# Feature Funktionen

Eine unbekannte Klassifizierung $f : X \to \{0, 1\}$ sei zu erlernen.

(1) Der Anwender hat stets die Möglichkeit, die Beispiele durch Zusatzinformationen (Features) aufzubereiten.

- Eine Funktion
$$\phi : X \to \mathbb{R}^N$$
heißt eine **Feature-Funktion**.
- $\phi(x) = (\phi_1(x), \ldots, \phi_N(x))$ ist der **Feature-Vektor** des Beispiels $x$ und $\phi(X) = \{\phi(x) \mid x \in X\}$ ist der **Feature-Raum**.

(2) Statt dem Lernalgorithmus klassifizierte Beispiele $x_1, \ldots x_s$ zu liefern, werden die Beispiele $\phi(x_1), \ldots, \phi(x_s)$ geliefert.

Die große Gefahr: Der Feature-Raum $\phi(X)$ ist unter Umständen hochdimensional: In der Trainingsphase wird auswendig gelernt ohne gute Verallgemeinerungen zu erlauben (Overfitting).

# Feature-Funktionen: Der Bag-of-Words Ansatz

In der Textklassifizierung werden häufig Stammsilben (sinntragende Silben) eingesetzt.

- Für eine Stammsilbe $s$ und ein Dokument $x \in D$ aus einer Menge $D$ von Dokumenten sei
  - $h_s(x)$ die Häufigkeit der Stammsilbe $s$ im Dokument $x$,
  - $D(s)$ die Anzahl untersuchter Dokumente mit mindestens einem Auftreten der Stammsilbe $s$ und
  - $\kappa(x)$ eine Normalisierungskonstante.
- Für ein Dokument $x$ ist $\phi(x) = (\phi_s(x) \mid s)$ ein Feature-Vektor mit

$$\phi_s(x) = \frac{h_s(x) \cdot \log_2(\frac{|D|}{|D(s)|})}{\kappa(x)}$$

$\phi_s(x)$ gewichtet die Häufigkeit der Stammsilbe $s$ im Dokument $x$ mit dem Informationsgehalt $\log_2(\frac{|D|}{|D(s)|})$ der Stammsilbe.

# Feature-Funktionen: Interpolation

Wir erhalten Punkte $x \in X \subseteq \mathbb{R}^3$, die Nullstellen eines unbekannten Polynoms vom Grad höchstens zwei sind.

- Wir wählen die Feature-Funktion

$$\phi(x_1, x_2, x_3) = (x_1^2, x_2^2, x_3^2, x_1 \cdot x_2, x_1 \cdot x_3, x_2 \cdot x_3, x_1, x_2, x_3, 1).$$

- Wir suchen also nach Koeffizienten $c_i$, so dass

$$(c_1 \cdot x_1^2 + c_2 \cdot x_2^2 + c_3 \cdot x_3^2) + (c_4 \cdot x_1 \cdot x_2 + c_5 \cdot x_1 \cdot x_3 + c_6 \cdot x_2 \cdot x_3) + (c_7 \cdot x_1 + c_8 \cdot x_2 + c_9 \cdot x_3) + c_0 = 0$$

  für alle $x \in X$ gilt.
- Das unbekannte Polynom wird zu einer linearen Funktion, die durch ein lineares Gleichungssystem bestimmt werden kann.

# Die Grundidee: Klassifikation mit großem Margin

Bestimme eine trennende Hyperebene, die die positiven von den negativen Beispielen mit **möglichst großem Margin** trennt.

(1) Der Perzeptron-Algorithmus benötigt höchstens

$$(\frac{2R}{\gamma})^2$$

Gegenbeispiele, wenn alle Beispiele zur Kugel um 0 mit Radius $R$ gehören.

- Wenige Gegenbeispiele bei entsprechend großem Margin $\gamma$.
- Abhängigkeit von der Dimension des Feature Raums nur indirekt, nämlich über den Margin $\gamma$ und den Radius $R$!

(2) Kein Overfitting für relativ kleines $R$ und relativ großes $\gamma$?

# Der Kernel-Trick I

Was passiert im Beispielraum, wenn wir die positiven von den negativen Beispielen durch eine lineare Funktion

$$f(x) = \langle c, \phi(x) \rangle + b,$$

im Feature-Raum trennen?

Der Perzeptron Algorithmus bestimmt $c = \sum_{i=1}^{s} \alpha_i \cdot \phi(x_i)$ als Linearkombination der Gegenbeispiele $x_1, \ldots, x_s$. Dann ist

$$\begin{aligned} f(x) &= \langle c, \phi(x) \rangle + b = \langle \sum_{i=1}^{s} \alpha_i \cdot \phi(x_i), \phi(x) \rangle + b \\ &= \sum_{i=1}^{s} \alpha_i \cdot \langle \phi(x_i), \phi(x) \rangle + b. \end{aligned}$$

Wir definieren den **Kern** $K$ durch $K(x, z) = \langle \phi(x), \phi(z) \rangle$.

# Der Kernel-Trick II

(1) Wir wissen, dass $f(x) = \sum_{i=1}^{s} \alpha_i \cdot \langle \phi(x_i), \phi(x) \rangle + b$ und $K(x, z) = \langle \phi(x), \phi(z) \rangle$. Deshalb ist

$$f(x) = \sum_{i=1}^{s} \alpha_i \cdot K(x_i, x) + b.$$

(2) Eine schnelle Berechnung von $f(x)$ gelingt im Beispielraum, falls
- die Anzahl $s$ der Beispiele nicht zu groß ist und
- $K(x_i, x)$ schnell berechnet werden kann.

(3) Welche Funktionen $K$ kommen in Frage?
- Offensichtlich ist $K(x, z) = \langle x, z \rangle$ eine Möglichkeit.
- Später: **polynomielle Kerne** $K(x, z) = (\langle x, z \rangle)^d$,

  der **Gauß Kern** $K(x, z) = e^{-\frac{||x-z||^2}{2\sigma^2}}$ und viele weitere Funktionen.

Im Beispielraum trennen wir also unter Umständen mit komplexen Hyperflächen $\{x \mid \sum_{i=1}^{s} \alpha_i \cdot K(x_i, x) = -b\}$.

# Die Rolle der Feature-Vektoren

Das Ziel: Bestimme eine lineare Trennung der positiven und negativen Beispiele mit großem Margin.

(1) Wähle zuerst eine **problem-spezifische** Feature-Funktion $\phi$, um eine Trennung der positiven von den negativen Beispielen vorzubereiten.

(2) Wähle einen passenden Kern $K$ aus, um eine Trennung mit großem Margin zu erhalten.

- Der Kern wird durch eine Feature-Funktion $\psi$ definiert, die im Allgemeinen sehr viele Features besitzt: Für den Gauß-Kern $K(x,z) = e^{-\frac{||x-z||^2}{2\sigma^2}}$ werden unendliche viele Features benötigt!
- $\psi \circ \phi$ ist die endgültige Feature Funktion.

Für die Anwendungen sind nur $\phi$ und $K$ von Interesse.

# Support-Vektor Maschinen: Das Verfahren I

(1) Für den Beispielraum $X$ bestimme eine Feature-Funktion $\phi : X \to \mathbb{R}$.

(2) Fordere genügend viele klassifizierte Beispiele $(\phi(x_1), b_1), \ldots, (\phi(x_s), b_s)$ an.

(3) Trenne positive und negative Beispiele $y_i = \phi(x_i)$ mit möglichst großem Margin:

- Wenn die lineare Funktion $\langle w, y \rangle + t$ mit $||w|| = 1$ den Margin $\gamma$ erreicht, dann ist $b_i \cdot ( \langle w, y_i \rangle + t ) \geq \gamma$ für alle $i$.
- Deshalb löse das Minimierungsproblem

$$\text{maximiere}_{w,t} \quad \gamma \quad \text{so dass } ||w|| = 1 \text{ und für jedes } i, (1 \leq i \leq s): \quad b_i \cdot ( \langle w, \phi(x_i) \rangle + t ) \geq \gamma.$$

# Support Vektor Maschinen: Das Verfahren II

Das Optimierungsproblem hat nur lineare Nebenbedingungen bis auf die Bedingung $||w|| = 1$.

- Wir können die Bedingung $||w|| = 1$ durch $||w|| \leq 1$ ersetzen: Die Bedingung ist zumindest konvex.
- Statt den Margin $\gamma$ unter der Nebenbedingung $||w|| \leq 1$ zu maximieren, minimiere $||w||$ unter der Nebenbedingung $\gamma \geq 1$:

$$\text{minimiere}_{w,t} \quad ||w||^2 \quad \text{so dass für jedes } i \ (1 \leq i \leq s): \quad b_i \cdot ( \langle w, \phi(x_i) \rangle + t ) \geq 1.$$

Das Optimierungsproblem ist gutartig: Eine konvexe quadratische Form ist unter linearen Nebenbedingungen zu minimieren.

- Aber das Optimierungsproblem „spielt“ im hochdimensionalen Feature-Raum!

# Support-Vektor Maschinen: Soft Margin

Wir haben bisher die vollständige lineare Trennbarkeit, den sogenannten **Hard Margin** Fall angenommen.
Im realistischerem **Soft Margin** Fall erlauben wir, dass eine wenige Beispiele nicht trennbar sind.

Füge Slack Variablen $\xi_i$ für jedes Beispiel hinzu und löse

$$\text{minimiere}_{w,t,\xi} \quad ||w||^2 + C \cdot \sum_{i=1}^{m} \xi_i^2 \tag{1}$$
$$\text{so dass für jedes } i\text{:} \quad b_i \cdot ( \langle w, \phi(x_i) \rangle + t ) \geq 1 - \xi_i.$$

- Je größer $\xi_i$ umso schlechter die Klassifizierung von $\phi(x_i)$.
- Der Parameter $C$ definiert wie stark falsche Klassifizierungen bestraft werden.

# Support-Vektor Maschinen: Die zentralen Fragen

- Wir nehmen an, dass die Trainingsmenge durch eine freundliche Verteilung $D$ bestimmt wird
- und dass deshalb eine trennende Hyperebene mit großem Margin $\gamma$ existiert.

(1) Welche Aussagen können wir für den Lernerfolg machen? Können wir wieder ein Gütesiegel vergeben, wenn wir wissen, dass $\gamma$ groß ist?

(2) Sind denn Hyperebenen nicht zu schwach, um Beispiele mit komplexer Klassifikation zu trennen?

> Die Hyperebene im Feature-Raum übersetzt sich in eine möglicherweise sehr komplexe „Hyperfläche“ im Beispielraum!.

(3) Bei sehr vielen Features ist die Lösung der Optimierungsprobleme sehr zeitaufwändig. Wann gelingt eine schnelle Berechnung direkt, in Umgehung des Feature-Raums?

# Text Klassifikation: Der Bag-of-Words Ansatz

Zwei Anwendungen:

- Klassifiziere Nachrichten einer Nachrichtenagentur in verschiedene Kategorien.
- Ordne medizinische Dokumente einer von 23 Krankheiten zu.

(1) Der Bag-of-Words Ansatz besitzt für jede Stammsilbe $s$ und jedes Dokument $x$ das Feature

$$\phi_s(x) = \frac{h_s(x) \cdot \log_2(\frac{|D|}{|D(s)|})}{\kappa(x)}$$

(2) Das innere Produkt

$$\langle \phi(x), \phi(z) \rangle = \sum_s \phi_s(x) \cdot \phi_s(z)$$

wird benutzt. Auch polynomielle Kerne und der Gauß Kern erreichen eine ähnliche Leistung:

Die Wahl des Kerns ist nicht entscheidend.

# Experimentbedingungen

(1) Nachrichtenklassifikation:
- Training mit 9603 Nachrichten der Agentur Reuters, Evaluierung mit 3299 Nachrichten.
- Durchschnittliche Nachrichtenlänge ungefähr 200 Worte.
- Die 10,000 Stammsilben mit gröstem Informationsgehalt (auf der Trainingsmenge) werden ausgewählt.

(2) Krankheitszuodnung:
- 10,000 medizinische Dokumente werden zum Training und 10,000 Dokumente zur Evaluierung benutzt.
- 15,561 Stammsilben werden ausgewählt, wobei jede Stammsilbe in mindestens drei Dokumenten vorkommt.

In beiden Fällen ist die Lernleistung konventionellen Ansätzen (Bayes-Verfahren, Rocchio, C4.5 und $k$-nearest Neighbor) überlegen.

# Bilderkennung I

100 Objekte werden aus 72 verschiedenen Blickwinkeln aufgenommen. Jedes Bild ist einem der Objekte zuzuordnen.

- Die Feature Funktion $\phi$:
  - Die 7200 Bilder wurden, nach Durchschnittsbildung auf $4 \times 4$ Gittern, von einer Auflösung von $128 \times 128$ Pixel auf $32 \times 32$ Pixel reduziert.
  - Die Feature-Funktion weist jedem Bild also einen Vektor von 1024 Graustufen zu.
- Die Objekte sind aufgrund der großen Zahl der Features bereits mit dem linearen Kern trennbar.

# Bilderkennung II

1400 Photos der Corel Stock Photo Collection sind verschiedenen Kategorien zuzuweisen.
2/3 der Photos werden zum Training und 1/3 zum Test verwandt.

(1) Die Feature Funktion:
- Jedem Pixel wird sein HSV-Wert (Hue Saturation Value) zugewiesen. HSV-Werte sind „biologisch plausibler“ als RGB-Werte, da auch Helligkeitsinformation benutzt wird.
- Der 3-dimensionale Würfel der HSV-Werte wird in 4096 Farbbereiche zerlegt.
- Jedem Bild wird ein Histogramm als Feature-Vektor zugeodnet, das die Häufigkeit für jeden Farbbereich wiedergibt.
- Der Feature-Vektor ist translations-invariant.

(2) Der Kern $K(x, z) = \exp^{-||x-z||_1}$ mit $||x - z||_1 = \sum_i |x_i - y_i|$ wird erfolgreich eingesetzt.
- Die Wahl des Kerns ist kritisch.

# Handschrifterkennung

- In einem Benchmark-Datensatz des US Postal Service für die Zifferenerkennung treten 7291 Trainings- und 2007 Testbeispiele auf.
- Jede Ziffer des Datensatzes wird als $16 \times 16$ Pixelmatrizen mit 256 Graustufen repräsentiert.

(1) Die Feature-Funktion: Es ist $\phi(x) = x$. Die Features entsprechen somit den Grauwerten der einzelnen Pixel.

(2) Es werden polynomielle Kerne und der Gauß Kern benutzt:
   - Für polynomielle Kerne ergibt sich lineare Separierbarkeit erst für den Grad $d \geq 3$.

(3) Der Ansatz der Support-Vektormaschinen wird somit ohne problem-spezifische Aufbereitung eingesetzt.

Die Lernleistung ist nur wenig schwächer als die der besten maßgeschneiderten Methoden.

# Das Hidden Markoff Modell

Ein Hidden Markov Modell (HMM) wird beschrieben durch

(1) ein Alphabet $\Sigma$,

(2) die Zustandsmenge $Q$,

(3) den Anfangszustand $q_0 \in Q$,

(4) die stochastische Matrix $P$ der Zustandsübergänge und

(5) Emmissionswahrscheinlichkeiten $e_q(a)$ für einen Zustand $q \neq q_0$ und einen Buchstaben $a \in \Sigma$.

- Eine MM $M = (\Sigma, Q, q_0, P, e)$ und eine Sequenz $u$ sind gegeben. Was ist die Wahrscheinlichkeit $p_M[\ u\ ]$, dass $M$ $u$ erzeugt?
- Bestimme die Wahrscheinlichkeit $p_M[\ W, u\ ]$, dass die Zustandsfolge $W = (q_0, \ldots, q_n)$ durchlaufen und $u$ erzeugt wird.

$$p_M[\ W, u\ ] = \Pi_{i=1}^{n} P[q_{i-1}, q_i] \cdot e_{q_i}(u_i).$$

# Der Feature-Vektor

Für die Wahrscheinlichkeit $p_M[\, u \,]$ der Erzeugung von $u$ gilt

$$p_M[\, u \,] = \sum_W p_M[\, W, u \,].$$

(1) Angenommen, wir haben eine gute HMM $M$ für die Protein Superfamilie $\{u_1, \ldots, u_k\}$: $M$ ist gut, wenn

$$\Pi_{i=1}^{k} p_M[\, u_i \,]$$

fast größtmöglich ist.

(2) Sei $\Theta_M$ der Vektor, der den Emissions- und Übergangswahrscheinlichkeiten entspricht.

(3) Die Sequenz $u$ erhält den Gradienten

$$\nabla \log p_M[\, u \,]$$

als Feature-Vektor.

# Der Kern

- Wenn $M$ die Sequenzen $u_1, \ldots, u_k$ gut modelliert, dann ist $\nabla \log p_M[\, u_i \,] \approx 0$:

  Sonst führt Gradientenaufstieg zu einem besseren Modell.
- Hoffnung: Wenn $u$ zur Superfamilie gehört, dann unterscheidet sich der Gradient von $u$ nicht stark von den Gradienten der Superfamilie.

- Der Gauß Kern wird benutzt:

$$K(u, v) = \exp^{-\frac{||\nabla \log p_M[\, u \,] - \nabla \log p_M[\, v \,]||_2^2}{2 \cdot \sigma^2}}$$

- Positive Beispiele waren Sequenzen einer Superfamilie, negative Beispiele waren Sequenzen anderer Superfamilien.

Die Klassifizierungsleistung des zugrunde liegenden HMM's wird wesentlich verbessert.

# Kern(-Funktionen)

Ein Kern (Englisch: Kernel) für eine Menge $X$ ist eine Funktion $K : X^2 \to \mathbb{R}$, so dass

$$K(x, z) = \langle \phi(x), \phi(z) \rangle$$

für eine Funktion $\phi : X \to \mathbb{R}^N$ und für ein inneres Produkt $\langle -, - \rangle$.

Welche Eigenschaften sollte ein Kern besitzen?
Setze $\alpha(x, z) = \frac{K(x,z)}{\sqrt{K(x,x)} \cdot \sqrt{K(z,z)}}$. Dann

$$\begin{aligned} \alpha(x, z) &= \frac{K(x, z)}{\sqrt{K(x, x)} \cdot \sqrt{K(z, z)}} = \frac{\langle \phi(x), \phi(z) \rangle}{\sqrt{\langle \phi(x), \phi(x) \rangle} \cdot \sqrt{\langle \phi(z), \phi(z) \rangle}} \\ &= \langle \frac{\phi(x)}{||\phi(x)||}, \frac{\phi(z)}{||\phi(z)||} \rangle = \text{Kosinus}(\phi(x), \phi(z)) \in [-1, 1]. \end{aligned}$$

# Kerne als Ähnlichkeitsmaße

Es ist $\frac{K(x,z)}{\sqrt{K(x,x)}\cdot\sqrt{K(z,z)}} = \text{Kosinus}(\phi(x),\phi(z)) \in [-1,1]$.

- Die Vektoren $\phi(x)$ und $\phi(z)$ (und damit die Beispiele $x$ und $z$) sind „ähnlich“, wenn $\alpha(x,z)$ nahe bei Eins liegt.
- Positive und negative Beispiele sollten jeweils untereinander ähnlich sein, um eine lineare Trennung zu erlauben.

Nach Skalierung wird ein Kern zu einem Ähnlichkeitsmaß, das Ähnlichkeit jeweils unter den positiven wie auch den negativen Beispielen „herstellen“ sollte.

# Nützliches aus der linearen Algebra

## Orthogonalisierung

- Die Transponierte $U^T$ einer orthogonalen Matrix $U$ ist die zu $U$ inverse Matrix. Es gilt also $U \cdot U^T = E$ für die Einheitsmatrix $E$.
- Für jede symmetrische Matrix $A$ gibt es eine orthogonale Matrix $U$ und eine Diagonalmatrix $D$ mit

$$U^T \cdot A \cdot U = D.$$

## Positiv semi-definite Matrizen

Eine symmetrische quadratische Matrix $A$ heißt positiv semi-definit, wenn für alle Vektoren $x$

$$x^T \cdot B \cdot x \geq 0.$$

Die Matrix $B$ ist genau dann positiv semi-definit, wenn alle Eigenwerte von $B$ nicht-negativ sind.

# Der Satz von Mercer I

**Wann ist eine Funktion $K : X^2 \to \mathbb{R}$ ein Kern?**

(1) Die Funktion $K$ sei symmetrisch und $Y \subseteq X$ sei endlich. Definiere die Matrix $M_{K,Y}$ durch $M_{K,Y}[y_1, y_2] = K(y_1, y_2)$ für $y_1, y_2 \in Y$.

(2) $M_{K,Y}$ ist eine symmetrische Matrix. Also gilt $U^T \cdot M_{K,Y} \cdot U = D$ für eine orthogonale Matrix $U$.

Es sei $U = [u_1, \ldots, u_n]$ und $D[i, i] = \lambda_i$. Dann $M_{K,Y} \cdot u_i = \lambda_i \cdot u_i$.

(3) Angenommen, $Y = X$ und $M_{K,X}$ ist positiv semi-definit: Definiere die Feature Funktion $\phi(x_i) = (\sqrt{\lambda_j} \cdot u_{i,j} \mid j)$. Dann

$$\langle \phi(x_i), \phi(x_k) \rangle = \sum_j \lambda_j u_{i,j} \cdot u_{k,j} = (U \cdot D \cdot U^T)_{i,k} = M_{K,X}[x_i, x_k].$$

(4) Aber $M_{K,X}[x_i, x_k] = K(x_i, x_k)$ und $K$ ist ein Kern.

# Der Satz von Mercer II

Die Menge $X$ sei endlich und $K$ sei symmetrisch. Wir behaupten:
$K : X \times X \to \mathbb{R}$ ist ein Kern $\Leftrightarrow$ $M_{K,X}$ ist positiv semi-definit.

(1) Wenn $M_{K,X}$ positiv semi-definit ist, dann ist $M_{K,X}$ ein Kern.

(2) Angenommen, $K$ ist ein Kern, also $K(x,z) = \langle \phi(x), \phi(z) \rangle$

- Wenn ein Eigenwert $\lambda_i$ negativ ist, setze $z = \sum_j u_{i,j} \cdot \phi(x_j)$:

$$\begin{aligned} \langle z, z \rangle &= \langle \sum_j u_{i,j} \cdot \phi(x_j), \sum_k u_{i,k} \cdot \phi(x_k) \rangle \\ &= \sum_{j,k} u_{i,j} \cdot M_{K,X}(x_j, x_k) \cdot u_{i,k} = u_i^T \cdot M_{K,X} \cdot u_i = \lambda_i < 0. \end{aligned}$$

- Wir erhalten somit einen Widerspruch zur Definition des inneren Produkts. Also ist jeder Eigenwert nicht-negativ und $M_{K,X}$ ist positiv semi-definit.

# Der Satz von Mercer III

Und wenn $X$ nicht endlich ist?

Sei $X$ eine kompakte Teilmenge des $\mathbb{R}^n$. Die Funktion $K : X^2 \to \mathbb{R}$ sei stetig und symmetrisch. Dann gilt

$$K \text{ ist ein Kern} \quad \Leftrightarrow \quad M_{K,Y} \text{ ist für jede endliche Teilmenge } Y \text{ von } X \text{ positiv semi-definit.}$$

**Konsequenz**: Angenommen, $K_1$ und $K_2$ sind Kerne. Dann ist auch $K = K_1 + K_2$ ein Kern.

Warum? Wir wenden den Satz von Mercer an. Sei $Y$ eine endliche Teilmenge von $X$. Es ist

$$\alpha^T \cdot M_{K,Y} \cdot \alpha = \alpha^T \cdot M_{K_1,Y} \cdot \alpha + \alpha^T \cdot M_{K_2,Y} \cdot \alpha \geq 0$$

und das war zu zeigen.

# Weitere Konsequenzen

Angenommen, $K_1$ und $K_2$ sind Kerne und $\alpha \geq 0$. Dann sind auch $K_1 + K_2, \alpha \cdot K_1$ und $K_1 \cdot K_2$ Kerne.

# Und noch mehr Kerne

(1) Sei $K$ ein Kern und sei $p$ ein Polynom mit nicht-negativen Koeffizienten. Dann ist auch $p(K)$ ist ein Kern.

(2) $f : \mathbb{R} \to \mathbb{R}$ sei durch eine Potenzreihe mit nicht-negativen Koeffizienten darstellbar. Dann ist auch $f(K)$ ein Kern.

(1) Kerne sind abgeschlossen unter Polynomen mit nicht-negativen Koeffizienten, da Kerne unter Addition und Multiplikation abgeschlossen sind.

(2) $f = \lim_{n\to\infty} p_n$ für Polynome $p_n$.

- Sei $Y$ eine beliebige endliche Teilmenge von $X$: Die Matrizen $M_{p_n(K),Y}$ sind positiv semi-definit.
- $M_{f(K),Y} = \lim_{n\to\infty} M_{p_n(K),Y}$ und $M_{f(K),Y}$ ist positiv semi-definit, da

$$x^T \cdot M_{f(K),Y} \cdot x = x^T \cdot (\lim_{n\to\infty} M_{p_n(K),Y}) \cdot x = \lim_{n\to\infty} x^T \cdot M_{p_n(K),Y} \cdot x \geq 0.$$

# „Basis Kerne“

Wir wissen jetzt, wie man aus bestehenden Kerne weitere Kerne baut.
Wir müssen diesen Prozess mit bestehenden Kernen starten!

- Sei $f : X \to \mathbb{R}$ eine beliebige Funktion. Dann ist

$$K(x, z) = f(x) \cdot f(z)$$

  ein Kern: Benutze $f$ als Feature-Funktion.
- Sei $A$ eine symmetrische, positiv semi-definite Matrix. Dann ist

$$K(x, z) = x^T \cdot A \cdot z \text{ ein Kern.}$$

  - Diagonalisiere $A$: Es ist $A = U^T \cdot D \cdot U$.
  - Setze $B = \sqrt{D} \cdot U$. Als Konsequenz

$$K(x, z) = x^T \cdot A \cdot z = x^T \cdot U^T \cdot \sqrt{D} \cdot \sqrt{D} \cdot U \cdot z = x^T \cdot B^T \cdot B \cdot z = \langle B \cdot x, B \cdot z \rangle.$$

# Der Gauß Kern

(1) Sei $K$ ein Kern. Dann ist auch $\exp^{K(x,z)}$ ein Kern.

(2) $\exp^{-\frac{||x-z||^2}{\sigma^2}}$ ist ein Kern und wird der Gauß Kern genannt.

(1) Die Expoentialfunktion hat die Potenzreihe $\sum_{i=0}^{\infty} \frac{x^n}{n!}$. Alle Koeffizienten sind nicht-negativ und deshalb ist $\exp^{K(x,z)}$ ein Kern.

(2) Es ist

$$\exp^{-\frac{||x-z||^2}{\sigma^2}} = (\exp^{-\frac{||x||^2}{\sigma^2}} \cdot \exp^{-\frac{||z||^2}{\sigma^2}}) \cdot \exp^{\frac{2\cdot\langle x,z\rangle}{\sigma^2}} .$$

- Das erste Produkt ist ein Kern. Warum?
- Der dritte Faktor ist nach (1) ein Kern und
- der Gauß Kern ist als Produkt von Kernen ein Kern.

# Fat-Shattering Dimension

Viele Features und Overfitting (Auswendiglernen ohne die Fähigkeit der Verallgemeinerung) droht.
Was passiert, wenn wir eine Hypothese mit großem Margin finden?

(1) Wir haben uns an der VC-Dimension einer Hypothesenklasse $\mathcal{H}$ orientiert, um eine ausreichende Beispielzahl für PAC-Algorithmen zu bestimmen.
   - $\mathrm{VC}(\mathcal{H})$ war die Größe der größten von $\mathcal{H}$ zertrümmerten Menge.

(2) Wir brauchen so etwas wie die Anzahl der Freiheitsgrade aller „Hypothesen mit großem Margin“.
   - Statt eine Menge $S$ zu zertrümmern, sollten wir fordern, dass $S$ mit großem Margin zertrümmert wird.

Wir werden auf die Fat-Shattering Dimension geführt.

# Die Fat-Shattering Dimension I

Sei $\mathcal{H}$ eine Menge von Funktionen $h : X \to \mathbb{R}$ und $\gamma \in \mathbb{R}$.

(1) $S = \{y_1, \ldots, y_s\} \subseteq X$ wird durch $\mathcal{H}$ $\gamma$-zertrümmert, falls:
es gibt $r \in \mathbb{R}^m$, so dass es für jeden Vektor $b \in \{-1, 1\}^m$ eine Funktion $h_b \in \mathcal{H}$ gibt mit

$$h_b(y_i) \begin{cases} \geq r_i + \gamma & \text{falls } b_i = 1, \\ \leq r_i - \gamma & \text{falls } b_i = -1. \end{cases}$$

(2) Die Fat-Shattering Dimension von $\mathcal{H}$ stimmt überein mit

$$\mathrm{fat}_{\mathcal{H}}(\gamma) = \max\{\ S \subseteq X \mid \mathcal{H}\ \gamma\text{-zertrümmert } S\ \}.$$

# Die Fat-Shattering Dimension II

Angenommen, $S = \{y_1, \ldots, y_m\}$ wird $\gamma$-zertrümmert.

- Die Vektoren $b \in \{-1, 1\}^m$ entsprechen allen möglichen Teilmengen von $S$.
- Für jeden Vektor $b$ muss $S$ „gemäß $b$ und Margin $\gamma$“ zertrümmert werden.

(1) Kann die Fat-Shattering Dimension eine ausreichende Beispielzahl ähnlich wie die VC-Dimension garantieren? Genügen

$$s = O(\frac{1}{\varepsilon} \cdot (\ln(\frac{1}{\varepsilon}) \cdot \mathrm{fat}_{\mathcal{H}}(\gamma) + \ln(\frac{1}{\delta}))$$

Beispiele für konsistente Hypothese mit großem Margin?

(2) Wenn ja, was ist die Fat-Shattering Dimension linearer Hypothesen?

# Eine ausreichende Beispielzahl

$\mathcal{H}$ sei eine Menge von Funktionen $h : \mathcal{B} \to [-1, 1]$ für eine Menge $\mathcal{B}$ von Beispielen. Weiterhin sei $\gamma \in (0, 1)$ und $d = \text{fat}_{\mathcal{H}}(\frac{\gamma}{8})$.
Dann gilt mit Wahrscheinlichkeit $1 - \delta$ (über die Auswahl einer Teilmenge $S \subseteq \mathcal{B}$ von $s$ Beispielen)

- für jede Verteilung $D$ auf der Menge $\mathcal{B}$ und
- für jede Hypothese $h \in \mathcal{H}$ mit Margin $m_S(h) \geq \gamma$ auf der Beispielmenge $S$, dass

$$\text{fehler}_D(\text{sign}(h)) = O\left( \frac{1}{s} \cdot ( d \cdot \log(\frac{s}{d \cdot \gamma}) \cdot \log(\frac{s}{\gamma^2}) + \log(\frac{1}{\delta}) ) \right),$$

solange $s \geq c \cdot d$ für eine hinreichend grosse Konstante $c$.

# Interpretation

Wenn wir logarithmische Faktoren vernachlässigen folgt

$$\text{fehler}_D(\text{sign}(h)) = O\left( \frac{1}{s} \cdot (\text{fat}_{\mathcal{H}}(\frac{\gamma}{8}) + \log(\frac{1}{\delta}) ) \right).$$

(1) Oder, wenn wir die Beispielzahl $s$ in Abhängigkeit vom Fehler $\varepsilon$, der Fat-Shattering Dimension und $\delta$ ausdrücken:

$$s = O\left( \frac{1}{\varepsilon} \cdot (\text{fat}_{\mathcal{H}}(\frac{\gamma}{8}) + \log(\frac{1}{\delta}) ) \right).$$

(2) Das Argument verläuft ähnlich wie im Fall der VC-Dimension.

Wesentlicher Unterschied: Die Wahl der Hypothesenklasse hängt von der Beispielmenge $S$ ab, denn nur Hypothesen mit Margin mindestens $\gamma$ auf der Menge $S$ werden betrachtet.

# Lineare Thresholdfunktionen

$$\mathcal{L}_R = \{\ x \mapsto \langle w, x \rangle \mid ||w|| \leq 1 \text{ und } ||x|| \leq R\ \}$$

beschreibt die Klasse aller Thresholdfunktionen für Argumente mit Norm höchsten $R$. Dann gilt

$$\text{fat}_{\mathcal{L}_R}(\gamma) \leq \left(\frac{R}{\gamma}\right)^2.$$

(1) Die Fat-Shattering Dimension hängt somit nur von der maximalen Norm $R$, nicht aber von der Dimension des Raums ab!

(2) Dann hängt aber auch die Beispielzahl nur von der maximalen Norm sowie vom erreichten Margin ab!

# Lineare Trennung mit großem Margin

## Erfolgreiches Hard-Margin Lernen bei großem Margin.

Jedes Beispiel aus einer Menge $\mathcal{B}$ möge eine Norm von höchstens $R$ besitzen. Ein beliebiges Konzept $c$ sei durch eine Klassifizierung $f : \mathcal{B} \to \{-1, 1\}$ gegeben.
Dann gilt mit Wahrscheinlichkeit mindestens $1 - \delta$ (über die Auswahl einer Menge $S \subseteq \mathcal{B}$ von $s$ Beispielen)

- für jede Verteilung $D$ auf der Menge $\mathcal{B}$ und
- für jede Thresholdfunktion $\text{sign}(h(y)) = \langle w, y \rangle + t$ mit $||w|| = 1$ und $\text{Margin}_{S,f}(h) \geq \gamma$, dass

$$\text{fehler}_D(c, h) = O\left( \frac{1}{s} \cdot \left( \frac{R^2}{\gamma^2} \cdot \log\left(\frac{s \cdot \gamma}{R^2}\right) \cdot \log\left(\frac{s}{\gamma^2}\right) + \log\left(\frac{1}{\delta}\right) \right) \right),$$

wobei $s \geq c \cdot \frac{R^2}{\gamma^2}$ für eine hinreichend große Konstante $c$ zu fordern ist.

# Interpretation

Wenn logarithmische Faktoren wieder vernachlässigt werden, folgt

$$\text{fehler}_D(c,h) = O\left(\frac{1}{s}\cdot\left(\frac{R^2}{\gamma^2}+\log\left(\frac{1}{\delta}\right)\right)\right),$$

beziehungsweise

$$s = O\left(\frac{1}{\varepsilon}\cdot\left(\frac{R^2}{\gamma^2}+\log\left(\frac{1}{\delta}\right)\right)\right).$$

(1) Nach $(2\cdot\frac{R}{\gamma})^2$ Gegenbeispielen hat der Perzeptron-Algorithmus erfolgreich gelernt. Support-Vektor Maschinen brauchen im Wesentlichen eine um den Faktor $\frac{1}{\varepsilon}$ größere Beispielzahl.

(2) Eine lineare Trennung mit großem Margin wird leichter bei großem Radius: Die Beispielzahl sollte mit $R$ wachsen.

# Lineare Thresholdfunktionen

Es ist $\mathcal{L}_R = \{\ x \mapsto \langle w, x\rangle \mid ||w|| \leq 1 \text{ und } ||x|| \leq R\ \}$.

Warum ist $\text{fat}_{\mathcal{L}}(\gamma) \leq \left(\frac{R}{\gamma}\right)^2$?

(1) Wir setzen $d = \text{fat}(\mathcal{L}_r)$ und wählen Beispiele $x_1, \ldots, x_d$ mit $||x_1||, \ldots, ||x_d|| \leq R$.

(2) Wenn $I = \{1, \ldots, d\}$, dann zeigen wir

$$2^{-d} \cdot \sum_{J \subseteq I} || \sum_{i \in J} x_i - \sum_{i \in I \setminus J} x_i||^2 \leq d \cdot R^2.$$

Insbesondere gibt es $J \subseteq I$ mit $|| \sum_{i \in J} x_i - \sum_{i \in I \setminus J} x_i|| \leq R \cdot \sqrt{d}$.

(3) Für jede Teilmenge $J \subseteq I$ zeigen wir

$$d \cdot \gamma \leq || \sum_{i \in J} x_i - \sum_{i \in I \setminus J} x_i||$$

(4) und $\sqrt{d} \leq R/\gamma$ folgt.

# Der Einfluß der Norm

Wir setzen $\alpha_i^J = 1$, wenn $i \in J$, und $\alpha_i^J = -1$, wenn $i \notin J$.

$$
\begin{aligned}
& 2^{-d} \cdot \sum_{J \subseteq I} || \sum_{i \in J} x_i - \sum_{i \in I \setminus J} x_i ||^2 = 2^{-d} \cdot \sum_{J \subseteq I} || \sum_{i=1}^{d} \alpha_i^J \cdot x_i ||^2 \\
= \quad & 2^{-d} \cdot \left( \sum_{J \subseteq I} \sum_{i=1}^{d} (\alpha_i^J)^2 \cdot ||x_i||^2 + 2 \cdot \sum_{J \subseteq I} \sum_{i \neq j} \alpha_i^J \alpha_j^J \cdot \langle x_i, x_j \rangle \right) \\
= \quad & 2^{-d} \cdot \left( \sum_{J \subseteq I} \sum_{i=1}^{d} ||x_i||^2 + 2 \cdot \sum_{i \neq j} \left[ \sum_{J \subseteq I} \alpha_i^J \alpha_j^J \cdot \langle x_i, x_j \rangle \right] \right) \\
= \quad & 2^{-d} \cdot \sum_{J \subseteq I} \sum_{i=1}^{d} ||x_i||^2 = \sum_{i=1}^{d} ||x_i||^2 \leq d \cdot R^2,
\end{aligned}
$$

denn $||x_i|| \leq R$.

# Der Einfluß des Margins I

- Angenommen, $\mathcal{L}_R$ $\gamma$-zertrümmert die Beispielmenge $\{x_1, \ldots, x_d\}$.
- Zeige $d \cdot \gamma \leq ||\sum_{i \in J} x_i - \sum_{i \in I \setminus J} x_i||$ für eine beliebige Teilmenge $J \subseteq I = \{1, \ldots, d\}$.

(1) Setze $b_i = 1$, wenn $i \in J$, und $b_i = -1$ sonst.

(2) Dann gibt es Schwellenwerte $r_1, \ldots, r_d \in \mathbb{R}$ und einen Gewichtsvektor $w_b$ mit

$$||w_b|| = 1 \text{ und } b_i \cdot (\langle w_b, x_i \rangle - r_i) \geq \gamma \text{ für } i = 1, \ldots, d.$$

(3) Als Konsequenz

$$\langle w_b, \sum_{i \in J} x_i \rangle \geq \sum_{i \in J} r_i + |J| \cdot \gamma \text{ und } \langle w_b, \sum_{i \in I \setminus J} x_i \rangle \leq \sum_{i \in I \setminus J} r_i - |I \setminus J| \cdot \gamma$$

# Der Einfluß des Margins II

## Fallannahme:

$\sum_{i,i\in J} r_i \geq \sum_{i,i\in I\setminus J} r_i$.

(4) Als Konsequenz von (3) und der Fallannahme:

$$\begin{aligned}\langle w_b, \sum_{i\in J} x_i - \sum_{i\in I\setminus J} x_i\rangle &\geq (\sum_{i\in J} r_i + |J|\cdot\gamma) - (\sum_{i\in I\setminus J} r_i - |I\setminus J|\cdot\gamma)\\ &\geq |I|\cdot\gamma = d\cdot\gamma.\end{aligned}$$

(5) Es ist $||w_b|| = 1$ und mit Cauchy-Schwartz ($\langle a,b\rangle \leq ||a||\cdot||b||$):

$$\langle w_b, \sum_{i\in J} x_i - \sum_{i\in I\setminus J} x_i\rangle \leq ||\sum_{i\in J} x_i - \sum_{i\in I\setminus J} x_i||.$$

- Als Konsequenz $d\cdot\gamma \leq ||\sum_{i\in J} x_i - \sum_{i\in I\setminus J} x_i||$.
- Analoge Argumentation im komplementären Fall $\sum_{i,i\in J} r_i < \sum_{i,i\in I\setminus J} r_i$.

# Soft Margin I

Was heißt es „im Wesentlichen“ einen großen Margin zu besitzen, obwohl sogar Missklassifikationen zugelassen werden?

(1) $\mathcal{H}$ sei eine Klasse von Funktionen $h : X \to \mathbb{R}$.

(2) Der Slack eines klassifizierten Beispiels $(x, b) \in X \times \{-1, 1\}$ bezüglich $h \in \mathcal{H}$ und dem Ziel-Margin $\gamma$ ist

$$\text{slack}(x, b, \gamma, h) = \max\{0, \gamma - b \cdot h(x)\}.$$

Wenn $\text{slack}(x, b, \gamma, h) > \gamma$, dann missklassifiziert $h$ Beispiel $x$.

(3) Der Slack-Vektor für $S = \{(x_1, b_1), \ldots, (x_m, b_m)\}$, $h \in \mathcal{H}$ und den Ziel-Margin $\gamma$ ist

$$\text{slack}(S, \gamma, h) = (\ \text{slack}(x_i, b_i, \gamma, h) \mid 1 \leq i \leq m\ ).$$

Wir sagen: $h$ erreicht auf der Beispielmenge $S$ „im wesentlichen“ den Margin $\gamma$, falls $\text{slack}(S, \gamma, h)$ ein Vektor mit kleiner Norm ist.

# Soft Margin II

Jedes Beispiel aus einer Menge $\mathcal{B}$ habe eine Norm höchstens $R$. Das Konzept $c$ entspreche der Klassifizierung $f : \mathcal{B} \to \{-1, 1\}$.
Dann gilt mit Wahrscheinlichkeit mindestens $1 - \delta$ (über die Auswahl einer Menge $S \subseteq \mathcal{B}$ von $s$ Beispielen)

- für jede Verteilung $D$ auf der Menge $\mathcal{B}$,
- für jede Thresholdfunktion $\text{sign}(h(y)) = sign(\langle w, y \rangle + t)$ mit $||w|| = 1$:

$$\text{fehler}_D(c, h) = O\left( \frac{1}{s} \cdot ( d \cdot \log(\frac{s}{d}) \cdot \log(s) + \log(\frac{s}{\delta}) ) \right),$$

falls $d = \frac{R^2 + ||\text{slack}(S,\gamma,h)||_2^2}{\gamma^2}$ und $s = \Omega(d)$.

# Zusammenfassung

## Was haben wir bisher erreicht?

(1) Wir haben große Klassen von Kernfunktionen konstruiert, die in vielen Anwendungen den Margin vergrößern.

(2) Wir wissen, dass ungefähr $s = \frac{1}{\varepsilon} \cdot \left( (\frac{R}{\gamma})^2 + \ln(\frac{1}{\delta}) \right)$ Beispiele im Hard Margin Fall genügen, wenn wir eine lineare Trennung mit Margin $\gamma$ erreicht haben.

(3) Im Soft Margin Fall genügen ungefähr $s = \frac{1}{\varepsilon} \cdot \left( (\frac{R+||\text{slack}(S,\gamma,h)||_2}{\gamma})^2 + \ln(\frac{1}{\delta}) \right)$ Beispiele.

## Was ist noch zu tun?

Wie können wir eine lineare Trennung mit großem Margin in einem hochdimensionalen Feature-Raum effizient berechnen?

- Der wichtige Gauß Kern benötigt unendlich viele Features!

# Konvexe Mengen und Funktionen

- $f : \mathbb{R}^N \to \mathbb{R}$ ist konvex, falls für alle $x, y \in \mathbb{R}^N$ und alle $\lambda \in [0, 1]$

$$f(\ \lambda \cdot x + (1 - \lambda) \cdot y)\ ) \leq \lambda \cdot f(x) + (1 - \lambda) \cdot f(y),$$

  falls also der Funktionsgraph stets unter der Sekante durch $(x, f(x))$ und $(y, f(y))$ liegt.
- Eine Menge $X \subseteq \mathbb{R}^N$ ist konvex, wenn für je zwei Punkte $x, y \in X$ die Gerade $\{\lambda \cdot x + (1 - \lambda) \cdot y \mid 0 \leq \lambda \leq 1\}$ von $x$ nach $y$ in $X$ liegt.

(1) $\sum_i w_i^2$ ist konvex:
   - Die Funktion $x \to x^2$ ist konvex
   - und eine Summe konvexer Funktionen ist konvex.

(2) Ein Durchschnitt konvexer Mengen ist konvex.

# Optimierungsprobleme

Das Ziel: $\text{minimiere}_{w,t} \ \ ||w||^2$

so dass für jedes $i$ $(1 \leq i \leq s)$: $b_i \cdot ( \langle w, y_i \rangle + t ) \geq 1$.

Stattdessen betrachten wir das allgemeine Optimierungsproblem

$$\text{minimiere}_x \ \ f(x), \ \ \text{so dass } h_j(x) \leq 0 \text{ für } j = 1, \ldots, m$$

$x$ heißt eine Lösung, falls $h_j(x) \leq 0$ für alle $j$.

Welche speziellen Eigenschaften hat unser Minimierungsproblem?

(1) Die Zielfunktion $||w||^2 = \sum_i w_i^2$ ist konvex.

(2) Alle Restriktionen sind lineare Ungleichungen.

(3) Eine konvexe Funktion ist über einer konvexen Menge, nämlich einem Durchschnitt von Halbräumen, zu minimieren.

# Konvexe Optimierungsprobleme

Ein Minimierungsproblem heißt konvex, wenn die Zielfunktion $f$ konvex ist und falls alle $h_j$ konvex sind.

(1) Damit ist eine konvexe Funktion über einer konvexen Menge zu minimieren.

(2) In diesem Fall ist jedes lokale Minimum $x$ auch ein globales Minimum:

- Ist $x$ kein globales Minimum, dann gibt es eine Lösung $y$ mit $f(y) < f(x)$.
- Für die $x$ und $y$ verbindende Gerade gilt für $\lambda < 1$

$$f(\ \lambda \cdot x + (1 - \lambda) \cdot y)\ ) \leq \lambda \cdot f(x) + (1 - \lambda) \cdot f(y) < f(x)$$

und $x$ ist kein lokales Minimum.

# Das duale Problem

Das Minimierungsproblem

$$\text{minimiere}_x \;\; f(x), \;\; \text{so dass } h_j(x) \leq 0 \text{ für } j = 1, \ldots, m$$

sei konvex.

(1) $L(x, \xi) = f(x) + \sum_j \xi_j \cdot h_j(x)$ heißt die verallgemeinerte Lagrange-Funktion des Minimierungsproblems.

(2) Das „duale“ Problem hat die Form

$$\text{maximiere}_{\xi \geq 0} \left( \text{minimiere}_x \; L(x, \xi) \right).$$

Im dualen Problem werden die Nebenbedingungen durch (nicht-negative) Lagrange Multiplikatoren in die Zielfunktion gehoben.

# Was ist der Zusammenhang zum dualen Problem?

Für jedes $\xi \geq 0$ und jede Lösung $y$ gilt

$$\min_x f(x) + \sum_j \xi_j \cdot h_j(x) \leq f(y) + \sum_j \xi_j \cdot h_j(y) \leq f(y),$$

denn $h_1(y), \ldots, h_m(y) \leq 0$.

Also gilt:

$$\begin{aligned} & \text{maximiere}_{\xi \geq 0} \left( \text{minimiere}_x \, L(x, \xi) \right) \\ \leq \quad & \text{minimiere}_{x \text{ ist Loesung}} f(x). \end{aligned}$$

Tatsächlich gilt sogar Gleichheit:

$$\begin{aligned} & \text{maximiere}_{\xi \geq 0} \left( \text{minimiere}_x \, L(x, \xi) \right) \\ = \quad & \text{minimiere}_{x \text{ ist Loesung}} f(x). \end{aligned}$$

# Dualität: Die Karush-Kuhn-Tucker Bedingungen

## Die KKT Bedingungen

Angenommen $x^*$ ist das globale Minimum von $f(x)$, wenn die Bedingungen $h_1(x) \leq 0, \ldots, h_m(x) \leq 0$ eingehalten werden müssen. Dann gibt es Lagrange-Multiplikatoren $\xi_1^*, \ldots, \xi_m^* \geq 0$ mit

(1) $\nabla f(x^*) + \sum_{j=1}^{m} \xi_j^* \cdot \nabla h_j(x^*) = \nabla L(x^*, \xi^*) = 0$ und

(2) $\xi_j^* \cdot h_j(x^*) = 0$ für alle $j$.

Wenn das globale Minimum $x^*$ ein innerer Punkt ist, d.h. wenn $h_j(x^*) < 0$ für alle $j$ gilt:

- Dann muss der Gradient von $f$ im Punkt $x^*$ verschwinden: Es ist $\nabla f(x^*) = 0$.
- Setze $\xi_1^* = \cdots = \xi_m^* = 0$ und die KKT-Bedingungen sind erfüllt.

# Beweis der Dualität

Sei $x^*$ ein globales Minimum mit den Lagrange-Multiplikatoren $\xi_1^*, \ldots, \xi_m^* \geq 0$.

(1) Dann ist $\nabla f(x^*) + \sum_{j=1}^{m} \xi_j^* \cdot \nabla h_j(x^*) = 0$.

(2) Es gilt

$$\min_y f(y) + \sum_j \xi_j^* \cdot h_j(y) = f(x^*) + \sum_j \xi_j^* \cdot h_j(x^*), \quad \text{da}$$

- $F(y) = f(y) + \sum_j \xi_j^* \cdot h_j(y)$ ist konvex, denn $\xi^* \geq 0$.
- Wegen (1) ist $x^*$ ein lokales Minimum von $F$ und damit auch ein globales Optimum.

(3) $\sum_j \xi_j^* \cdot h_j(x^*) = 0$, denn $\xi_j^* \cdot h_j^*(x) = 0$ für alle $j$ (KKT). Also

$$\min_y f(y) + \sum_j \xi_j^* \cdot h_j(y) = f(x^*) + \sum_j \xi_j^* \cdot h_j(x^*) = f(x^*).$$

# Hard Margin: Das duale Problem I

Die verallgemeinerte Lagrange Funktion für den Hard Margin Fall:

$$L(w,t,\xi) = \frac{1}{2} \cdot \langle w,w \rangle - \sum_{i=1}^{s} \xi_i \cdot ( \, b_i \cdot (\langle w, y_i \rangle + t) - 1 \, ).$$

Sei $(w^*, t^*)$ ein globales Minimum mit den Lagrange-Multiplikatoren $\xi_1^*, \ldots, \xi_m^* \geq 0$.

$$\begin{aligned} \nabla_w L(w^*, t^*, \xi^*) &= w^* - \sum_{i=1}^{s} \xi_i^* \cdot b_i \cdot y_i, \\ \nabla_t L(w^*, t^*, \xi^*) &= -\sum_{i=1}^{s} \xi_i^* \cdot b_i. \end{aligned}$$

Es ist $\nabla_w L(w^*, t^*) = 0$ und $\nabla_t L(w^*, t^*) = 0$: Wir können $w^*$ nach Nullsetzung bestimmen und $t^*$ taucht nicht mehr auf.

# Hard Margin: Das duale Problem II

Nach Rückeinsetzung ergibt sich

$$L(w^*, t^*, \xi^*) = -\frac{1}{2} \sum_{i,j=1}^{s} b_i \xi_i^* \cdot \langle y_i, y_j \rangle \cdot b_j \xi_j^* + \sum_{i=1}^{s} \xi_i^*.$$

Das duale Problem ist $\max_{\xi \geq 0}(\min_{w,t} L(w, t, \xi))$
oder äquivalent $\min_{\xi \geq 0}(\max_{w,t} - L(w, t, \xi))$.

## Das duale Problem:

$$\text{minimiere}_{\xi} \quad \frac{1}{2} \sum_{i,j=1}^{s} b_i \xi_i \cdot \langle y_i, y_j \rangle \cdot b_j \xi_j - \sum_{i=1}^{s} \xi_i$$

$$\text{so dass } \xi \geq 0 \quad \text{und} \quad \sum_{i=1}^{s} \xi_i \cdot b_i = 0.$$

# Hard Margin: Das duale Problem III

Die Beispiele $y_i$ liegen im Feature-Raum:
Es ist $y_i = \phi(x_i)$ und $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle = \langle y_i, y_j \rangle$.

Das duale Minimierungsproblem hat deshalb die Form

$$\text{minimiere}_\xi \ \frac{1}{2} \sum_{i,j=1}^{s} b_i \xi_i \cdot K(x_i, x_j) \cdot b_j \xi_j - \sum_{i=1}^{s} \xi_i$$

$$\text{so dass } \xi \geq 0 \ \text{ und } \sum_{i=1}^{s} \xi_i \cdot b_i = 0.$$

Wir können im niedrig-dimensionalen Beispielraum minimieren.

(1) Die Anzahl der Unbestimmten des Minimierungsproblems stimmt mit der Anzahl $s$ der Beispiele überein.

(2) Es gibt nur wenige, nämlich $s + 1$ Nebenbedingungen.

# Support-Vektoren I

Die KKT Bedingung $\xi_i^* \cdot (b_i \cdot (\langle w^*, y_i \rangle + t^*) - 1) = 0$ muss stets erfüllt sein.

(1) Setze $\text{SV} = \{\, i \mid b_i \cdot (\langle w^*, y_i \rangle + t^*) = 1 \,\}$.

(2) Ein Beispiel $x_i$ mit $i \in \text{SV}$ heißt ein Support-Vektor.

Es ist $\xi_i^* = 0$ für jedes $i \notin \text{SV}$ und deshalb

$$w^* = \sum_{i=1}^{s} b_i \xi_i^* \cdot y_i = \sum_{i \in \text{SV}} b_i \xi_i^* \cdot y_i.$$

**Nur** die Support-Vektoren $y_i$, also
die am nächsten an der trennenden Hyperebene liegenden Beispiele,
bestimmen den optimalen Gewichtsvektor $w^*$.

# Support-Vektoren II

Je weniger Support-Vektoren, umso wahrscheinlicher ist erfolgreiches Lernen.

(1) Warum?
- Die Hypothese kann durch wenige Beispiele beschrieben werden.
- Einfache Hypothesen machen erfolgreiches Lernen wahrscheinlicher: Occam.

(2) Eine Support-Vektor Hypothese erhält automatisch ein Gütesiegel bei wenigen Support-Vektoren.

# Zusammenfassung

(1) Der Anwender wählt eine Feature-Funktion $\phi$ und einen Kern $K(\phi(x), \phi(z)) = \langle \psi\phi(x), \psi\phi(z) \rangle$ auf den Feature-Vektoren.

Viele Kerne stehen zur Verfügung wie etwa polynomielle Kerne und der Gauß Kern.

(2) Die Bestimmung einer trennenden Hyperebene kann noch immer im niedrig-dimensionalen Beispielraum durchgeführt werden.

(3) Bei $s$ Beispielen aus dem Ball um 0 mit Radius $R$ und Margin $\gamma$ gilt im Wesentlichen

$$\varepsilon = O(\frac{1}{s} \cdot \frac{R^2}{\gamma^2} + \ln(\frac{1}{\delta}))$$

mit Wahrscheinlichkeit mindestens $1 - \delta$.

(4) Wenige Support-Vektoren machen erfolgreiches Lernen hochwahrscheinlich.

# Soft Margin

Im Feature-Raum zu lösen:

$$\text{minimiere}_{w,\beta,\xi} \quad ||w||^2 + C \cdot \sum_{i=1}^{m} \xi_i^2$$

$$\text{so dass für jedes } i\text{:} \quad b_i \cdot ( \langle w, y_i \rangle + \beta ) \geq 1 - \xi_i.$$

Nach Analyse des dualen Programms im Beispielraum zu lösen:

$$\text{minimiere}_{\xi} \quad \frac{1}{2} \sum_{i,j=1}^{s} b_i \xi_i \cdot \left( K(x_i, x_j) + \frac{1}{C} \cdot \delta_{i,j} \right) \cdot b_j \xi_j - \sum_{i=1}^{s} \xi_i$$

$$\text{so dass } \xi \geq 0 \quad \text{und} \quad \sum_{i=1}^{s} \xi_i \cdot b_i = 0.$$