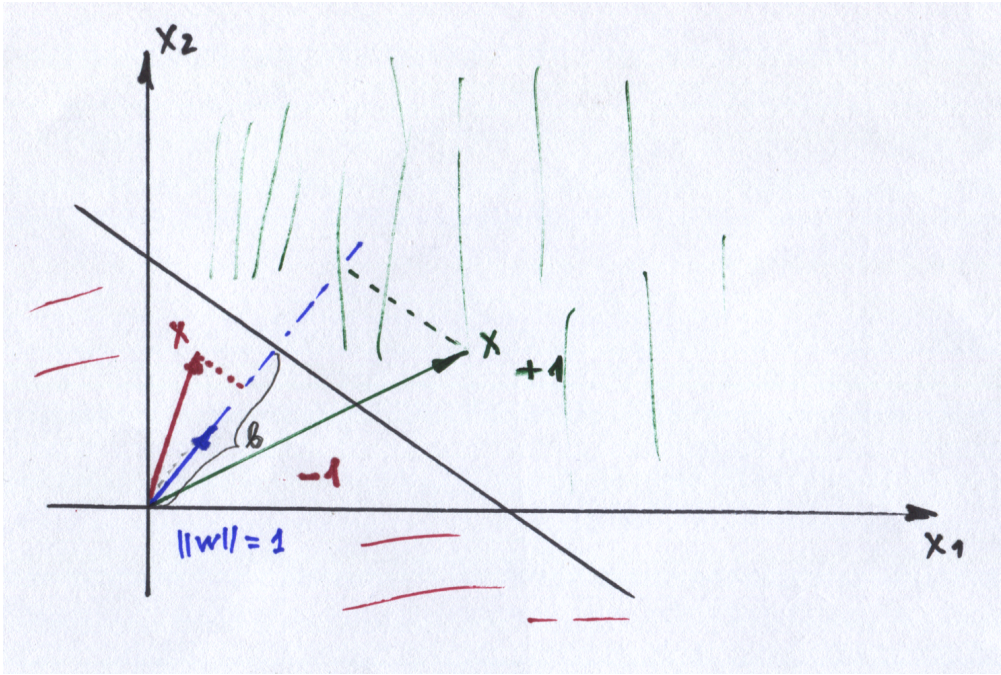
Mustererkennung: Support Vector Machines

Lineare Klassifikatoren

Baustein für so gut wie alles andere – Abbildung $f : \mathbb{R}^n \to \{+1, -1\}$
Aufteilung des Input-Raums auf zwei Halbräume, die zwei Klassen entsprechen



$$y = f(x) = \operatorname{sgn}(\langle x, w \rangle - b)$$

mit dem **Gewichtsvektor** $w \in \mathbb{R}^n$ und Schwellwert $b \in \mathbb{R}$.
Geometrische Interpretation: w ist der **Normalenvektor** der **Hyperebene**,
die die Halbräume trennt, für diese gilt $\langle x, w \rangle = b$.
Wenn $||w|| = 1$, dann ist b der Abstand der Hyperebene vom Koordinatenursprung.

Synonyme – Neuron-Modell, Perzeptron ...

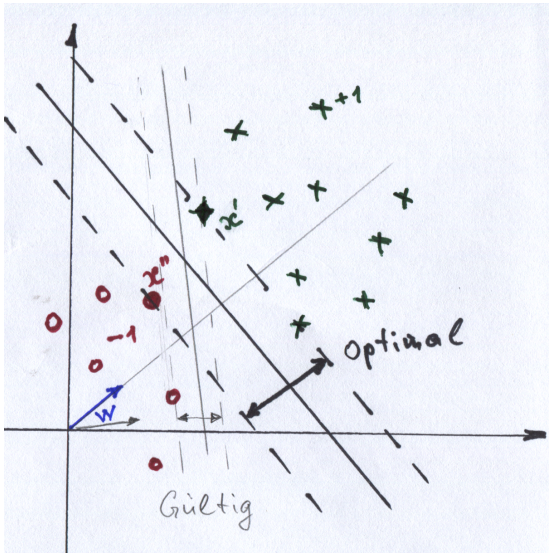
Lineare SVM

Gegeben sei eine Lernstichprobe $X = \Big((x_i, y_i) \ldots\Big)$ mit den
(i) Daten $x_i \in \mathbb{R}^n$ und (ii) Klassen $y_i \in \{-1, +1\}$

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

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

Gesucht wird die Hyperebene **maximaler Breite**, die die Daten richtig separiert:



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:

$$\|w\|^2 \to \min_{w,b}$$
$$\text{s.t.} \quad y_i \cdot [\langle w, x_i \rangle + b] \geq 1 \quad \forall i$$

Lineare SVM

Aufgabe:

$$\|w\|^2 \to \min_{w,b}$$
$$\text{s.t.} \quad y_i \cdot [\langle w, x_i \rangle + b] \geq 1 \quad \forall i$$

Lagrangian:

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_i \alpha_i \cdot (y_i \cdot [\langle w, x_i \rangle + b] - 1) \to \max_{\alpha} \min_{w,b}$$
$$\alpha_i \geq 0 \quad \forall i$$

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:
$\alpha_i \to \infty$, $L(w, b, \alpha) \to \infty$ – 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 $L(w, b, \alpha) < \infty$

Lineare SVM

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:

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

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

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

Man setzt diese **lineare Kombination** von x_i in die Entscheidungsregel ein

$$f(x) = \operatorname{sgn}\Big(\langle x, w \rangle + b\Big) = \operatorname{sgn}\bigg(\Big\langle x, \sum_i \alpha_i y_i x_i \Big\rangle + b\bigg) =$$

$$\operatorname{sgn}\bigg(\sum_i \alpha_i y_i \langle x, x_i \rangle + b\bigg)$$

und sieht, dass der Vektor w nicht explizit benötigt wird!!!
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:

$$\sum_i \alpha_i - \frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \to \max_\alpha$$

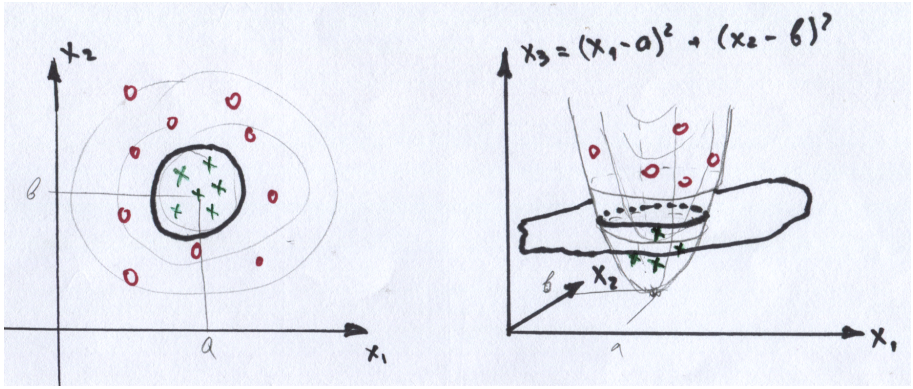
$$\text{s.t.} \quad \alpha_i \geq 0, \quad \sum_i \alpha_i y_i = 0$$

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

Merkmalsräume

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

Beispiel:



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

$\Phi : \mathbb{R}^2 \to \mathbb{R}^3$
$\Phi(x_1, x_2) = (x_1^2, \sqrt{2} x_1 x_2, x_2^2)$
(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 $\langle \Phi(x), \Phi(x') \rangle$.

$$\begin{aligned} \Big\langle \Phi(x_1, x_2), \Phi(x_1', x_2') \Big\rangle &= \Big\langle (x_1^2, \sqrt{2} x_1 x_2, x_2^2), (x_1'^2, \sqrt{2} x_1' x_2', x_2'^2) \Big\rangle = \\ & x_1^2 x_1'^2 + 2 x_1 x_2 x_1' x_2' + x_2^2 x_2'^2 = \\ & (x_1 x_1' + x_2 x_2')^2 = \langle x, x' \rangle^2 = \end{aligned}$$

$$k(x, x') \quad - \quad \textbf{Kernel}$$

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

Kernels

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

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

Der Merkmalsraum $\mathcal{H}$ und die Abbildung $\Phi : \mathcal{X} \to \mathcal{H}$ müssen dabei nicht unbedingt explizit definiert werden (Black Box).

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

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

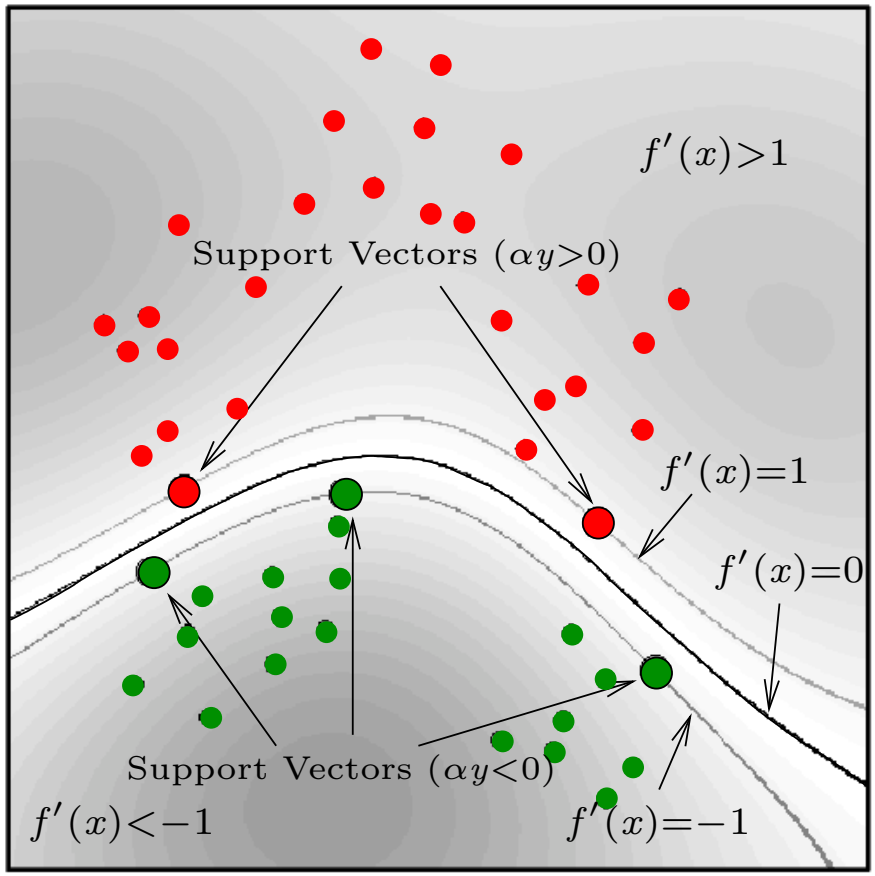
Sind k_1 und k_2 Kernels, dann sind αk_1, $k_1 + k_2$, $k_1 k_2$ auch Kernels
(es gibt noch weitere Operationen um Kernels aus Kernels zu bauen).

Beispiele (oft benutzte Kernels):

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

Beispiel

Entscheidungsregel mit dem Gausschen Kernel $k(x, x') = \exp\left[-\frac{\|x-x'\|^2}{2\sigma^2}\right]$:



$$f(x) = \operatorname{sgn}\Big(f'(x)\Big) = \operatorname{sgn}\left(\sum_i y_i \alpha_i \exp\left[-\frac{\|x - x_i\|^2}{2\sigma^2}\right]\right)$$

Zusammenfassung

- SVM ist ein Beispiel des diskriminativen Lernens
  $\Rightarrow$ mit allen entsprechenden Vor- und Nachteilen
- Baustein – lineare Klassifikatoren
  Formulierung mit Skalarprodukten – die Daten an sich werden nicht benötigt
- Merkmalsräume – machen nichtlineare Entscheidungen im Input-Raum möglich
- Kernel – Skalarprodukt im Merkmalsraum
  $\Rightarrow$ der Merkmalsraum muss nicht unbedingt explizit definiert werden

Literatur:

- Bernhard Schölkopf, Alex Smola ...
- Nello Cristianini, John Shawa-Taylor ...
- Internet ...