8. Support-Vektor-Maschinen

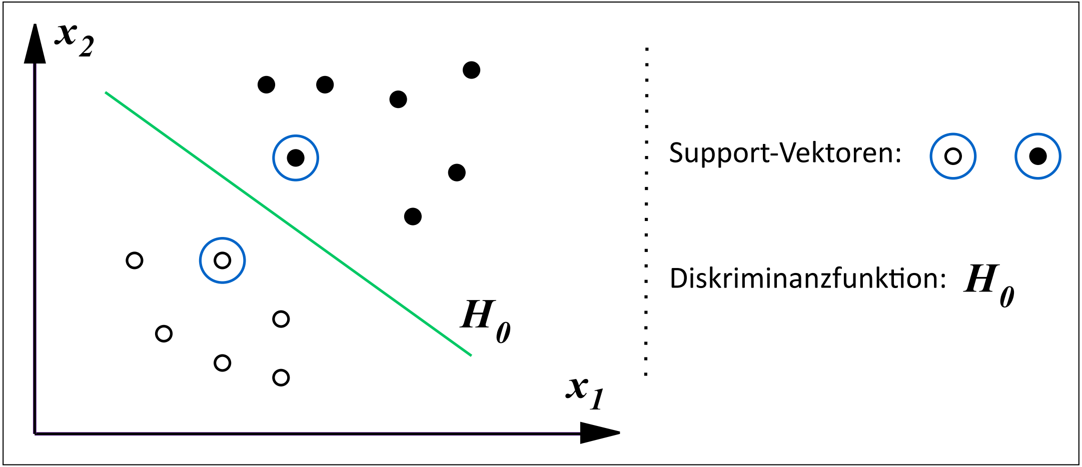
Synonym: Support-Vektor-Netze.
Verwendung: Klassifikation oder Funktionsapproximation.
Hier: Fokussierung auf Klassifikation (zwei Klassen).

8.1 SVM mit linearer Diskriminanzfunktion

Überblick:

- Support-Vektoren, Diskriminanzfunktion
- Trennband maximaler Breite
- Quadratische Optimierungsaufgabe
- Hinweis auf die optimale separierende Hyperebene
- Äquivalente quadratische Optimierungsaufgabe
- Berechnung der optimalen separierenden Hyperebene

Support-Vektoren, Diskriminanzfunktion



Trennband maximaler Breite

Aufgabe: Bestimmung einer Diskriminanzfunktion, so daß das dazu parallele Trennband maximale Breite hat.

Menge $\boldsymbol{\Omega_T} := \{(\boldsymbol{x}^1, \boldsymbol{r}^1), \cdots, (\boldsymbol{x}^M, \boldsymbol{r}^M)\}$, $\boldsymbol{r}^m \in \{-1, +1\}$

$\boldsymbol{\Omega_T}$ ist absolut trennbar, falls $\boldsymbol{w} := (\boldsymbol{w}_1, \cdots, \boldsymbol{w}_I)^T$ und $\boldsymbol{w}_0$ existieren, mit

$$\begin{aligned} \boldsymbol{w}^T \boldsymbol{x}^m + \boldsymbol{w}_0 &\geq \;\; 1, \quad \text{falls } \boldsymbol{r}^m = 1 \\ \boldsymbol{w}^T \boldsymbol{x}^m + \boldsymbol{w}_0 &\leq -1, \quad \text{falls } \boldsymbol{r}^m = -1 \end{aligned}$$

Trennband maximaler Breite

Zusammenfassung beider Fälle:

$$\boldsymbol{r^m(w^T x^m + w_0) \geq 1,\ \forall m \in \{1, \cdots, M\}} \tag{1}$$

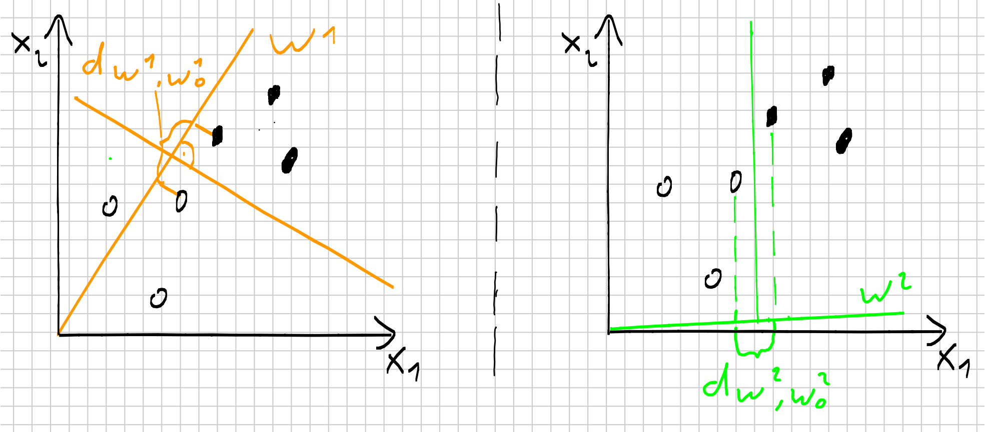
Optimale separierende Hyperebene ist definiert durch:

$$\boldsymbol{w^{*,T} x + w_0^* = 0},$$

wobei $\frac{\boldsymbol{w}^*}{\|\boldsymbol{w}^*\|}$ diejenige Richtung ist, bei welcher nach Projektion der Trainingsvektoren der Abstand zwischen den resultierenden Randpunkten der Klassen maximal ist. Genannter Abstand ist definiert durch:

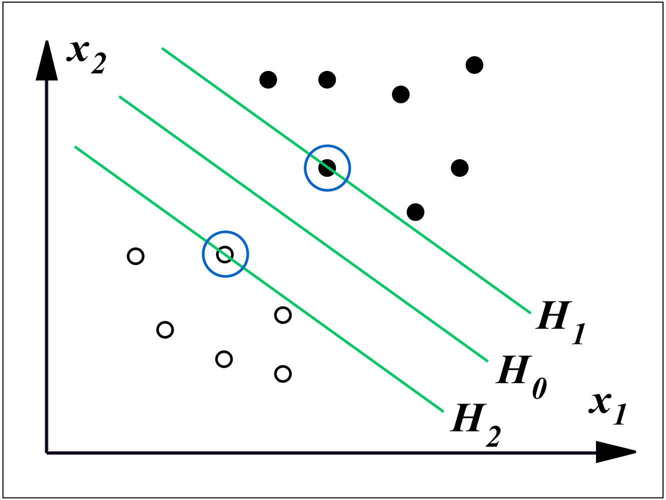
$$\boldsymbol{d_{w,w_o} := \min_{\{(x,r=1)\}} \frac{w^T x}{\|w^T\|} - \max_{\{(x,r=-1)\}} \frac{w^T x}{\|w^T\|}}$$

Trennband maximaler Breite



$d_{w^2,w_0^2} < d_{w^1,w_0^1}$, also w^1 besser

Trennband maximaler Breite



Trennband maximaler Breite

Für einen Trainingsvektor auf $\boldsymbol{H_1}$ gilt: $\boldsymbol{w^T x + w_0 = 1}$

$\Rightarrow$ orthogonaler Abstand zum Ursprung: $\frac{\boldsymbol{(1 - w_0)}}{\boldsymbol{\|w\|}}$

Für einen Trainingsvektor auf $\boldsymbol{H_2}$ gilt: $\boldsymbol{w^T x + w_0 = -1}$

$\Rightarrow$ orthogonaler Abstand zum Ursprung: $\frac{\boldsymbol{(-1 - w_0)}}{\boldsymbol{\|w\|}}$

Also $\boldsymbol{d_{w,w_o} := \frac{2}{\|w\|}}$.

Gesucht: Optimale separierende Hyperebene durch Minimierung von $\boldsymbol{w^T w}$ unter der Nebenbedingung gemäß Gleichung (1).

Quadratische Optimierungsaufgabe

Standardansatz mit der Lagrange-Funktion:

$\boldsymbol{f_{Lagr}(w, w_0, B)} :=$

$$\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} - \sum_{m=1}^{M} \beta^m(r^m(\boldsymbol{w}^T\boldsymbol{x}^m + w_0) - 1) \qquad (2)$$

Sattelpunktfunktion f_{Lagr} in $M + I + 1$ Parametern.

Lagrange-Multiplikatoren $B := (\beta^1, \cdots, \beta^M)^T$, nicht-negativ.

Quadratische Optimierungsaufgabe

Ausmultiplizieren von Gleichung (2) liefert:

$$\boldsymbol{f_{Lagr}(w, w_0, B)} := \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \sum_{m=1}^{M} \boldsymbol{\beta}^m - \sum_{m=1}^{M} \boldsymbol{\beta}^m \boldsymbol{r}^m \boldsymbol{w_0} - \sum_{m=1}^{M} \boldsymbol{\beta}^m \boldsymbol{r}^m \boldsymbol{w}^T \boldsymbol{x}^m \tag{3}$$

Quadratische Optimierungsaufgabe

Satz aus der Theorie der Quadratischen Optimierung:

Sattelpunkt ist Lösung der Optimierungsaufgabe.

Minimierung bezüglich $\boldsymbol{w}, \boldsymbol{w_0}$ und Maximierung bezüglich $\boldsymbol{B}$.

Lösung erhält man prinzipiell durch Anwendung einer mathematischen Standardmethode (z.B. in MATLAB).

Hinweis auf die optimale separierende Hyperebene

Bestimmung von $\boldsymbol{w}^*$:

$$\nabla_{\boldsymbol{w}} \boldsymbol{f}_{\boldsymbol{Lagr}}(\boldsymbol{w}, \boldsymbol{w}_0, \boldsymbol{B}) = \boldsymbol{w} - \sum_{m=1}^{M} \boldsymbol{\beta}^m \boldsymbol{r}^m \boldsymbol{x}^m \stackrel{!}{=} \vec{\mathbf{0}}$$

$$\Rightarrow \qquad \boldsymbol{w} = \sum_{m=1}^{M} \boldsymbol{\beta}^m \boldsymbol{r}^m \boldsymbol{x}^m \tag{4}$$

Hinweis: Optimale separierende Hyperebene basiert auf Linearkombination von denjenigen Trainingsvektoren mit $\boldsymbol{\beta}^m > \mathbf{0}$.

Bestimmung von $\boldsymbol{w}_0^*$:

Siehe später in diesem Unterkapitel.

Äquivalente quadratische Optimierungsaufgabe

Nachfolgend wird eine äquivalente quadratische Optimierungsaufgabe hergeleitet. Dabei zeigt sich, daß

- nur diejenigen Trainingsvektoren von $\boldsymbol{\Omega_T}$ eine Rolle spielen, die auf $\boldsymbol{H_1}$ und $\boldsymbol{H_2}$ liegen, welche somit als Support-Vektoren bezeichnet werden;
- beim Training und bei der Anwendung die Input-Vektoren aus $\boldsymbol{\Omega}$ immer nur paarweise als Skalarprodukte vorkommen, auf dessen Grundlage auch nicht-lineare Diskriminanzfunktionen lernbar sind (siehe später in Unterkapitel 8.2).

Äquivalente quadratische Optimierungsaufgabe

In der Lösung der quadratischen Optimierung gilt insbesondere:

$$\frac{\partial f_{Lagr}(\boldsymbol{w}, w_0, \boldsymbol{B})}{\partial w_0} = \sum_{m=1}^{M} \beta^m r^m \stackrel{!}{=} 0 \tag{5}$$

Einsetzen Gleichungen (5) und (4) in Gleichung (3) liefert:

$$\begin{aligned} f_{Lagr}(\boldsymbol{w}, w_0, \boldsymbol{B}) &= \sum_{m=1}^{M} \beta^m - \frac{1}{2} \underbrace{(\sum_{m=1}^{M} \beta^m r^m \boldsymbol{x}^m)}_{\boldsymbol{w}}{}^T \underbrace{(\sum_{l=1}^{M} \beta^l r^l \boldsymbol{x}^l)}_{\boldsymbol{w}} \\ &= \boldsymbol{B}^T \boldsymbol{U} - \frac{1}{2} \boldsymbol{B}^T \boldsymbol{D} \boldsymbol{B} =: f_{\ddot{A}qui}(\boldsymbol{B}) \end{aligned} \tag{6}$$

mit $\boldsymbol{U} := (1, \cdots, 1)^T$, $\boldsymbol{D} := (D_{ml})_{\substack{m=1\cdots M \\ l=1\cdots M}}$, $D_{ml} := r^m r^l \boldsymbol{x}^{m,T} \boldsymbol{x}^l$

Äquivalente quadratische Optimierungsaufgabe

Äquivalente Optimierungsaufgabe: Maximierung von $\boldsymbol{f_{\ddot{A}qui}(B)}$ bezüglich $\boldsymbol{B}$ unter den Nebenbedingungen $\boldsymbol{\beta^m \geq 0}$, $\forall \boldsymbol{m} \in \{\mathbf{1}, \cdots, \boldsymbol{M}\}$.

Am Sattelpunkt $\boldsymbol{w^*, w_0^*, B^*}$ gilt (nach Kuhn/Tucker):

$$\boldsymbol{\beta^{*,m}} \underbrace{\boldsymbol{(r^m(w^{*,T}x^m + w_0^*) - 1)}}_{(7')} = \mathbf{0}, \; \forall \boldsymbol{m} \in \{\mathbf{1}, \cdots, \boldsymbol{M}\} \qquad (7)$$

Falls $\boldsymbol{\beta^{*,m} \neq 0}$, dann muß Ausdruck (7') gleich null sein, d.h. $\boldsymbol{x^m}$ liegt auf $\boldsymbol{H_1}$ oder $\boldsymbol{H_2}$.

Falls $\boldsymbol{\beta^{*,m} = 0}$, dann kann $\boldsymbol{x^m}$ dahinter liegen.

Berechnung der optimalen separierenden Hyperebene

Berechnung von $\boldsymbol{w}^*$ basierend auf Gleichung (4):

$$\boldsymbol{w}^* := \sum_{\boldsymbol{m}=1}^{\boldsymbol{M}} \boldsymbol{\beta}^{*,\boldsymbol{m}} \boldsymbol{r}^{\boldsymbol{m}} \boldsymbol{x}^{\boldsymbol{m}}$$

Hinweis: zur Repräsentation von $\boldsymbol{w}^*$ tragen nur die auf $\boldsymbol{H}_1$ oder $\boldsymbol{H}_2$ liegenden Trainingsvektoren (Support-Vektoren) bei.

Berechnung von $\boldsymbol{w}_0^*$:

Sei $\boldsymbol{x}^{\boldsymbol{m}}$ ein Support-Vektor, z.B. derjenigen Klasse mit $\boldsymbol{r}^{\boldsymbol{m}} = \mathbf{1}$.

Dann gilt $\quad \boldsymbol{w}^{*,\boldsymbol{T}} \boldsymbol{x}^{\boldsymbol{m}} + \boldsymbol{w}_0 = \mathbf{1} \Rightarrow \boldsymbol{w}_0^* := \mathbf{1} - \boldsymbol{w}^{*,\boldsymbol{T}} \boldsymbol{x}^{\boldsymbol{m}}$

8.2 SVM mit nicht-linearer Diskriminanzfunktion

Überblick:

- Skalarprodukte von Vektoren
- Transformation des Eingaberaumes
- Implizite Raumtransformation via Kernfunktion
- Beispiele von Kernfunktionen

Skalarprodukte von Vektoren

- In zentraler Optimierungsfunktion (6) tauchen die Trainingsvektoren nur paarweise als Skalarprodukte auf.

- Auch in der Anwendung der Diskriminanzfunktion, d.h. bei Klassifikation eines Eingabevektors $\boldsymbol{x}$, taucht dieser Vektor nur in Skalarprodukten mit Trainingsvektoren $\boldsymbol{x}^m$ auf.

  Diskriminanzfunktion: $\boldsymbol{f_{Disk}}(\boldsymbol{x}) := \mathrm{sign}(\boldsymbol{w}^{*,T}\boldsymbol{x} + \boldsymbol{w}_0^*)$

  $$\text{mit} \quad \boldsymbol{w}^{*,T}\boldsymbol{x} := \sum_{m=1}^{M} \boldsymbol{\beta}^{*,m}\boldsymbol{r}^m\boldsymbol{x}^{m,T}\boldsymbol{x}$$

- Die Skalarprodukte werden im Folgenden durch sogenannte Kernfunktionen ersetzt. Dadurch sind auch nicht-lineare Diskriminanzfunktionen erlernbar.

Transformation des Eingaberaumes

Transformation eines $\boldsymbol{I_1}$-dimensionalen Eingabevektors in einen $\boldsymbol{I_2}$-dimensionalen Vektor eines anderen Raumes.

$$\boldsymbol{f_{Tran}} : \mathbb{R}^{\boldsymbol{I_1}} \rightarrow \mathbb{R}^{\boldsymbol{I_2}}$$

$$\boldsymbol{f_{Tran}(x)} := \begin{pmatrix} \boldsymbol{f^1_{Tran}(x)} \\ \vdots \\ \boldsymbol{f^{I_2}_{Tran}(x)} \end{pmatrix}$$

Transformation des Eingaberaumes

Konstruktion des Gewichtsvektors im transformierten Raum:

$$\boldsymbol{w}^* := \sum_{m=1}^{M} \beta^{*,m} r^m \boldsymbol{f}_{Tran}(\boldsymbol{x}^m)$$

Nur diejenigen Summanden mit $\beta^{*,m} > 0$ sind relevant.

Diskriminanzfunktion im transformierten Raum:

$$f_{Disk}(\boldsymbol{x}) := \operatorname{sign}(\boldsymbol{w}^{*,T} \boldsymbol{f}_{Tran}(\boldsymbol{x}) + w_0^*)$$

Transformation des Eingaberaumes

Paarweises Auftreten von $\boldsymbol{f}_{\boldsymbol{Tran}}$ bei der Anwendung:

$$\boldsymbol{w}^{*,T}\boldsymbol{f}_{\boldsymbol{Tran}}(\boldsymbol{x}) := \sum_{m=1}^{M} \beta^{*,m} r^m \boldsymbol{f}_{\boldsymbol{Tran}}(\boldsymbol{x}^m)^T \boldsymbol{f}_{\boldsymbol{Tran}}(\boldsymbol{x})$$

Paarweises Auftreten von $\boldsymbol{f}_{\boldsymbol{Tran}}$ bei Optimierung von $\boldsymbol{B}$:

$$\boldsymbol{D}_{m\ell} := r^m r^\ell \boldsymbol{f}_{\boldsymbol{Tran}}(\boldsymbol{x}^m)^T \boldsymbol{f}_{\boldsymbol{Tran}}(\boldsymbol{x}^\ell)$$

Es gibt Kombinationen aus bestimmter Transformationsfunktion $\boldsymbol{f}_{\boldsymbol{Tran}}$ und einer zugehörigen, sogenannten Kernfunktion $\boldsymbol{f}_{\boldsymbol{Kern}}$, für die gilt:

$$\boldsymbol{f}_{\boldsymbol{Tran}}(\boldsymbol{x}^i)^T \boldsymbol{f}_{\boldsymbol{Tran}}(\boldsymbol{x}^j) \stackrel{!}{=} \boldsymbol{f}_{\boldsymbol{Kern}}(\boldsymbol{x}^i, \boldsymbol{x}^j), \quad \forall \boldsymbol{x}^i, \boldsymbol{x}^j$$

Implizite Raumtransformation via Kernfunktion

Beispiel:

$$f_{Tran} : \mathbb{R}^2 \to \mathbb{R}^3,\ f_{Kern} : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$$

$$f_{Tran}(x) := \begin{pmatrix} (x_1)^2 \\ \sqrt{2}x_1x_2 \\ (x_2)^2 \end{pmatrix}$$

$$f_{Kern}(x^i, x^j) := (x^{i,T}x^j)^2$$

$$\Rightarrow f_{Tran}(x^i)^T f_{Tran}(x^j) \overset{!}{=} f_{Kern}(x^i, x^j), \quad \forall x^i, x^j$$

Implizite Raumtransformation via Kernfunktion

Beim Lernen muß dann nur Kernfunktion $\boldsymbol{f_{Kern}}$ herangezogen werden:

$$\boldsymbol{D_{m\ell} := r^m r^\ell f_{Kern}(x^m, x^\ell)}$$

Diskriminanzfunktion:

$$\boldsymbol{f_{Disk}(x)} := \mathsf{sign}\left(\sum_{m=1}^{M} \boldsymbol{\beta^{*,m} r^m f_{Kern}(x, x^m) + w_0^*}\right)$$

Nur diejenigen Summanden mit $\boldsymbol{\beta^{*,m} > 0}$ sind relevant.

In Artikeln von *Mercer/Courant* findet man Zusammenhänge von Transformations- und Kernfunktionen.

Beispiele von Kernfunktionen

Gaußfunktion:

$$\boldsymbol{f_{Kern}(x^i, x^j) := e^{-\frac{\|x^i - x^j\|^2}{2\sigma^2}}}$$

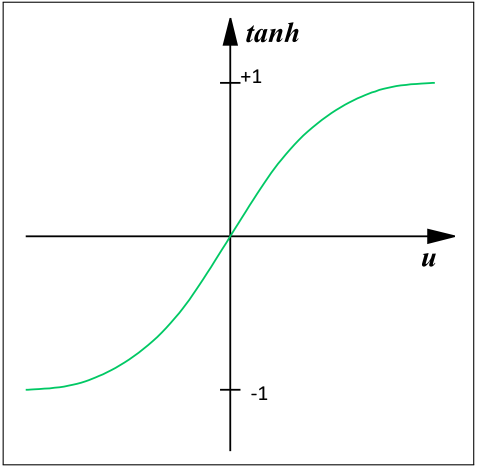
- Lokalisierende Gauß-(Basis)funktion als Kernfunktion.
- Diskriminanzfunktion ist durch dieselbe Art von Basisfunktionen wie beim RBF-Netz definiert.
- Im Gegensatz zur Clusterung und Verwendung von Cluster-Zentren werden die Zentren der Gauß-Funktionen hier durch Support-Vektoren definiert.

Beispiele von Kernfunktionen

Hyperbolischer Tangens:

$$f_{Kern}(x^i, x^j) := \tanh(x^{i,T} x^j)$$

$$\tanh(u) := \frac{\sinh(u)}{\cosh(u)} = \frac{\frac{e^u - e^{-u}}{2}}{\frac{e^u + e^{-u}}{2}}$$



Beispiele von Kernfunktionen

Hyperbolischer Tangens (forts.):

- Nicht-Lokalisierende Sigmoid-ähnliche Kernfunktion, als kleine Gemeinsamkeit mit MLPs.
- Jedoch wird bei MLP der Sigmoid in mehreren, aufeinander folgenden Schichten jeweils auf die linearen Assoziationen von vorherigem Schicht-Output und Gewichtsvektoren für nachfolgende Knoten angewendet.
- Dagegen wird der Hyperbolischer Tangens bei SVM auf nur einer Schicht auf die linearen Assoziationen von Eingabevektor und ausgewählten Trainingsvektoren angewendet.

Beispiele von Kernfunktionen

Polynom: $$\boldsymbol{f_{Kern}(x^i, x^j) := (x^{i,T} x^j + 1)^d}$$

Entspricht polynomiale Klassifikation, hier z.B mit Polynomgrad $\boldsymbol{d}$.

Literatur

- B. Schölkopf, A. Smola: Learning with Kernels - Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning), MIT Press, Cambridge, MA, 2002, ISBN 0-262-19475-9.