Einführung in die KI – WS2012/2013

Support Vector Machines

Prof. Dr. Ulf Brefeld

brefeld@kma.informatik.tu-darmstadt.de

TECHNISCHE UNIVERSITÄT DARMSTADT

Empirische Risikominimierung

TECHNISCHE UNIVERSITÄT DARMSTADT

- Binäre Klassifikation: $\mathcal{Y} = \{+1, -1\}$
- Trainingsmenge $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n) \in \mathcal{X} \times \mathcal{Y}$
- Gesucht: $f : \mathcal{X} \to \{+1, -1\}$
- Lineares Modell $f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b$
- Verlust $\ell : \mathcal{Y} \times \mathcal{X} \times \mathcal{F} \to \mathbb{R}$

$$\ell_{0,1}(y, \mathbf{x}, f) = \begin{cases} 1 & : \quad \text{sign}[f(\mathbf{x})] = y \\ 0 & : \quad \text{sonst} \end{cases}$$

- Vorläufige Annahme: Daten sind linear separierter

Empirische Risikominimierung

TECHNISCHE UNIVERSITÄT DARMSTADT

- Minimiere den Verlust
- Optimimale Funktion f^{opt}:

$$f^{opt} = \operatorname*{argmin}_f \frac{1}{n} \sum_{i=1}^{n} \ell_{0/1}(y_i, \mathbf{x}_i, f)$$

- ∞-viele ununterscheidbare/gleich gute Lösungen!
- Schlechte Konvergenzrate

Beispiel

TECHNISCHE UNIVERSITÄT DARMSTADT

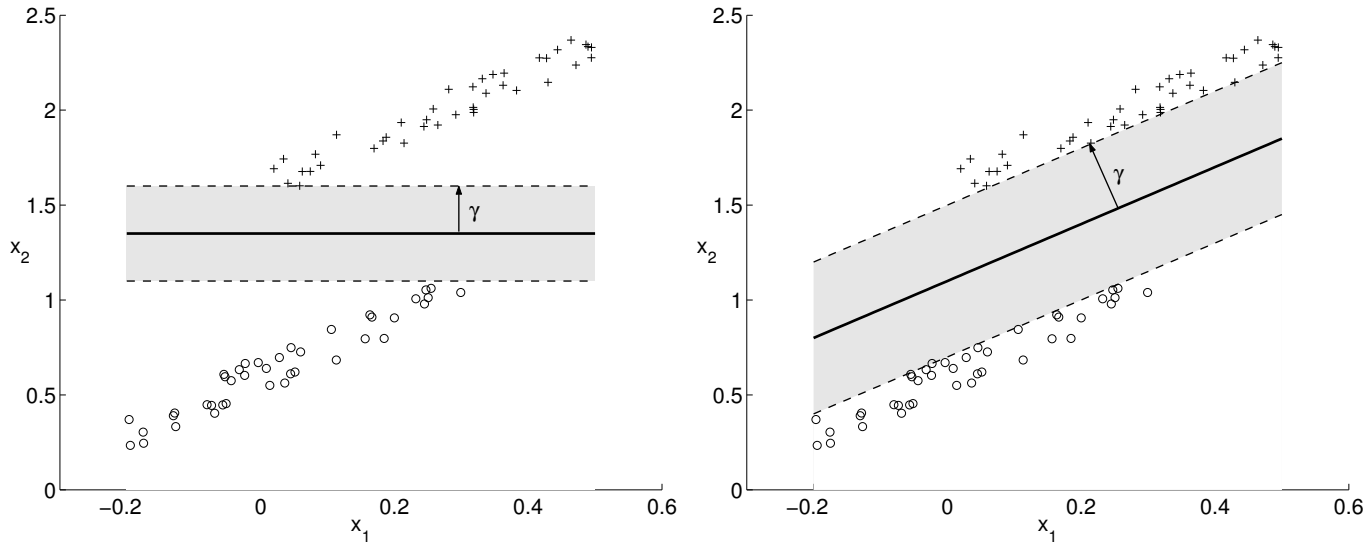

Abbildung: *Links: Eine suboptimale Trennung. Rechts: Die optimale Trennhyperebene mit maximalem geometrischem Margin γ. Die Geradengleichung ist gegeben durch $\langle \mathbf{w}, \mathbf{x} \rangle + b = 0$ (durchgezogene Linie). Die (gestrichelten) Korridorgeraden sind gegeben durch $\langle \mathbf{w}, \mathbf{x} \rangle + b = \pm\tilde{\gamma}$.*

Idee

TECHNISCHE UNIVERSITÄT DARMSTADT

- Linear separierbare Trainingsmenge
  $S = \{(\mathbf{x}_i, y_i) : i = 1, \ldots, n\} \subset \mathcal{X} \times \{+1, -1\}$
- Bilde Klassenmengen S_{+1} und S_{-1}, so dass
  $S_y = \{(\mathbf{x}_i, y_i) : (\mathbf{x}_i, y_i) \in S,\ y_i = y\}, \quad y = \pm 1$
- Es gilt $S = S_{+1} \bigcup S_{-1}$ und $S_{+1} \bigcap S_{-1} = \emptyset$.
- Annahme: Hyperebene $(\mathbf{w}, b)$ trennt S_{+1} und S_{-1}
- Lege Korridor um Hyperebene $(\mathbf{w}, b)$, so dass kein Beispiel innerhalb des Korridors liegt
- Je breiter der Korridor, desto besser die Separierung
- Neues Optimierungsproblem: Maximiere die Breite des Korridors

Der Margin

TECHNISCHE UNIVERSITÄT DARMSTADT

- Ein Maß für den Abstand zwischen einer Trennebene und einem Beispiel $(\mathbf{x}_i, y_i) \in S$ ist der Funktionswert $\tilde{\gamma} = y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)$
- $\tilde{\gamma}$ nennen wir den funktionalen Margin
- $\tilde{\gamma}$ hängt jedoch von $\mathbf{w}$ ab
  - $\mathbf{w}$ kann (nahezu) beliebig skaliert werden ohne die Klassentrennung zu verändern
- Den normierten Funktionswert $\gamma = \frac{\tilde{\gamma}}{\|\mathbf{w}\|}$ nennen wir den geometrischen Margin
- γ ist gleich der euklidschen Distanz eines Punktes zur Trennhyperebene
- Die Hyperebene, die die größte Distanz zu den Datenpunkten aufweist, also den breitesten Korridor realisiert, nennen wir die optimale Trennhyperebene

Definition Margin

TECHNISCHE UNIVERSITÄT DARMSTADT

Definition (Margins)

Gegeben sei eine Trainingsmenge $S = \{(\mathbf{x}^{(i)}, y^{(i)})\} \subset \mathcal{X} \times \{+1, -1\}$, $\mathcal{X} \subseteq \mathbb{R}^d$ sowie eine Hyperebene mit Normalenvektor $\mathbf{w} \in \mathbb{R}^d$ und der Schwelle b.

- Der funktionale Margin $\tilde{\gamma}_i(\mathbf{w}, b)$ eines Beispiels $(\mathbf{x}^{(i)}, y^{(i)}) \in S$ ist definiert als $\tilde{\gamma}_i(\mathbf{w}, b) := y^{(i)} \left(\langle \mathbf{w}, \mathbf{x}^{(i)} \rangle_{\mathcal{X}} + b \right)$
- Der funktionale Margin $\tilde{\gamma}_S(\mathbf{w}, b)$ einer Trainingsmenge S ist definiert als $\tilde{\gamma}_S(\mathbf{w}, b) := \min_{(\mathbf{x}^{(i)}, y^{(i)}) \in S} \tilde{\gamma}_i(\mathbf{w}, b)$
- Der geometrische Margin $\gamma_i(\mathbf{w}, b)$ eines Beispiels $(\mathbf{x}^{(i)}, y^{(i)}) \in S$ ist definiert als $\gamma_i(\mathbf{w}, b) := \frac{\tilde{\gamma}_i(\mathbf{w}, b)}{||\mathbf{w}||}$
- Der geometrische Margin $\gamma_S(\mathbf{w}, b)$ einer Trainingsmenge S ist definiert als $\gamma_S(\mathbf{w}, b) := \frac{\tilde{\gamma}_S(\mathbf{w}, b)}{||\mathbf{w}||}$

Folgerungen

TECHNISCHE UNIVERSITÄT DARMSTADT

- Aus der Definition folgt, daß für den linear separierbaren Fall, jedes Beispiel aus S einen positiven funktionalen Margin besitzt und dadurch der funktionale Margin der Trainingsmenge ebenfalls positiv ist.
- Offensichtlich gilt dies analog für den geometrischen Margin, da $\|\mathbf{w}\| > 0$. Für $\|\mathbf{w}\| = 1$ sind funktionaler und geometrischer Margin identisch.
- Wir bilden die funktionalen Margins $\tilde{\gamma}_{+1}$ und $\tilde{\gamma}_{-1}$ der Mengen S_{+1} und S_{-1} bzgl. der Hyperebene $(\mathbf{w}, b)$ durch

$$\tilde{\gamma}_{+1} := \tilde{\gamma}_{+1}(\mathbf{w}, b) = \min_{(\mathbf{x}^{(i)}, y^{(i)}) \in S_{+1}} y_i \left(\langle \mathbf{w}, \mathbf{x}^{(i)} \rangle + b \right)$$

$$\tilde{\gamma}_{-1} := \tilde{\gamma}_{-1}(\mathbf{w}, b) = \min_{(\mathbf{x}^{(i)}, y^{(i)}) \in S_{-1}} y_i \left(\langle \mathbf{w}, \mathbf{x}^{(i)} \rangle + b \right)$$

Konstruktion des Korridors

TECHNISCHE UNIVERSITÄT DARMSTADT

Daraus konstruieren wir zwei neue Hyperebenen, die den Korridor beschreiben:

$$\langle \mathbf{w}, \mathbf{x} \rangle + b = +\tilde{\gamma}_{+1} \quad (1)$$

$$\langle \mathbf{w}, \mathbf{x} \rangle + b = -\tilde{\gamma}_{-1} \quad (2)$$

Die Breite des entstandenen Korridors ist $\gamma_{+1} + \gamma_{-1}$. Um die Schreibweise zu vereinfachen bestimmen wir die gemeinsame Mittelebene durch

$$\langle \mathbf{w}, \mathbf{x} \rangle + b = \frac{\gamma_{+1} - \gamma_{-1}}{2} =: b_0$$

Sie realisiert den funktionalen Margin $\tilde{\gamma}$ von

$$\tilde{\gamma} = \frac{\tilde{\gamma}_{+1} + \tilde{\gamma}_{-1}}{2} \quad (3)$$

Die optimale Trennhyperebene

TECHNISCHE UNIVERSITÄT DARMSTADT

Die Hyperebenen (1) und (2) können wir jetzt relativ zur Mittelebene schreiben als

$$\begin{aligned} \langle \mathbf{w}, \mathbf{x} \rangle_{\mathcal{X}} + b &= b_0 + \tilde{\gamma} \\ \langle \mathbf{w}, \mathbf{x} \rangle_{\mathcal{X}} + b &= b_0 - \tilde{\gamma}. \end{aligned}$$

Die Gleichungen sind nicht eindeutig, da sie mit einer beliebigen positiven Konstante multipliziert werden können, ohne die Ebene oder die durch sie definierte Trennung zu verändern. Teile durch $\tilde{\gamma}$, da aus Gleichung (3) und Definition (1) $\tilde{\gamma} > 0$ folgt. Mit $\mathbf{w}' = \frac{1}{\tilde{\gamma}}\mathbf{w}$ und $b' = \frac{b - b_0}{\tilde{\gamma}}$ erhalten wir dann die vereinfachte Form der Korridorebenen

$$\langle \mathbf{w}', \mathbf{x} \rangle_{\mathcal{X}} + b' = +1 \quad \text{und} \quad \langle \mathbf{w}', \mathbf{x} \rangle_{\mathcal{X}} + b' = -1,$$

die einen funktionalen Margin von 1 realisiert,

$$\langle \mathbf{w}', \mathbf{x} \rangle + b' = 0. \tag{4}$$

Neues Optimierungsproblem

TECHNISCHE UNIVERSITÄT DARMSTADT

Für den geometrischen Margin folgt direkt

$$\gamma = \frac{1}{\|\mathbf{w}'\|}$$

Der geometrische Margin wird maximal, wenn die Norm von $\mathbf{w}$ bei festem $\tilde{\gamma} = 1$ minimal wird. Wir können das folgende Optimierungsproblem angeben

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \|\mathbf{w}\|^2 \\ \text{u.d.N.} \quad & y^{(i)} \left(\langle \mathbf{w}, \mathbf{x}^{(i)} \rangle + b \right) \geq 1 \qquad \forall i. \end{aligned} \tag{5}$$

Das Minimum realisiert die optimale Trennhyperebene mit dem geometrischen Margin $\gamma = \frac{1}{\|\mathbf{w}\|}$.

Satz von Lagrange

TECHNISCHE UNIVERSITÄT DARMSTADT

- Optimierungsprobleme mit Nebenbedingungen
- Jede NB wird mit einem nicht-negativem Lagrangemultiplikator gewichtet und zur Zielfunktion addiert/subtrahiert (je nach Ungleichungstyp)
- Die primale Lagrangefunktion mit den Lagrangemultiplikatoren $\alpha_i \geq 0$ ist gegeben durch

$$L_P(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N} \alpha_i \left[y^{(i)} \left(\langle \mathbf{w}, \mathbf{x}^{(i)} \rangle + b \right) - 1 \right] \tag{6}$$

- Die Lagrangefunktion soll bezüglich $\mathbf{w}$ und b minimiert und bezüglich α_i maximiert werden.
- Das Optimierungsproblem ist ein konvexes, quadratisches Programm
- Das Optimum ist daher ein Sattelpunkt.

Sattelpunkt Beispiel

TECHNISCHE UNIVERSITÄT DARMSTADT

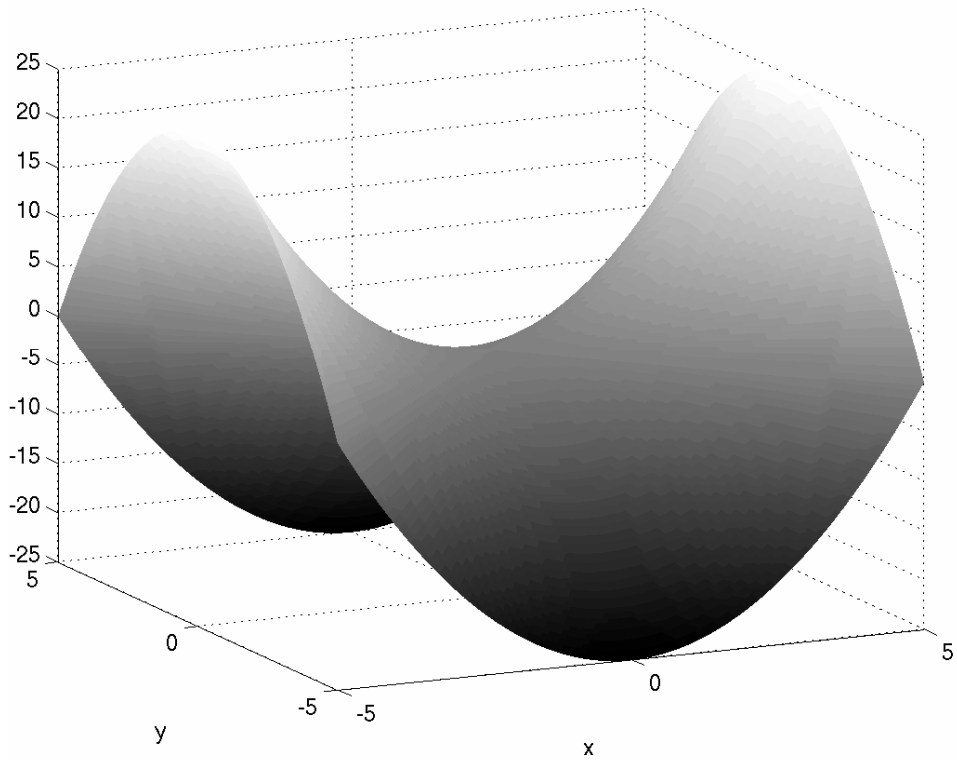

Abbildung: *Sattelpunkt der Funktion* $f(x, y) = x^2 - y2$ *an der Stelle* $(0, 0)$*. Es gilt:* $f(0, y) \leq f(0, 0) \leq f(x, 0)$.

Optimierung

TECHNISCHE UNIVERSITÄT DARMSTADT

- Ziel: Eliminiere die primalen Variablen ($\mathbf{w}$, b), so dass nur noch die (dualen) Lagrangevariablen über bleiben.
- Maximiere L_P bzgl. α unter der Nebenbedingung $\alpha_i \geq 0$ und

$$\frac{\partial L_P}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{N} \alpha_i \, y^{(i)} \mathbf{x}^{(i)} \stackrel{!}{=} \mathbf{0} \quad \Rightarrow \quad \mathbf{w}^{opt} = \sum_{i=1}^{N} \alpha_i \, y^{(i)} \mathbf{x}^{(i)}$$

$$\frac{\partial L_P}{\partial b} = -\sum_{i=1}^{N} \alpha_i \, y^{(i)} \stackrel{!}{=} 0 \quad \Rightarrow \quad 0 = \sum_{i=1}^{N} \alpha_i \, y^{(i)}$$

Duale Repräsentation

TECHNISCHE UNIVERSITÄT DARMSTADT

Setzen wir jetzt die Ableitungen wieder in die primale Lagrangefunktion ein, folgt die duale Repräsentation, das sogenannte Wolfe-Dual, welches nur noch von den Lagrangemultiplikatoren $\boldsymbol{\alpha}$ abhängt und diesbezüglich maximiert werden soll:

$$
\begin{aligned}
L_P(\mathbf{w}, b, \boldsymbol{\alpha}) &= \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N} \alpha_i \left[y^{(i)} \left(\langle \mathbf{w}, \mathbf{x}^{(i)} \rangle + b \right) - 1 \right] \\
&= \frac{1}{2} \left(\sum_{i=1}^{N} \alpha_i \, y^{(i)} \mathbf{x}^{(i)} \right)^2 - \left(\sum_{i}^{N} \alpha_i \, y^{(i)} \mathbf{x}^{(i)} \right)^2 + \sum_{i=1}^{N} \alpha_i \\
L_D(\boldsymbol{\alpha}) &= \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \, \alpha_j \, y^{(i)} y^{(j)} \langle \mathbf{x}^{(i)}, \mathbf{x}^{(j)} \rangle . \qquad (7)
\end{aligned}
$$

Achtung: Das duale Optimierungsproblem hängt nur noch von Linearkombinationen von Trainingspunkten ab.

Eigenschaften

TECHNISCHE UNIVERSITÄT DARMSTADT

- Gleichung (7) charakterisiert die einfachste Support-Vector-Machine, die für linear separierbare Trainingsdaten die optimale Trennhyperebene durch die Maximierung des Margins berechnet.
- Verfahren dieser Art werden in der Literatur auch Maximum-Margin-Methoden genannt.
- Im Sattelpunkt gelten die KKT-Bedingungen für alle $i = 1, \ldots, n$

$$\begin{aligned}
\frac{\partial L(\mathbf{w}^{opt}, b^{opt}, \boldsymbol{\alpha}^{opt})}{\partial \mathbf{w}} &= 0 \\
\frac{\partial L(\mathbf{w}^{opt}, b^{opt}, \boldsymbol{\alpha}^{opt})}{\partial b} &= 0 \\
\alpha_i^{opt} \left[y_i \left(\left\langle \mathbf{w}^{opt}, \mathbf{x}^{(i)} \right\rangle + b^{opt} \right) - 1 \right] &= 0 \qquad (8) \\
y_i \left(\left\langle \mathbf{w}^{opt}, \mathbf{x}^{(i)} \right\rangle + b^{opt} \right) - 1 &\geq 0 \qquad (9) \\
\alpha_i^{opt} &\geq 0
\end{aligned}$$

Supportvektoren

TECHNISCHE UNIVERSITÄT DARMSTADT

- Nach der KKT-Komplementär-Bedingung ist im optimalen Punkt $(\mathbf{w}^{opt}, b^{opt}, \boldsymbol{\alpha}^{opt})$ jeder Lagrangemultiplikator α_i^{opt} mit der zugehörigen Nebenbedingung über Gleichung (8) gebunden
- Für alle $\alpha_i^{opt} > 0$ muß folglich die Gleichheit in (9) gelten
- Die Beispiele, für die die Lagrangemultiplikatoren α_i^{opt} einen Wert ungleich 0 annehmen, müssen gemäß (8) auf einer der beiden Korridorebenen liegen
- Da diese Punkte den kleinsten Abstand zur Trennhyperebene haben (siehe Definition 1), würde sich die Lösung verändern, wenn wir diese Punkte weglassen würden.
- Diese Punkte nennen wir Supportvektoren: $S_V = \{i : \alpha_i > 0\}$

Gewichtsvektor

TECHNISCHE UNIVERSITÄT DARMSTADT

Da für alle Beispiele $(\mathbf{x}_i, y_i)$ mit $i \notin S_V$ das zugehörige $\alpha_i^{opt} = 0$ ist, gilt aber

$$\mathbf{w}^{opt} = \sum_{i=1}^{N} \alpha_i \, y^{(i)} \mathbf{x}^{(i)} = \sum_{i \in S_V} \alpha_i \, y^{(i)} \mathbf{x}^{(i)}. \tag{10}$$

Die gefundene Lösung hängt also nur von einigen (wenigen) Trainingspunkten, den Supportvektoren ab. Ohne diese Eigenschaft wären große Datenmengen für eine SVM nicht handhabbar.

Schwellwert

TECHNISCHE UNIVERSITÄT DARMSTADT

Im Gegensatz zu $\mathbf{w}^{opt}$ wird die Schwelle b^{opt} nicht explizit durch die Optimierung bestimmt. Gleichung (8) gibt jedoch die Möglichkeit für ein $\alpha_i > 0$ den Wert von b^{opt} direkt auszurechnen, eine exaktere Lösung erhält man durch die folgende Mittelung

$$\begin{aligned} b^{opt} &= \frac{1}{|S_V|} \sum_{i \in S_V} \left(y^{(i)} - \langle \mathbf{w}^{opt}, \mathbf{x}^{(i)} \rangle_{\mathcal{X}} \right) \\ &= \frac{1}{|S_V|} \sum_{i \in S_V} \left[y^{(i)} - \sum_{j=1}^{N} \alpha_j \, y^{(j)} \langle \mathbf{x}^{(j)}, \mathbf{x}^{(i)} \rangle_{\mathcal{X}} \right] . \end{aligned} \tag{11}$$

Klassifikation neuer Beispiele

TECHNISCHE UNIVERSITÄT DARMSTADT

Ist die Lösung gefunden, kann ein neues Beispiel **x** über die Linearkombination der Supportvektoren klassifiziert werden durch

$$
\begin{aligned}
f(\mathbf{x}) &= \text{sign}\left[\langle \mathbf{w}^{opt}, \mathbf{x}\rangle + b^{opt}\right] \\
&= \text{sign}\left[\sum_{i \in S_V} \alpha_i^{opt}\, y^{(i)} \langle \mathbf{x}^{(i)}, \mathbf{x}\rangle + b^{opt}\right]
\end{aligned}
$$

Zusammenfassung

TECHNISCHE UNIVERSITÄT DARMSTADT

Gegeben eine linear separable Trainingsmenge findet der Maximum Margin Klassifizierer immer die optimale Trennhyperebene. In der Realität sind die Daten jedoch oft verrauscht und nur in den seltensten Fällen linear trennbar. Probleme solcher Art sind für die hier vorgestellte Maschine nicht lösbar. Der nächste Abschnitt beschreibt die nichtlineare Erweiterung der Maximum Margin SVM.

Die Wirklichkeit

TECHNISCHE UNIVERSITÄT DARMSTADT

- Daten sind oft verrauscht
- Daten sind selten linear separierbar
- Generalisieren bedeutet auch kleine Fehler in Kauf nehmen

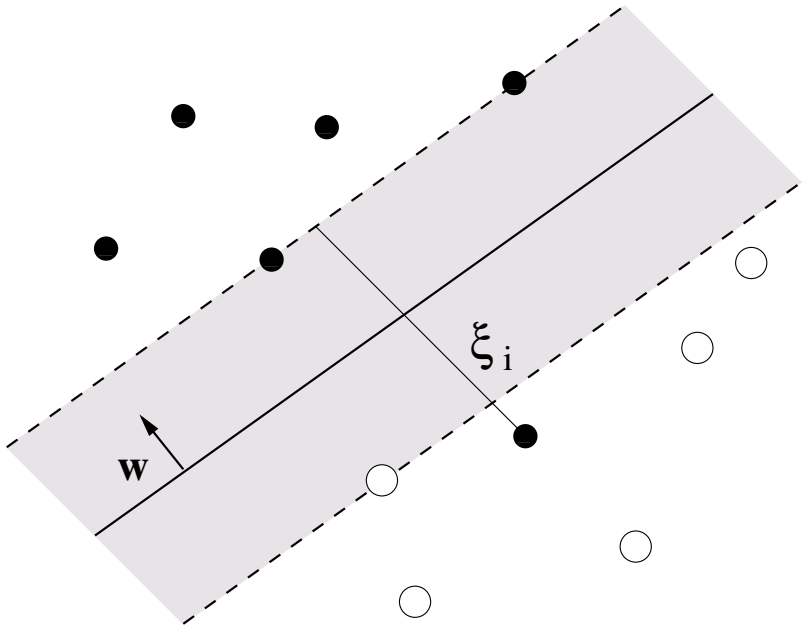

Abbildung: *Slackvariablen kompensieren Fehler*

Slackvariablen

TECHNISCHE UNIVERSITÄT DARMSTADT

- Slackvariablen $\xi_i \geq 0$ kompensieren lokale Fehler
- Margin darf an einigen Stellen verletzt werden
- Erweitere die Nebenbedingungen zu:

$$\begin{aligned} \langle \mathbf{w}, \mathbf{x}_i \rangle + b &\geq +1 - \xi_i & y_i &= +1 \\ \langle \mathbf{w}, \mathbf{x}_i \rangle + b &\geq -1 + \xi_i & y_i &= -1 \\ \xi_i &\geq 0 & \forall i & \end{aligned} \tag{12}$$

- Anhand von ξ_i kann die Lage des zugehörigen Beispiels $(\mathbf{x}_i, y_i)$ bestimmt werden (siehe nächste Folie)

Fallunterscheidung

TECHNISCHE UNIVERSITÄT DARMSTADT

Wir unterscheiden die folgenden Fälle:

- Für $\xi_i = 0$ wird das Muster richtig klassifiziert, ebenso
- für $0 < \xi_i < 1$, das Beispiel liegt hier aber im Korridor
- Ist $\xi_i = 1$ liegt der Punkt auf der Trennhyperebene und wird
- im Fall $\xi_i > 1$ falsch klassifiziert

Wir suchen eine Lösung, die einerseits den Margin maximiert, andererseits die Fehlklassifizierungen minimiert. Da für die ξ_i der falsch klassifizierten Beispiele $\xi_i > 1$ gilt, können wir $\sum \xi_i$ als obere Schranke für das empirische Risiko mit dem $0/1$-Verlust setzen:

$$\sum_{i=1}^{n} \xi_i \geq \sum_{i=1}^{n} \ell_{0/1}(y_i, \mathbf{x}_i, f)$$

Soft-Margin Support Vector Machine: Primales Optimierungsproblem

TECHNISCHE UNIVERSITÄT DARMSTADT

Zusammenfassen der Nebenbedingungen aus (12) zu

$$y^{(i)} \left[\langle \mathbf{w}, \phi(\mathbf{x}^{(i)}) \rangle_{\mathcal{H}_K} + b \right] \quad \geq \quad 1 - \xi_i$$

ergibt folgendes konvexes, quadratisches Optimierungsproblem:

$$\begin{array}{lll} \min\limits_{\mathbf{w},b,\boldsymbol{\xi}} & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum\limits_{i=1}^{N}\xi_i & \\ \text{u.d.N.} & y^{(i)}\left(\langle \mathbf{w}, \mathbf{x}_i \rangle + b\right) \geq 1 - \xi_i & \\ & \xi_i \geq 0 & i = 1, \ldots, n \end{array} \tag{13}$$

Trade-off Variable $C > 0$:

- Große Werte von C gewichten die Minimierung des Fehlers
- Kleine C maximieren den Margin
- In der Praxis wird zur Bestimmung des optimalen C ein Modellselektionsverfahren benötigt

Optimierung

TECHNISCHE UNIVERSITÄT DARMSTADT

Die Optimierungsaufgabe aus (13) führt zur primalen Lagrangefunktion $L_P(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\mu})$ mit:

$$L_P = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{N}\alpha_i\left[y^{(i)}\left(\langle\mathbf{w}, \mathbf{x}_i\rangle + b\right) - 1 + \xi_i\right] - \sum_{i=1}^{N}\mu_i\,\xi_i,$$

mit den Lagrangemultiplikatoren α_i und μ_i. Im Sattelpunkt gelten die KKT-Bedingungen: (folgende Folie)

KKT-Bedingungen

TECHNISCHE UNIVERSITÄT DARMSTADT

$$\frac{\partial L_P}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{n} \alpha_i \, y_i \mathbf{x}_i = \mathbf{0} \tag{14}$$

$$\frac{\partial L_P}{\partial b} = -\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{15}$$

$$\frac{\partial L_P}{\partial \xi_i} = C - \alpha_i - \mu_i = 0 \tag{16}$$

$$y_i \left(\langle \mathbf{w}, \mathbf{x}_i \rangle + b \right) - 1 + \xi_i \geq 0 \tag{17}$$

$$\xi_i \geq 0 \tag{18}$$

$$\alpha_i \geq 0 \tag{19}$$

$$\mu_i \geq 0 \tag{20}$$

$$\alpha_i \left[y_i \left(\langle \mathbf{w}, \mathbf{x}_i \rangle + b \right) - 1 + \xi_i \right] = 0 \tag{21}$$

$$\xi_i \left(C - \alpha_i \right) = 0 \tag{22}$$

Soft-Margin Support-Vector-Machine

TECHNISCHE UNIVERSITÄT DARMSTADT

Die duale Lagrangefunktion ergibt sich durch Einsetzen der Relationen (14)-(16) in die primale Funktion:

$$L_D(\boldsymbol{\alpha}) \quad = \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \, \alpha_j \, y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle$$

- Die Gleichung $C - \alpha_i - \mu_i = 0$ zusammen mit der Nichtnegativitätsbedingung $\mu_i \geq 0$ verlangt $\alpha_i \leq C$, während $\xi_i > 0$ nur für $\mu_i = 0$ auftritt und $\alpha_i = C$ verlangt.
- Die daraus entstehende, implizite Nebenbedingung $0 \leq \alpha_i \leq C$ beschränkt α_i von oben durch C.
- Der Vektor $\boldsymbol{\alpha}$ liegt folglich in einer Box mit der Seitenlänge C im positiven Halbraum, weshalb diese Nebenbedingung auch Box-Constraint genannt wird.

Primale Parameter

TECHNISCHE UNIVERSITÄT DARMSTADT

Fassen wir die Indizes aller Supportvektoren in der Menge $S_V = \{i : 0 < \alpha_i < C\}$ zusammen, kann die optimale Lösung $\mathbf{w}^{opt}$ durch (14) wieder als Linearkombination der Supportvektoren geschrieben werden

$$\mathbf{w}^{opt} = \sum_{i \in S_V} \alpha_i \, y^{(i)} \mathbf{x}_i, \tag{23}$$

und die Schwelle b^{opt} durch (21) berechnet werden. Auch hier ist die Mittelung über alle b numerisch sicherer

$$b^{opt} = \frac{1}{|S_V|} \sum_{i \in S_V} \left[y_i - \sum_{j=1}^{n} \alpha_j \, y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \right]. \tag{24}$$

Literatur

TECHNISCHE UNIVERSITÄT DARMSTADT

- K. P. Bennet, E. J. Bredensteiner. Duality and Geometry in SVM Classifiers. In *Proceedings of the Seventeenth International Conference on Machine Learning*, 2000.
- C. J. C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. In *Data Mining and Knowledge Discovery*, Vol. 2, Nr. 2, Seite 121-167, 1998.
- C. Cortes, V. Vapnik. Support Vector Networks. In *Machine Learning*, Vol. 20, Seite 273-297, 1995.
- N. Cristianini and J. Shaw-Taylor. *Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, Camebridge, 2000.
- K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda und B. Schölkopf. An Introduction to Kernel-Based Learning Algorithms. In *IEEE Transactions On Neural Networks*. Vol. 12, Nr. 2, 2001
- V. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, New York 1998