

11 Support Vektor Maschinen

11.1 Optimale Trennebene für linear separable Probleme

- betrachte (separables) Zweiklassenproblem
(auf das jedes Mehrklassenproblem zurückgeführt werden kann)
- Stichprobe: $\{(\vec{c}_i, y_i) \mid i = 1, \dots, I\}$, mit: $y_i \in \{+1, -1\}$
- gesucht ist Trennebene: $\vec{w}^T \vec{c} + b = 0$, sodaß gilt:

$$\vec{w}^T \vec{c}_i + b \geq 0 \text{ falls } y_i = +1 \quad (11.1)$$

$$\vec{w}^T \vec{c}_i + b < 0 \text{ falls } y_i = -1 \quad (11.2)$$

$$(11.3)$$

- mit Umskalierung erhalten wir:

$$y_i(\tilde{w}^T \vec{c}_i + \tilde{b}) \geq 1, \text{ für } i = 1, \dots, I \quad (11.4)$$

11.1 Optimale Trennebene für linear separable Probleme

1. wähle $\epsilon = \min_{i:y_i=-1} |\vec{w}^T \vec{c}_i + b|$

(beachte: für $y_i = +1$ könnte $\epsilon = 0$ werden)

2. sei $y_i = -1$ $\vec{w}^T \vec{c}_i + b < 0$

$$\vec{w}^T \vec{c}_i + b \leq -\epsilon$$

$$\vec{w}^T \vec{c}_i + b + \frac{\epsilon}{2} \leq -\frac{\epsilon}{2} \quad \left| +\frac{\epsilon}{2} \right.$$

$$\frac{2}{\epsilon} \vec{w}^T \vec{c}_i + \frac{2}{\epsilon} \left(b + \frac{\epsilon}{2} \right) \leq -1 \quad \left| \cdot \frac{2}{\epsilon} \right.$$

$$\vec{w}^T \vec{c}_i + \tilde{b} \leq -1,$$

$$\text{mit } \vec{w} := \frac{2}{\epsilon} \vec{w}^T, \tilde{b} := \frac{2}{\epsilon} \left(b + \frac{\epsilon}{2} \right)$$

$$y_i (\vec{w}^T \vec{c}_i + \tilde{b}) \geq 1$$

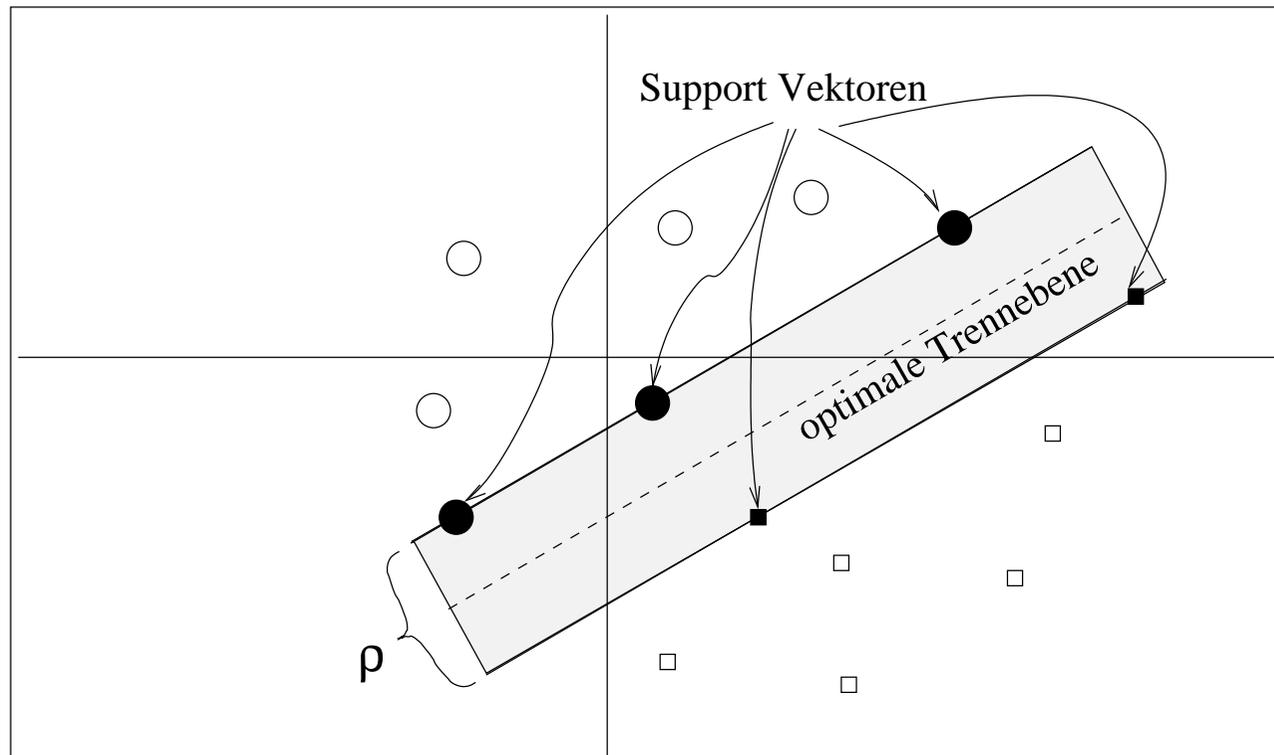
11.1 Optimale Trennebene für linear separable Probleme

3. nun sei $y_i = +1$

$$\begin{aligned} \vec{w}^T \vec{c}_i + b &\geq 0 \\ \vec{w}^T \vec{c}_i + b + \frac{\epsilon}{2} &\geq \frac{\epsilon}{2} && | +\frac{\epsilon}{2} \\ \frac{2}{\epsilon} \vec{w}^T \vec{c}_i + \frac{2}{\epsilon} \left(b + \frac{\epsilon}{2} \right) &\geq 1 && | \cdot \frac{2}{\epsilon} \\ \vec{\tilde{w}}^T \vec{c}_i + \tilde{b} &\geq 1 \\ y_i (\vec{\tilde{w}}^T \vec{c}_i + \tilde{b}) &\geq 1 \end{aligned}$$

11.1 Optimale Trennebene für linear separable Probleme

- **margin of separation** ρ ist die Breite des “Schlauches”, der die Stichprobenelemente \vec{c}_i beider Klassen trennt:

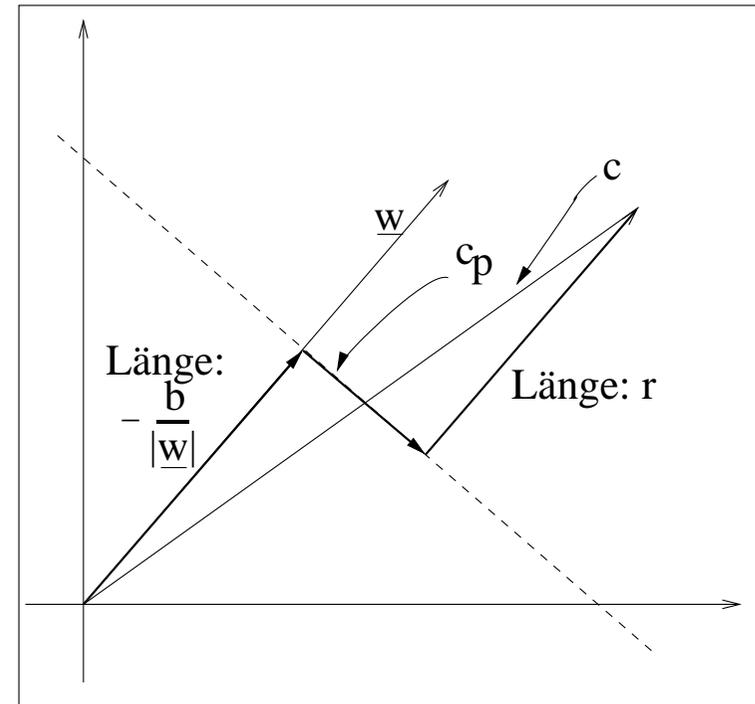


- die Trennebene mit maximalem ρ ist die **optimale Trennebene**
- die \vec{c}_i , die diesen Abstand einnehmen, sind die **Support Vektoren**
es gilt dann in (11.4) die Gleichheit

11.1 Optimale Trennebene für linear separable Probleme

- Abstand r von der Trennebene (\vec{w}, b) ,
mit der Funktion: $g(\vec{c}) := \vec{w}^T \vec{c} + b$ ausgedrückt

$$\begin{aligned}\vec{c} &= \vec{c}_p + \left(r - \frac{b}{|\vec{w}|} \right) \frac{\vec{w}}{|\vec{w}|} \\ g(\vec{c}) &= \vec{w}^T \left(\frac{r}{|\vec{w}|} - \frac{b}{|\vec{w}|^2} \right) \vec{w} + b \\ &= |\vec{w}|^2 \left(\frac{r}{|\vec{w}|} - \frac{b}{|\vec{w}|^2} \right) + b \\ &= r |\vec{w}| - b + b = r |\vec{w}| \\ r &= \frac{g(\vec{c})}{|\vec{w}|}\end{aligned}$$



- für jeden Support Vektoren \tilde{c} gilt

$$g(\tilde{c}) = \pm 1 \quad (11.5)$$

$$\tilde{r} = \frac{\pm 1}{|\vec{w}|} \quad (11.6)$$

$$\Rightarrow \rho = 2 \cdot \tilde{r} = \frac{2}{|\vec{w}|} \quad (11.7)$$

- Folgerung:

- der Vektor \vec{w} der optimalen Trennebene ist derjenige Vektor minimaler Norm, der die Bedingung (11.4) erfüllt
- die optimale Trennebene ist möglichst “flach” ($|\vec{w}|$ minimale)

- Bemerkung/Definition:

1. für jede Entscheidungsfunktion $g(\vec{c}; \vec{w}, b)$, die korrekt die gesamte Stichprobe trennt, ist auch $g(\vec{c}; \lambda\vec{w}, \lambda b)$, $\lambda > 0$, eine korrekte Trennebene
wir betrachten gerade diejenigen Trennebenen, für die die Support Vektoren c gilt: $g(\vec{c}) = \pm 1$
diese nennen wir **kanonische Trennebenen**

11.1 Optimale Trennebene für linear separable Probleme

2. von allen kanonischen Trennebenen wählen wir diejenige, mit minimalem $|\vec{w}|$

11.2 Strukturelle Risikominimierung

Satz (Vapnik) für eine Menge $g(\vec{c}; \vec{w}, b)$ von kanonischen Trennebenen in \mathbb{R}^d

- für eine Stichprobe der Größe N , sodaß R der Radius der kleinsten Hyperkugel ist, die alle Datenpunkte enthält,
- mit $|\vec{w}| \leq A$, gilt

die VC-Dimension ist $h \leq \min(R^2 A^2, d) + 1$

- alle kanonischen Trennebenen liefern $R_{\text{emp}}[g(\vec{c}; \vec{w}, b)] = 0$
- von all diesen hat diejenige mit minimalem $|\vec{w}|$ die kleinste VC-Dimension
- formal definieren wir also die geschachtelte Menge von Funktionsklassen:

$$H_l := \{g(\vec{c}; \vec{w}, b) \mid g(\vec{c}; \vec{w}, b) \text{ ist kanonische Trennebene mit } |\vec{w}| \leq A_l\}$$

mit $A_1 \leq A_2 \leq \dots \leq A_l$

- und die optimale kanonische Trennebene ist daher Lösung für SRM-Prinzip

11.2 Strukturelle Risikominimierung

Bemerkung es könnte (natürlich) passieren, daß eine Trennebene, die die Stichprobe nicht korrekt trenne, d.h. $R_{\text{emp}}[g(\vec{c}; \vec{w}, b)] > 0$ liefert, noch kleineres $|\vec{w}|$ und daher kleinere VC-Dimension hat, und u.U. noch kleineres strukturelles Risiko resultiert

11.3 Optimierung

Primäres Optimierungsproblem

- gesucht sind die optimalen Werte für \vec{w} und b , die
 - $\Phi(\vec{w}) = \frac{1}{2}\vec{w}^T \vec{w}$ minimieren und
 - $y_i(\vec{w}^T \vec{c}_i + b) \geq 1$, für $i = 1, \dots, I$ erfüllen
- Minimierung der Lagrange-Funktion:

$$J(\vec{w}, b, \vec{\alpha}) = \frac{1}{2}\vec{w}^T \vec{w} - \sum_{i=1}^I \alpha_i [y_i(\vec{w}^T \vec{c}_i + b) - 1]$$

mit den Lagrange-Multiplikatoren $\alpha_i \geq 0$

- leichtere Lösung mit identischem Optimum mit:

Duales Optimierungsproblem

- primäres und duales Problem haben identische Lösung
- gesucht sind die optimalen Werte für α_i , die

$$- Q(\vec{\alpha}) = \sum_{i=1}^I \alpha_i - \frac{1}{2} \sum_{i=1}^I \sum_{j=1}^I \alpha_i \alpha_j y_i y_j \vec{c}_i^T \vec{c}_j \text{ minimieren und}$$

$$- \sum_{i=1}^I \alpha_i y_i = 0 \text{ sowie}$$

$$- \alpha_i \geq 0 \text{ erfüllen}$$

- das duale Problem enthält \vec{w} und b **nicht** ! Optimum für

$$- \vec{w} = \sum_{i=1}^I \alpha_i y_i \vec{c}_i$$

wobei nur für Support Vektoren \vec{c}_i gelten kann: $\alpha_i \neq 0$

(d.h. es gehen ausschließlich Support Vektoren in die Lösung ein)

$$- b = y_i - \vec{w}^T \vec{c}_i, \text{ für ein beliebiges } i \text{ mit } \alpha_i \neq 0$$

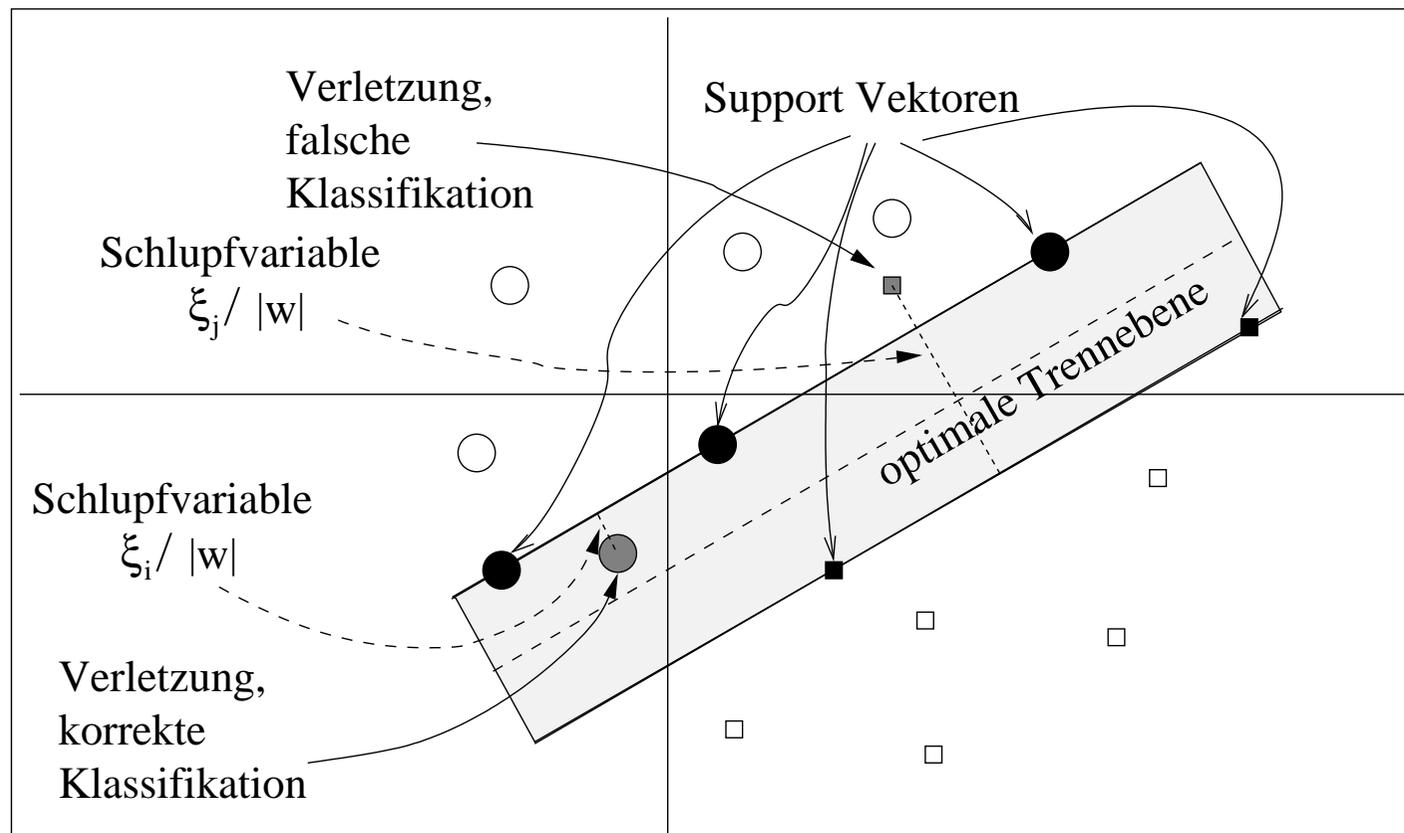
- liefert konvexes quadratisches Optimierungsproblem (nicht trivial)

11.4 Optimale Trennebene für nicht-linear separable Probleme

- erlaube Verletzung der optimalen Trennung um ξ_i :

$$y_i(\vec{w}^T \vec{c}_i + b) \geq 1 - \xi_i, \text{ für } i = 1, \dots, I \quad (11.8)$$

mit den **Schlupfvariablen (slack variables)** $\xi_i \geq 0$



11.4 Optimale Trennebene für nicht-linear separable Probleme

- neues Optimierungsziel:
optimale Trennebene, die den mittleren Klassifikationsfehler minimiert
(d.h. mittlere Anzahl ξ_i , mit $\frac{\xi_i}{|\vec{w}|} > 1$)
führt auf NP-vollständiges Problem, daher:
- als Approximation suche optimalen Werte für \vec{w} , die

- $\Phi(\vec{w}, \vec{\xi}) = \frac{1}{2} \vec{w}^T \vec{w} + C \sum_{i=1}^I \xi_i$ minimieren und

- $y_i(\vec{w}^T \vec{c}_i + b) \geq 1 - \xi_i$, für $i = 1, \dots, I$ sowie

- $\xi_i \geq 0$ erfüllen

wobei C ein Kontrollparameter ist

- Optimierung über ähnliches duales Problem wie im separablen Fall

11.5 Nicht-lineare Einbettung

- linear separable Probleme sind nicht sehr spannend (und real)
- durch nicht-lineare Einbettung (= Abbildung) in einen hoch-dimensionalen Merkmalsraum wird jedoch ein nicht-linear separables Problem mit hoher Wahrscheinlichkeit linear separabel (vgl. Polynom-Klassifikator!)

- wir betrachten $M > N$ nicht-lineare Funktionen $\phi_i : \mathbb{R}^N \rightarrow \mathbb{R}$
Diese liefern zusammen:

$$\vec{\phi} : \mathbb{R}^N \rightarrow \mathbb{R}^M, \quad \vec{\phi}(\vec{c}) = (\phi_1(\vec{c}), \dots, \phi_M(\vec{c}))$$

- Optimierung wie bisher, aber in \mathbb{R}^M

11.5 Nicht-lineare Einbettung

- zur Klassifikation müssen wir auswerten:

$$\vec{w}^T \vec{\phi}(\vec{c}) + b = \left(\sum_{i=1}^I \alpha_i y_i \vec{\phi}^T(\vec{c}_i) \right) \vec{\phi}(\vec{c}) + b = \sum_{i=1}^I \alpha_i y_i \underbrace{\left(\vec{\phi}^T(\vec{c}_i) \vec{\phi}(\vec{c}) \right)}_{K(\vec{c}_i, \vec{c})} + b \quad (11.9)$$

(wegen $\vec{w} = \sum_{i=1}^I \alpha_i y_i \vec{c}_i$) mit dem (symmetrischen) **Kernoperator** $K(\vec{c}_i, \vec{c})$

11.5 Nicht-lineare Einbettung

- was haben wir gewonnen?

für gewisse Klassen von Kernoperator können wir die Operation im ursprünglichen Eingaberaum durchführen, und müssen nicht explizit in den hoch-dimensionalen Merkmalsraum gehen

(muß also doch nicht in \mathbb{R}^M erfolgen)

- Beispiel:

$$\begin{aligned}\vec{\phi}(c_1, c_2) &= (c_1^2, \sqrt{2} c_1 c_2, c_2^2) && \Rightarrow K(\vec{c}, \vec{c}_i) = (\vec{c}^T \vec{c}_i)^2 \\ \vec{\phi}(c_1, c_2) &= (c_1^2, c_2^2, \sqrt{2} c_1 c_2, \sqrt{2} c_1, \sqrt{2} c_2, 1) && \Rightarrow K(\vec{c}, \vec{c}_i) = (\vec{c}^T \vec{c}_i + 1)^2\end{aligned}$$

- auch das (duale) Optimierungsproblem – im linear separablen wie nicht-linear separablen Fall – kann damit im ursprünglichen Eingaberaum bearbeitet werden !

11.5 Nicht-lineare Einbettung

Wichtige Kernoperatoren

Kernoperator	Klassifikator	Bemerkung
$(\vec{c}^T \vec{c}_i + 1)^G$	Polynomklassifikator	Polynomgrad G apriori vorgegeben
$e^{-\frac{1}{2}\left(\frac{\vec{c}-\vec{c}_i}{\sigma}\right)^2}$	Radiale Basisfunktionen	gemeinsame, vorgegebene Breite σ^2
$\tanh(\beta_0 \vec{c}^T \vec{c}_i + \beta_1)$	zweilagiges Perzeptron	nur für einige Werte von β_0 und β_1

Bemerkungen

- die Dimensionalität des Merkmalsraumes wird durch die Optimierung als Anzahl der Support Vektoren vorgegeben (nicht durch den Designer)