Vorlesung 17 Support vector machines

Erinnerung: Wir haben das Modell

$$f_\Theta : \mathbb{R}^D \to \{-1, 1\}$$

mit $f_\Theta(x) = \operatorname{sgn}(\langle a, x\rangle + b)$, $\Theta = (a,b) \in \mathbb{R}^D \times \mathbb{R}$.

Soft-Margin SVM:

$$\begin{cases} \min\limits_{a, b, \xi} & \|a\|^2 + C \sum\limits_{k=1}^{n} \xi_k \\ \text{s.t.} & y_k(\langle a, x_k\rangle + b) \geq 1 - \xi_k, \end{cases}$$

wobei $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^D \times \{-1, 1\}$ Trainingsdaten sind.

Dual SVM

$$\begin{cases} \max\limits_{\alpha} & -u^T G u + \sum\limits_{k=1}^{n} \alpha_k \\ \text{s.t.} & \sum\limits_{k=1}^{n} \alpha_k y_k = 0 \quad \text{und} \quad 0 \leq \alpha_i \leq C \quad i = 1, \dots, n. \end{cases}$$

mit $u = \left(\frac{1}{2} \alpha_k y_k\right)_{k=1}^{n}$, $G = \left(\langle x_k, x_\ell\rangle\right)_{k,\ell=1}^{n}$.

Proposition 17.1

Sei $\alpha \in \mathbb{R}^n$ eine Lösung für Dual SVM. Dann haben optimale Parameter $(a^*, b^*) \in \mathbb{R}^D \times \mathbb{R}$ für Soft-Margin-SVM:

1) $a^* = \sum\limits_{k=1}^{n} u_k x_k$, $u_k = \frac{1}{2} \alpha_k y_k$.
2) b^* ist der Median-Wert von $y_k - \langle a^*, x_k\rangle$ für $\alpha_k \neq 0$.

Beweis

1) folgt aus Gleichung (xxx) von VL 16.
2) Erinnerung: Soft Margin SVM ist äquivalent zu

$$\min_{a,b,\xi} \; \max_{\alpha,\beta} \; \mathcal{L}(a,b,\xi,\alpha,\beta) \qquad (\alpha_k, \beta_k \geq 0)$$

wobei $$\mathcal{L} = \|a\|^2 + C \sum_{k=1}^{n} \xi_k - \sum_{k=1}^{n} \alpha_k \big(y_k (\langle a, x_k \rangle + b) - (1 - \xi_k) \big) - \sum^{n} \beta_k \xi_k .$$

Im optimalen Fall haben wir $\alpha + \beta = C$ (siehe VL 16). Daher,

$$\mathcal{L} = \|a\|^2 - \sum_{k=1}^{n} \alpha_k \big(y_k (\langle a, x_k \rangle + b) - 1 \big).$$

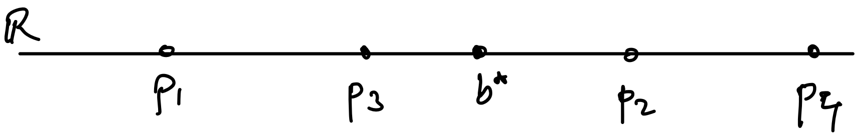
Da wir über α maximieren gilt:

$$y_k (\langle a, x_k \rangle + b) - 1 \leq 0 \implies \alpha_k = 0$$

so dass:

$$\mathcal{L} = \|a\|^2 - \sum_{k: \alpha_k > 0} \big| y_k (\langle a, x_k \rangle + b) - 1 \big|$$
$$= \|a\|^2 - \sum_{k: \alpha_k > 0} \big| \langle a, x_k \rangle + b - y_k \big|$$

Cartoon: $\mathbb{R}$



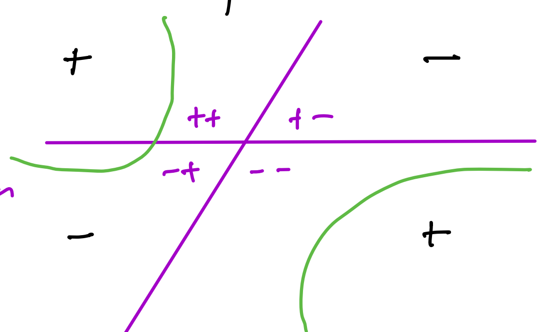
D.h. $b^* = \underset{k:\, \alpha_k \neq 0}{\mathrm{median}} \, (\underbrace{y_k - \langle a^*, x_k \rangle}_{p_k})$. (Beweis: Übung). ☐

<u>Beispiel</u> 17.2

Nicht alle Datenmengen lassen sich gut mit Hyperebenen trennen, auch nicht per Soft-Margin-SVM:

Lösung 1:

Mehrere Hyperebenen



Lösung 2:

Nicht-Lineare Separation

Hyperbel anstatt Linie.

Für nicht-lineare Separation führen wir wieder eine <u>Feature Map</u> ein

$$\phi: \mathbb{R}^D \to \mathbb{R}^P.$$

Bsp für Hyperbelsuche: $(D=2)$ $\quad \phi\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = (1, x_1, x_2, x_1 x_2, x_1^2, x_2^2)$.

<u>Beachte</u> In Dual SVM müssen wir nur $\langle \phi(x_k), \phi(x_l) \rangle$ kennen, nicht unbedingt ϕ selbst.

<u>Definition</u> 17.3

Eine Funktion der Form

$$\mathcal{K}(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle$$

für $\phi: \mathbb{R}^D \to \mathbb{R}^P$ nicht-linear, heißt Kernel-Map.

Eine Kernel-Map gibt uns die <u>Kernel-Matrix</u>

$$G = \left(\mathcal{K}(x_k, x_l) \right)_{k,l=1}^{n} \in \mathbb{R}^{n \times n}$$

für Dual-SVM.

<u>Lemma</u> 17.4

Sei $G \in \mathbb{R}^{n \times n}$. Dann ist G positiv-semidefinit, genau dann wenn P existiert und $z_1, \dots, z_n \in \mathbb{R}^P$ mit

$$G = \left(\langle z_k, z_l \rangle \right)_{k,l=1}^{n}.$$

<u>Beweis</u>

Sei $Z = [z_1 \dots z_n] \in \mathbb{R}^{P \times n}$.

1) Falls $G = Z^T Z$, dann gilt für alle $a \in \mathbb{R}^n$:

$$a^T G a = a^T Z^T Z a = (Za)^T (Za) \geq 0.$$

2) Falls G positiv-semidefinit ist, finden wir eine Cholesky-Zerlegung $G = Z^T Z$. Die Spalten von Z sind dann die z_i.

Nach Prop. 17.1. haben wir die Optimalwerte (für nichtlineares ϕ)

$$a^* = \sum_{k=1}^{n} u_k \, \phi(x_k)$$

$$b^* = \text{median} \quad y_k - \langle a^*, x_k \rangle, \; \alpha_k \neq 0.$$

D.h. wenn wir folgende Funktion definieren

$$\psi(x) = \sum_{k=1}^{n} u_k \, \mathcal{K}(x_k, x). = \langle a^*, \phi(x) \rangle$$

Dann: $b^* = \text{median} \; |y_k - \psi(x_k)| \quad \alpha_k \neq 0$ und

$$f_\Theta(x) = \operatorname{sgn}(\psi(x) + b^*).$$

Dies führt zu folgendem Algorithmus:

<u>Input:</u>
- Training data $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^D \times \{-1, 1\}$
- Kernel map $\mathcal{K}(x_1, x_2)$
- Regularisierungsparameter C

<u>Output:</u> Funktion der Form $f_\Theta(x) = \operatorname{sgn}(\langle a, x \rangle + b)$

1) Berechne die Kernel-Matrix $G = (\mathcal{K}(x_k, x_\ell))_{k,\ell=1}^{n}$
2) Löse Dual-SVM mit G und C, erhalte α
3) Definiere $\psi(x) = \sum_{k=1}^{n} \frac{1}{2} \alpha_k y_k \mathcal{K}(x_k, x)$.
4) $b =$ median von $y_k - \psi(x_k)$, $\alpha_k \neq 0$.
5) Return $f_\Theta(x) = \operatorname{sgn}(\psi(x) + b)$.

Das folgende Lemma illustriert, warum Kernel maps hilfreich sind, um ϕ nicht direkt auswerten zu müssen.

Lemma 17.3

Sei für $x \in \mathbb{R}^D$ $\quad \phi(x) = \left(x_1^{i_1} \cdots x_D^{i_D} \right) \big|_{i_1 + \dots + i_D \leq d}$

Dann:

$$\langle \phi(u), \phi(v) \rangle = \left(\langle u, v \rangle + 1 \right)^d .$$

Beweis

$$\langle \phi(u), \phi(v) \rangle = \sum_{\substack{i_0, i_1, \dots, i_D \\ i_0 + i_1 + \dots + i_D = d}} u_0^{i_0} u_1^{i_1} \cdots u_D^{i_D} v_0^{i_0} v_1^{i_1} \cdots v_D^{i_D} \quad , \; u_0 = v_0 = 1$$

$$= \sum_{i_0 + i_1 + \dots + i_D = d} (u_0 v_0)^{i_0} (u_1 v_1)^{i_1} \cdots (u_D v_D)^{i_D}$$

$$= \left(u_0 v_0 + u_1 v_1 + \dots + u_D v_D \right)^d = \left(\langle u, v \rangle + 1 \right)^d .$$

Da $\quad (x_0 + \dots + x_D)^d = \sum_{i_0 + \dots + i_D = d} x_0^{i_0} \cdots x_D^{i_D}$.

$\square$