10 Support Vektor Maschinen

10.1 Optimale Trennebene für linear separable Probleme

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

$$\vec{w}^T \vec{c}_i + b \geq 0 \text{ falls } \omega_i = +1 \quad (10.1)$$
$$\vec{w}^T \vec{c}_i + b < 0 \text{ falls } \omega_i = -1 \quad (10.2)$$
$$(10.3)$$

- mit Umskalierung erhalten wir:

$$\omega_i(\vec{w}^T \vec{c}_i + b) \geq 1, \text{für } i = 1, \ldots, I \quad (10.4)$$

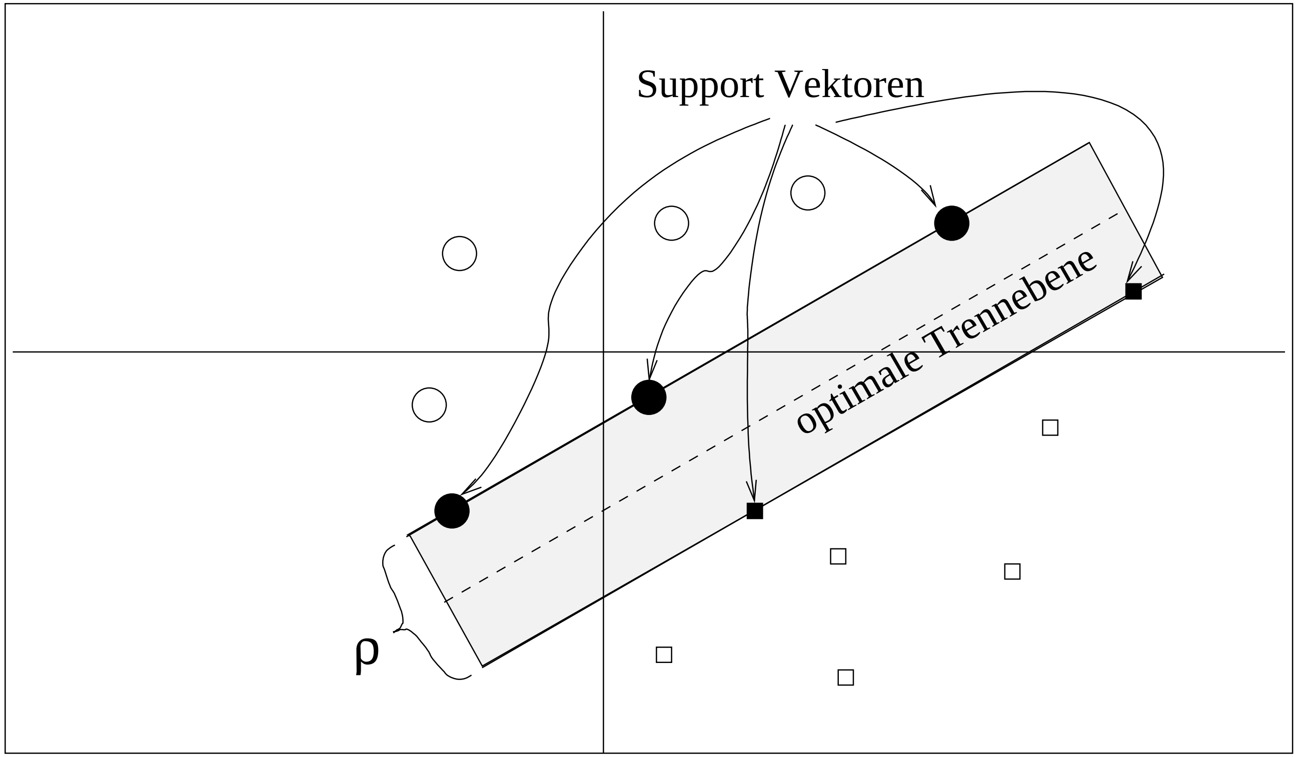
1. sei $\omega_i = -1$

$$\begin{aligned}
\vec{w}^T\vec{c}_i + b &< 0 \\
\vec{w}^T\vec{c}_i + b &\leq -\epsilon, && \text{für ein geeignetes } \epsilon > 0 \\
\vec{w}^T\vec{c}_i + b + \frac{\epsilon}{2} &\leq -\frac{\epsilon}{2} && \mid +\frac{\epsilon}{2} \\
\frac{2}{\epsilon}\vec{w}^T\vec{c}_i + \frac{2}{\epsilon}\left(b + \frac{\epsilon}{2}\right) &\leq -1 && \mid \cdot\frac{2}{\epsilon} \\
\tilde{\vec{w}}^T\vec{c}_i + \tilde{b} &\leq -1, && \text{mit } \tilde{\vec{w}} := \frac{2}{\epsilon}\vec{w}^T, \tilde{b} := \frac{2}{\epsilon}\left(b + \frac{\epsilon}{2}\right) \\
\omega_i(\tilde{\vec{w}}^T\vec{c}_i + \tilde{b}) &\geq 1
\end{aligned}$$

2. nun sei $\omega_i = +1$

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

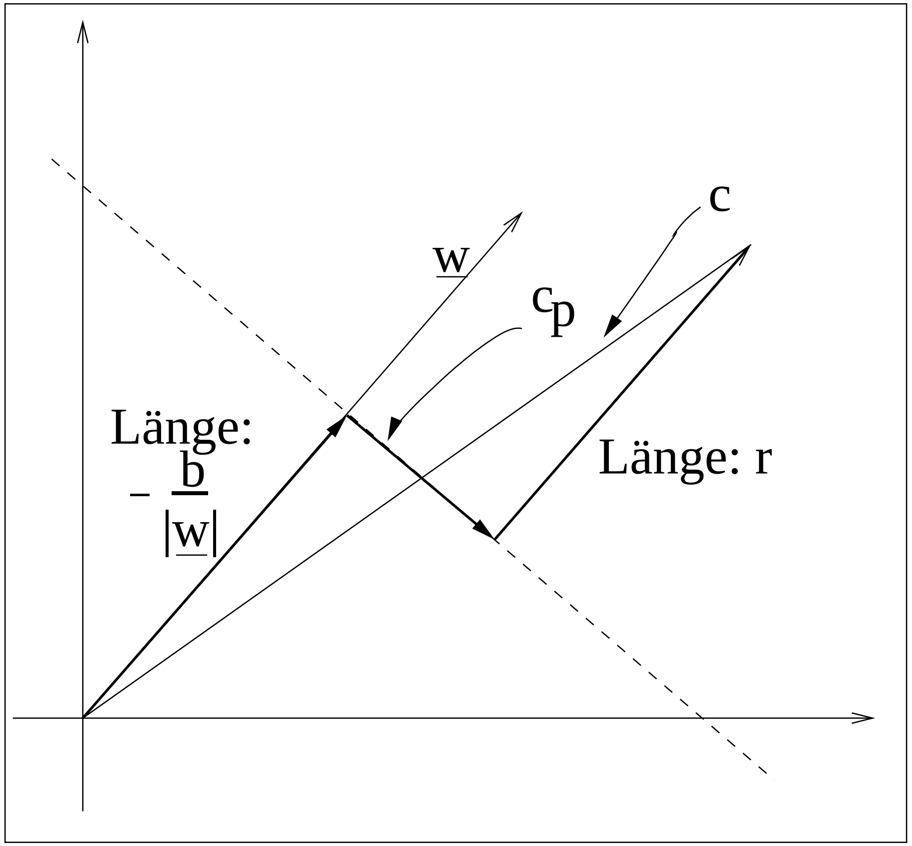
- **margin of separation** ρ ist die Breite des “Schlauches”, der die Stichprobenelemente $\vec{c}_i$ beider Klassen trennt:



- 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 (10.4) die Gleichheit

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

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



- für jeden Support Vektoren $\tilde{\vec{c}}$ gilt

$$
\begin{aligned}
g(\tilde{\vec{c}}) &= \pm 1 && (10.5)\\
\tilde{r} &= \frac{\pm 1}{\mid \vec{w} \mid} && (10.6)\\
\Rightarrow \rho &= 2 \cdot \tilde{r} = \frac{2}{\mid \vec{w} \mid} && (10.7)
\end{aligned}
$$

- Folgerung:
  - der Vektor $\vec{w}$ der optimalen Trennenebe ist derjenige Vektor minimaler Norm, der die Bedingung (10.4) erfüllt
  - die optimale Trennebene ist möglichst “flach” ($\mid \vec{w} \mid$ minimale)

10.2 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
  - $\omega_i(\vec{w}^T\vec{c}_i + b) \geq 1,$ für $i = 1, \ldots, I$ erfüllen
- Minimierung der Lagrange-Funktion:

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

  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
  - $Q(\vec{\alpha}) = \sum_{i=1}^{I} \alpha_i - \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{I} \alpha_i \alpha_j \omega_i \omega_j \vec{c}_i^T \vec{c}_j$ minimieren und
  - $\sum_{i=1}^{I} \alpha_i \omega_i = 0$ sowie
  - $\alpha_i \geq 0$ erfüllen
- das duale Problem enthält $\vec{w}$ und b **nicht** ! Optimum für
  - $\vec{w} = \sum_{i=1}^{I} \alpha_i \omega_i \vec{c}_i$

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

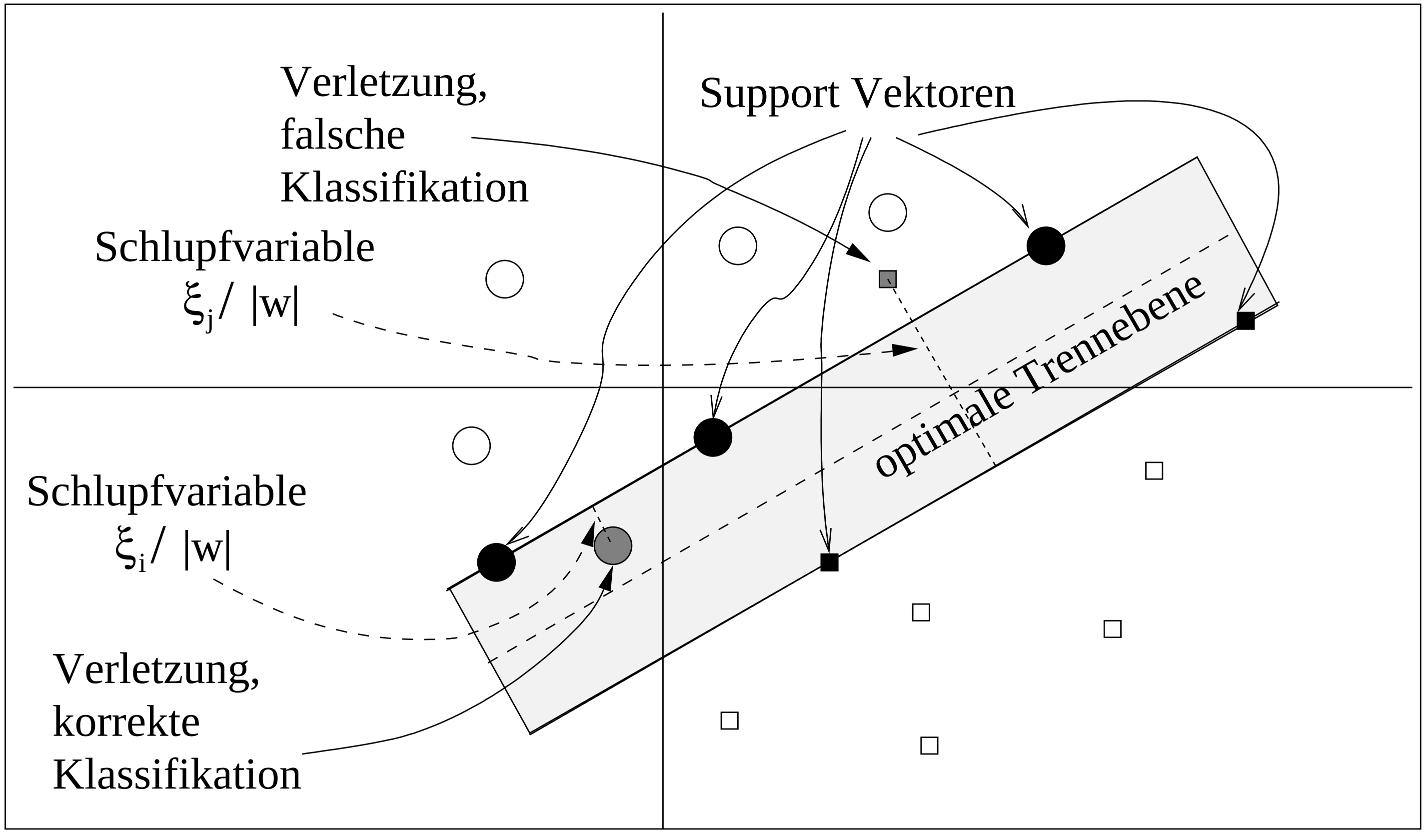
    (d.h. es gehen ausschließlich Support Vektoren in die Lösung ein)
  - $b = \omega_i - \vec{w}^T \vec{c}_i$, für ein beliebiges i mit $\alpha_i \neq 0$
- liefert konvexes quadratisches Optimierungsproblem (nicht trivial)

10.3 Optimale Trennebene für nicht-linear separable Probleme

- erlaube Verletzung der optimalen Trennung um ξ_i:

$$\omega_i(\vec{w}^T\vec{c}_i + b) \geq 1 - \xi_i, \text{ für } i = 1, \ldots, I \tag{10.8}$$

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



- neues Optimierungsziel:
  optimale Trennebene, die den mittleren Klassifikationsfehler minimiert
  (d.h. mittlere Anzahl ξ_i, mit $\xi_i > 1$)
  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
  - $\omega_i(\vec{w}^T\vec{c}_i + b) \geq 1 - \xi_i,$ für $i = 1, \ldots, I$ sowie
  - $\xi_i \geq 0$ erfüllen

  wobei C ein Kontrollparameter ist
- Optimierung über ähnliches duales Problem wie im separablen Fall

10.4 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:

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

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

- zur Klassifikation müssen wir auswerten:

$$\vec{w}^T \vec{\phi}(\vec{c}) + b = \left(\sum_{i=1}^{I} \alpha_i \omega_i \vec{\phi}^T(\vec{c}_i) \right) \vec{\phi}(\vec{c}) + b = \sum_{i=1}^{I} \alpha_i \omega_i \underbrace{\left(\vec{\phi}^T(\vec{c}_i) \vec{\phi}(\vec{c}) \right)}_{K(\vec{c}_i, \vec{c})} + b \tag{10.9}$$

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

- was haben wir gewonnen?

  für eine Klasse 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:

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

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

Wichtige Kernoperatoren

Kernoperator	Klassifikator	Bemerkung
$\left(\vec{c}^T \vec{c}_i + 1\right)^G$	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)