

Mustererkennung: Support Vector Mashines

D. Schlesinger – TUD/INF/KI/IS

Kosinez Algorithmus – lineare SVM

Gegeben sei eine Lernstichprobe $X = ((x_i, y_i) \dots)$ mit den

(i) Daten $x_i \in \mathbb{R}^n$ und (ii) Klassen $y_i \in \{-1, +1\}$ (± 1 Kodierung)

Gesucht wird die Hyperebene maximaler Breite, die die Daten richtig separiert, d.h.

$$y_i \cdot [\langle w, x_i \rangle + b] \geq 0$$

Die bezüglich der Lernstichprobe X **kanonische** Form:

$$\min_i |\langle w, x_i \rangle + b| = 1$$

Margin:

$$\langle w, x' \rangle + b = +1, \quad \langle w, x \rangle + b = -1$$

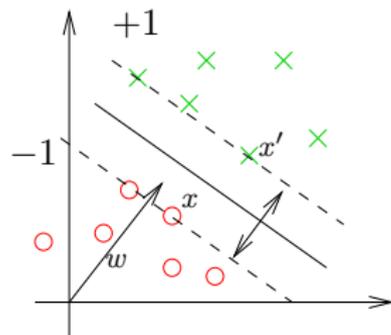
$$\langle w, x' - x \rangle = 2$$

$$\langle w / \|w\|, x' - x \rangle = 2 / \|w\|$$

Aufgabe:

$$\|w\|^2 \rightarrow \min_{w, b}$$

$$\text{s.t. } y_i \cdot [\langle w, x_i \rangle + b] \geq 1 \quad \forall i$$



Lagrangian:

$$L(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_i \alpha_i \cdot (y_i \cdot [\langle w, x_i \rangle + b] - 1) \rightarrow \max_{\alpha} \min_{w, b}$$

$$\alpha_i \geq 0 \quad \forall i$$

Bedeutung der dualen Variablen α -s:

- (a) $y_i \cdot [\langle w, x_i \rangle + b] - 1 < 0$ (die Nebenbedingungen sind verletzt)
Maximierung bezüglich α_i :
 $\alpha_i \rightarrow \infty, L(w, b, \alpha) \rightarrow \infty$ – kein Minimum möglich
- (b) $y_i \cdot [\langle w, x_i \rangle + b] - 1 > 0$
Maximierung bezüglich α_i gibt $\alpha_i = 0$ – kein Einfluss auf den Lagrangian
- (c) $y_i \cdot [\langle w, x_i \rangle + b] - 1 = 0$
 α_i ist egal, der Vektor x_i liegt am Rande des Streifen – **Support Vektor**

Ist die Lernstichprobe linear separierbar, so $\alpha_i < \infty$ und $L(w, b, \alpha) < \infty$

Lagrangian:

$$L(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_i \alpha_i \cdot (y_i \cdot [\langle w, x_i \rangle + b] - 1)$$

Ableitungen:

$$\frac{\partial L}{\partial b} = \sum_i \alpha_i y_i = 0$$

$$\frac{\partial L}{\partial w} = w - \sum_i \alpha_i y_i x_i = 0$$

$$w = \sum_i \alpha_i y_i x_i$$

Die Entscheidungsregel $f : \mathbb{R}^n \rightarrow \{-1, +1\}$ ist:

$$\begin{aligned} f(x) &= \operatorname{sgn}(\langle x, w \rangle + b) = \operatorname{sgn}\left(\left\langle x, \sum_i \alpha_i y_i x_i \right\rangle + b\right) = \\ &\operatorname{sgn}\left(\sum_i \alpha_i y_i \langle x, x_i \rangle + b\right) \end{aligned}$$

(!!!) Der Vektor w ist nicht explizit benötigt,
die Entscheidungsregel lässt sich als lineare Kombination der Skalarprodukte ausdrücken.

Man setzt

$$\sum_i \alpha_i y_i = 0$$
$$w = \sum_i \alpha_i y_i x_i$$

in den Lagrangian

$$L(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_i \alpha_i \cdot (y_i \cdot [\langle w, x_i \rangle + b] - 1)$$

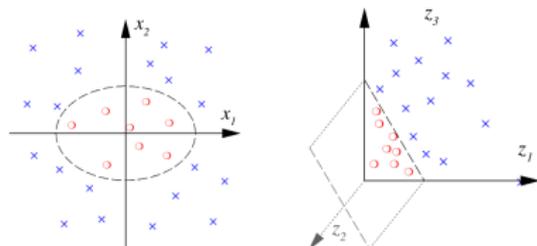
und erhält die duale Aufgabe:

$$\sum_i \alpha_i - \frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \rightarrow \max_{\alpha}$$
$$\text{s.t. } \alpha_i \geq 0, \quad \sum_i \alpha_i y_i = 0$$

(!!!) Die zu lösende Aufgabe lässt sich mithilfe der Skalarprodukte ausdrücken, die Daten x_i an sich sind nicht relevant.

Der Input-Raum \mathcal{X} wird mithilfe einer (nichtlinearen) Transformation $\Phi : \mathcal{X} \rightarrow \mathcal{H}$ auf den **Merkmalsraum** \mathcal{H} abgebildet. Danach werden die Daten im Merkmalsraum klassifiziert.

Beispiel:



Quadratischer Klassifikator
 $f(x) = \text{sgn}(a \cdot x_1^2 + b \cdot x_1 x_2 + c \cdot x_2^2)$

$\Phi : \mathbb{R}^2 \rightarrow \mathbb{R}^3$

$\Phi(x_1, x_2) = (x_1^2, \sqrt{2}x_1 x_2, x_2^2)$

(Abbilder $\Phi(x)$ sind linear separierbar)

Um die Daten im \mathcal{H} zu trennen bzw. den linearen Klassifikator im \mathcal{H} zu lernen, werden nicht die Abbilder $\Phi(x)$ benötigt, sondern nur Skalarprodukte $\langle \Phi(x), \Phi(x') \rangle$.

$$\begin{aligned} \langle \Phi(x_1, x_2), \Phi(x'_1, x'_2) \rangle &= \langle (x_1^2, \sqrt{2}x_1 x_2, x_2^2), (x_1'^2, \sqrt{2}x_1' x_2', x_2'^2) \rangle = \\ &= x_1^2 x_1'^2 + 2x_1 x_2 x_1' x_2' + x_2^2 x_2'^2 = \\ &= (x_1 x_1' + x_2 x_2')^2 = \langle x, x' \rangle^2 = \\ &= k(x, x') \quad - \quad \text{Kernel} \end{aligned}$$

Skalarprodukt im \mathcal{H} lässt sich im \mathcal{X} berechnen!!!

Kernel ist eine Funktion $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, die Skalarprodukt im Merkmalsraum realisiert

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle$$

Der Merkmalsraum \mathcal{H} und die Abbildung $\Phi : \mathcal{X} \rightarrow \mathcal{H}$ müssen dabei nicht unbedingt explizit definiert werden.

Ist eine Funktion $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ ein Kernel, so existiert eine Abbildung $\Phi : \mathcal{X} \rightarrow \mathcal{H}$ so dass (siehe oben).

Gegeben sei eine Funktion $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$. Ist das ein Kernel? \rightarrow Mercer's Theorem.

Sind k_1 und k_2 Kernels, dann sind αk_1 , $k_1 + k_2$, $k_1 k_2$ auch Kernels
(es gibt noch weitere Operationen um Kernels aus Kernels zu bauen).

Beispiele (oft benutzte Kernels):

- Polynomial: $k(x, x') = (\langle x, x' \rangle + c)^d$
- Sigmoid: $k(x, x') = \tanh(\kappa \langle x, x' \rangle + \Theta)$
- Gaussian: $k(x, x') = \exp(-\|x - x'\|^2 / (2\sigma^2))$ (Interessant: $\mathcal{H} = \mathbb{R}^\infty$)

Gaussian: $k(x, x') = \exp(-\|x - x'\|^2 / (2\sigma^2))$

