunktes $\tilde{x}$ ergibt sich dann durch

$$\tilde{\zeta} := \text{sign}(\tilde{x}^T a - \beta) = \text{sign}(\tilde{x}^T(\sum_{i\in\mathcal{B}} u_i\zeta_i x^{(i)}) - \beta) = \text{sign}((\sum_{i\in\mathcal{B}} u_i\zeta_i \tilde{x}^T x^{(i)}) - \beta).$$

Zur Herleitung der Kernel-SVM sei noch eine Umformulierung der Zielfunktion in (11) aufgezeigt. Dazu sei $Z := \text{Diag}(\zeta)$ die Diagonalmatrix mit Diagonale $\zeta \in \mathbb{R}^m$. Der quadratische Term in der Zielfunktion von (11) lässt sich dann schreiben als

$$\|\sum_{i=1}^{m} u_i\zeta_i x^{(i)}\|_2^2 = u^T(ZQZ)u \tag{13}$$

mit der Matrix Q mit Einträgen $Q_{i,j} = (x^{(i)})^T x^{(j)}$. Als Gramsche Matrix d.h.

$$Q = (x^{(1)}, \ldots, x^{(m)})^T(x^{(1)}, \ldots, x^{(m)}) \in \mathbb{R}^{m\times m}$$

ist Q dabei symmetrisch positiv semidefinit, $u^T Qu \geq 0\ \forall u \in \mathbb{R}^m$, in Zeichen,

$$Q \succeq 0.$$

4. Kernel SVM

Nun gibt es auch viele Anwendungen, bei denen die Daten in der gegebenen Form grundsätzlich nicht durch eine Hyperebene getrennt werden können, d.h., dass der "Rand" der die beiden Klassen trennt, keine Hyperebene ist, sondern ein etwas komplizierterer nichtlinearer Rand. In diesen Fällen sucht man eine nichtlineare Abbildung

$$\phi : \Omega \to \mathcal{W},$$

die die $x^{(i)} \in \Omega$ so in einen meist höherdimensionalen Skalarproduktraum $\mathcal{W}$ (der z.B. auch ein Funktionenraum sein kann) mit Skalarprodukt $\langle ., . \rangle_{\mathcal{W}}$ und der induzierten Norm $\| \, . \, \|_{\mathcal{W}}$ abbildet, $x^{(i)} \mapsto \phi(x^{(i)})$, dass die Bilder der beiden Klassen durch eine Hyperebene getrennt werden können. Mit der Wahl von ϕ soll einerseits die eingangs angesprochene "Ähnlichkeit" von Daten bewahrt werden, d.h. dass ϕ gewisse Stetigkeitseigenschaften erfüllen möge, und zum anderen die unbekannte "Trennung" in zwei Klassen besser erfassbar werden. Unter Beibehaltung der "soft margin" erhält man anstelle von (7) dann das Problem

$$\min_{\tilde{a},\ \beta,\ s \geq 0} \{ \ \tfrac{1}{2} \|\tilde{a}\|_{\mathcal{W}}^2 + Ce^T s \ \mid \ \zeta_i \left(\langle \tilde{a}, \phi(x^{(i)}) \rangle_{\mathcal{W}} - \beta \right) \geq 1 - s_i, \quad \forall \ 1 \leq i \leq m \} \tag{14}$$

wobei die Vektoren $\tilde{a}$ nun in dem höher-dimensionalen Bildraum $\mathcal{W}$ von ϕ liegen. (Dass das Minimum tatsächlich existiert wird nachfolgend begründet.) Durch die Nichtlinearität der Abbildung ϕ übersetzt sich die lineare Trennung im Bildraum von ϕ in eine nichtlineare Trennung im ursprünglichen Datenraum mit den Daten $x^{(i)}$.

Sei nun die Optimallösung $\tilde{a}, \beta$ aus (14) gegeben. Für einen neuen Datenpunkt $\tilde{x}$ wird das Label $\tilde{\zeta}$ dann geschätzt durch

$$\tilde{\zeta} := \operatorname{sign} \left(\langle \, \tilde{a}, \phi(\tilde{x}) \rangle_{\mathcal{W}} - \beta \right). \tag{15}$$

Auch wenn die Dimension von $\mathcal{W}$ unendlich sein sollte, so ist Problem (14) trotzdem ein endlichdimensionales Optimierungsproblem wie kurz begründet werden soll:

Dazu sei $M := \operatorname{Spann}\{\phi(x^{(i)})\}_{1 \leq i \leq m} \subset \mathcal{W}$ (die lineare Hülle der $\phi(x^{(i)})$ für $1 \leq i \leq m$). Dann ist $\langle ., . \rangle_{\mathcal{W}}$ auch ein Skalarprodukt auf M und somit ist für beliebig fest gewähltes $\tilde{a} \in \mathcal{W}$ die Funktion $\psi : M \to \mathbb{R}$ mit $\psi(x) := \|\tilde{a} - x\|_{\mathcal{W}}^2$ streng konvex. Als konvexe Funktion auf dem endlich-dimensionalen Raum M ist ψ auch stetig. Ferner ist für $\bar{x} \in M$ die Niveaumenge $\{x \in M \mid \psi(x) \leq \psi(\bar{x})\}$ beschränkt. Also hat ψ eine eindeutige Minimalstelle auf M, die mit $\tilde{a}_M$ bezeichnet sei. (Für diese Aussage wird die Vollständigkeit von $\mathcal{W}$ nicht benötigt.) Für $\lambda \in \mathbb{R}$ und festes $i \in \{1, \ldots, m\}$ folgt aus der Definition von M und $\tilde{a}_M$ dass

$$\|\tilde{a} - \tilde{a}_M\|_{\mathcal{W}}^2 \leq \|\tilde{a} - \tilde{a}_M + \lambda \phi(x^{(i)})\|_{\mathcal{W}}^2 = \|\tilde{a} - \tilde{a}_M\|_{\mathcal{W}}^2 + 2\lambda \langle \tilde{a} - \tilde{a}_M, \phi(x^{(i)}) \rangle_{\mathcal{W}} + \lambda^2 \|\phi(x^{(i)})\|_{\mathcal{W}}^2,$$

d.h. $0 \leq 2\lambda \langle \tilde{a} - \tilde{a}_M, \phi(x^{(i)}) \rangle_{\mathcal{W}} + \lambda^2 \|\phi(x^{(i)})\|_{\mathcal{W}}^2$ für $\lambda \in \mathbb{R}$. Setzt man

$$\lambda := \begin{cases} -\langle \tilde{a} - \tilde{a}_M, \phi(x^{(i)}) \rangle_{\mathcal{W}} & \text{falls} \quad \|\phi(x^{(i)})\|_{\mathcal{W}}^2 = 0, \\ -\langle \tilde{a} - \tilde{a}_M, \phi(x^{(i)}) \rangle_{\mathcal{W}} / \|\phi(x^{(i)})\|_{\mathcal{W}}^2 & \text{sonst}, \end{cases}$$

so folgt aus $\langle \tilde{a} - \tilde{a}_M, \phi(x^{(i)}) \rangle_{\mathcal{W}} \neq 0$, ein Widerspruch. Also ist $\langle \tilde{a} - \tilde{a}_M, \phi(x^{(i)}) \rangle_{\mathcal{W}} = 0$.

Setzt man $\tilde{a}_{M^\perp} := \tilde{a} - \tilde{a}_M$ so folgt

$$\langle \tilde{a}_{M^\perp}, \phi(x^{(i)}) \rangle_{\mathcal{W}} = 0 \ \ (1 \leq i \leq m), \qquad \text{und deswegen auch} \qquad \langle \tilde{a}_{M^\perp}, \tilde{a}_M \rangle_{\mathcal{W}} = 0. \tag{16}$$

Sei nun $\tilde{a}$ zulässig für (14), so folgt aus (16)

$$\|\tilde{a}\|_{\mathcal{W}}^2 = \|\tilde{a}_{M^\perp} + \tilde{a}_M\|_{\mathcal{W}}^2 = \|\tilde{a}_M\|_{\mathcal{W}}^2 + \|\tilde{a}_{M^\perp}\|_{\mathcal{W}}^2$$

und $\langle \tilde{a}, \phi(x^{(i)})\rangle_{\mathcal{W}} = \langle \tilde{a}_M, \phi(x^{(i)})\rangle_{\mathcal{W}}$ d.h. $\tilde{a}_M$ ist ebenfalls zulässig für (14) und der Zielfunktionswert ist allenfalls kleiner, sodass (14) äquivalent ist zu

$$\min_{\tilde{a}_M \in M,\ \beta,\ s\geq 0} \{\ \tfrac{1}{2}\|\tilde{a}_M\|^2_{\mathcal{W}} + Ce^T s\ \mid\ \zeta_i \left(\langle \tilde{a}_M, \phi(x^{(i)})\rangle_{\mathcal{W}} - \beta\right) \geq 1 - s_i,\ \ \forall\ 1 \leq i \leq m\}. \quad (17)$$

Dieses Problem ist endlichdimensional und hat dieselbe Struktur wie Problem (7).[2] Es lässt sich daher analog wie in (11) mit der Zielfunktion aus (13) umformen, wobei sich das duale Problem

$$\inf_{u\in\mathbb{R}^m} \left\{\tfrac{1}{2}u^T ZKZu - e^T u\ \mid\ u^T\zeta = 0,\ Ce \geq u \geq 0\right\} \quad (18)$$

ergibt, mit dem Term $u^T ZKZu$ anstelle des Terms $u^T ZQZu$ in (13). Dabei sind die Einträge von K gegeben durch $K_{i,j} = \langle \phi(x^{(i)}), \phi(x^{(j)})\rangle_{\mathcal{W}}$. Genau wie die obige Matrix $Q \in \mathbb{R}^{m\times m}$ ist auch $K \in \mathbb{R}^{m\times m}$ als Gramsche Matrix positiv semidefinit, $K \succeq 0$, und hängt nicht von der Dimension des Bildraums $\mathcal{W}$ ab — wohl aber von der Anzahl der Stützpunkte m.

Zur Festlegung einer geeigneten Funktion ϕ wird der sogenannte Kernel-Trick angewendet, indem man anstelle der Transformation ϕ nur eine symmetrische stetige Abbildung

$$\kappa : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$$

so definiert, dass man $\kappa(x, y)$ als $\kappa(x, y) = \langle \phi(x), \phi(y)\rangle_{\mathcal{W}}$ für eine Funktion ϕ interpretieren könnte. Dazu fordert man, dass für alle $m \in \mathbb{N}$ und alle $x^{(1)}, \ldots, x^{(m)}$ aus Ω die Matrizen $K \in \mathbb{R}^{m\times m}$ mit den Einträgen $K_{i,j} = \kappa(x^{(i)}, x^{(j)})$ für $1 \leq i, j \leq m$ stets $K \succeq 0$ erfüllen. In diesem Fall nennt man κ einen[3]

$$\textit{positiv definiten Kern.} \quad (19)$$

Zusammenfassend bildet man also die Matrix K mit $K_{i,j} := \kappa(x^{(i)}, x^{(j)})$ für $1 \leq i, j \leq m$ und löst (18). Mit der Optimallösung u setzt man $\mathcal{B} := \{i \mid u_i > 0\}$ und definiert wie in (12) die Optimallösung $\tilde{a} = \sum_{i\in\mathcal{B}} u_i\zeta_i\phi(x^{(i)})$ von (14). Diese wird nicht explizit benötigt (d.h. die Funktion ϕ nicht explizit ausgewertet). Die Klassifizierung eines neuen Datenpunktes $\tilde{x}$ erfolgt nämlich nach der Vorschrift

$$\tilde{\zeta} = \operatorname{sign}(\langle \tilde{a}, \phi(\tilde{x})\rangle_{\mathcal{W}} - \beta) = \operatorname{sign}(\langle \sum_{i\in\mathcal{B}} u_i\zeta_i\phi(x^{(i)}), \phi(\tilde{x})\rangle_{\mathcal{W}} - \beta) = \operatorname{sign}(\sum_{i\in\mathcal{B}} u_i\zeta_i\kappa(x^{(i)}, \tilde{x}) - \beta),$$

wobei sich β unter Ausnutzung von $\langle \tilde{a}, \phi(x^{(i)})\rangle_{\mathcal{W}} = \sum_{j\in\mathcal{B}} u_j\zeta_j\kappa(x^{(i)}, x^{(j)})$ wieder wie bei der soft-margin SVM ergibt: Man setzt $\zeta \in \mathbb{R}^m$ den Vektor mit Komponenten ζ_i und $b \in \mathbb{R}^m$ als den Vektor mit Komponenten $1 - \zeta_i \sum_{j\in\mathcal{B}} u_j\zeta_j\kappa(x^{(i)}, x^{(j)})$. Dann ist wieder

$$\beta = \operatorname{argmin}\ \{e^T \max\{b + \hat{\beta}\zeta,\ 0\} \mid \hat{\beta} \in \mathbb{R}\}.$$

[2] Weil das Skalarprodukt $\langle\ .\ ,\ .\ \rangle_{\mathcal{W}}$ in (15) aber auf beliebige $\tilde{x} \in \Omega$ angewendet wird ist es sinnvoll, das Problem (14) über dem allgemeinen Raum $\mathcal{W}$ zu formulieren und sich nicht von vorne herein auf die endlich-dimensionale Formulierung (17) einzuschränken.

[3] Zunächst wäre hier die Notation "positiv semidefiniter Kern" angebracht. Dass der Kern in einem gewissen Raum eine Norm definiert, die die Notation "positiv definiter Kern" rechtfertigt, wird nachfolgend in (24) begründet.

Im nächsten Absatz wird die Frage betrachtet, ob es zu gegebenem positiv definiten Kern $\kappa : \Omega \times \Omega \to \mathbb{R}$ denn auch eine Funktion $\phi : \Omega \to \mathcal{W}$ gibt, so dass zu gegebenen $x^{(i)} \in \Omega$ für $1 \leq i \leq m$ stets gilt

$$\kappa(x^{(i)}, x^{(j)}) = \langle \phi(x^{(i)}), \phi(x^{(j)}) \rangle_{\mathcal{W}} \quad \text{für} \quad 1 \leq i, j \leq m. \tag{20}$$

Da die Anzahl der Gleichungen in (20), die von ϕ zu erfüllen sind, mit m wächst und für m keine obere Schranke festgelegt ist, ist es naheliegend, für die Freiheitsgrade von ϕ, d.h. für die Dimension von $\mathcal{W}$, einen unendlichen Wert zuzulassen.

4.1 Interpretation im Feature Space $\mathcal{W}$

Der Bildraum $\mathcal{W}$ obiger Funktion ϕ wird auch "feature space" (frei übersetzt: Merkmals-Raum) genannt; in ihm wird die lineare Trennung der beiden Klassen vorgenommen. Dabei ist zu gegebenem κ weder die Abbildung ϕ noch deren Bildraum $\mathcal{W}$ eindeutig. Nachfolgend werden die Existenz und wünschenswerte Eigenschaften von ϕ betrachtet.

4.1.1 Existenz

Wenn die Reihenfolge der Argumentation zu (20) umgekehrt wird, und zu gegebenem κ und zu ("zuerst") gegebenen Punkten $x^{(i)}$ eine Funktion ϕ gesucht ist, so lässt sich die Existenz von ϕ einfach begründen: Für $K \succeq 0$ gibt es eine Eigenwertzerlegung, $K = U^T D U$ mit einer orthogonalen Matrix U mit Spalten $u^{(i)}$ und einer Diagonalmatrix D. Setzt man $\tilde{u}^{(i)} := D^{1/2} u^{(i)}$, so gilt $K_{i,j} = (\tilde{u}^{(i)})^T \tilde{u}^{(j)}$. Man kann also eine beliebige Abbildung $\phi : \mathbb{R}^n \to \mathbb{R}^m =: \mathcal{W}$ so definieren, dass für $1 \leq i \leq m$ die Interpolationsbedingungen $\phi(x^{(i)}) = \tilde{u}^{(i)}$ erfüllt sind und damit auch die gewünschte Beziehung $K_{i,j} = \langle \phi(x^{(i)}), \phi(x^{(j)}) \rangle_{\mathcal{W}}$. Hier wird ausgenutzt, dass die Dimension von $\mathcal{W}$ frei wählbar vorausgesetzt wurde. Im Vergleich zu (20) ist hier aber die Reihenfolge umgekehrt: Die Punkte $x^{(i)}$ werden benutzt um die Funktion ϕ zu definieren. Warum dieses Vorgehen problematisch ist wird nachfolgend an einem einfachen Beispiel erläutert.

4.1.2 Überanpassung

Wenn z.B. viele Messwerte (x_i, y_i) einer Funktion $f : \mathbb{R} \to \mathbb{R}$, $x \mapsto y$ gegeben sind, die näherungsweise auf einer Geraden liegen und der Wert von f an einer "neuen" Stelle $\tilde{x}$ geschätzt werden soll so könnte man zum Einen eine Gerade $g : x \mapsto ax + b$ so legen, dass g die Messwerte in einem gewissen Sinn möglichst gut approximiert. Das führt z.B. zum sogenannten Ausgleichsproblem oder "least squares problem" zur Bestimmung der beiden Parameter $a, b \in \mathbb{R}$. Damit lässt sich $f(\tilde{x}) \approx g(\tilde{x})$ approximieren. Man könnte aber auch ein Polynom p von hohem Grad so bestimmen dass alle Messwerte exakt interpoliert werden und dann $f(\tilde{x}) \approx p(\tilde{x})$ approximieren. Typischerweise oszilliert ein solches Polynom aber sehr stark, und liefert daher eine sehr unzuverlässige Vorhersage von $f(\tilde{x})$. Die höhere Anzahl an regelbaren Parametern in p im Vergleich zu nur zwei Parametern in g liefert keine

zuverlässigere Approximation. Dieser altbekannte Sachverhalt wird auch als Overfitting oder Überanpassung bezeichnet.

Bei den Support Vector Machines liegen auch viele Messwerte vor und die Trennung in zwei Klassen soll anhand von unbekannten "Ähnlichkeitseigenschaften" vorgenommen werden. Wie in obigem Beispiel gilt auch hier, dass die Konstuktion einer Funktion ϕ so dass alle Trainingsdaten korrekt getrennt werden können für sich alleine keine zuverlässige Klassifizierung garantiert. Gesucht ist im Folgenden eine Abbildung ϕ, die zum Einen nicht von der konkreten Wahl der $x^{(i)}$ abhängt (diese bestimmen nur die trennende Hyperebene im Raum $\mathcal{W}$) und zum Anderen so gewählt ist, dass ϕ sich nicht "zu chaotisch" verhält (nicht "zu sehr oszilliert") sondern gewisse Stetigkeitseigenschaften besitzt, die die vorausgesetzten aber unbekannten "Ähnlichkeitseigenschaften" der ursprünglichen Datenpunkte bewahrt.

4.1.3 Kreuzvalidierung

Die sogenannte Kreuzvalidierung liefert einen Ansatz um die Zuverlässigkeit der Klassifizierung abzuschätzen. Ein einfacher Ansatz der Kreuzvalidierung sei nachfolgend geschildert: Wenn die Trainingsdaten zufällig und unabhängig voneinander erzeugt wurden, so kann man die Trainingsdaten – wiederum zufällig – in zwei Teile aufteilen, z.B. 70% im einen Teil und der Rest im anderen, und die Trennung nur unter Nutzung der 70% berechnen. Die 30% nutzt man dann um zu prüfen, wie viele der Daten korrekt klassifiziert werden, und dies kann als geschätzte Fehlerrate für die Trennung genutzt werden. Die tatsächliche Trennung kann dann anhand aller Trainingsdaten erfolgen; der "70%-Schätzer" dient nur zur Abschätzung der Fehlerrate – und zwar nur als Fehlerrate für neue Daten, die aus derselben Verteilung generiert werden. Mit Modifizierungen dieses Ansatzes lassen sich auch gewisse Parameter anpassen, die die Qualität der Klassifizierung bestimmen.

4.1.4 Beispiel Gauß-Kern

Als Beipiel betrachten wir die Funktion $\kappa : \Omega \times \Omega \to \mathbb{R}$ mit $\kappa(x, y) := e^{-c\|x-y\|_2^2}$ mit einer Konstanten $c > 0$ bzw. die Matrix K mit den Einträgen $K_{i,j} = \kappa(x^{(i)}, x^{(j)}) := e^{-c\|x^{(i)}-x^{(j)}\|_2^2}$. In Anlehnung an die Gauß'sche Verteilungskurve wird diese Funktion κ als Gauß-Kern bezeichnet. In Abschnitt 5.1 im Anhang wird kurz begründet, dass obiges K stets positiv semidefinit ist. Es ist aber nicht offensichtlich, wie das Skalarprodukt $\langle ., . \rangle_{\mathcal{W}}$ und eine Abbildung ϕ zu bestimmen wären, für die stets gilt

$$\langle \phi(x), \phi(y) \rangle_{\mathcal{W}} \equiv e^{-c\|x-y\|_2^2}.$$

Falls eine solche Abbildung ϕ existiert, so folgt aber sofort, dass

$$\|\phi(x)\|_{\mathcal{W}}^2 = \langle \phi(x), \phi(x) \rangle_{\mathcal{W}} = \kappa(x, x) = e^0 = 1$$

für alle $x \in \Omega$. Wir nennen κ mit $\|\phi(x)\|_{\mathcal{W}}^2 = \kappa(x, x) \equiv 1$ im Folgenden einen

isonormierten Kern.

Dabei kann jeder Kern, für den $\kappa(x,x) > 0 \quad \forall x \in \Omega$ gilt, wie in Abschnitt 5.1 derart diagonal skaliert werden, dass er isonormiert ist. Die Skalierung ist in der Regel nichtlinear und ändert somit auch die Trennungseigenschaften.

4.2 Bestimmung von ϕ und $\langle ., . \rangle_{\mathcal{W}}$

In Anlehnung an [3] sei ein stetiger, symmetrisch positiv definiter Kern κ auf der kompakten konvexen Menge $\Omega \subset \mathbb{R}^n$ gegeben. Für fest gewähltes $x \in \Omega$ definieren wir die Abbildung $\phi(x) := K_x : \Omega \to \mathbb{R}$ mittels

$$K_x := \kappa(x, \ . \ ), \qquad \text{d.h. } \phi(x)[z] \equiv K_x(z) \equiv \kappa(x,z) \text{ für } z \in \Omega.$$

Um im Folgenden nicht mit der verwirrenden Tatsache durcheinander zu kommen, dass $\phi(x)$ selbst eine Funktion ist, nutzen wir die intuitivere Schreibweise K_x anstelle von $\phi(x)$. Die endlichen Linearkombinationen solcher Funktionen K_x bilden dann den Raum

$$\mathcal{W} := \text{Spann}(\{K_x \mid x \in \Omega\}) = \{f \mid \exists k \in \mathbb{N},\ x^{(i)} \in \Omega,\ \alpha_i \in \mathbb{R}\ (1 \le i \le k) :\ f = \sum_{i=1}^{k} \alpha_i K_{x^{(i)}}\}.$$

Ferner sei die Abbildung $\langle ., . \rangle : \quad \mathcal{W} \times \mathcal{W} \to \mathbb{R}$ durch

$$\langle f, g \rangle := \sum_{i=1}^{k} \sum_{j=1}^{\ell} \alpha_i \beta_j \kappa(x^{(i)}, x^{(j)}) \tag{21}$$

für $f := \sum_{i=1}^{k} \alpha_i K_{x^{(i)}}$ und $g := \sum_{j=1}^{\ell} \beta_j K_{x^{(j)}}$ definiert. Obige Notation legt nahe, dass $\langle ., . \rangle$ ein Skalarprodukt ist. Zunächst soll aber begründet werden, dass $\langle ., . \rangle$ wohldefiniert ist: Da nicht vorausgesetzt wurde, dass alle $K_{x^{(i)}}$ linear unabhängig sind, könnte es verschiedene Darstellungen für eine gegebene Funktion $g \in \mathcal{W}$ geben. Die obige Abbildung ist aber unabhängig von der Darstellung von g, denn nach (21) ist

$$\langle f, g \rangle = \sum_{i=1}^{k} \alpha_i \underbrace{\left(\sum_{j=1}^{\ell} \beta_j \kappa(x^{(i)}, x^{(j)}) \right)}_{= \sum_{j=1}^{\ell} \beta_j K_{x^{(j)}}(x^{(i)}) = g(x^{(i)})} = \sum_{i=1}^{k} \alpha_i g(x^{(i)}).$$

Die rechte Seite hängt dabei nicht von den gewählten Koeffizienten $x^{(j)}$ und β_j für die Darstellung von g ab, nur von den Funktionswerten $g(x^{(i)})$. Außerdem sieht man, dass die rechte Seite linear in g ist. Analog folgt, dass die Abbildung auch unabhängig von der Darstellung von f und linear in f ist, d.h. $\langle f, g \rangle$ ist bilinear – und wie κ auch symmetrisch. Schließlich vererbt sich auch die positive Semidefinitheit:

$$\langle f, f \rangle = \sum_{i=1}^{k} \sum_{j=1}^{k} \alpha_i \alpha_j \kappa(x^{(i)}, x^{(j)}) \ge 0$$

nach der Definition eines positiv definiten Kerns. Daher gilt auch die Cauchy-Schwarz'sche Ungleichung, $\langle f,g\rangle^2 \leq \langle f,f\rangle\langle g,g\rangle$ (mit dem üblichen Beweis[4]).

Nun gilt für $x \in \Omega$, $f = \sum_{i=1}^{k} \alpha_i K_{x^{(i)}} \in \mathcal{W}$ und $g := K_x$ in (21) auch

$$\langle f, \kappa(.,x)\rangle = \langle f, K_x\rangle \overbrace{=}^{(21)} \sum_{i=1}^{k} \alpha_i \kappa(x^{(i)},x) = \sum_{i=1}^{k} \alpha_i K_{x^{(i)}}(x) = f(x), \tag{22}$$

eine zentrale Eigenschaft, die als "reproducing Kernel"-Eigenschaft bezeichnet wird. Sei nun $x, z \in \Omega$ und $g := \kappa(.,z)$, dann folgt aus (22) auch

$$\langle K_x, K_z\rangle = \langle \kappa(.,x), \kappa(.,z)\rangle = \langle \kappa(.,x), g\rangle = \langle g, \kappa(.,x)\rangle = g(x) = \kappa(x,z). \tag{23}$$

Mit (22) und der Cauchy-Schwarz'schen Ungleichung folgt für $f \in \mathcal{W}$ und $x \in \Omega$ weiter

$$f(x)^2 = (\langle \kappa(.,x), f\rangle)^2 \leq \langle \kappa(.,x), \kappa(.,x)\rangle\langle f,f\rangle = \kappa(x,x)\langle f,f\rangle,$$

wobei die letzte Gleichung aus (23) folgt. Wenn also $\langle f,f\rangle = 0$ gelten sollte so folgt $f(x) \equiv 0$ für $x \in \Omega$, d.h.

$$\langle .,.\rangle_{\mathcal{W}} := \langle .,.\rangle \text{ ist ein Skalarprodukt,} \tag{24}$$

das eine Norm $\| \, . \, \|$ auf $\mathcal{W}$ erzeugt[5]. Allerdings ist $\mathcal{W}$ nicht vollständig bezüglich dieser Norm. Nachfolgend wird die Existenz der Trennung angesprochen.

4.2.1 Trennbarkeit im Feature Space

Für den Fall, dass κ so gewählt ist, dass die Funktionen $\phi(x^{(1)}), \ldots, \phi(x^{(m)})$ linear unabhängig sind[6], lässt sich die Existenz einer trennenden Hyperebene auch ohne soft margin explizit begründen:

Dazu betrachten wir das Problem (17) und nutzen die Notation

$$\tilde{a}_M = \sum_{i=1}^{m} \alpha_i \phi(x^{(i)}) = \left[\phi(x^{(1)}), \ldots, \phi(x^{(m)})\right] \alpha$$

[4]Aufgrund der Bilinearität und Semidefinitheit ist $0 \leq \langle f - \lambda g, f - \lambda g\rangle = \langle f,f\rangle - 2\lambda\langle f,g\rangle + \lambda^2\langle g,g\rangle$ für $\lambda \in \mathbb{R}$. Also $2\lambda\langle f,g\rangle \leq \langle f,f\rangle + \lambda^2\langle g,g\rangle$ für alle λ. Wenn $\langle g,g\rangle = 0$ impliziert dies $\langle f,g\rangle = 0$, d.h. $\langle f,g\rangle^2 = 0 \leq \langle f,f\rangle\langle g,g\rangle$, und wenn $\langle g,g\rangle > 0$ so folgt mit der Wahl $\lambda := \langle f,g\rangle/\langle g,g\rangle$ dass

$$2\tfrac{\langle f,g\rangle^2}{\langle g,g\rangle} \leq \langle f,f\rangle + \langle g,g\rangle\tfrac{\langle f,g\rangle^2}{\langle g,g\rangle^2} \qquad \text{d.h.} \qquad \tfrac{\langle f,g\rangle^2}{\langle g,g\rangle} \leq \langle f,f\rangle \qquad \text{bzw.} \qquad \langle f,g\rangle^2 \leq \langle f,f\rangle\langle g,g\rangle.$$

[5]Die Normeigenschaft scheint zunächst überraschend, da die $K_{x^{(i)}}$ nicht als linear unabhängig vorausgesetzt wurden und lediglich $K \succeq 0$ gefordert wurde, aber bei linear abhängigen $K_{x^{(i)}}$ wird auch der Raum $\mathcal{W}$ kleiner. Auf $\mathcal{W}$ ist das Skalarprodukt (strikt) positiv definit.

[6]Diese Vorausseztung ist z.B. für den Gauß-Kern stets erfüllt.

und die Festlegung $C = \infty$, d.h. $s = 0$. Dann ist (17) gegeben durch

$$\begin{aligned} & \min_{\tilde{a}_M \in M,\ \beta} \{\ \tfrac{1}{2}\|\tilde{a}_M\|_{\mathcal{W}}^2\ \mid\ \zeta_i \left(\langle \tilde{a}_M, \phi(x^{(i)})\rangle_{\mathcal{W}} - \beta\right) \geq 1,\ \ \forall\ 1 \leq i \leq m\} \qquad (25) \\ = & \min_{\tilde{a}_M \in M,\ \beta} \{\ \tfrac{1}{2}\alpha^T K \alpha\ \mid\ ZK\alpha - \beta\zeta \geq e\}, \end{aligned}$$

wobei die Einträge der Matrix K wieder durch $K_{i,j} = \langle \phi(x^{(i)}), \phi(x^{(i)})\rangle_{\mathcal{W}}$ gegeben seien. Da K nach Voraussetzung invertierbar ist und $Z = Z^{-1}$, besitzt letzteres Problem die zulässige Lösung $\beta = 0$ und $\alpha = K^{-1}Ze$, die zusammen mit $s := 0$ auch zulässig ist für (17) mit Zielfunktionswert $\frac{1}{2}e^T ZK^{-1}Ze$. Diese zulässige Lösung ist in der Regel nicht optimal, doch sieht man, dass der zugehörige Zielfunktionswert im Allgemeinen wächst, wenn die kleinsten Eigenwerte von K gegen Null streben, eine Beoachtung, die auch häufig für die Optimallösung von (25) zutrifft.

Wählt man einen Kern, für den die $\phi(x^{(i)})$ zu paarweise verschiedenen $x^{(i)}$ stets linear unabhängig sind, d.h. für den stets eine exakte Trennung der Daten erreichbar ist, so ergibt sich wieder das Problem der Überanpassung aus Abschnitt 4.1.2. Es ist daher auch für solche Kerne in der Regel sinnvoll, einen Ansatz mit Soft Margin zu wählen.

4.2.2 Gauß-Kern

Für den Gauß-Kern mit $c > 0$ kann man zeigen, dass zu beliebigen paarweise verschiedenen $x^{(i)}$ die Funktionen $K_{x^{(i)}}$ linear unabhängig sind, d.h., dass die Dimension von $\mathcal{W}$ nicht endlich ist. Aber auch hier hängt der Raum $\mathcal{W}$ von der Wahl von c ab. Für $x \in \Omega$ hat die Funktion $\phi(x)$ die Form

$$\phi(x)[y] \equiv e^{-c\|x-y\|_2^2} \quad \text{für } y \in \Omega.$$

Beim Übergang $c \to \infty$ konvergiert $\phi(x)$ gegen die charakteristische Funktion des Punktes $x \in \Omega$ und die Matrix K in Abschnitt 4.2.1 konvergiert für jede Wahl der (paarweise verschiedenen) Datenpunkte $x^{(i)}$ gegen die Einheitsmatrix. Dabei ist mit der Wahl $\alpha = \zeta$ und $\beta = 0$ der Zielfunktionswert von (25) durch $m/2$ gegeben, also nicht sehr groß. Für $c \to 0$ konvergiert $\phi(x)$ gegen die konstante Funktion 1 auf Ω und die Matrix K strebt dann gegen die Matrix ee^T vom Rang 1. Für kleine $c > 0$ strebt der Optimalwert von (25) gegen unendlich.

Neben dem Parameter c im Gauß-Kern ist auch die Konstante C aus dem Soft-Margin-Ansatz in (6), die die Verletzung der Trennungseigenschaften bestraft, ein frei wählbarer Parameter beim Einsatz von Gauß-Kernen. Dabei kann es auch für große endliche Werte von C sein, dass die Optiallösung von (17) zu einer Trennung führt, die alle Trainingsdaten korrekt trennt, aber einige der Trainingsdaten näher bei der trennenden Hyperebene liegen als andere Trainingsdaten, die zu Stützvektoren gehören. (Im Fall einer Trennung mit $C = \infty$ sind alle Stützvektoren gleich weit von der trennenden Hyperebene entfernt und näher ligende Datenpunkte gibt es nicht.) Eine Aufgabe bei der Festlegung einer SVM mit Gauß-Kern besteht daher darin, "gute" Parameter c und C zu identifizieren, was häufig über eine Kreuz-Validierung erfolgt.

Im folgenden Beispiel wurden die zu trennenden Mengen entlang eines 3×3-Schachbrettmusters gewählt mit 500 Trainingspunkten ohne Klassifizierungsfehler. Die Datenpunkte sind in Abbildung 1 durch schwarze Sterne markiert. Die Abbildung zeigt die Regionen, die durch die SVM identifiziert wurden in grün und rot. Ein größerer Wert von γ ist hier mit "mehr Krümmung" im Rand verbunden.

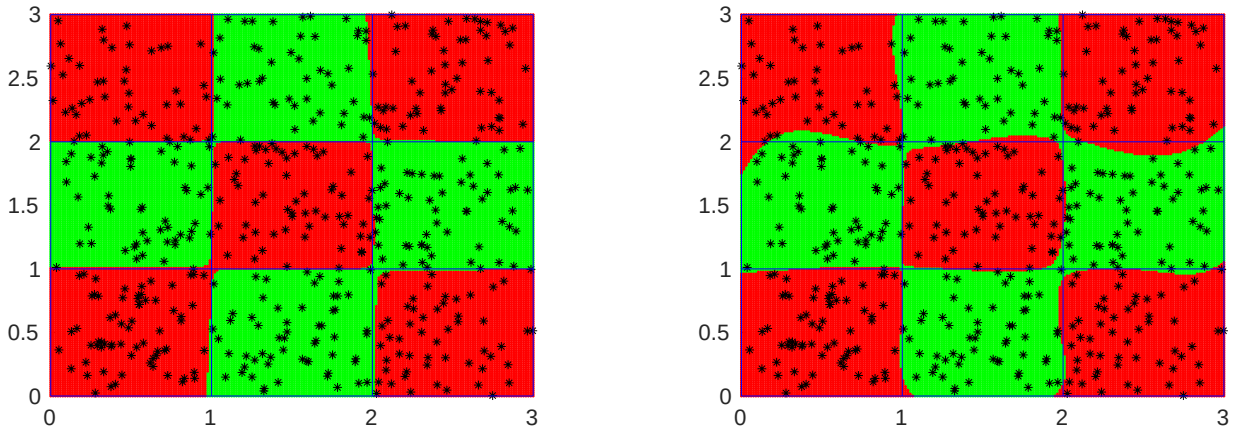

Abbildung 1, Gauß-Kern mit $C = \infty$ und $\gamma = 0.03$ (links) sowie $\gamma = 3$ (rechts).

Auch im rechten Plot sind alle Trainings-Punkte korrekt identifiziert; d.h. der rechte Plot könnte grundsätzlich näher an der wirklichen Einteilung liegen. Eine passende Wahl von c und C wird daher in der Regel über passende Kreuz-Validierungs-Strategien ermittelt.

5. Anhang

5.1 Positive Kerne

Das Hadamardprodukt zweier positiv semidefiniter $n \times n$ Matrizen ist wieder positiv semidefinit, denn aus den Zerlegungen $A = \sum_i a^{(i)}(a^{(i)})^T \succeq 0$ und $B = \sum_j b^{(j)}(b^{(j)})^T \succeq 0$ folgt

$$A \circ B = \sum_{i,j} (a^{(i)} \circ b^{(j)})(a^{(i)} \circ b^{(j)})^T \succeq 0.$$

Nach Konstrution ist die Matrix $(x^{(1)} \dots x^{(m)})^T (x^{(1)} \dots x^{(m)})$ mit den Einträgen $(x^{(i)})^T x^{(j)}$ an den Stellen (i, j) eine positiv semidefinite Gram-Matrix.

Somit sind die Hadamard-Produkte obiger Gram-Matrizen positiv semidefinit und damit auch die Exponentialfunktion als Summe solcher Produkte d.h. die Matrix mit den Einträgen $e^{c(x^{(i)})^T x^{(j)}}$ mit $c > 0$ ist positiv semidefinit. Damit zeigt man, dass K mit

$$K_{i,j} := e^{-c\|x^{(i)} - x^{(j)}\|^2} = e^{2c(x^{(i)})^T x^{(j)} - c\|x^{(i)}\|^2 - c\|x^{(j)}\|^2}$$

positiv semidefinit ist, denn die Terme $e^{-c\|x^{(i)}\|^2}$ und $e^{-c\|x^{(j)}\|^2}$ bewirken nur eine symmetrische

diagonale Skalierung der Matrix K, d.h. einen Wechsel von $K \succeq 0$ zu $DKD \succeq 0$ mit einer Diagonalmatrix D.

Analog folgt, dass $\kappa(x,y) \equiv ((x^T y)+1)^p \equiv \sum_{j=0}^{p} \binom{p}{j} (x^T y)^j$ mit $p \in \mathbb{N}$ ein positiv definiter Kern ist.

Und falls $a \in \mathbb{R},\ a > 0$ gegeben ist und $g : [-a,a] \to \mathbb{R}$ eine Potenzreihe mit nichtnegativen Koeffizienten ist, so folgt für Ω mit $\sup_{x\in\Omega}\{\|x\|_2^2\} \leq a$ wie oben, dass der Kern $\kappa(x,y) \equiv g(x^T y)$ ein positiv definiter Kern ist. Und da mit $K \succeq 0$ auch stets $DKD \succeq 0$ für jede Diagonalmatrix D gilt, folgt sogar, dass auch der Kern $\kappa(x,y) \equiv h(x)h(y)g(x^T y)$ ein positiv definiter Kern ist.

Und falls $g(x^T x) > 0$ für alle $x \in \Omega$, so ist κ mit der Wahl $h(x) := g(x^T x)^{-1/2}$ ein isonormierter Kern (wie in Abschnitt 4.1.4 definiert).

5.2 Exakte Trennbarkeit mit Gauß-Kernen

Zum Nachweis, dass die Funktionen $x \mapsto e^{-c\|x-x^i\|_2^2}$ für paarweise verschiedene x^i (mit $1 \leq i \leq m$) stets linear unabhängig sind betrachte man die Gleichung, einen Vektor $\alpha \in \mathbb{R}^m$ zu bestimmen mit

$$0 \equiv \sum_{i=1}^{m} \alpha_i e^{-c\|x-x^i\|_2^2} = \sum_i \alpha_i e^{-c\|x\|_2^2} e^{-c\|x^i\|_2^2} e^{2cx^T x^i} = e^{-c\|x\|_2^2} \sum_i \alpha_i e^{-c\|x^i\|_2^2} e^{2cx^T x^i} \quad \forall x \in \mathbb{R}^n$$

Setzt man $\beta_i := \alpha_i e^{-c\|x^i\|_2^2}$ so ist das obige System wegen $e^{-c\|x\|_2^2} > 0$ gleichwertig zu

$$0 = \sum_i \beta_i e^{2cx^T x^i} \quad \forall x \in \mathbb{R}^n.$$

(Insbesondere ist $\beta_i = 0 \Longleftrightarrow \alpha_i = 0$.) Man wähle jetzt einen Vektor $\bar{x}$ mit $\bar{x}^T(x^i - x^j) \neq 0$ für alle $i \neq j$ und setze dann $x := j\bar{x}$ für $0 \leq j \leq m-1$ in das obige System ein. Es ergibt sich

$$0 = \sum_i \beta_i e^{2cj\bar{x}^T x^i} = \sum_i \beta_i (a_i)^j \quad \text{für } 0 \leq j \leq m-1$$

mit $a_i := e^{2c\bar{x}^T x^i}$. Nach Wahl von $\bar{x}$ sind die a_i paarweise verschieden. Das obige System hat nur die Lösung $\beta = 0$ falls auch das transponierte Sytsem

$$0 = \sum_j \tilde{\beta}_j (a_i)^j$$

nur die Lösung $\tilde{\beta} = 0$ besitzt. Das letztere System besagt, dass das Polynom $t \mapsto \sum \tilde{\beta}_j t^j$ die Null an den Stellen a_i interpoliert. Somit muss $\tilde{\beta} = 0$ gelten. (Eindeutigkeit der Polynominterpolation.) Wie oben hergleitet impliziert dies die gesuchte Forderung $\alpha = 0$.

Literatur

[1] J Cervantes, F Garcia-Lamont, L Rodríguez-Mazahua, A comprehensive survey on support vector machine classification: Applications, challenges and trends. Neurocomputing 408, 189-215, Elsevier. 2020.

[2] P-H. Chen, C-J. Lin, B. Schölkopf, A tutorial on ν-support vector machines, Appl. Stochastic Models Bus. Ind., 21:111-136, 2005.

[3] T. Hofmann, B. Schölkopf, und A.J. Smola, Kernel methods in machine learning. The Annals of Statistics, pp. 1171–1220, 2008.

[4] F. Jarre und J. Stoer, *Optimierung*, Lehrbuch, Springer Verlag, Zweite Auflage 2019.

[5] Nesterov, Yurii und Arkadii Nemirovskii, *Interior-point polynomial algorithms in convex programming* Society for industrial and applied mathematics, 1994.

[6] W.S. Noble, What is a support vector machine? Nature biotechnology, 24(12), 1565-1567, 2006.

[7] C. O'Neil, Weapons of Math Destruction Crown Books, 2016.

[8] DA Pisner, DM Schnyer, Support vector machine. Chapter 6 in Machine learning, Methods and Applications to Brain Disorders, Academic Press, Elsevier, 101-121, 2020.

[9] Bernhard Schölkopf und Alexander J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning Series) 2001

[10] S. Suthaharan, Support Vector Machine. In: Machine Learning Models and Algorithms for Big Data Classification. Integrated Series in Information Systems, vol 36. Springer, Boston, 2016.

[11] V.N. Vapnik and A.Y. Chervonenkis, The necessary and suffcient conditions for consistency in the empirical risk minimization method. Pattern Recognition and Image Analysis 1(3) 283-305. 1991.