

MUSTERERKENNUNG

Vorlesung im Sommersemester 2017

Prof. E.G. Schukat-Talamazzini

Stand: 3. April 2017

Teil IX

Supportvektormaschinen

Grundidee

Separierbare Daten

Nichtseparierbare Daten

Kernel Trick

Beispiele

Grundidee

Separierbare Zwei-Klassen-Probleme

Nichtseparierbare Zwei-Klassen-Probleme

Nichtlineare Einbettungen des Merkmalraums

Beispiel: IRIS-Datensatz

Grundidee

Separierbare Daten

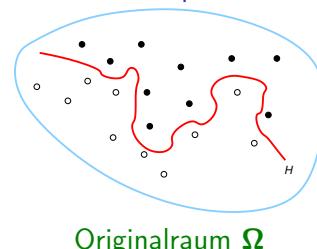
Nichtseparierbare Daten

Kernel Trick

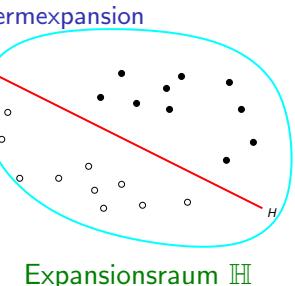
Beispiele

Supportvektormaschine

Optimale Trennfläche & Nichtlineare Termexpansion



Originalraum Ω



Expansionsraum \mathbb{H}

Kernel Trick

- Nichtlineare Abbildung $\phi : \Omega \rightarrow \mathbb{H}$ in hochdimensionalen Vektorraum
- Dort ist **jede** Musterkonstellation separierbar!

Maximaler Sicherheitsabstand

- Perfekte Klassentrennung ($K = 2$) durch eine Hyperfläche
- Maximaler lotrechte Distanz zu allen Datenpunkten

Lineare Trennung zweier Musterklassen

Problem

Separiere die Lerndatenprobe

$$\omega = \omega_1 \cup \omega_2 = \{(\mathbf{x}_t, y_t) \mid t = 1, \dots, T\} \subseteq \mathbb{R}^D \times \{-1, +1\}$$

mit einer linearen (affinen!) Diskriminantenfunktion

$$u(\mathbf{x}) = \mathbf{x}^\top \mathbf{w} + b,$$

d.h., finde Normalenvektor \mathbf{w} und Offset b mit der Eigenschaft

$$\begin{cases} \mathbf{x}_t^\top \mathbf{w} + b > 0 & \forall t : y_t = +1 \\ \mathbf{x}_t^\top \mathbf{w} + b < 0 & \forall t : y_t = -1 \end{cases}.$$

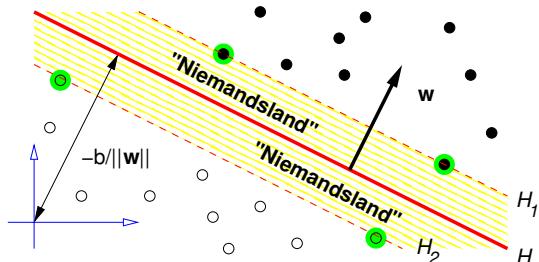
Fakt

Die Separierbarkeit liegt vor gdw. es $\mathbf{w} \in \mathbb{R}^D$, $b \in \mathbb{R}$ gibt mit

$$y_t \cdot (\mathbf{x}_t^\top \mathbf{w} + b) - 1 \geq 0 \quad (\forall t)$$

(Reskalierungsargument; T endlich).

Separierung mit maximalem Sicherheitsabstand



Lemma

Die Hyperebene $H : \mathbf{x}^\top \mathbf{w} + b$, welche die Lernbeispiele

$$\omega = \{(\mathbf{x}_t, y_t) \mid t = 1, \dots, T\}$$

mit einem Niemannsland maximaler Breite separiert, gehorcht den Restriktionen

$$y_t \cdot (\mathbf{x}_t^\top \mathbf{w} + b) - 1 \geq 0 \quad (t = 1, \dots, T)$$

und besitzt dabei die kleinstmögliche Norm $\|\mathbf{w}\|$ des Normalenvektors \mathbf{w} .

Beweis.

Positive Beispiele mit Gleichheit in der Restriktionsungleichung liegen auf der Hyperebene

$$H_1 : \mathbf{x}^\top \mathbf{w} + b = +1$$

mit dem Normalenvektor \mathbf{w} und dem lotrechten Abstand $|1 - b|/\|\mathbf{w}\|$ zum Ursprung.

Negative Beispiele mit Gleichheit in der Restriktionsungleichung liegen auf der Hyperebene

$$H_2 : \mathbf{x}^\top \mathbf{w} + b = -1$$

mit dem Normalenvektor \mathbf{w} und dem lotrechten Abstand $|-1 - b|/\|\mathbf{w}\|$ zum Ursprung.

Die Hyperebenen H , H_1 und H_2 sind also parallel zueinander, und die Abstände von H zu H_1 und H_2 betragen beide $1/\|\mathbf{w}\|$.

Das Niemannsland hat folglich die Breite $2/\|\mathbf{w}\|$. □

Primäre Optimierungsaufgabe

Lagrange-Formulierung für den separierbaren Fall

Minimiere das Funktional

$$\ell_p(\omega) = \frac{1}{2} \|\omega\|^2 - \sum_t \alpha_t y_t \cdot (\omega^\top \mathbf{x}_t + b) + \sum_t \alpha_t$$

bezüglich ω, b unter den Restriktionen

$$\forall t : \begin{cases} \alpha_t & \geq 0 \\ \partial \ell_p / \partial \alpha_t & = 0 \end{cases}$$

für die Lagrangemultiplikatoren α_t .

Duale Optimierungsaufgabe

Wolfe-Formulierung für den separierbaren Fall

Maximiere das Funktional

$$\ell_d(\omega) = \frac{1}{2} \|\omega\|^2 - \sum_t \alpha_t y_t \cdot (\mathbf{x}_t^\top \omega + b) + \sum_t \alpha_t$$

bezüglich α unter den Restriktionen

$$\forall t : \begin{cases} \alpha_t & \geq 0 \\ \partial \ell_d / \partial \alpha_t & = 0 \\ \partial \ell_d / \partial b & = 0 \end{cases}$$

und

$$\sum_t \alpha_t y_t = 0.$$

Konvexe quadratische Optimierung

Lemma

Für den gesuchten Normalenvektor ω der optimalen Hyperebene für eine separable Stichprobe finden wir nach Maximierung der Wolfe'schen dualen Zielfunktion

$$\ell_d(\omega) = \sum_t \alpha_t - \frac{1}{2} \cdot \sum_{s,t} \alpha_s \alpha_t \cdot y_s y_t \cdot \mathbf{x}_s^\top \mathbf{x}_t$$

unter den Restriktionen

$$\forall t : \alpha_t \geq 0 \quad \text{und} \quad \sum_t \alpha_t y_t = 0$$

die Darstellung

$$\omega = \sum_t \alpha_t \cdot y_t \cdot \mathbf{x}_t.$$

Bemerkung

Die infolge $\alpha_t \neq 0$ explizit in dieser Linearkombination auftretenden Stichprobenvektoren \mathbf{x}_t heißen **Supportvektoren** des Klassifikationsproblems.

Beweis.

Aus der dualen Formulierung folgen wegen

$$\frac{\partial \ell_d}{\partial b} = - \sum_t \alpha_t y_t = 0$$

die Bedingung

$$\sum_t \alpha_t y_t = 0$$

und wegen

$$\nabla_\omega \ell_d = \nabla_\omega \left(\frac{1}{2} \omega^\top \omega \right) - \sum_t \nabla_\omega \alpha_t y_t \mathbf{x}_t^\top \omega = \omega - \sum_t \alpha_t y_t \mathbf{x}_t = 0$$

die Bedingung

$$\omega = \sum_t \alpha_t y_t \mathbf{x}_t$$

Die Substitution dieser Bedingungen in die duale Zielgröße ergibt

$$\begin{aligned} \ell_d(\omega) &= \frac{1}{2} \|\omega\|^2 - \|\omega\|^2 - 0 \cdot b + \sum_t \alpha_t \\ &= \sum_t \alpha_t - \frac{1}{2} \cdot \left(\sum_t \alpha_t y_t \mathbf{x}_t \right)^\top \left(\sum_s \alpha_s y_s \mathbf{x}_s \right) \\ &= \sum_t \alpha_t - \frac{1}{2} \cdot \sum_{s,t} \alpha_s \alpha_t \cdot y_s y_t \cdot \mathbf{x}_s^\top \mathbf{x}_t \end{aligned}$$

Die Bedingung $\sum_t \alpha_t y_t = 0$ entfällt im übrigen, sofern wir $b = 0$ gefordert haben; in sehr hochdimensionalen Räumen bedeutet dieses Vorgehen keine ernste Einschränkung.

Karush-Kuhn-Tucker Bedingungen

$$\frac{\partial \ell}{\partial w_i} = w_i - \sum_t \alpha_t y_t x_{t,i} = 0 \quad (\forall i)$$

$$\frac{\partial \ell}{\partial b} = - \sum_t \alpha_t y_t = 0$$

$$y_t \cdot (x_t^\top w + b) - 1 \geq 0 \quad (\forall t)$$

$$\alpha_t \cdot (y_t \cdot (x_t^\top w + b) - 1) = 0 \quad (\forall t)$$

Grundidee

Separierbare Zwei-Klassen-Probleme

Nichtseparierbare Zwei-Klassen-Probleme

Nichtlineare Einbettungen des Merkmalraums

Beispiel: IRIS-Datensatz

Bemerkung

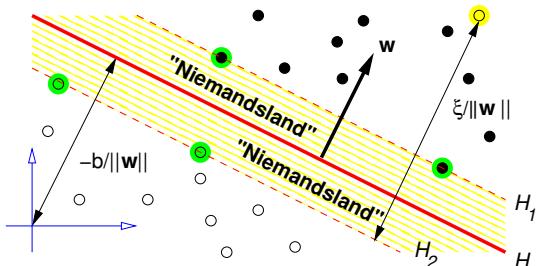
Jedes Datenelement mit $\alpha_t \neq 0$ liefert uns eine Bestimmungsgleichung

$$b = y_t^{-1} - x_t^\top w$$

für den Offsetparameter b .

Regularisierte Separierung

Klassentrennung toleriert einige wenige Ausnahmepositionen



Schlupfvariable $\xi_t, t = 1, \dots, T$

Berücksichtigung zusätzlicher Kosten für „Grenzüberschreitungen“

$$x_t^\top w + b \geq +1 - \xi_t \quad (\forall t : y_t = +1)$$

$$x_t^\top w + b \leq -1 + \xi_t \quad (\forall t : y_t = -1)$$

$$\xi_t \geq 0 \quad (\forall t)$$

Primäre Optimierungsaufgabe

Lagrange-Formulierung für den nichtseparierbaren Fall

Minimiere das Funktional

$$\ell_p(w) = \frac{1}{2} \|w\|^2 + C \cdot \sum_t \xi_t - \sum_t \alpha_t (y_t(x_t^\top w + b) - 1 + \xi_t) - \sum_t \mu_t \xi_t$$

bezüglich w, b unter den Restriktionen

$$\forall t : \begin{cases} \alpha_t & \geq 0 \\ \mu_t & \geq 0 \\ \frac{\partial \ell_p}{\partial \alpha_t} & = 0 \\ \frac{\partial \ell_p}{\partial \mu_t} & = 0 \end{cases}$$

für die Lagrangemultiplikatoren α_t .

Duale Optimierungsaufgabe

Wolfe-Formulierung für den nichtseparierbaren Fall

Maximiere das Funktional

$$\ell_d(\omega) = \frac{1}{2} \|\omega\|^2 + C \cdot \sum_t \xi_t - \sum_t \alpha_t (y_t (\mathbf{x}_t^\top \omega + b) - 1 + \xi_t) - \sum_t \mu_t \xi_t$$

bezüglich α unter den Restriktionen

$$\forall t : \begin{cases} \alpha_t & \geq 0 \\ \mu_t & \geq 0 \\ \partial \ell_p / \partial w_t & = 0 \\ \partial \ell_p / \partial b & = 0 \end{cases}$$

und

$$\sum_t \alpha_t y_t = 0.$$

Konvexe quadratische Optimierung

Lemma

Für den gesuchten Normalenvektor ω der optimalen Hyperebene für eine separable Stichprobe finden wir nach Maximierung der Wolfe'schen dualen Zielfunktion

$$\ell_d(\omega) = \sum_t \alpha_t - \frac{1}{2} \cdot \sum_{s,t} \alpha_s \alpha_t \cdot y_s y_t \cdot \mathbf{x}_s^\top \mathbf{x}_t$$

unter den Restriktionen

$$\forall t : 0 \leq \alpha_t \leq C \quad \text{und} \quad \sum_t \alpha_t y_t = 0$$

die Darstellung

$$\omega = \sum_t \alpha_t \cdot y_t \cdot \mathbf{x}_t.$$

Bemerkung

Die \mathbf{x}_t mit $\alpha_t \neq 0$ sind wieder die **Supportvektoren**.

Die \mathbf{x}_t mit $\alpha_t = C$ liegen auf der falschen Seite der Grenze!

Karush-Kuhn-Tucker Bedingungen

$$\begin{aligned} w_i - \sum_t \alpha_t y_t \mathbf{x}_{t,i} &= \frac{\partial \ell}{\partial w_i} = 0 & \xi_t &\geq 0 \\ - \sum_t \alpha_t y_t &= \frac{\partial \ell}{\partial b} = 0 & \alpha_t &\geq 0 \\ C - \alpha_t - \mu_t &= \frac{\partial \ell}{\partial \xi_t} = 0^* & \mu_t &\geq 0 \\ y_t \cdot (\mathbf{x}_t^\top \omega + b) - 1 + \xi_t &\geq 0^{**} & \mu_t \xi_t &= 0^{****} \\ \alpha_t \cdot (y_t \cdot (\mathbf{x}_t^\top \omega + b) - 1 + \xi_t) &= 0^{***} \end{aligned}$$

Bemerkung

Jedes Datenelement mit $0 \neq \alpha_t \neq C$ liefert uns eine Bestimmungsgleichung für den Offsetparameter:

$$b = y_t^{-1} - \mathbf{x}_t^\top \omega$$

Bemerkungen

1. Wegen KKT (**) liegen alle Vektoren \mathbf{x}_t auf der richtigen Seite der Grenze; allerdings kann u.U. **Schlupf** vorliegen.
2. Für alle Supportvektoren ($\alpha_t > 0$) wird wegen (***) die Ungleichung (**) sogar zu einer Gleichung; diese Vektoren \mathbf{x}_t liegen also genau auf der „Grenze des Erlaubten“.
3. Für alle **unkritischen** Supportvektoren ($\alpha_t < C$) gilt aber $\mu_t \neq 0$ wegen (*), also $\xi_t = 0$ wegen (****); es liegt mithin **kein Schlupf** vor und \mathbf{x}_t liegt auf der „sicheren Seite“.
4. Die durch oben genannte KKT-Bedingungen spezifizierte **quadratische Optimierungsaufgabe** (QOP) kann mit Hilfe von Standardsoftwarepaketen gelöst werden.
5. Für Lerndatensätze realistischer Größenordnung sind allerdings **spezialisierte Lösungsverfahren** zu entwickeln um zu einem noch **praktikablen Berechnungsaufwand** zu gelangen.
6. In das QOP gehen ausschließlich die paarweisen **Skalarprodukte**

$$\mathbf{x}_s^\top \mathbf{x}_t \in \mathbb{IR}, \quad s, t \in \{1, \dots, T\}$$

der Lerndatenvektoren ein.

Der **Berechnungsaufwand** ist also — abgesehen von einer initialen Last von $O(T^2 D)$ — **unabhängig von der Dimension** der Merkmalvektoren.

Grundidee

Separierbare Zwei-Klassen-Probleme

Nichtseparierbare Zwei-Klassen-Probleme

Nichtlineare Einbettungen des Merkmalraums

Beispiel: IRIS-Datensatz

Nichtlineare Einbettungen des Merkmalraumes

Genuß ohne Reue für Prüfgrößen auf Basis innerer Musterprodukte

MAK-Prüfgröße

SVM-Prüfgröße

$$\begin{aligned}
 u_\kappa(\mathbf{x}) &= \|\mathbf{x} - \boldsymbol{\mu}_\kappa\|^2 \\
 &= \mathbf{x}^\top \mathbf{x} - 2 \cdot \mathbf{x}^\top \boldsymbol{\mu}_\kappa + \boldsymbol{\mu}_\kappa^\top \boldsymbol{\mu}_\kappa \\
 &= \langle \mathbf{x}, \mathbf{x} \rangle - 2 \sum_{\mathbf{z} \in \omega_\kappa} \langle \mathbf{x}, \mathbf{z} \rangle + c_\kappa \\
 u(\mathbf{x}) &= \mathbf{x}^\top \mathbf{w} + b \\
 &= \mathbf{x}^\top \left(\sum_t \alpha_t y_t \mathbf{x}_t \right) + b \\
 &= \sum_t \alpha_t y_t \cdot \langle \mathbf{x}, \mathbf{x}_t \rangle + b
 \end{aligned}$$

Fakt

Die Prüfgrößen zahlreicher Klassifikatortypen nutzen die Merkmalvektoren der Lern- und Testmuster ausschließlich in Gestalt innerer Produkte!

Kernoperatoren

Schnelle Operatoren zur Berechnung hochdimensionaler Skalarprodukte

Expansion in einen Hilbertraum $(\mathbb{H}, \langle \cdot, \cdot \rangle)$

$$\phi : \begin{cases} \mathbb{R}^D & \rightarrow \mathbb{H} \\ \mathbf{x} & \mapsto \tilde{\mathbf{x}} = \phi(\mathbf{x}) \end{cases}$$

SVM für expandierte Merkmalvektoren

$$u(\phi \mathbf{x}) = \left\langle \phi \mathbf{x}, \sum_t \alpha_t y_t \cdot \phi \mathbf{x}_t \right\rangle + b = \sum_t \alpha_t \cdot y_t \cdot \underbrace{\langle \phi \mathbf{x}, \phi \mathbf{x}_t \rangle}_{K(\mathbf{x}, \mathbf{x}_t)} + b$$

Problem

Können wir Produkte $\langle \mathbf{v}, \mathbf{w} \rangle_{\mathbb{H}}$ schneller als in $O(\dim \mathbb{H})$ berechnen?

Vielleicht wenigstens dann, wenn $\mathbf{v}, \mathbf{w} \in \phi(\Omega)$ sind?

Beispiele mit quadratischen Merkmaltermen

Beispiel

Betrachte Datenvektoren im \mathbb{R}^2 und den Kernoperator $K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^\top \mathbf{y})^2$. Dann gilt für die Abbildung

$$\phi : \begin{cases} \mathbb{R}^2 & \rightarrow \mathbb{H} = \mathbb{R}^3 \\ (\mathbf{x}_1, \mathbf{x}_2)^\top & \mapsto (\mathbf{x}_1^2, \sqrt{2} \cdot \mathbf{x}_1 \mathbf{x}_2, \mathbf{x}_2^2)^\top \end{cases}$$

die gewünschte Beziehung $\langle \mathbf{x}, \mathbf{y} \rangle = K(\mathbf{x}, \mathbf{y})$.
Gleiches gilt aber auch für

$$\phi' : \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} \mapsto \begin{pmatrix} (\mathbf{x}_1^2 - \mathbf{x}_2^2) \\ 2\mathbf{x}_1 \mathbf{x}_2 \\ (\mathbf{x}_1^2 + \mathbf{x}_2^2) \end{pmatrix} \quad \text{oder} \quad \phi'' : \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} \mapsto \begin{pmatrix} \mathbf{x}_1^2 \\ \mathbf{x}_1 \mathbf{x}_2 \\ \mathbf{x}_1^2 \mathbf{x}_2 \\ \mathbf{x}_2^2 \end{pmatrix}$$

Beweis.

$$\underbrace{(\mathbf{x}_1 \mathbf{y}_1 + \mathbf{x}_2 \mathbf{y}_2)^2}_{K(\mathbf{x}, \mathbf{y})} = \begin{pmatrix} \mathbf{x}_1^2 \\ \sqrt{2} \mathbf{x}_1 \mathbf{x}_2 \\ \mathbf{x}_2^2 \end{pmatrix}^\top \begin{pmatrix} \mathbf{y}_1^2 \\ \sqrt{2} \mathbf{y}_1 \mathbf{y}_2 \\ \mathbf{y}_2^2 \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^2 - \mathbf{x}_2^2 \\ 2\mathbf{x}_1 \mathbf{x}_2 \\ \mathbf{x}_1^2 + \mathbf{x}_2^2 \end{pmatrix}^\top \begin{pmatrix} \mathbf{y}_1^2 - \mathbf{y}_2^2 \\ 2\mathbf{y}_1 \mathbf{y}_2 \\ \mathbf{y}_1^2 + \mathbf{y}_2^2 \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^2 \\ \mathbf{x}_1 \mathbf{x}_2 \\ \mathbf{x}_2^2 \end{pmatrix}^\top \begin{pmatrix} \mathbf{y}_1^2 \\ \mathbf{y}_1 \mathbf{y}_2 \\ \mathbf{y}_2^2 \end{pmatrix}$$

□

Der Trick mit den Kernen

Termexpansionen und Kernoperatoren

- Unterschiedliche Expansionen $\phi(\cdot)$ besitzen u.U. denselben Kernoperator $K(\cdot, \cdot)$.
- Der Kernoperator sagt nichts über die Dimension $\dim(\mathbb{H})$ der Termexpansion aus.
- Wir müssen $\phi(\cdot)$ und \mathbb{H} nicht kennen um eine ϕ -expandierte Supportvektormaschine zu bauen.

Entwurf kerngesteuerter Verfahren

? Welche Funktionen $\mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{R}$ sind nachweislich Kernoperatoren?

Demonstration der Beweisidee im Endlichen

Lineare Algebra  Funktionalanalysis

Konstruktion des RKHS ('reproducing kernel hilbert space')

1. Betrachte endliche Teilmenge: $\{x_1, \dots, x_T\} = \omega \subset \Omega = \mathbb{R}^D$
2. Gramsche Matrix $G \in \mathbb{R}^{T \times T}$ mit Einträgen $g_{st} := K(x_s, x_t)$
3. Kernoperator $K(\cdot, \cdot)$ symmetrisch und positiv-semidefinit:

$$G = UD^2U^\top =: A^\top A$$

4. partielle Expansionsvorschrift (a_t Spalten von A):

$$\phi : \begin{cases} \omega & \rightarrow \mathbb{R}^T \\ x_t & \mapsto a_t \end{cases}$$

5. Inneres Produkt in $\phi(\omega) \subset \mathbb{R}^T$ reproduziert Kernoperator:

$$\langle \phi x_s, \phi x_t \rangle_{\mathbb{R}^T} = \langle a_s, a_t \rangle_{\mathbb{R}^T} = (A^\top A)_{st} = g_{st} = K(x_s, x_t)$$

Für $\omega = \mathbb{R}^D$ sind die a_t „unendlich lange“ Vektoren und A eine unendliche „Matrix“.

Die Mercer-Bedingung (1953)

Hinreichende Bedingung für die Existenz einer Expansion

Satz

Gegeben sei der Kernoperator $K : \mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{R}$.

Dann sind die folgenden Aussagen äquivalent:

- Für alle Abbildungen $g : \mathbb{R}^D \rightarrow \mathbb{R}$ mit $\int g^2(x) dx < \infty$ gilt

$$\int g(x) \cdot K(x, y) \cdot g(y) dx dy \geq 0.$$

(Der Kernoperator ist „positiv-semidefinit“.)

- Es gibt einen (un)endlichdimensionalen Vektorraum \mathbb{H} und eine Injektion $\phi : \mathbb{R}^D \rightarrow \mathbb{H}$ mit

$$K(x, y) = \langle \phi x, \phi y \rangle = \sum_i (\phi x)_i \cdot (\phi y)_i \quad (\forall x, y \in \mathbb{R}^D)$$

(Der Kernoperator „reproduziert“ das innere Expansionsprodukt.)

Der Zoo der Kernoperatoren

Lemma

1. Jeder Kernoperator der **polynomialen** Gestalt

$$K : \begin{cases} \mathbb{R}^D \times \mathbb{R}^D & \rightarrow \mathbb{R} \\ (x, y) & \mapsto (x^\top y)^p \end{cases}, \quad p \in \mathbb{N}$$

erfüllt das Mercer-Kriterium.

2. Mit den Kernoperatoren $K_n : \mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{R}$, $n \in \mathbb{N}$ erfüllt auch jede **uniform konvergente Reihe**

$$K(x, y) \stackrel{\text{def}}{=} \sum_{n=0}^{\infty} c_n \cdot K_n(x, y), \quad c_n > 0 (\forall n)$$

das Mercer-Kriterium.

3. **Summen, Produkte, positive Taylorreihen und Normen:**

$$K^*(x, y) = K(x, y) / \sqrt{K(x, x) \cdot K(y, y)}$$

Beweis.

Wir zeigen die Ungleichung

$$\int \left(\sum_{d=1}^D x_d y_d \right)^p \cdot g(x)g(y) dx dy \geq 0 .$$

Obiger Ausdruck bildet eine Summe von Multinomialausdrücken der Gestalt

$$\frac{p!}{r_1! r_2! \dots r_D!} \int x_1^{r_1} x_2^{r_2} x_3^{r_3} \dots x_D^{r_D} \cdot y_1^{r_1} y_2^{r_2} y_3^{r_3} \dots y_D^{r_D} \cdot g(x)g(y) dx dy ,$$

wobei über alle Exponentenvektoren

$$\mathbf{r} \in \mathbb{N}^D \quad \text{mit} \quad \sum_{i=1}^D r_i = p$$

iteriert wird. Jeder Summand kann ohne Schwierigkeiten in folgende quadratische Form gebracht werden:

$$\frac{p!}{r_1! r_2! \dots r_D!} \left(\int x_1^{r_1} x_2^{r_2} x_3^{r_3} \dots x_D^{r_D} \cdot g(x) dx \right)^2 \geq 0$$

□

Polynomialkern

Prüfgröße $\hat{=}$ Polynom in x vom Grad p

$$\dim(\mathbb{H}) = \binom{D+p-1}{p}$$

$$K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^\top \mathbf{y} + 1)^p, \quad p \in \mathbb{N}$$

Z.B. Grad 4 für (16×16) -Muster ergibt $256 + 4 - 1$ über 4, also $\dim(\mathbb{H}) = 183, 181, 376$

Gaußkern

Prüfgröße $\hat{=}$ Radialbasisfunktion-Klassifikator (RBF)

$$\dim(\mathbb{H}) = \infty$$

$$K(\mathbf{x}, \mathbf{y}) = e^{-\|\mathbf{x}-\mathbf{y}\|^2/2\sigma^2}, \quad \sigma > 0$$

Tangens-Hyperbolicus-Kern

Prüfgrößen $\hat{=}$ $(D-\tilde{T}-K)$ -MLP mit Sigmoidfunktion

$$\dim(\mathbb{H}) = \infty$$

$$K(\mathbf{x}, \mathbf{y}) = \tanh(\kappa \cdot \mathbf{x}^\top \mathbf{y} - \theta), \quad \kappa, \delta \in \mathbb{R}$$

Nur für ausgewählte Parameterkombinationen κ, θ ; \tilde{T} = Anzahl Supportvektoren

Sukzessive Konstruktion neuer Kernoperatoren

Konstruiere Expansion $\phi(\cdot)$ oder verwende Verknüpfungsregel

1. Summen
 $K(\mathbf{x}, \mathbf{y}) = K_1(\mathbf{x}, \mathbf{y}) + K_2(\mathbf{x}, \mathbf{y})$

Konkateniere $\phi_1(\mathbf{x})$ und $\phi_2(\mathbf{x})$

2. Produkte
 $K(\mathbf{x}, \mathbf{y}) = K_1(\mathbf{x}, \mathbf{y}) \cdot K_2(\mathbf{x}, \mathbf{y})$

Bilde alle $\phi_1^{(i)}(\mathbf{x}) \cdot \phi_2^{(j)}(\mathbf{x})$

3. Positive Vielfache
 $K(\mathbf{x}, \mathbf{y}) = c_0 \cdot K_0(\mathbf{x}, \mathbf{y})$

Skaliere $\phi_0(\mathbf{x})$ mit $\sqrt{c_0}$

4. (In-)homogene Polynome
 $K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^\top \mathbf{y} + \alpha)^p, \alpha \in \{0, 1\}$

wegen $\mathbf{x}^\top \mathbf{y}$ und Punkt (2)

5. Normierte Kerne
 $K(\mathbf{x}, \mathbf{y}) = K_0(\mathbf{x}, \mathbf{y}) / \sqrt{K_0(\mathbf{x}, \mathbf{x}) \cdot K_0(\mathbf{y}, \mathbf{y})}$

Skaliere $\phi_0(\mathbf{x})$ mit $\sqrt{K_0(\mathbf{x}, \mathbf{x})}$

6. RBF- oder Gauß-Kerne
 $K(\mathbf{x}, \mathbf{y}) = e^{-(\mathbf{x}-\mathbf{y})^2} = \frac{e^{2\mathbf{x}\mathbf{y}}}{e^{\mathbf{x}^2} \cdot e^{\mathbf{y}^2}} = \frac{e^{2\mathbf{x}\mathbf{y}}}{\sqrt{e^{2\mathbf{x}^2} \cdot e^{2\mathbf{y}^2}}} = \frac{K_e(\mathbf{x}, \mathbf{y})}{\sqrt{K_e(\mathbf{x}, \mathbf{x}) \cdot K_e(\mathbf{y}, \mathbf{y})}}$

folgt wegen (2) aus skalarem Fall

$$K(\mathbf{x}, \mathbf{y}) = e^{-(\mathbf{x}-\mathbf{y})^2} = \frac{e^{2\mathbf{x}\mathbf{y}}}{e^{\mathbf{x}^2} \cdot e^{\mathbf{y}^2}} = \frac{e^{2\mathbf{x}\mathbf{y}}}{\sqrt{e^{2\mathbf{x}^2} \cdot e^{2\mathbf{y}^2}}} = \frac{K_e(\mathbf{x}, \mathbf{y})}{\sqrt{K_e(\mathbf{x}, \mathbf{x}) \cdot K_e(\mathbf{y}, \mathbf{y})}}$$

und $K_e(\mathbf{x}, \mathbf{y})$ schreibt sich als uniform konvergente Potenzreihe

Kernoperatoren für nichtnegative Merkmalsvektoren

Häufigkeit · Wahrscheinlichkeit · Proportion · Energie · Intensität

$\phi(m)$	=	1	1	1	1	0	0	0
$\phi(n)$	=	1	1	0	0	0	0	0

Histogram Intersection Kernel (HIK)

$$K_{\text{HIK}}(\mathbf{x}, \mathbf{y}) \stackrel{\text{def}}{=} \sum_i \min(x_i, y_i)$$

Kernoperatoreigenschaft?

Expansion $\phi_{\text{HIK}}(\cdot)$ mit

$$\phi_z(x) = \begin{cases} 1 & x \geq z \\ 0 & x < z \end{cases}$$

ergibt das Produkt:

$$\langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle = \underbrace{\int_0^{\infty} \phi_z(x) \cdot \phi_z(y) dz}_{\min(x, y)}$$

multivariat \Leftrightarrow skalar

Induziert Betragssnorm!

$$\|\phi\mathbf{x} - \phi\mathbf{y}\|_{\mathbb{H}}^2$$

$$= \langle \phi\mathbf{x}, \phi\mathbf{x} \rangle + \langle \phi\mathbf{y}, \phi\mathbf{y} \rangle - 2 \cdot \langle \phi\mathbf{x}, \phi\mathbf{y} \rangle$$

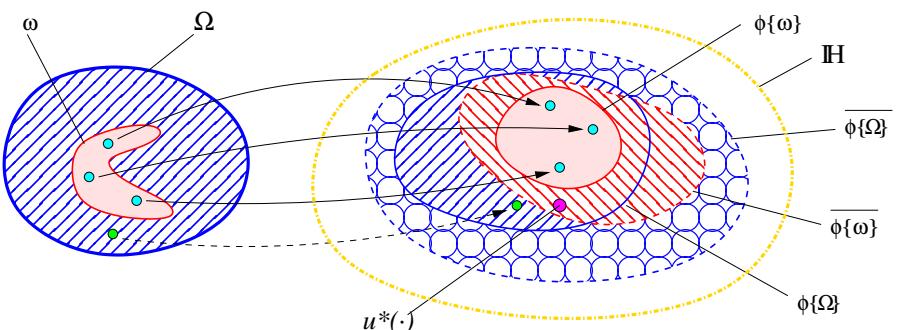
$$= K(\mathbf{x}, \mathbf{x}) + K(\mathbf{y}, \mathbf{y}) - 2 \cdot K(\mathbf{x}, \mathbf{y})$$

$$= \sum_i x_i + \sum_i y_i - 2 \cdot \sum_i \min(x_i, y_i)$$

$$= \sum_i |\mathbf{x}_i - \mathbf{y}_i| = \|\mathbf{x} - \mathbf{y}\|_1$$

$$\|\cdot\|_1\text{-Optimierung} \Leftrightarrow \|\cdot\|_2\text{-Optim.}$$

Originärer & virtueller Merkmalraum



Die Zimmer des Hilbertraums \mathbb{H}

- Expandierte Probenvektoren
- Expandierte Merkmalvektoren
- Lernbare Hyperebenen des Expansionsraums
- Kerreproduzierender Hilbertraum (RKHS)

$\phi\{\omega\}$
 $\phi\{\Omega\}$
 $\overline{\phi\{\omega\}}$
 $\overline{\phi\{\Omega\}}$

Die gelernte Prüfgröße $u^*(\cdot)$ liegt nicht notwendig in $\phi\{\Omega\}$ oder gar in $\phi\{\omega\}$.

Supportvektoren

SVM $\hat{=}$ semi-parametrischer Klassifikator

Wieviele Supportvektoren gibt es?

Ihre Anzahl $2 \leq \tilde{T} \leq T$ liegt erst nach Lösung des QOP vor.

Wen interessiert ihre Anzahl?

In kerngesteuerten SV-Maschinen ist der Rechenaufwand zur Klassifikation eines Musters meistens von der Ordnung $O(\tilde{T} \cdot D)$.

Wen interessiert ihre Anzahl noch?

Der Speicheraufwand für eine kerngesteuerte SVM ist ebenfalls von der Ordnung $O(\tilde{T} \cdot D)$.

Semiparametrische Klassifikationsverfahren

Aufwand unabhängig von $T = |\omega|$

Aufwand linear abhängig von $T = |\omega|$

Aufwand sublinear abhängig von $T = |\omega|$

parametrisch
nichtparametrisch
semiparametrisch

Supportvektorklassifikation mit linearem Kern

Beispiel: IRIS-Datensatz ($3 \times 50 \times 4$)

Grundidee

Separierbare Zwei-Klassen-Probleme

Nichtseparierbare Zwei-Klassen-Probleme

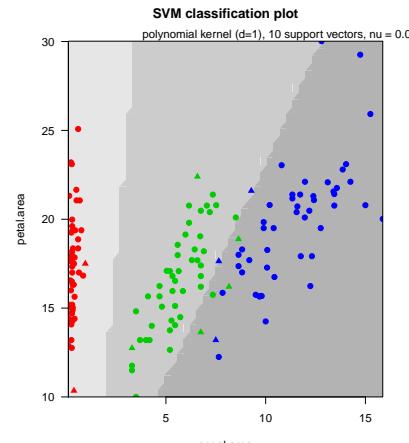
Nichtlineare Einbettungen des Merkmalraums

Beispiel: IRIS-Datensatz

Einfache Klassentrennung

X_1 = petale Blattfläche

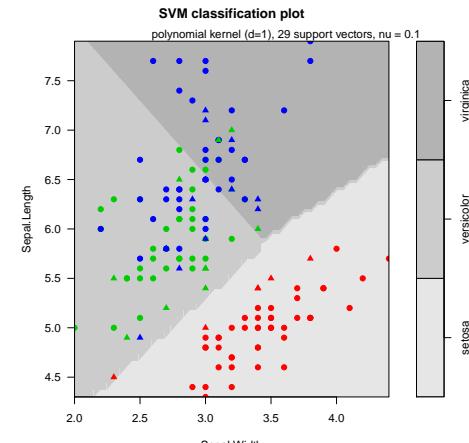
X_2 = sepale Blattfläche



Schwierige Klassentrennung

X_1 = petale Blattlänge

X_2 = petale Blattbreite

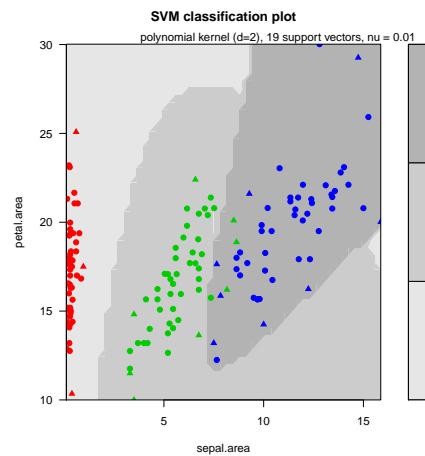


Supportvektorklassifikation mit quadratischem Kern

Beispiel: IRIS-Datensatz ($3 \times 50 \times 4$)

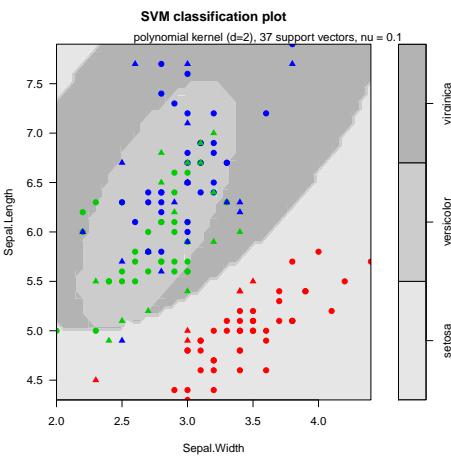
Einfache Klassentrennung

X_1 = petale Blattfläche
 X_2 = sepale Blattfläche



Schwierige Klassentrennung

X_1 = petale Blattlänge
 X_2 = petale Blattbreite

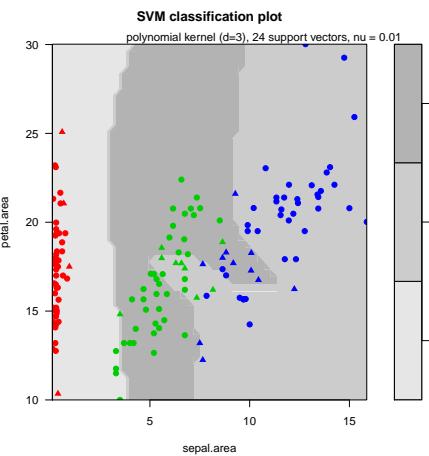


Supportvektorklassifikation mit kubischem Kern

Beispiel: IRIS-Datensatz ($3 \times 50 \times 4$)

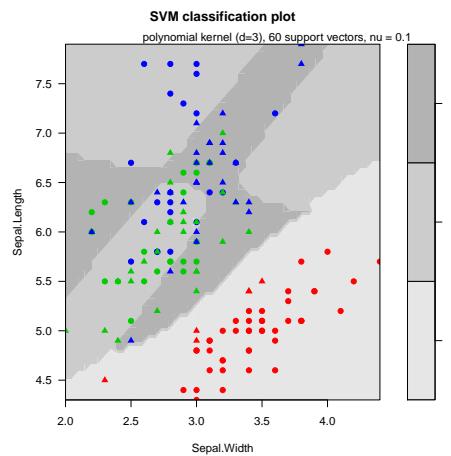
Einfache Klassentrennung

X_1 = petale Blattfläche
 X_2 = sepale Blattfläche



Schwierige Klassentrennung

X_1 = petale Blattlänge
 X_2 = petale Blattbreite

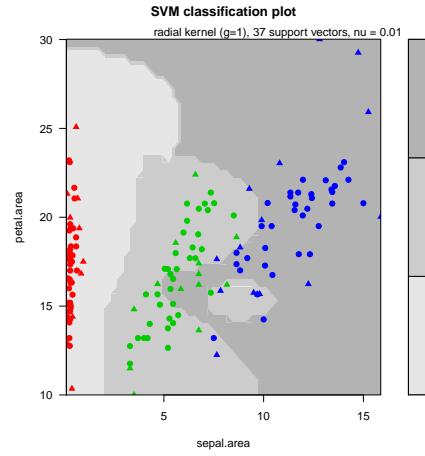


Supportvektorklassifikation mit radialem Kern

Beispiel: IRIS-Datensatz ($3 \times 50 \times 4$)

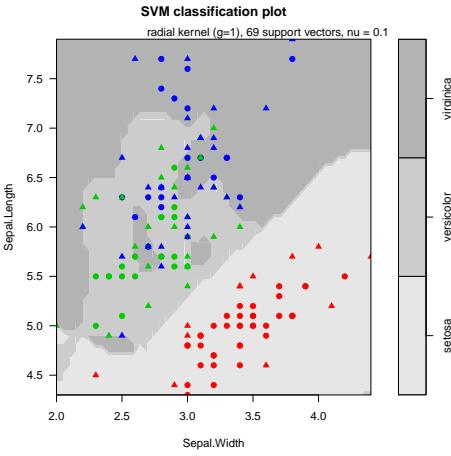
Einfache Klassentrennung

X_1 = petale Blattfläche
 X_2 = sepale Blattfläche



Schwierige Klassentrennung

X_1 = petale Blattlänge
 X_2 = petale Blattbreite



Zusammenfassung (9)

1. Eine **Supportvektormaschine** löst grundsätzlich nur Zwei-Klassen-Aufgaben.
2. Die Supportvektormaschine trennt die Muster zweier **separierbarer** Klassen durch eine **lineare Hyperebene** mit maximalem **Sicherheitsabstand**.
3. Sind die beiden Klassen **nicht separierbar**, so wird jede Verletzung der Sicherheitszone durch einen **Strafterm** geahndet; diese Regularisierung wird durch geeignete **Schlupfvariable** realisiert.
4. Die Berechnung der optimalen Trennfläche erfordert die Lösung einer **quadratischen Optimierungsaufgabe** mit **linearen Nebenbedingungen**.
5. Ihre Flächennormale ist eine Linearkombination aus den **Supportvektoren** — das sind genau diejenigen Lernvektoren mit **Sicherheitszonenkontakt**.
6. In der Lernphase wie auch der Klassifikationsphase werden ausschließlich die wechselseitigen **inneren Produkte** der Merkmalvektoren benötigt.
7. Die Operation der SVM kann daher auch in hochdimensionale **Termexpansionsräume** verlagert werden (**Kern-Trick**), wenn nur die Produkte $K(x, y) = \langle \phi x, \phi y \rangle$ effizient berechnet werden können.
8. Zahlreiche **Kernoperatoren** sind bekannt, etwa für **polomiale** und **dünne Parzen-artige** Prüfgrößen.