Stützvektormethode (SVM)

- Maximieren der Breite einer separierenden Hyperebene – maximum margin method
- Transformation des Datenraums durch Kernfunktion
- Strukturelle Risikominimierung

- Vladimir Vapnik „The Nature of Statistical Learning Theory“ Springer Vg. 1995
- W.N. Wapnik, A. Tscherwonenkis „Theorie der Zeichenerkennung“ Akademie Vg. 1979
- Christopher Burges "A Tutorial on Support Vector Machines for Pattern Recognition" in: Data Mining and Knowledge Discovery2, 1998, 121-167

Erinnerung: Funktionslernen

Gegeben:

Beispiele X in LE

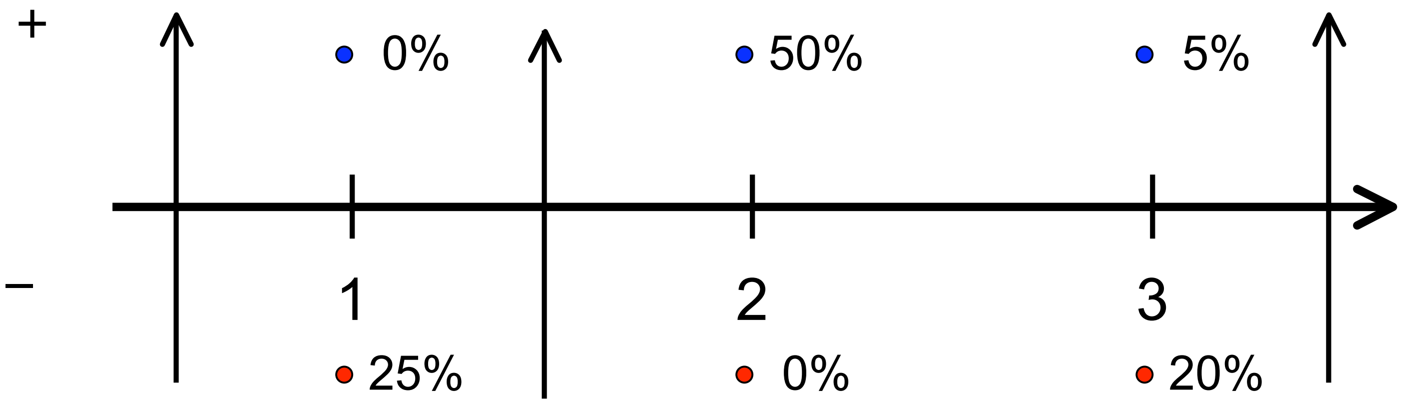
- die anhand einer Wahrscheinlichkeitsverteilung P auf X erzeugt wurden und
- mit einem Funktionswert Y = t(X) versehen sind (alternativ: Eine Wahrscheinlichkeitsverteilung P(Y|X) der möglichen Funktionswerte).

H die Menge von Funktionen in LH.

Ziel: Eine Hypothese h(X) $\in$ H, die das erwartete Fehlerrisiko R(h) minimiert.

Risiko: $$R(h) = \sum_x Q(x,h)P(x)$$

Beispiel: Funktionenlernen



- $H = \{ f_a \mid f_a(x) = 1 \text{ für } x \geq a,\ f_a(x) = -1 \text{ sonst},\ a \in \Re \}$
- $R(f_0) = 0{,}25 + 0 + 0{,}20 = 0{,}45$
- $R(f_{1,5}) = 0 + 0 + 0{,}20 = 0{,}20$
- $R(f_{3,5}) = 0 + 0{,}5 + 0{,}05 = 0{,}55$

Reale Beispiele

- Klassifikation: $Q(x,h) = 0$, falls $t(x) = h(x)$,
  1 sonst
  - Textklassifikation (x = Worthäufigkeiten)
  - Handschriftenerkennung (x = Pixel in Bild)
  - Vibrationsanalyse in Triebwerken (x = Frequenzen)
  - Intensivmedizinische Therapie (x = Vitalzeichen)
- Regression: $Q(x,h) = (t(x)-h(x))^2$
  - Zeitreihenprognose (x = Zeitreihe, $t(x)$ = nächster Wert)

Erinnerung: Minimierung des beobachteten Fehlers

Funktionslernaufgabe nicht direkt lösbar. Problem:

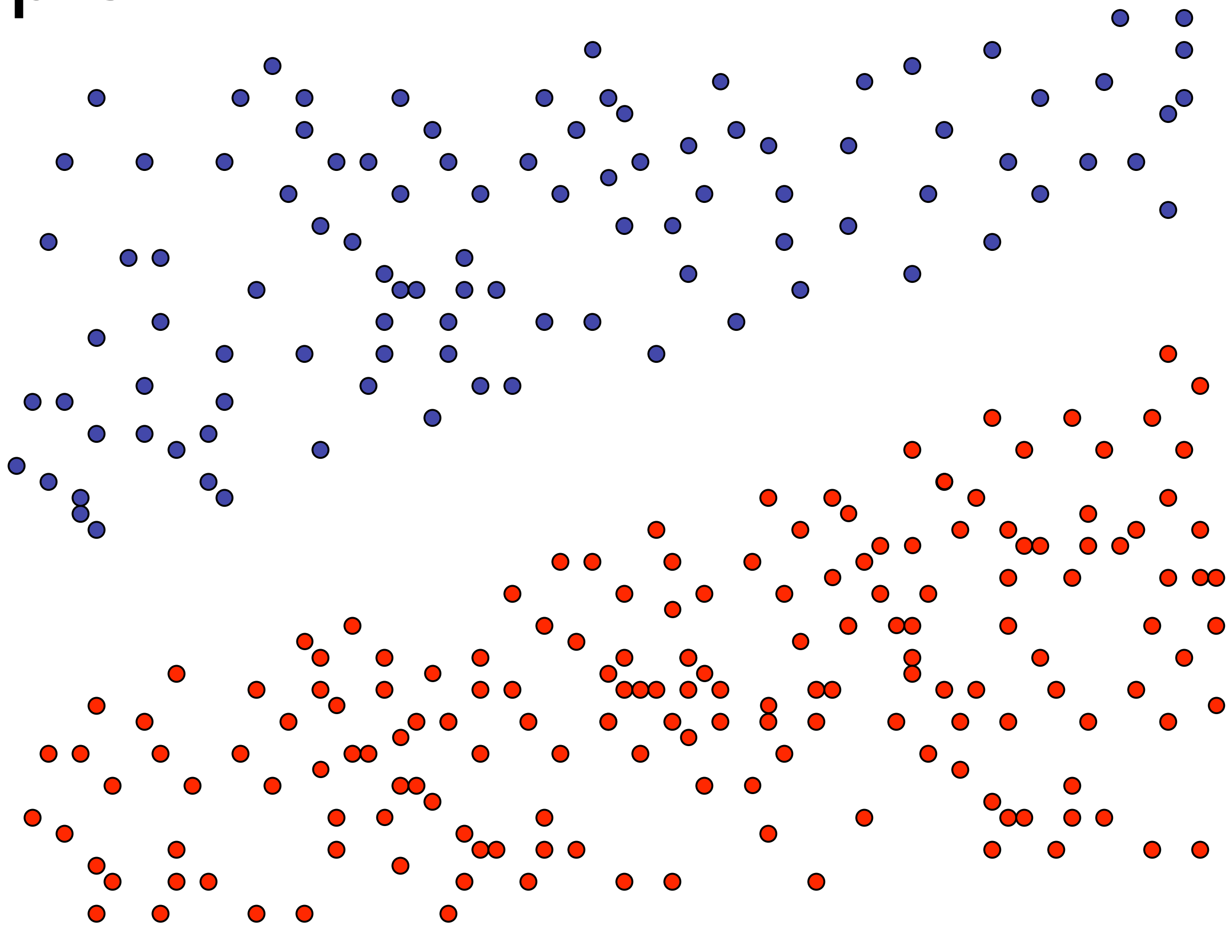
- Die tatsächliche Funktion t(X) ist unbekannt.
- Die zugrunde liegende Wahrscheinlichkeit ist unbekannt.

Ansatz:

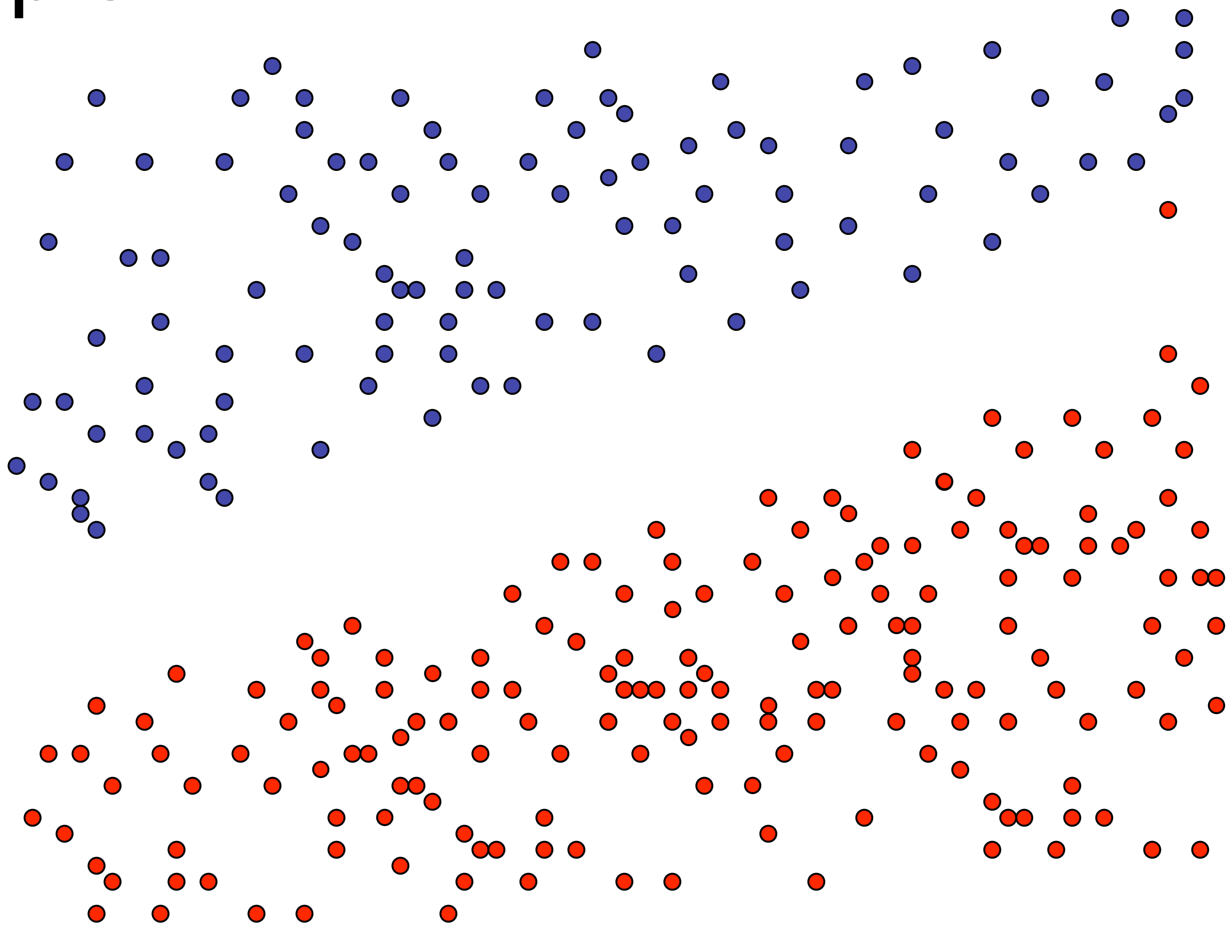
- eine hinreichend große Lernmenge nehmen und für diese den Fehler minimieren.

$\Rightarrow$ Empirical Risk Minimization

Beispiel



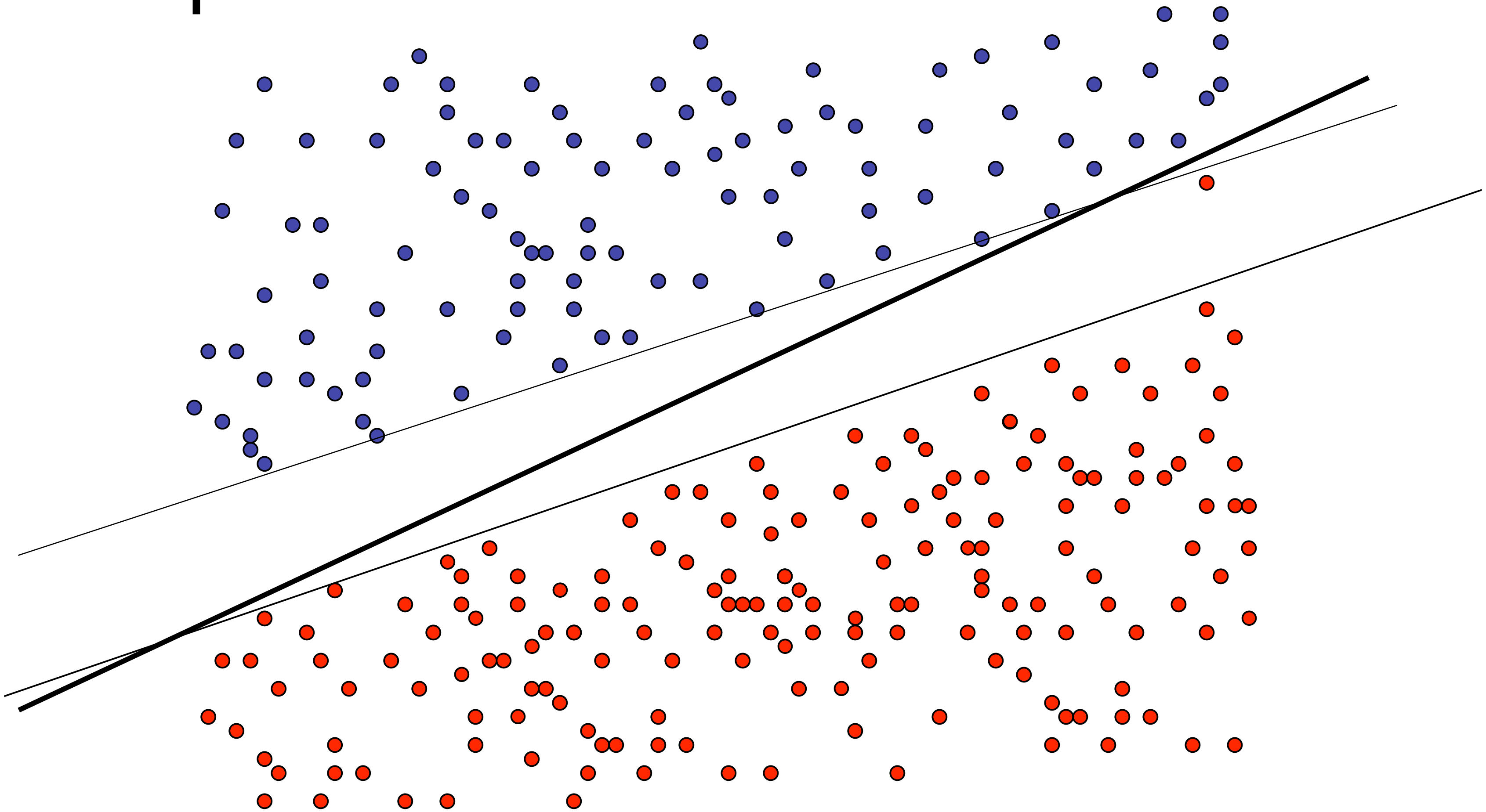
Beispiel II



Probleme der ERM

- Aufgabe ist nicht eindeutig beschrieben: Mehrere Funktionen mit minimalem Fehler existieren. Welche wählen?
- Overfitting: Verrauschte Daten und zu wenig Beispiele führen zu falschen Ergebnissen.

Beispiel III



Einführung

- Bernhard Schölkopf, Alexander Smola „Learning with Kernels“ MIT Press 2002
- Zwei-Klassen-Problem:
  - Trainingsdaten $(x_1, y_1), ..., (x_m, y_m), x_m \in X, y_m \in \{+1, -1\}$
  - Ähnlichkeit eines neuen x_i bestimmt y_i
  - Ähnlichkeitsmaß $k: X \times X \rightarrow \Re$
    $(x, x') \rightarrow k(x, x')$
    z.B. Skalarprodukt $x^*x' := (... + [x]_i \, [x']_i + ...)$

Grundbegriffe

- Skalarprodukt x*y: Seien x und y Vektoren aus $\Re^p$

$$x * y = \sum_{i=1}^{p} [x][y]$$

- Euklidsche Länge (Betrag) eines Vektors ||x||:

$$\|x\| = \sqrt{x * x} = \left(\sum_{i=1}^{p} [x]^2 \right)^{\frac{1}{2}}$$

- Hyperebene H: Sei $w \neq 0$ der Normalenvektor und $b \in \Re$ der bias

$$H(w,b) = \{ x \mid w * x + b = 0 \}$$

Warum Skalarprodukt?

- Cosinus des Winkels zwischen x und x‘, wenn beide Vektoren auf die Länge 1 normiert sind.
- Abstand zwischen x und x‘ ist Länge des Differenzvektors.
- Voraussetzung: Beispiele sind Vektoren.
- Überführung in einen Raum mit Skalarprodukt
  $\Phi : X \rightarrow \mathcal{H}$
- Wenn X bereits ein Raum mit Skalarprodukt ist, kann nicht-lineare Abbildung Φ auch sinnvoll sein.

Einfachster Lernalgorithmus

- Beispiele in einem Raum mit Skalarprodukt.
- Durchschnitt einer Klasse:

  in der Mitte liegt Punkt
  $c:=(c_+ + c_-)/2$
  Vektor x-c verbindet
  neues Beispiel und c

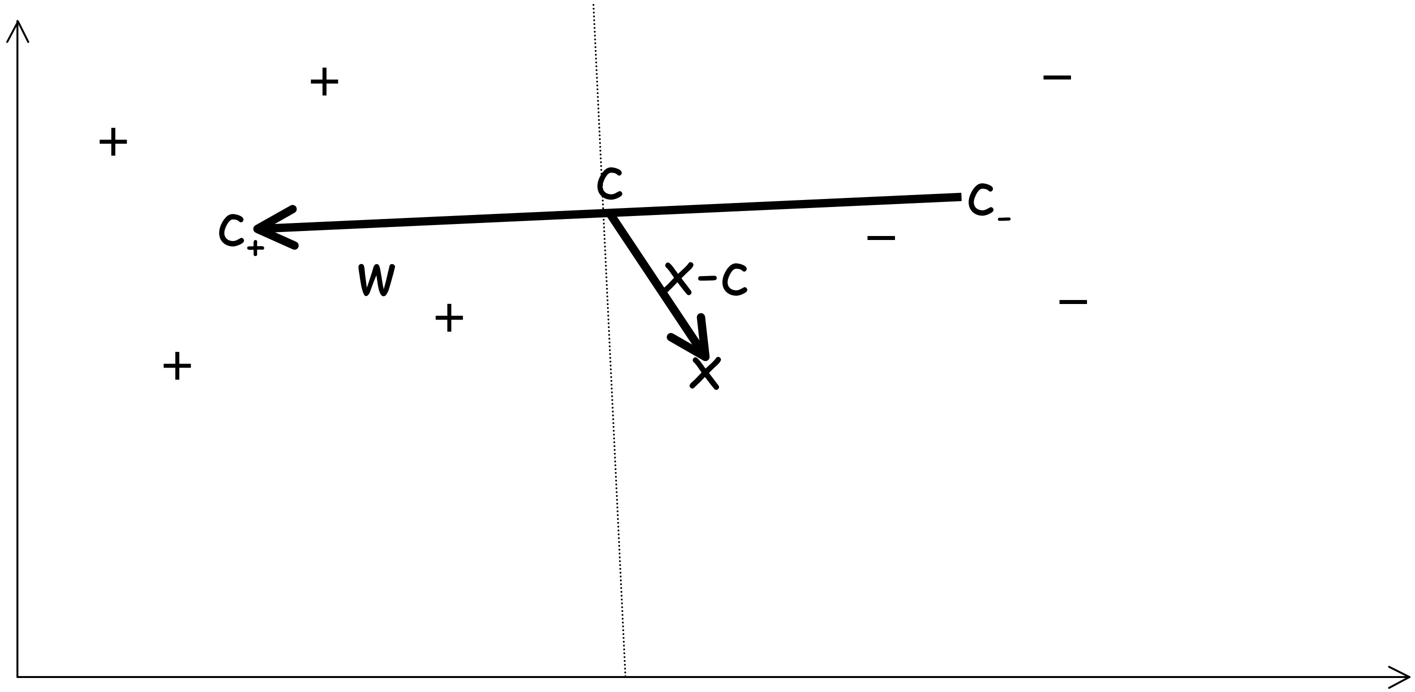
$$c_+ = \frac{1}{m_+} \sum_{\{i|y_i=+1\}} x_i$$

$$c_- = \frac{1}{m_-} \sum_{\{i|y_i=-1\}} x_i$$

Anzahl positiver Beispiele: m_+

- Ähnlichkeit zum Durchschnitt einer Klasse:
  Winkel zwischen $w:=c_+ - c_-$ und x-c
- Berechnen über Skalarprodukt!

Lernalgorithmus im Bild



Lernalgorithmus in Formeln

$$
\begin{aligned}
y &= sign\left(\left(x - c\right) * w\right) \\
&= sign\left(\left(x - \left(c_+ + c_-\right)/2\right) * \left(c_+ - c_-\right)\right) \\
&= sign\left(\left(x * c_+\right) - \left(x * c_-\right) - \frac{1}{2}c_+^2 - \frac{1}{2}c_+ * c_- + \frac{1}{2}c_-^2 + \frac{1}{2}c_+ * c_-\right) \\
&= sign\left(\left(x * c_+\right) - \left(x * c_-\right) + \frac{1}{2}\left(\left\|c_-\right\|^2 - \left\|c_+\right\|^2\right)\right) \\
&= sign\left(\left(x * c_+\right) - \left(x * c_-\right) + b\right)
\end{aligned}
$$

Entscheidungsfunktion

Wir setzen nun die Mittelwerte für c_+ und c_- ein:

$$y = sign\left(\frac{1}{m_+} \sum_{\{i|y_i=+\}} x * x_i - \frac{1}{m_-} \sum_{\{i|y_i=-\}} x * x_i + b \right)$$

$$= sign\left(\frac{1}{m_+} \sum_{\{i|y_i=+\}} k(x, x_i) - \frac{1}{m_-} \sum_{\{i|y_i=-\}} k(x, x_i) + b \right)$$

Das neue Beispiel wird also mit allen Trainingsbeispielen verglichen.

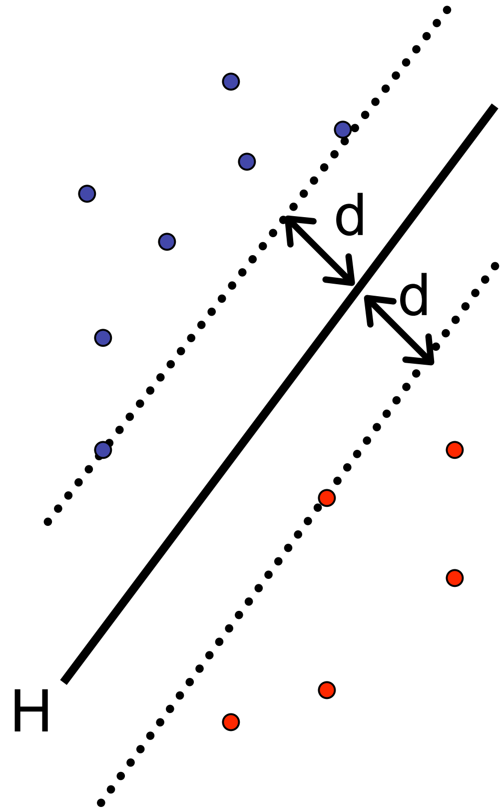
Fast...

... wäre das schon die Stützvektormethode. Aber:

- Einfach den Mittelpunkt der Beispiele einer Klasse zu berechnen ist zu einfach, um ein ordentliches w zu bekommen.
- Man erhält so nicht die optimale Hyperebene.

Die optimale Hyperebene

- Beispiele heißen <u>linear trennbar</u>, wenn es eine Hyperebene H gibt, die die positiven und negativen Beispiele voneinander trennt.
- H heißt <u>optimale Hyperebene</u>, wenn ihr Abstand d zum nächsten positiven und zum nächsten negativen Beispiel maximal ist.
- Satz: Es existiert eine eindeutig bestimmte optimale Hyperebene.

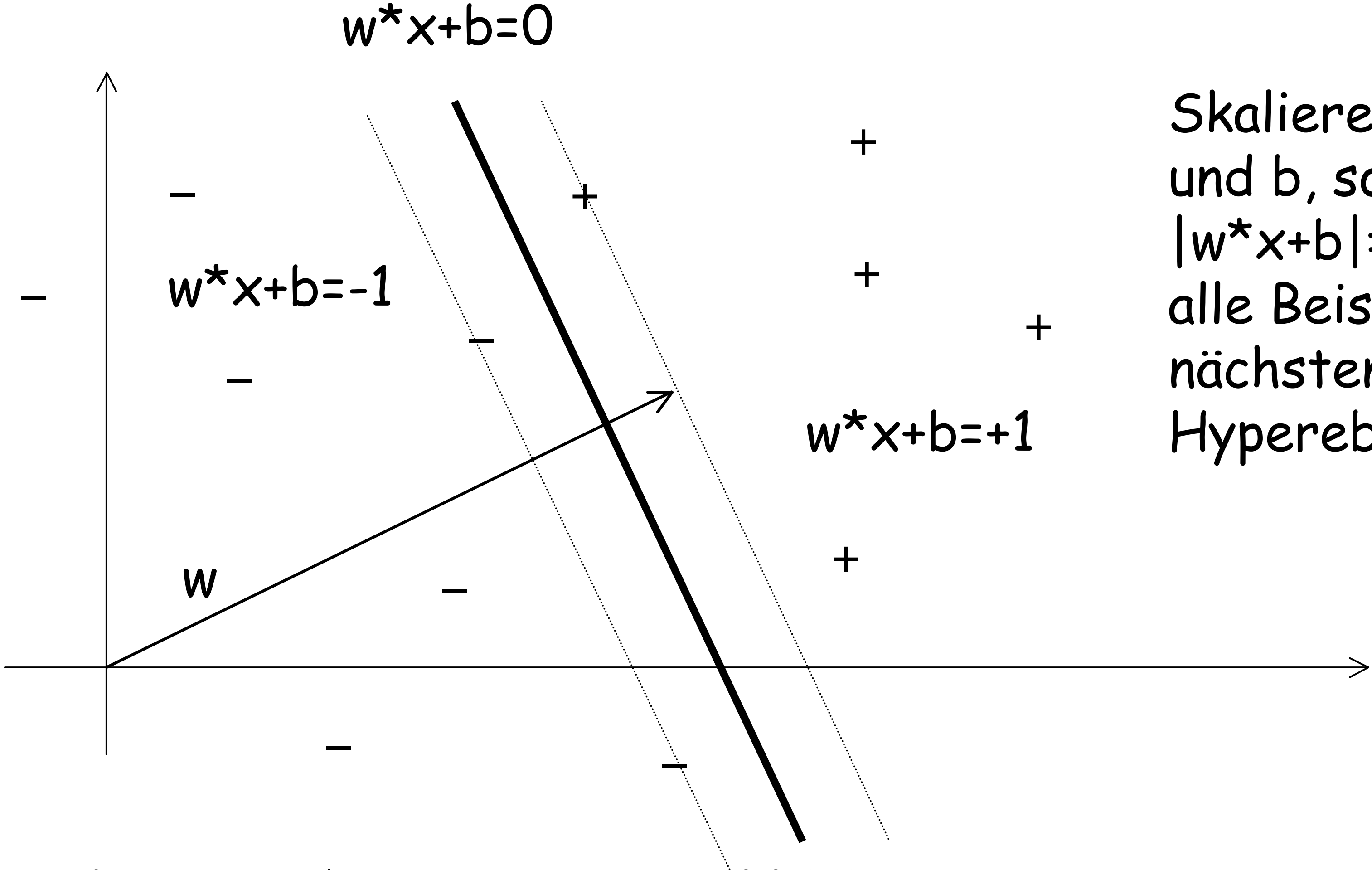


Grundbegriffe II

- Der Normalenvektor steht senkrecht auf allen Vektoren der Hyperebene. Es gilt:

$$w * x + b \begin{cases} > 0 \; falls\; x\; im\; positiven\; Raum \\ = 0 \; falls\; x\; auf\; H \\ < 0 \; falls\; x\; im\; negativen\; Raum \end{cases}$$

Bild



Skalieren von w und b, so dass |w*x+b|=1 für alle Beispiele am nächsten zur Hyperebene.

Separierende Hyperebene

- Beispiele in Form von Vektoren x aus $\Re^p$ und Klassifikation
  y=+1 (positive Beispiele) oder
  y=-1 (negative Beispiele)
  $E=\{ [x_1,y_1], [x_2,y_2], ..., [x_m,y_m]\}$
- Separierende Hyperebene H:
  positive Beispiele im positiven Halbraum,
  negative Beispiele im negativen Halbraum,
  $x*w+b=0$ für Punkte auf der Hyperebene.
- Der Abstand von H zum Ursprung ist $b / ||w||$
- Die Separierbarkeit erfüllen viele Hyperebenen.

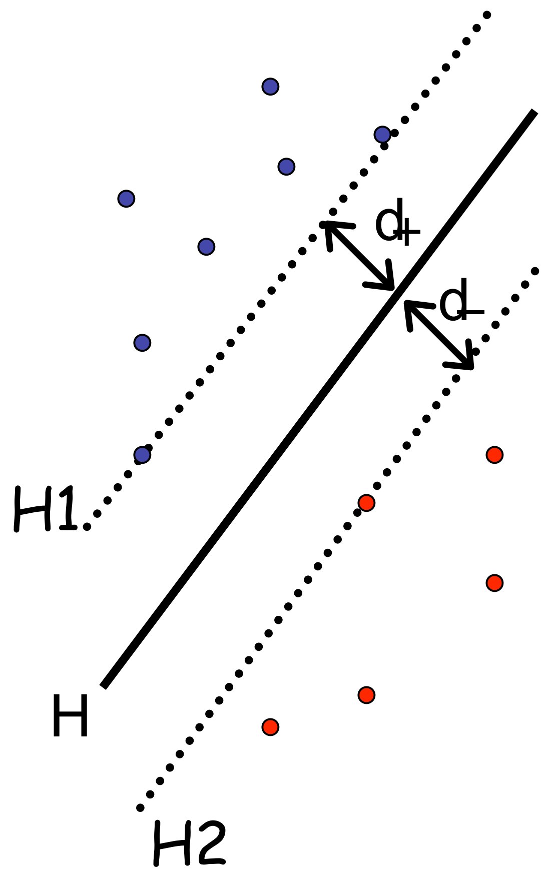
Margin für separierbare Beispiele

- Abstand d_+ von H zum nächsten positiven Beispiel
- Abstand d_- von H zum nächsten negativen Beispiel
- Margin: $d_+ + d_-$
- H1 $x_i * w + b \geq +1 \; bei \; y_i = +1$
- H2 $x_i * w + b \leq -1 \; bei \; y_i = -1$
  zusammengefasst: $\forall x_i : y_i (w * x_i + b) - 1 > 0$
- Der Abstand von H1 zum Ursprung ist |1-b | / ||w||
- Der Abstand von H2 zum Ursprung ist |–1-b | / ||w||
- $d_+ = d_- = 1 / ||w||$ und margin = 2 / ||w||

Margin

- H1 und H2 sind parallel, haben denselben Normalenvektor w.
- Per Konstruktion liegt kein Beispiel zwischen H1 und H2.
- Um 2 / ||w|| zu maximieren, müssen wir ||w|| minimieren.
- Die Nebenbedingungen müssen eingehalten werden:

$$\forall i : y_i \left(x_i * w + b \right) - 1 \geq 0$$



Minimieren der Länge

- Um die geometrische Breite $\frac{1}{\|w\|}$ zu maximieren, müssen wir die Länge von w minimieren.
  Wir können genauso gut w*w minimieren.
- So finden wir nun eine <u>eindeutige</u> Hyperebene aus den vielen möglichen trennenden.
- Für alle Beispiele ist sie richtig: $f(x_i)>0$ gdw. $y_i>0$
- Wir können sie anwenden, um neue unklassifizierte Beobachtungen zu klassifizieren:
  $f(x)=w*x+b$
  das Vorzeichen gibt die Klasse an.

Optimierungsaufgabe

- Minimiere $||w||^2$
- so dass für alle i gilt:

  $f(x_i) = w^*x_i+b \geq 1$ für $y_i = 1$ und

  $f(x_i) = w^*x_i+b \leq -1$ für $y_i = -1$
- Äquivalente Nebenbedingungen: $y_i{}^*f(x_i) - 1 \geq 0$
- Konvexes, quadratisches Optimierungsproblem $\Rightarrow$ eindeutig in $O(n^3)$ für n Beispiele lösbar.
- Satz: $||w|| = 1/d$, d = Breite der optimalen Hyperebene bzgl. der Beispiele.

Lagrange-Funktion

- Sei das Optimierungsproblem gegeben, f(w) zu minimieren unter der Nebenbedingung $g_i(w) \geq 0$ $i=1,...,m$, dann ist die Lagrange-Funktion

$$L(w,\alpha) = f(w) - \sum_{i=1}^{m} \alpha_i g_i(w)$$

- Dabei muss gelten $\alpha_i \geq 0$
- Für Ungleichheitsbedingungen werden α-Multiplikatoren eingeführt, Gleichheitsbedingungen werden direkt eingesetzt.
- Es ist leichter, Vektor α zu bestimmen, als direkt nach der Erfüllung der Bedingungen zu suchen.

Optimierungsfunktion als Lagrange

- Minimiere L(w,b,α)!

$$L(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{m} \alpha_i \left(y_i \left(x_i * w + b\right) - 1\right)$$

- Eine optimale Lösung zeichnet sich durch die folgenden notwendigen Bedingungen an α aus:

$$w = \sum_{i=1}^{m} \alpha_i y_i x_i \qquad \sum_{i=1}^{m} \alpha_i y_i = 0$$

- L soll bezüglich w und b minimiert, bezüglich α maximiert werden.

Karush-Kuhn-Tucker Bedingungen

- Für das primale Optimierungsproblem gelten die KKT Bedingungen gdw. w, b, α die Lösung ist.

$$\frac{\partial}{\partial w_v} L(w,b,\alpha) = w_v - \sum_i \alpha_i y_i x_{i,v} = 0 \qquad v = 1,\dots,d$$

$$\frac{\partial}{\partial b} L(w,b,\alpha) = -\sum_i \alpha_i y_i = 0$$

$$y_i\left(x_i * w + b\right) - 1 \geq 0$$

$$\forall i : \alpha_i \geq 0$$

$$\forall i : \alpha_i\left(y_i\left(w * x_i + b\right) - 1\right) = 0$$

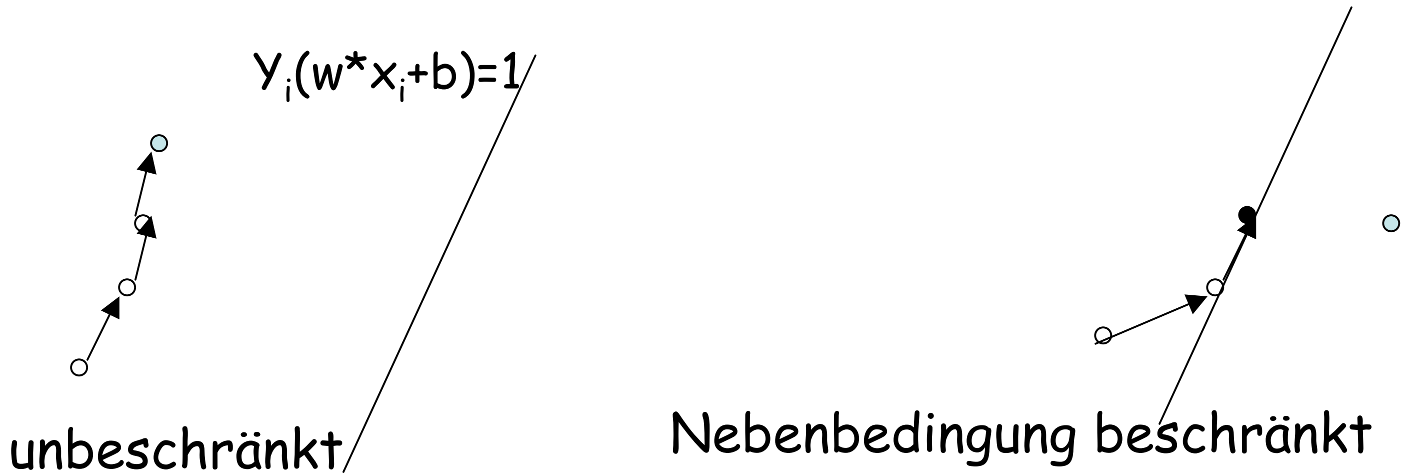
i Beispiele, v Attribute der Beispiele=Komponenten der Vektoren

Duales Problem

- Die Gleichheitsbedingungen werden in $L(w,b,\alpha)$ eingesetzt.
- Der duale Lagrange-Ausdruck $L(\alpha)$ soll maximiert werden.
- Das Minimum des ursprünglichen Optimierungsproblems tritt genau bei jenen Werten von w,b,α auf wie das Maximum des dualen Problems.

Anschaulich?

- Wir wollen w minimieren, also Δw=0, also Minimum von w in Richtung des Gradienten suchen.
- Die Nebenbedingungen sind entweder weit ab oder der auf ihnen liegende nächste Punkt zum Minimum gibt das Minimum unter Einhaltung der Nebenbedingungen an.



Umformung

$$\frac{1}{2} w * w - \sum_{i=1}^{m} \alpha_i \left[y_i \left(x_i * w + b \right) - 1 \right]$$

$$= \frac{1}{2} w * w - \sum_{i=1}^{m} \alpha_i y_i \left(x_i * w + b \right) + \sum_{i=1}^{m} \alpha_i$$

$$= \frac{1}{2} w * w - \sum_{i=1}^{m} \alpha_i y_i x_i * w - \sum_{i=1}^{m} \alpha_i y_i b + \sum_{i=1}^{m} \alpha_i$$

$$= \frac{1}{2} w * w - \sum_{i=1}^{m} \alpha_i y_i x_i * w + \sum_{i=1}^{m} \alpha_i$$

Bei gutem α muss gelten $0 = \sum_{i=1}^{m} \alpha_i y_i$

Umformung II

- Es gilt für optimalen Vektor **α** $w = \sum_{i=1}^{m} \alpha_i y_i x_i$ wir ersetzen

$$\frac{1}{2} w * w \qquad - \sum_{i=1}^{m} \alpha_i y_i x_i * w \qquad + \sum_{i=1}^{m} \alpha_i$$

$$= \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i * x_j \qquad - \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i * x_j \qquad + \sum_{i=1}^{m} \alpha_i$$

$$= + \sum_{i=1}^{m} \alpha_i \qquad - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i * x_j$$

- Mit den Nebenbedingungen:

$$0 = \sum_{i=1}^{m} \alpha_i y_i \text{ und } \alpha_i \geq 0$$

SVM Optimierungsproblem

- Maximiere
  unter $0 \leq \alpha_i$ für alle i und $\sum \alpha_i y_i = 0$

$$L(\alpha) = \sum_{i=1}^{n} \alpha_i - \tfrac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j \left(x_i * x_j \right)$$

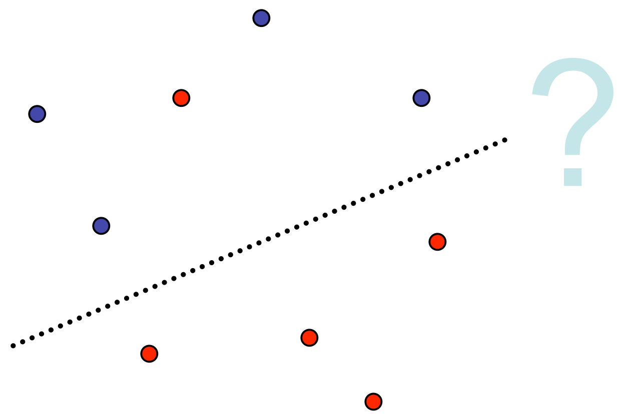
- Für jedes Beispiel gibt es ein α in der Lösung.
  - $0 = \alpha_i$ heißt, dass das Beispiel x_i im passenden Halbraum liegt.
  - $0 < \alpha_i$ heißt, dass das Beispiel x_i auf H1 oder H2 liegt (Stützvektor).
- Es gilt $w = \sum \alpha_i y_i x_i$,
  - Also $f(x) = \sum \alpha_i y_i (x_i * x) + b$
  - Also ist der beste Normalenvektor w eine Linearkombination von Stützvektoren ($\alpha_i \neq 0$).

Was wissen wir jetzt?

- Maximieren des Margins einer Hyperebene ergibt eine eindeutige Festlegung der optimalen trennenden Hyperebene.
- Dazu minimieren wir die Länge des Normalenvektors w.
  - Formulierung als Lagrange-Funktion
  - Formulierung als duales Optimierungsproblem
- Das Lernergebnis ist eine Linearkombination von Stützvektoren.
- Mit den Beispielen müssen wir nur noch das Skalarprodukt rechnen.

Nicht linear trennbare Daten

- In der Praxis sind linear trennbare Daten selten.
- 1. Ansatz: Entferne eine minimale Menge von Datenpunkten, so dass die Daten linear trennbar werden (minimale Fehlklassifikation).
- Problem: Algorithmus wird exponentiell.

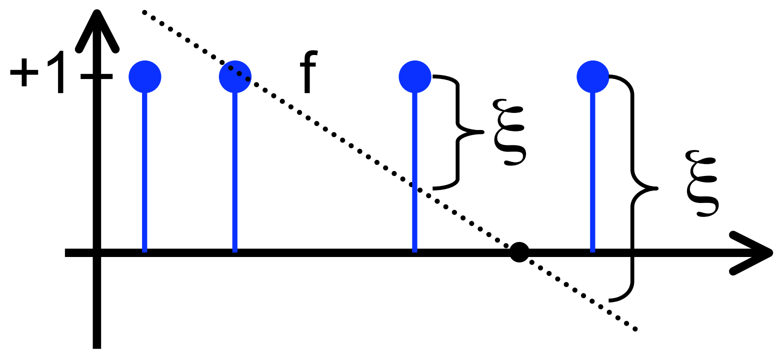


Weich trennende Hyperebene

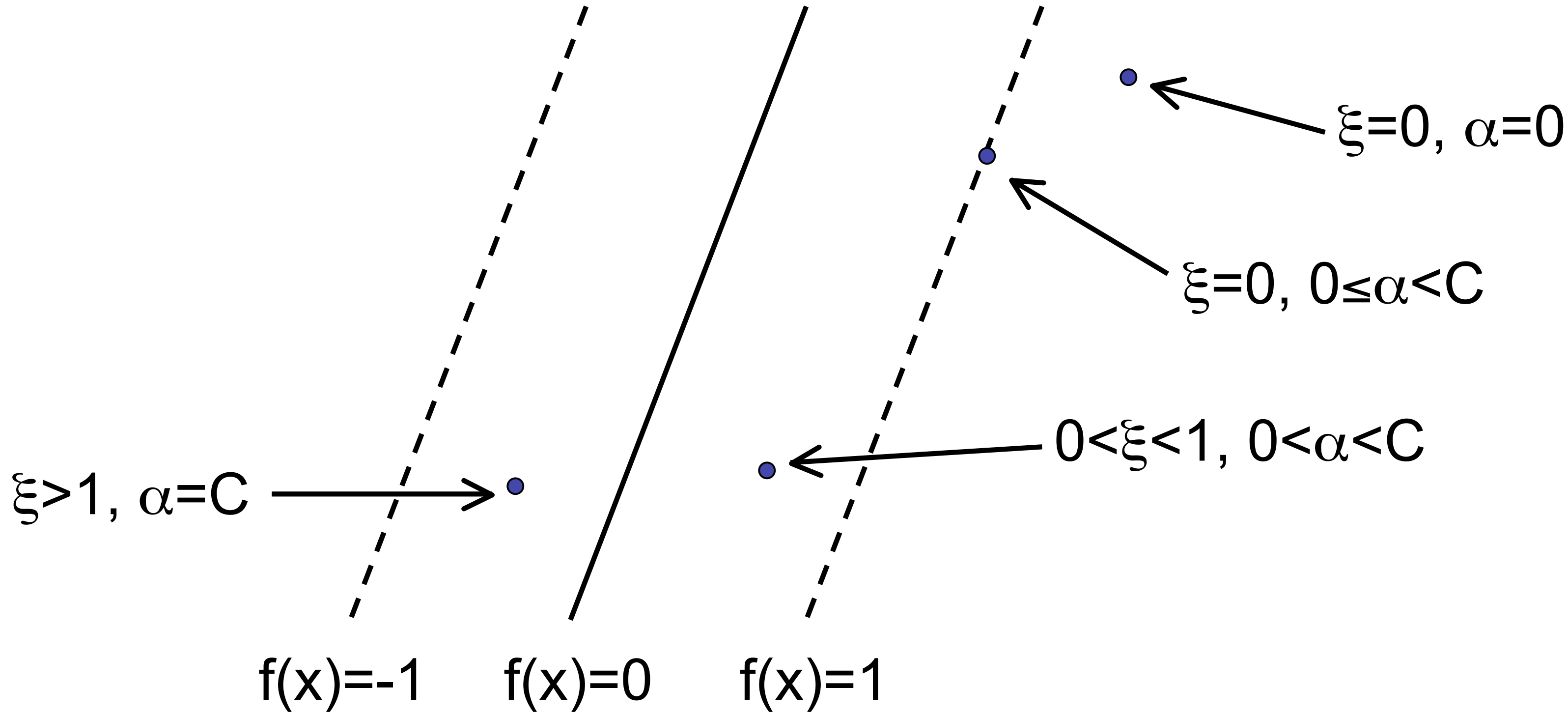
- Wähle $C \in \Re_{>0}$ und minimiere $\|w\|^2 + C\sum_{i=1}^{n} \xi_i$
- so dass für alle i gilt:

  $f(x_i) = w*x_i+b \geq 1-\xi_i$ für $y_i = 1$ und

  $f(x_i) = w*x_i+b \leq -1+\xi_i$ für $y_i = -1$
- Äquivalent: $y_i*f(x_i) \geq 1-\xi_i$



Bedeutung von ξ und α



Beispiele x_i mit $\alpha_i > 0$ heißen Stützvektoren $\Rightarrow$ SVM