Vorlesung 16 Support vector machines

In den letzten Vorlesungen haben wir Regressionsprobleme behandelt. Dabei war die Response-Variable $y \in \mathbb{R}^N$ kontinuierlich.

Heute: $y \in \{-1, 1\}$ ist binär.
Wir wollen wieder ein Modell $f_\Theta : \mathbb{R}^D \to \{-1, 1\}$ finden, dass die Trainingsdaten $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^D \times \{-1, 1\}$ beschreibt.

Support vector machines sind ein Modell für dieses Setting:

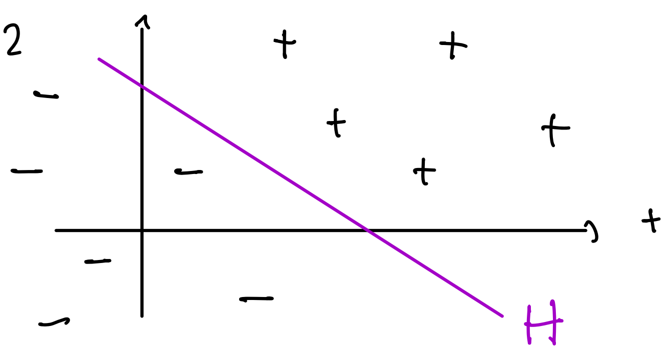
$$f_\Theta(x) = \operatorname{sgn}(\langle a, x \rangle + b), \quad \Theta = (a, b) \in \mathbb{R}^D \times \mathbb{R}.$$

Beachte Wir behandeln hier ein deterministisches Modell.

Die Geometrie dieses Modells ist wie folgt.
Für alle $\Theta = (a, b) \in \mathbb{R}^D \times \mathbb{R}$ haben wir die Hyperebene

$$H = \{ x \in \mathbb{R}^D \mid \langle a, x \rangle + b = 0 \}$$

Beispiel $D = 2$



Ziel Daten durch eine Hyperebene trennen.

Es bleibt zu klären, wie wir entsprechende Parameter für unsere Daten finden.

Dazu brauchen wir zwei Lemmata.

Lemma 16.1

Sei $x \in \mathbb{R}^D$ und $\theta = (a,b) \in \mathbb{R}^D \times \mathbb{R}$. Für alle $y \in \{-1, 1\}$ gilt:

$$y = f_\theta(x) \iff y(\langle a, x\rangle + b) > 0.$$

Beweis

Möglichkeit 1): $y = f_\theta(x) = -1$. Dann ist $\langle a, x\rangle + b < 0$

$$\rightarrow \quad y(\langle a, x\rangle + b) > 0$$

Möglichkeit 2) $y = f_\theta(x) = 1$. Dann ist $\langle a, x\rangle + b > 0$

$$\rightarrow \quad y(\langle a, x\rangle + b) > 0.$$

$\square$

Lemma 16.2

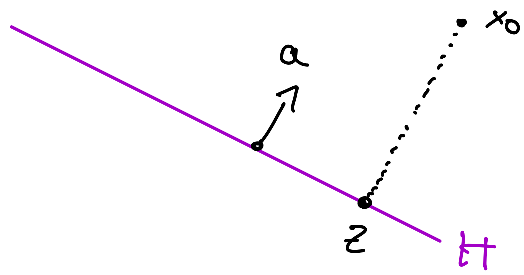
Sei $x_0 \in \mathbb{R}^D$ und $H = \{x \in \mathbb{R}^D \mid \langle a, x\rangle + b = 0\}$.

Sei $\langle a, a\rangle = 1$. Der Euklidische Abstand von x_0 zu H ist

$$|\langle a, x_0\rangle + b| = y_0(\langle a, x_0\rangle + b), \qquad y_0 = f_\theta(x_0).$$

Beweis

Sei $z \in H$ der Punkt auf H, der den Abstand zu x_0 minimiert.



Dann können wir

$$x_0 = z + ra \quad \text{für } r \in \mathbb{R}$$

schreiben, wobei $|r| = \|z - x_0\| =$ Abstand von x_0 zu H.

Beachte: $x_0 - z = ra$ ist orthogonal zu H.

Daraus folgt:

$$\langle a, x_0\rangle + b = \langle a, z + ra\rangle + b$$

$$= \langle a, z\rangle + r\langle a, a\rangle + b = r, \text{ da } z \in H.$$

$\Rightarrow \quad |\langle a, x_0\rangle + b| = |r| =$ Abstand von x_0 zu H..

Eine geeignete Parameterwahl ist es nun den Abstand der Input-Daten $x_1, \ldots, x_n \in \mathbb{R}^D$ zur Hyperebene H zu maximieren unter den Bedingungen, dass $f_\theta(x_i) = y_i$. Dies führt zu folgendem Optimierungsproblem:

$\|a\|^2 = \langle a, a\rangle = 1$

$$(*) \quad \max_{\Theta = (a,b) \in \mathbb{R}^D \times \mathbb{R}} \; r \qquad \text{Abstand zu } H.$$

$$\text{s.t.:} \quad y_k(\langle a, x_k\rangle + b) \geq r. \quad \text{für } k = 1, \ldots, n,$$

$$r \geq 0.$$

Nebenbedingungen

Das Optimierungsproblem $(*)$ wird oft in einer äquivalenten Form beschrieben: Sei

$$a' = \frac{a}{r}, \qquad b' = \frac{b}{r}$$

Dann erhalten wir die Constraints

$$y_k(\langle a', x\rangle + b') \geq 1.$$

und es gilt:

$$\|a'\| = \frac{\|a\|}{r} = \frac{1}{r}$$

Definition 16-3.

Die Hard-Margin-SVM ist durch folgendes Optimierungsproblem gegeben:

$$\min_{\Theta = (a,b) \in \mathbb{R}^D \times \mathbb{R}} \|a\|^2$$

$$\text{s.t.} \quad y_k(\langle a, x_k\rangle + b) \geq 1 \quad \text{für } k = 1, \ldots, n.$$

Das Problem mit Hard-Margin SVM ist, dass Ausreißer in den Daten dafür sorgen, dass das Optimierungsproblem keine Lösung hat. Soft-Margin SVM löst dieses Problem wie folgt:

Definition 16.4
Soft-Margin SVM ist durch folgendes Optimierungsproblem gegeben.

$$(**) \qquad \begin{aligned} \min_{a,b,\xi} \quad & \|a\|^2 + C \sum_{k=1}^{n} \xi_k \\ \text{s.t.} \quad & y_k(\langle a, x_k\rangle + b) \geq 1 - \xi_k, \quad k = 1, \ldots, n \\ & \xi_k \geq 0, \end{aligned}$$

wobei $C > 0$ ein Regularisierungsparameter ist.

Proposition 16.5
Der Hinge-Loss ist $\ell(y, \hat{y}) = \max\{0, 1 - y \cdot \hat{y}\}$.
Soft-Margin SVM ist durch ERM bzgl. eines regularisierten Hinge-Loss gegeben.

Beweis
Übung.

Hard-Margin-SVM und Soft-Margin-SVM werden Primal SVMs genannt. Eine andere Formulierung ist die Dual SVM.

Wir betrachten dazu die Lagrange-Funktion

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

(**) ist dann äquivalent zu

$$\min_{a, b, \xi} \max_{\alpha, \beta} \mathcal{L}(a, b, \xi, \alpha, \beta).$$

$$\text{s.t.} \quad \alpha_k, \beta_k \geq 0, \quad k = 1, \dots, n.$$

Die KKT-Bedingungen sagen, dass ein optimaler Wert folgendes erfüllen muss:

$$\frac{\partial \mathcal{L}}{\partial a} = \frac{\partial \mathcal{L}}{\partial b} = \frac{\partial \mathcal{L}}{\partial \xi} = 0 .$$

Sei jetzt $u := \left(\frac{1}{2} \alpha_k y_k \right)_{k=1}^{n}$, $v = \left(\langle a, x_k \rangle + b \right)_{k=1}^{n}$, $e = (1, \dots, 1)^T$. Dann:

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

$$= \langle a, a \rangle + C \cdot \langle e, \xi \rangle - 2 \langle u, v \rangle - \langle \alpha + \beta, \xi \rangle + \langle \alpha, e \rangle$$

Damit folgt:

(***)
$$\begin{cases} \frac{\partial \mathcal{L}}{\partial a} = 2a - 2 \sum_{k=1}^{n} u_k x_k = 0 & \Rightarrow \quad a = \sum_{k=1}^{n} u_k x_k . \\ \frac{\partial \mathcal{L}}{\partial b} = 2 \sum_{k=1}^{n} u_k = 0 . & \Rightarrow \quad \sum_{k=1}^{n} u_k = 0 \end{cases}$$

$$\frac{\partial \mathcal{L}}{\partial \xi} = Ce - (\alpha+\beta) = 0 \quad \Rightarrow \quad \alpha+\beta = Ce \Rightarrow \alpha_k \leq C, \quad k=1,\dots,n.$$

<u>Bemerkung</u> Der Name "Support Vector Machine" kommt von der Gleichung $a = \sum_{k=1}^{n} u_k x_k = \frac{1}{\lambda} \sum_{k=1}^{n} \alpha_k y_k x_k$.

Die x_k mit $\alpha_k \neq 0$ heißen "Support Vectors".

Wir setzen die Optimalitätsbedingungen (xxx) in $\mathcal{L}$ ein:

$$\begin{aligned}
\mathcal{L} &= \langle a,a\rangle + C\cdot\langle e,\xi\rangle - \lambda\langle u,v\rangle - \langle \alpha+\beta, \xi\rangle + \langle \alpha, e\rangle \\
&= \langle a,a\rangle - \lambda\langle u,v\rangle + \langle \alpha,e\rangle \quad (\text{weil } Ce-(\alpha+\beta)=0) \\
&= \langle \sum_{k=1}^{n} u_k x_k, \sum_{\ell=1}^{n} u_\ell x_\ell\rangle - \lambda\langle u,v\rangle + \langle \alpha,e\rangle \\
&= \sum_{k=1}^{n}\sum_{\ell=1}^{n} u_k u_\ell \langle x_k, x_\ell\rangle - \lambda\langle u,v\rangle + \langle \alpha,e\rangle \\
&= u^T G u - \lambda\langle u,v\rangle + \langle \alpha,e\rangle,
\end{aligned}$$

wobei $G = (\langle x_k, x_\ell\rangle)_{k,\ell=1}^{n} \in \mathbb{R}^{n\times n}$.

Weiterhin:

$$\begin{aligned}
\langle u,v\rangle &= \langle u, (\langle a, x_k\rangle + b)_{k=1}^{n}\rangle \\
&= \sum_{k=1}^{n} u_k \langle a, x_k\rangle + b \underbrace{\sum_{k=1}^{n} u_k}_{=0} \\
&= \sum_{k=1}^{n} u_k \langle \sum_{\ell=1}^{n} u_\ell x_\ell, x_k\rangle \\
&= \sum_{k=1}^{n}\sum_{\ell=1}^{n} u_k u_\ell \langle x_k, x_\ell\rangle = u^T G u
\end{aligned}$$

$$\Rightarrow \quad \mathcal{L} = -u^T G u + \sum_{k=1}^{n} \alpha_k$$

<u>Definition</u> 16.6

Die <u>Dual SVM</u> ist gegeben durch das Optimierungsproblem

$$\max_{\alpha} \; -u^T G u + \sum_{k=1}^{n} \alpha_k$$

s.t.

$$\sum_{k=1}^{n} \alpha_k y_k = 0, \qquad 0 < \alpha_k \leq C$$

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