11 Support Vektor Maschinen

11.1 Optimale Trennebene für linear separable Probleme

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

$$\vec{w}^T\vec{c}_n + b \geq 0 \text{ falls } y_n = +1 \quad (11.1)$$
$$\vec{w}^T\vec{c}_n + b < 0 \text{ falls } y_n = -1 \quad (11.2)$$
$$(11.3)$$

- mit Umskalierung erhalten wir:

$$y_n(\tilde{\vec{w}}^T\vec{c}_n + \tilde{b}) \geq 1, \text{für } n = 1, \ldots, N \quad (11.4)$$

1. wähle $\epsilon = \min_{n:y_n=-1} \left|\vec{w}^T\vec{c}_n + b\right|$

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

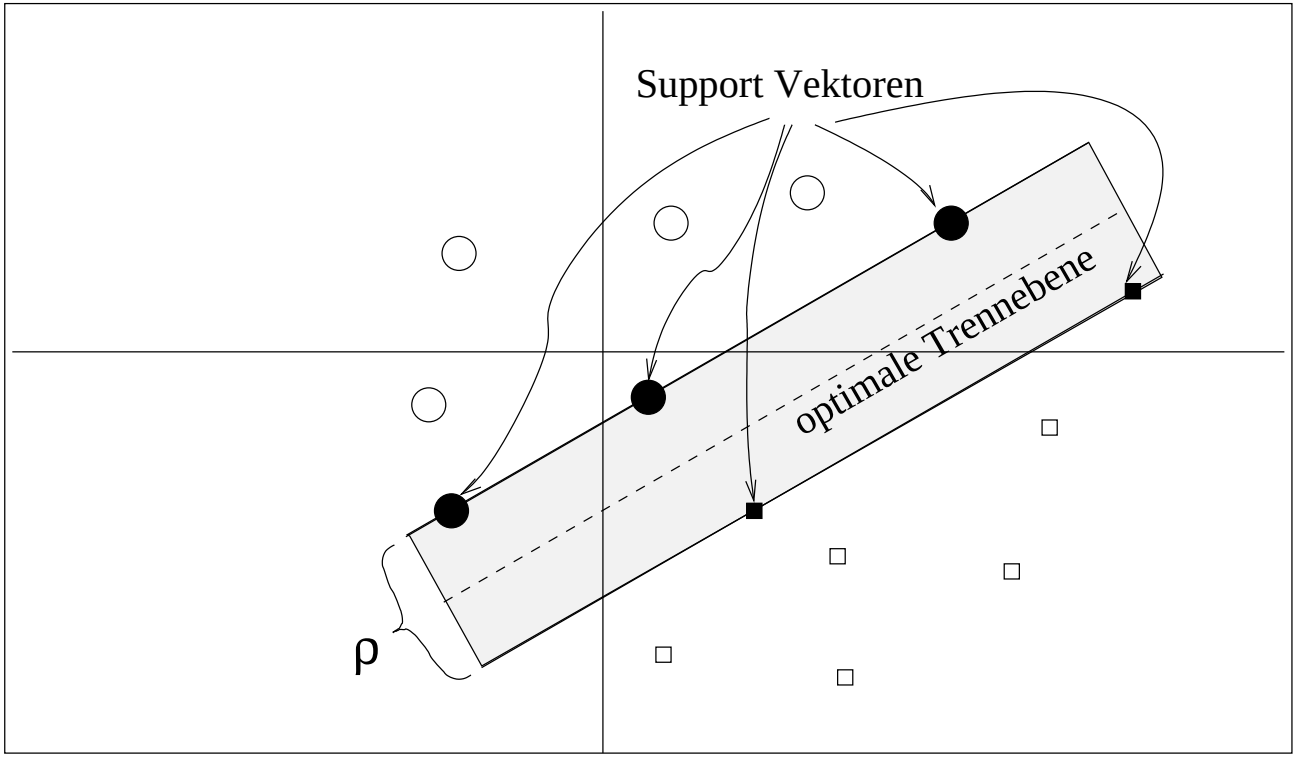
2. sei $y_n = -1$

$$\begin{aligned}
\vec{w}^T\vec{c}_n + b &< 0 \\
\vec{w}^T\vec{c}_n + b &\leq -\epsilon \\
\vec{w}^T\vec{c}_n + b + \frac{\epsilon}{2} &\leq -\frac{\epsilon}{2} && \left| + \frac{\epsilon}{2}\right. \\
\frac{2}{\epsilon}\vec{w}^T\vec{c}_n + \frac{2}{\epsilon}\left(b + \frac{\epsilon}{2}\right) &\leq -1 && \left| \cdot \frac{2}{\epsilon}\right. \\
\vec{\tilde{w}}^T\vec{c}_n + \tilde{b} &\leq -1, && \text{mit } \vec{\tilde{w}} := \frac{2}{\epsilon}\vec{w}^T, \tilde{b} := \frac{2}{\epsilon}\left(b + \frac{\epsilon}{2}\right) \\
y_n(\vec{\tilde{w}}^T\vec{c}_n + \tilde{b}) &\geq 1
\end{aligned}$$

3. nun sei $y_n = +1$

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

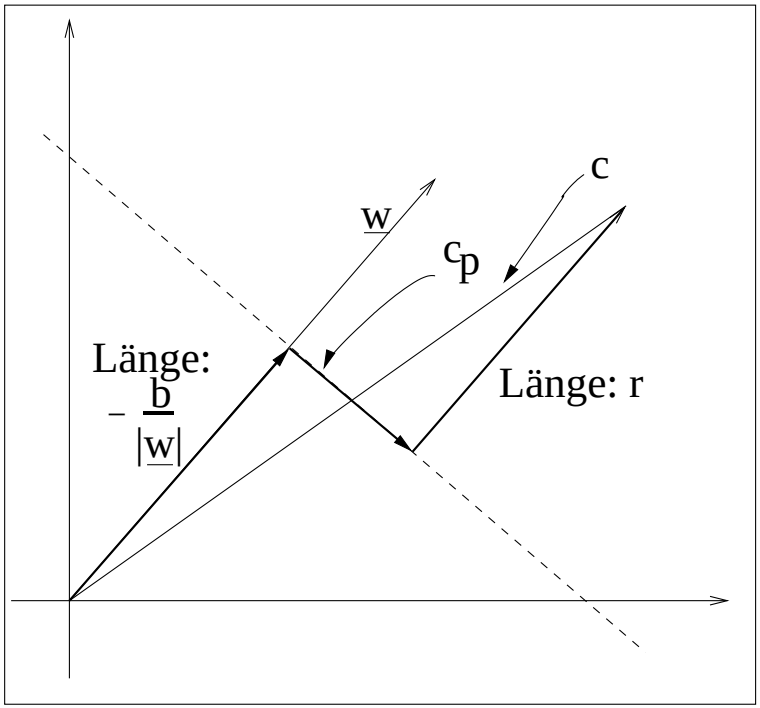
- **margin of separation** ρ ist die Breite des “Schlauches”, der die Stichprobenelemente $\vec{c}_n$ beider Klassen trennt:



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

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

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



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

$$d(\tilde{\vec{c}}) = \pm 1 \tag{11.5}$$

$$\tilde{r} = \frac{\pm 1}{\mid \vec{w} \mid} \tag{11.6}$$

$$\Rightarrow \rho = 2 \cdot \tilde{r} = \frac{2}{\mid \vec{w} \mid} \tag{11.7}$$

- Folgerung:
  - der Vektor $\vec{w}$ der optimalen Trennenebe ist derjenige Vektor minimaler Norm, der die Bedingung (11.4) erfüllt
  - die optimale Trennebene ist möglichst “flach” ($\mid \vec{w} \mid$ minimale)
- Bemerkung/Definition:
  1. für jede Entscheidungsfunktion $d(\vec{c}; \vec{w}, b)$, die korrekt die gesamte Stichprobe trennt, ist auch $d(\vec{c}; \lambda\vec{w}, \lambda b)$, $\lambda > 0$, eine korrekte Trennebene
     wir betrachten gerade diejenigen Trennebenen, für die die Support Vektoren c gilt: $d(\vec{c}) = \pm 1$
     diese nennen wir **kanonische Trennebenen**

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

11.2 Struktukturelle Risikominimierung

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

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

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

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

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

mit $A_1 \leq A_2 \leq \cdots \leq A_l$

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

Bemerkung es könnte (natürlich) passieren, daß eine Trennebene, die die Stichprobe nicht korrekt trenne, d.h. $R_{\text{emp}}[d(\vec{c}; \vec{w}, b)] > 0$ liefert, noch kleineres $\mid \vec{w} \mid$ 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
  - $y_n(\vec{w}^T\vec{c}_n + b) \geq 1$, für $n = 1, \ldots, N$ erfüllen
- Minimierung der Lagrange-Funktion:

$$J(\vec{w}, b, \vec{\alpha}) = \frac{1}{2}\vec{w}^T\vec{w} - \sum_{n=1}^{N} \alpha_n \left[y_n(\vec{w}^T\vec{c}_n + b) - 1\right]$$

  mit den Lagrange-Multiplikatoren $\alpha_n \geq 0$
- leichtere Lösung mit identischem Optimum mit:

Duales Optimierungsproblem

- primäres und duales Problem haben identische Lösung
- gesucht sind die optimalen Werte für α_n, die
  - $Q(\vec{\alpha}) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{j=1}^{I} \alpha_n \alpha_j y_n y_j \vec{c}_n^T \vec{c}_j$ minimieren und
  - $\sum_{n=1}^{N} \alpha_n y_n = 0$ sowie
  - $\alpha_n \geq 0$ erfüllen
- das duale Problem enthält $\vec{w}$ und b **nicht** ! Optimum für
  - $\vec{w} = \sum_{n=1}^{N} \alpha_n y_n \vec{c}_n$

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

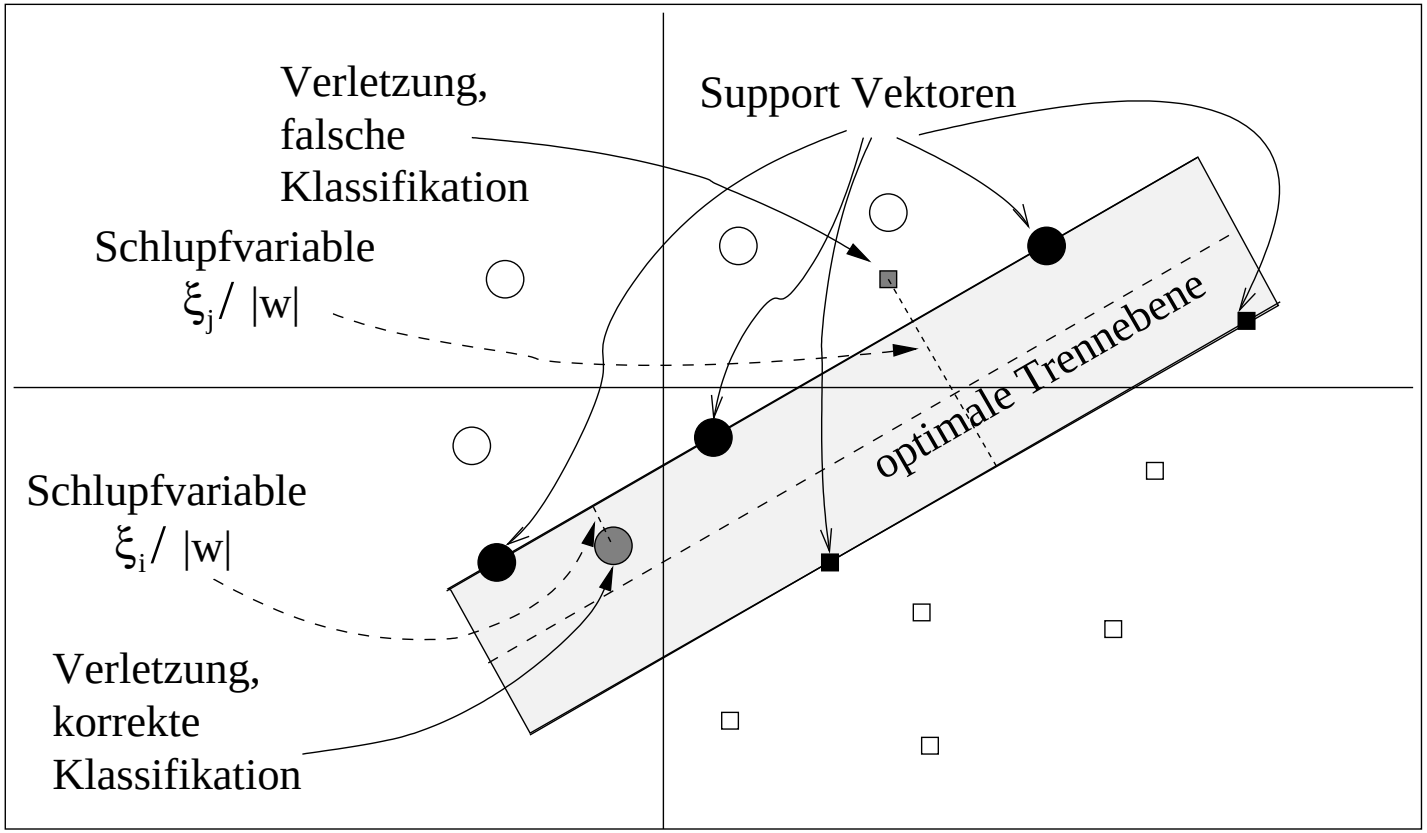
    (d.h. es gehen ausschließlich Support Vektoren in die Lösung ein)
  - $b = y_n - \vec{w}^T \vec{c}_n$, für ein beliebiges i mit $\alpha_n \neq 0$
- liefert konvexes quadratisches Optimierungsproblem (nicht trivial)

11.4 Optimale Trennebene für nicht-linear separable Probleme

- erlaube Verletzung der optimalen Trennung um ξ_n:

$$y_n(\vec{w}^T\vec{c}_n + b) \geq 1 - \xi_n, \text{ für } n = 1, \ldots, N \tag{11.8}$$

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



- neues Optimierungsziel:
  optimale Trennebene, die den mittleren Klassifikationsfehler minimiert
  (d.h. mittlere Anzahl ξ_n, mit $\frac{\xi_n}{|\vec{w}|} > 1$)
  führt auf NP-vollständiges Problem, daher:
- als Approximation suche optimalen Werte für $\vec{w}$, die
  - $\Phi(\vec{w}, \vec{\xi}) = \frac{1}{2}\vec{w}^T\vec{w} + C\sum_{n=1}^{N}\xi_n$ minimieren und
  - $y_n(\vec{w}^T\vec{c}_n + b) \geq 1 - \xi_n,$ für $n = 1, \ldots, N$ sowie
  - $\xi_n \geq 0$ 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 $\phi_n : \mathbb{R}^N \to \mathbb{R}$

  Diese liefern zusammen:

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

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

- zur Klassifikation müssen wir auswerten:

$$\vec{w}^T\vec{\phi}(\vec{c}) + b = \left(\sum_{n=1}^{N} \alpha_n y_n \vec{\phi}^T(\vec{c}_n)\right) \vec{\phi}(\vec{c}) + b = \sum_{n=1}^{N} \alpha_n y_n \underbrace{\left(\vec{\phi}^T(\vec{c}_n)\vec{\phi}(\vec{c})\right)}_{K\left(\vec{c}_n, \vec{c}\right)} + b \qquad (11.9)$$

(wegen $\vec{w} = \sum_{n=1}^{N} \alpha_n y_n \vec{c}_n$) mit dem (symmetrischen) **Kernoperator** $K\left(\vec{c}_n, \vec{c}\right)$

- 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:

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

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

Wichtige Kernoperatoren

Kernoperator	Klassifikator	Bemerkung
$\left(\vec{c}^T\vec{c}_n + 1\right)^G$	Polynomklassifikator	Polynomgrad G apriori vorgegeben
$e^{-\frac{1}{2}\left(\frac{\vec{c}-\vec{c}_n}{\sigma}\right)^2}$	Radiale Basisfunktionen	gemeinsame, vorgegebene Breite σ^2
$\tanh(\beta_0\vec{c}^T\vec{c}_n + \beta_1)$	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)