Support-Vektor Maschinen: Feature-Funktionen

Feature Funktionen

Beispiele werden durch Zusatzinformationen, **Features**, aufbereitet.

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**.

Der Lernalgorithmus erhält die klassifizierten Feature-Vektoren $\phi(x_1), \ldots, \phi(x_s)$ anstatt der ursprünglichen Beispiele $x_1, \ldots, x_s \in X$.

Feature-Funktionen: Der Bag-of-Words Ansatz

Atomarer Text (oder **Atome**) eines Textes T sind sinntragende Teilworte von T wie *Stammsilben* oder ganze *Worte*.

- Für ein Atom s und ein Dokument $x \in D$ aus einer Menge D von Dokumenten ist
  - $h_s(x)$ die Häufigkeit des Atoms s im Dokument x,
  - $D(s)$ die Anzahl untersuchter Dokumente mit mindestens einem Auftreten des Atoms 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 des Atoms s im Dokument x mit dem Informationsgehalt $\log_2(\frac{|D|}{|D(s)|})$ des Atoms.

Feature-Funktionen: Interpolation

Wir erhalten Paare, die aus einem Punkt $x \in X \subseteq \mathbb{R}^3$ und einem Wert $y = p(x) \in \mathbb{R}$ bestehen. Gesucht ist ein Polynoms p vom Grad $\leqslant 2$.

- 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

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

für alle $x \in X$ gilt.

Das unbekannte Polynom wird zu einer linearen Funktion, die durch ein lineares Gleichungssystem bestimmt werden kann.

Support-Vektor Maschinen: Die Grundidee

Support-Vektor Maschinen: Der Ansatz

Eine binäre Klassifizierung $f : X \to \{0, 1\}$ ist zu lernen.

1. Der Entwickler konstruiert eine Feature Funktion $\phi : X \to \mathbb{R}^N$.
2. In $\{\phi(x_1), \ldots, \phi(x_s)\}$ sind die positiven von den negativen Beispielen mit Hilfe einer Hyperebene im $\mathbb{R}^N$ „bestmöglich“ zu trennen.

\+ Halbräume sind die mächtigste, effizient lernbare Hypothesenklasse.

✓ Benutze die lineare Programmierung.

? Für großes N ist der Feature-Raum $\phi(X)$ **hochdimensional**:

! Die Bestimmung einer trennenden Hyperebene ist aufwändig.

! Sind viele Beispiele notwendig? Overfitting droht!

Die Grundidee: Klassifikation mit großem Margin

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

(1) Der Perzeptron-Algorithmus benötigt höchstens

$$\left(\frac{R}{\rho}\right)^2$$

Gegenbeispiele, wenn alle Beispiele die Norm $\leqslant R$ besitzen.

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

(2) Kein Overfitting für relativ kleines R und relativ großes ρ?

Aber der Aufwand in der Bestimmung einer trennenden Hyperebene im hochdimensionalen Feature-Raum ist doch hoch?!

Der Kernel-Trick

Der Kernel-Trick (1/3)

Was passiert *im Beispielraum* X bei der Trennung der positiven von den negativen Feature-Vektoren durch

$$f(x) = \langle w, \phi(x) \rangle + t?$$

Der Perzeptron Algorithmus bestimmt

$$w = \sum_{i=1}^{s} \alpha_i \cdot \phi(x_i)$$

als Linearkombination der Gegenbeispiele $\phi(x_1), \ldots, \phi(x_s) \implies$

$$\begin{aligned} f(x) &= \langle w, \phi(x) \rangle + t = \langle \sum_{i=1}^{s} \alpha_i \cdot \phi(x_i), \phi(x) \rangle + t \\ &= \sum_{i=1}^{s} \alpha_i \cdot \underbrace{\langle \phi(x_i), \phi(x) \rangle}_{=:K(x_i,x)} + t. \end{aligned}$$

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

Aus $f(x) = \sum_{i=1}^{s} \alpha_i \cdot \langle \phi(x_i), \phi(x) \rangle + t$ folgt

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

? Wann kann $K(x_i, x)$ schnell berechnet werden?

? 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 $\{x \mid \sum_{i=1}^{s} \alpha_i \cdot K(x_i, x) = -t\}$.

Der Kernel-Trick (3/3)

Wir möchten eine Signatur für einen Virus erstellen.

Für ein Alphabet Σ sei X eine Menge $X \subseteq \Sigma^*$ von Dateien, von denen wir wissen ob sie befallen sind oder nicht.

Für eine Zahl d wähle die Eigenschaft

„ein Wort $u \in \Sigma^{\leqslant d}$ als Teilwort zu haben“

als Feature und definiere die Feature-Funktion

$$\phi(x) := (b_v \mid v \in \Sigma^{\leqslant d}),$$

wobei $b_v \in \{0, 1\}$ und $b_v = 1 \iff v$ ist ein Teilwort von x.

:-(Der Feature-Raum hat die viel zu große Dimension $\approx |\Sigma|^d$.

:-) $\langle \phi(x), \phi(x') \rangle =$ Anzahl gemeinsamer Teilstrings für x und $x' \Longrightarrow$ der Kern lässt sich in Zeit $\mathcal{O}(\min\{|x|, |y|\}^2)$ berechnen.

Lernen mit großem Margin

Support-Vektor Maschinen: Das Verfahren (1/3)

Wenn positive und negative Feature-Vektoren linear trennbar sind:

1. Bestimme eine Feature-Funktion $\phi : X \to \mathbb{R}$.
2. Fordere klassifizierte Beispiele $(\phi(x_1), f(x_1)), \ldots, (\phi(x_s), f(x_s))$ an.
3. Trenne positive und negative Beispiele mit möglichst großem Margin:
   - Wenn die lineare Funktion $\langle w, z \rangle + t$ mit $||w|| = 1$ den Margin ρ erreicht, dann ist $f(x_i) \cdot (\langle w, \phi(x_i) \rangle + t) \geqslant \rho$ für alle i.
   - Deshalb löse das Maximierungsproblem

$$\begin{aligned} &\text{maximiere}_{w,t} \quad \rho \quad \text{sodass} \\ &\qquad ||w|| = 1 \text{ und} \\ &\qquad \text{für jedes } i, (1 \leqslant i \leqslant s)\text{: } f(x_i) \cdot (\langle w, \phi(x_i) \rangle + t) \geqslant \rho. \end{aligned}$$

Support-Vektor Maschinen: Das Verfahren (2/3)

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

? Wir können die Bedingung $||w|| = 1$ durch $||w|| \leqslant 1$ ersetzen: Die Bedingung ist zumindest konvex.

✓ Statt den Margin ρ unter der Nebenbedingung $||w|| \leqslant 1$ zu maximieren, minimiere $||w||$ unter der Nebenbedingung $\rho \geqslant 1$:

$$\text{minimiere}_{w,t} \ \ ||w||^2 \ \ \text{sodass}$$
$$\text{für jedes } i, (1 \leqslant i \leqslant s): f(x_i) \cdot (\langle w, \phi(x_i) \rangle + t) \geqslant 1.$$

\+ Das Optimierungsproblem ist gutartig: Eine konvexe quadratische Form ist unter linearen Nebenbedingungen zu minimieren,

? aber im hochdimensionalen Feature-Raum!

Support-Vektor Maschinen: Das Verfahren (3/3)

Bisherige Annahme: Vollständige Trennbarkeit (**Hard Margin**). Jetzt:

Nur **partielle** Trennbarkeit (**Soft Margin**)

Füge Slack Variablen ξ_i für jedes Beispiel hinzu und löse

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

$$\text{für jedes } i: \quad f(x_i) \cdot (\langle w, \phi(x_i) \rangle + t) \geqslant 1 - \xi_i$$

$$\text{und } \xi \geqslant 0.$$

- Je größer ξ_i umso schlechter die Klassifizierung von $\phi(x_i)$.
- Der Parameter C definiert wie stark falsche Klassifizierungen bestraft werden: Bestimme C mit **Validierung**.

Beispielkomplexität

Der empirische Margin-Loss: Vorbereitung

Wird der Margin ρ für ein Beispiel $\phi(x)$

- eingehalten.
- unterschritten,
- bzw. wird $\phi(x)$ sogar falsch klassifiziert?

Für eine positive reelle Zahl ist

$$\Phi_\rho(z) := \begin{cases} 0 & \rho < z \\ 1 - \frac{z}{\rho} & 0 \leqslant z \leqslant \rho \\ 1 & z < 0. \end{cases}$$

Im Folgenden: $h(x) = \langle w, x \rangle$ mit $||w|| \leqslant 1$ sei eine lineare Funktion.

Der empirische Margin-Loss

(a) Für eine Hypothese h mit $h : \mathbb{R}^N \to \mathbb{R}$, eine Verteilung D auf X und eine Beispielmenge $S = \{ x_1, \ldots, x_s \} \subseteq X$ ist

$$\text{Loss}^{S,\rho}(h) := \frac{1}{s} \cdot \sum_{i=1}^{s} \Phi_\rho \Big(f(x_i) h(\phi(x_i)) \Big)$$

der **empirische Margin-Loss**.

(b) Des weiteren ist

$$\text{Loss}_D(h) = \mathbb{E}_{x \sim D}[\, \Phi_\rho \Big(f(x) h(\phi(x)) \Big) \,]$$

der **erwartete Hinge[a] Loss**.

[a] „hinge“ (engl.) steht für „Scharnier“ oder „Gelenk“.

Der erwartete 0-1 Loss ist nicht größer als der erwartete Hinge Loss.

Der Verallgemeinerungsfehler für SVMs

Es gelte $\|\phi(x)\| \leqslant R$ für alle $x \in X$ und es sei $\rho > 0$.

Bei s Beispielen gilt für jedes $\delta > 0$ und jede lineare Funktion $h(x) = \langle w, x \rangle$ mit $||w|| \leqslant 1$,

$$\underbrace{\text{Loss}_D(h)}_{\text{erwarteter Hinge Loss}} \leqslant \underbrace{\text{Loss}^{S,\rho}(h)}_{\text{empirischer Margin-Loss}} + \frac{1}{\sqrt{s}} \cdot \left(2\sqrt{\frac{R^2}{\rho^2}} + \sqrt{\frac{\ln \frac{1}{\delta}}{2}} \right)$$

mit Wahrscheinlichkeit mindestens $1 - \delta$.

- $\frac{R}{\rho} << \sqrt{s}$, sonst ist Schranke bedeutungslos.
- 0-1 Loss $\leqslant$ Hinge-Loss
  - Wir erhalten eine obere Schranke für den wahren Fehler.
- Neue Perspektive: Keine Forderung an die Beispielzahl, sondern eine Aussage über den Fehler.
  - Ist die Aussage gut oder schlecht?

Gute oder schlechte Fehlerschranke?

$$\underbrace{\text{Loss}_D(h)}_{\text{erwarteter Hinge Loss}} \leqslant \underbrace{\text{Loss}^{S,\rho}(h)}_{\text{empirischer Margin-Loss}} + \frac{1}{\sqrt{s}} \cdot \left(2\sqrt{\frac{R^2}{\rho^2}} + \sqrt{\frac{\ln \frac{1}{\delta}}{2}} \right)$$

Eine heuristische Rechnung führt auf $\varepsilon \leqslant \frac{1}{\sqrt{s}} \cdot \left(2\sqrt{\frac{R^2}{\rho^2}} + \sqrt{\frac{\ln \frac{1}{\delta}}{2}} \right) \Longrightarrow$

$$\begin{aligned} \varepsilon^2 &\leqslant \frac{1}{s} \cdot \left(2\sqrt{\frac{R^2}{\rho^2}} + \sqrt{\frac{\ln \frac{1}{\delta}}{2}} \right)^2 \\ &\approx \frac{1}{s} \cdot \left(\frac{4R^2}{\rho^2} + \frac{\ln \frac{1}{\delta}}{2} \right) = \mathcal{O}\left(\frac{1}{s} \cdot \left(\frac{R^2}{\rho^2} + \ln \frac{1}{\delta} \right) \right) \end{aligned}$$

- Fehlerschranke wie im agnostischen Lernmodell.
- Schlechte Schranke, wenn R^2/ρ^2 sehr viel größer als die Dimension des Feature-Raums ist.

Fehlerschranke im Hard-Margin-Fall

1. $\text{LINEAR}_\rho(\text{Def})$ ist die Klasse aller Threshold-Funktionen mit Definitionsbereich Def und Margin ρ auf Def.
2. $\text{VC}(\text{LINEAR}_\rho(\text{Def})) \leqslant \frac{R^2}{\rho^2}$.

Im Idealfall gehört die Zielfunktion f zur Klasse $\text{LINEAR}_\rho(\text{Def}) \Longrightarrow$

$$s \approx \frac{1}{\varepsilon} \cdot \left(\frac{R^2}{\rho^2} \ln(\frac{1}{\varepsilon}) + \ln(\frac{1}{\delta}) \right) \text{ bzw } \varepsilon \approx \frac{1}{s} \cdot \left(\frac{R^2}{\rho^2} \ln(\frac{1}{\varepsilon}) + \ln(\frac{1}{\delta}) \right)$$

Bis auf logarithmische Faktoren gilt

$$\varepsilon \approx \frac{1}{s} \cdot \left(\frac{R^2}{\rho^2} + \ln(\frac{1}{\delta}) \right).$$

Was ist getan, was bleibt zu tun?

Ein vorläufiges Fazit:

- ✓ Effizientes Training
  - aber nicht für sehr große Beispielzahlen.
- ✓ ERM-Hypothesen werden bestimmt.
- ✓ Hochwahrscheinlicher Lernerfolg ist bei kleinem empirischen Margin-Loss garantiert.

1. Wie sehen typische Anwendungen aus?
2. Welche Funktionen können als Kern verwandt werden?
3. Wie findet man im hochdimensionalen Feature-Raum **schnell** eine lineare Trennung mit großem Margin?

Anwendungen

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 größtem Informationsgehalt (auf der Trainingsmenge) werden ausgewählt.

(2) Krankheitszuordnung:
- 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 (1/2)

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

- Die Feature Funktion ϕ:
  - Die 7200 Bilder werden, nach Durchschnittsbildung auf 4×4 Gittern, von einer Auflösung von 128×128 Pixel auf 32×32 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 (2/2)

- 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 $||x - z||_1 = \sum_i |x_i - z_i|$ wird erfolgreich eingesetzt.
   - Die Wahl des Kerns ist kritisch.

Ziffernerkennung

- In einem Benchmark-Datensatz des US Postal Service für die Ziffererkennung treten 7291 Trainings- und 2007 Testbeispiele auf.
- Jedes Beispiel des Datensatzes wird als 16×16 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 $d \geqslant 3$.

(3) Support-Vektor Maschinen werden somit ohne problem-spezifische Aufbereitung eingesetzt.

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

Kerne

Kerne

Sei V ein Vektorraum.

(a) $\langle -, - \rangle : V^2 \to \mathbb{R}$ ist ein **inneres Produkt** für V, falls

($*$) $\langle -, - \rangle$ bilinear ist, falls also

$$\begin{aligned} \langle \alpha \cdot x + \beta \cdot y, u \rangle &= \alpha \cdot \langle x, u \rangle + \beta \cdot \langle y, u \rangle \text{ und} \\ \langle x, \gamma \cdot u + \delta \cdot v \rangle &= \gamma \cdot \langle x, u \rangle + \delta \cdot \langle x, v \rangle \end{aligned}$$

für alle $\alpha, \beta, \delta, \gamma \in \mathbb{R}$ und Vektoren $x, y, u, v \in V$ gilt,

($*$) $\langle x, u \rangle = \langle u, x \rangle$ für alle Vektoren $x, u \in V$ gilt und falls

($*$) für alle $x \in V$: $\langle x, x \rangle \geqslant 0$ und $\langle x, x \rangle = 0 \iff x = 0$.

(b) Ein **Kern** für eine Menge X ist eine Funktion $K : X^2 \to \mathbb{R}$, wobei

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

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

Der Kern als Ähnlichkeitsmaß

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

- 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.

Wann ist K ein Kern? Der Satz von Mercer

Kerne: Auf welche Eigenschaften kommt es an?

Sei $S = \{ x_1, \ldots, x_s \} \subseteq X$ eine Beispielmenge und $K : X \times X \to \mathbb{R}$ eine Funktion.

Die $s \times s$-**Gram-Matrix** G_K^S wird definiert durch

$$G_K^S[i,j] := K(x_i, x_j).$$

Wenn K ein Kern ist, dann ist K **symmetrisch** und die Gram-Matrix G_K^S ist **positiv semidefinit**, d.h. $u^T \cdot G_K^S \cdot u \geqslant 0$ gilt für alle Vektoren u:

$$\begin{aligned} u^T \cdot G_K^S \cdot u &= \sum_{i,j} u_i \cdot G_K^S[i,j] \cdot u_j = \sum_{i,j} u_i \cdot \langle \phi(x_i), \phi(x_j) \rangle \cdot u_j \\ &= \langle \underbrace{\sum_i u_i \cdot \phi(x_i)}_{=x}, \underbrace{\sum_j u_j \cdot \phi(x_j)}_{=x} \rangle = \langle x, x \rangle \geqslant 0. \end{aligned}$$

Der Satz von Mercer

Die Funktion $K : X \times X \to \mathbb{R}$ sei symmetrisch. Dann gilt:

$$K \text{ ist ein Kern} \iff G_K^S \text{ ist für jede endliche Teilmenge } S \subseteq X \text{ positiv semidefinit.}$$

$\Longrightarrow$ ✓

$\Longleftarrow$ Definiere $\mathbb{R}^X := \{ f \,:\, f : X \to \mathbb{R} \}$ und $\psi(x) := K(*, x)$.

1. Der Vektorraum V werde von den Funktionen $\psi(x) \in \mathbb{R}^X$ aufgespannt. Dann ist

$$\langle \sum_i \alpha_i \cdot \psi(x_i), \sum_i \beta_i \cdot \psi(x_i') \rangle^* := \sum_{i,j} \alpha_i \beta_j \cdot K(x_i, x_j')$$

   ein inneres Produkt auf V.
2. Es ist $\langle \psi(x), \psi(x') \rangle^* = K(x, x') \Longrightarrow K$ ist ein Kern. □

Der Satz von Mercer: Erste Konsequenzen

α, β seien nicht-negative reelle Zahlen.

Wenn K_1 und K_2 Kerne sind, dann ist auch $K = \alpha \cdot K_1 + \beta \cdot K_2$ ein Kern.

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

$$\begin{aligned} u^T \cdot G_K^S \cdot u &= u^T \cdot \left(\alpha \cdot G_{K_1}^S + \beta \cdot G_{K_2}^S \right) \cdot u \\ &= \alpha \cdot \underbrace{\left(u^T \cdot G_{K_1}^S \cdot u \right)}_{\geqslant 0} + \beta \cdot \underbrace{\left(u^T \cdot G_{K_2}^S \cdot u \right)}_{\geqslant 0} \geqslant 0 \end{aligned}$$

und das war zu zeigen.

Symmetrische, positiv semidefinite Matrizen

1. Symmetrische Matrizen B sind diagonalisierbar, d.h. es gibt eine orthogonale Matrix U und eine Diagonalmatrix D mit

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

2. Wenn B positiv semidefinit ist, dann sind alle Eigenwerte nichtnegativ.

$$B = (U^T \cdot \sqrt{D}) \cdot (\sqrt{D} \cdot U).$$

B ist genau dann symmetrisch und positiv semidefinit, wenn es eine Matrix M gibt mit

$$B = M^T \cdot M.$$

Weitere Konsequenzen

Wenn K_1 und K_2 Kerne sind, dann ist auch $K_1 \cdot K_2$ ein Kern.

$G^S_{K_1}$ ist positiv semidefinit $\Longrightarrow G^S_{K_1} = M^T \cdot M$ für eine Matrix M.

$$\sum_{i,j} u_i \Big(G^S_{K_1}[i,j] \cdot G^S_{K_2}[i,j] \Big) u_j = \sum_{i,j} u_i \Big((M^T M)[i,j] \cdot G^S_{K_2}[i,j] \Big) u_j$$

$$= \sum_{i,j} u_i \Big(\sum_k M^T[i,k] \cdot M[k,j] \cdot G^S_{K_2}[i,j] \Big) u_j$$

$$= \sum_k \Big(\sum_{i,j} u_i u_j M[k,i] \cdot M[k,j] \cdot G^S_{K_2}[i,j] \Big)$$

$$= \sum_k \Big(\underbrace{\sum_{i,j} u_i M[i,k] \cdot G^S_{K_2}[i,j] \cdot u_j M[j,k]}_{\geqslant 0} \Big) \quad \checkmark$$

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 Addition und Multiplikation. ✓

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

- Für $S \subseteq X$ sind die Matrizen $G^S_{p_n(K)}$ positiv semidefinit.
- Es gilt $G^S_{f(K)} = \lim_{n\to\infty} G^S_{p_n(K)}$ und $G^S_{f(K)}$ ist positiv semidefinit, da

$$u^T \cdot G^S_{f(K)} \cdot u = u^T \cdot (\lim_{n\to\infty} G^S_{p_n(K)}) \cdot u = \lim_{n\to\infty} u^T \cdot G^S_{p_n(K)} \cdot u \geqslant 0.$$

„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 semidefinite Matrix. Dann ist

$$K(x, z) = x^T \cdot A \cdot z$$

ein Kern, denn

$$K(x, z) \quad = \quad x^T \cdot A \cdot z = x^T \cdot M^T \cdot M \cdot z = \langle M \cdot x, M \cdot z \rangle \geqslant 0.$$

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 Exponentialfunktion 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

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

- 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.

Zusammenfassung

Was haben wir bisher erreicht?

(1) Wir haben große Klassen von Kernfunktionen konstruiert, die in vielen Anwendungen Fast-Trennbarkeit garantieren.

(2) Wir wissen, dass – *bis auf logarithmische Faktoren* – ungefähr

$$s \approx \frac{1}{\varepsilon^2} \cdot \left((\frac{R}{\rho})^2 + \ln(\frac{1}{\delta}) \right) \text{ bzw } s \approx \frac{1}{\varepsilon} \cdot \left((\frac{R}{\rho})^2 + \ln(\frac{1}{\delta}) \right)$$

Beispiele im Soft-Margin-Fall bzw. Hard-Margin-Fall genügen $\Longrightarrow$ Erfolgreiches Lernen bei kleinem Margin-Loss.

Was ist noch zu tun? Die effiziente Berechnung einer lineare Trennung mit großem Margin in einem hochdimensionalen Feature-Raum!

- Der wichtige Gauß-Kern benötigt unendlich viele Features!

Konvexe Minimierung

Konvexe Mengen und Funktionen

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

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

  falls also der Funktionsgraph stets unter oder auf 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 Strecke $\{\lambda \cdot x + (1 - \lambda) \cdot y \mid 0 \leqslant \lambda \leqslant 1\}$ 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) Der Durchschnitt konvexer Mengen ist konvex.

Optimierungsprobleme

Das Ziel: $\text{minimiere}_{w,t} \ \frac{1}{2} \cdot ||w||^2$

so dass für jedes i $(1 \leqslant i \leqslant s)$: $f(x_i) \cdot (\langle w, \phi(x_i) \rangle + t) \geqslant 1$.

Welche speziellen Eigenschaften hat unser Minimierungsproblem?

(1) Die Zielfunktion $f(w) := \frac{1}{2} \cdot ||w||^2 = \sum_i w_i^2$ 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**, falls die Zielfunktion konvex ist und falls alle Nebenbedingungen konvex sind.

(1) Wir müssen eine konvexe Funktion über einer konvexen Menge minimieren.

(2) Jedes lokale Minimum x ist 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 $\lambda < 1$

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

und x ist kein lokales Minimum.

Das duale Problem

Das duale Problem

Unser Minimierungsproblem hat die allgemeine Form

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

wobei $f, h_1, \ldots, h_m$ konvex sind.

(a) Hebe die Nebenbedingungen durch die nicht-negativen **Lagrange-Multiplikatoren** ξ_i in die Zielfunktion. Dann heißt

$$L(x, \xi) = f(x) + \sum_j \xi_j \cdot h_j(x)$$

die **verallgemeinerte Lagrange-Funktion** von (1).

(b) Das „duale“ Problem hat die Form

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

Was ist der Zusammenhang zum dualen Problem?

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

$$\min_x \left\{ f(x) + \sum_j \xi_j \cdot h_j(x) \right\} \leqslant f(y) + \sum_j \xi_j \cdot h_j(y) \leqslant f(y),$$

denn $h_1(y), \ldots, h_m(y) \leqslant 0$ und $\xi \geqslant 0$.

Also gilt die **schwache Dualität**:

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

Wir zeigen, dass sogar Gleichheit gilt, d.h.

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

Die Karush-Kuhn-Tucker (KKT) Bedingungen

Die Funktionen $f, h_1, \ldots, h_m$ seien konvex und es gelte $h_1(x^*) \leqslant 0, \ldots, h_m(x^*) \leqslant 0$. Dann sind äquivalent:

(a) $f(x^*) = \min\{ f(x) \,:\, h_1(x) \leqslant 0, \ldots, h_m(x) \leqslant 0 \}$.

(b) Es gibt Lagrange-Multiplikatoren $\xi_1^*, \ldots, \xi_m^* \geqslant 0$ 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.

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

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

Und wenn nicht?

(b) $\Longrightarrow$ (a) Siehe Skript.

Beweis der Dualität

Die Funktion f sei konvex mit globalem Minimum x^*.

Die Lagrange-Multiplikatoren seien $\xi_1^*, \ldots, \xi_m^* \geqslant 0$.

(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 $\xi^* \geqslant 0$.
- 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

$$\min_y f(y) + \sum_j \xi_j^* \cdot h_j(y) = \underbrace{f(x^*) + \sum_j \xi_j^* \cdot h_j(x^*)}_{=L(x^*,\xi^*)} = f(x^*).$$

Das primale und duale Problem

Aus den KKT-Bedingungen folgt

$$\begin{aligned} \max_{\xi \geqslant 0} \min_{y} L(y,\xi) &\geqslant \min_{y} L(y,\xi^*) = L(x^*,\xi^*) = f(x^*) \\ &\underbrace{\geqslant}_{\text{schwache Dualität}} \max_{\xi \geqslant 0} \min_{y} L(y,\xi). \end{aligned}$$

Das **primale** Problem

$$\min_{x \text{ ist Loesung}} f(x)$$

werde von der Lösung x^* minimiert. Dann folgt für das **duale** Problem

$$\max_{\xi \geqslant 0} \left(\min_{x} L(x,\xi) \right) = L(x^*,\xi^*) = f(x^*)$$

und wir erhalten den **Dualitätssatz**. Wir benötigen später „nur“

$$\max_{\xi \geqslant 0} L(x^*,\xi) = f(x^*).$$

Dualitätssatz

Die Funktionen f und $h_1, \ldots, h_m$ seien konvex. Dann gilt

$$\max_{\xi \geqslant 0} \left(\min_x f(x) + \sum_j \xi_j \cdot h_j(x) \right) = \min_{x : h_j(x) \leqslant 0 \text{ für alle } j} f(x).$$

Es ist $\min_x \left(\max_{\xi \geqslant 0} f(x) + \sum_j \xi_j \cdot h_j(x) \right) = \min_{x : h_j(x) \leqslant 0 \text{ für alle } j} f(x) \Longrightarrow$

Eine äquivalente Form des Dualitätssatzes:

$$\max_{\xi \geqslant 0} \left(\min_x f(x) + \sum_j \xi_j \cdot h_j(x) \right) = \min_x \left(\max_{\xi \geqslant 0} f(x) + \sum_j \xi_j \cdot h_j(x) \right)$$

Der Hard-Margin-Fall

Ist das duale Problem wirklich so kompliziert?

Es ist

$$\max_{\xi \geqslant 0} L(w^*, t^*, \xi) = \min_{w,t:\ h_1(w,t),\ldots,h_m(w,t) \leqslant 0} \langle w, w \rangle.$$

1. Wende die KKT-Bedingungen an, um x^* mit Hilfe von ξ^* zu eliminieren und
2. bestimme die Lagrange-Multiplikatoren ξ^* mit dem dualen Maximierungsproblem.

Hard Margin: Das duale Problem (1/3)

Die verallgemeinerte Lagrange Funktion für den Hard Margin Fall:

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

Wende die KKT-Bedingungen an: Sei (w^*,t^*) ein globales Minimum mit den Lagrange-Multiplikatoren $\xi_1^*,\ldots,\xi_m^* \geqslant 0$.

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

Es ist $\nabla_w L(w^*,t^*) = 0$ und $\nabla_t L(w^*,t^*) = 0 \Longrightarrow$

Bestimme w^* nach Nullsetzung und t^* taucht nicht mehr auf.

Hard Margin: Das duale Problem (2/3)

Nach Rückeinsetzung ergibt sich

$$L(w^*, t^*, \xi^*) = -\frac{1}{2} \sum_{i,j=1}^{s} f(x_i)\xi_i^* \cdot \langle \phi(x_i), \phi(x_j) \rangle \cdot f(x_j)\xi_j^* + \sum_{i=1}^{s} \xi_i^*.$$

Das duale Problem $D := \max_{\xi \geqslant 0}(\min_{w,t} L(w, t, \xi))$ ist äquivalent zu

$$-\max_{\xi \geqslant 0} \left(\min_{w,t} L(w, t, \xi) \right) = -\max_{\xi \geqslant 0} L(w^*, t^*, \xi) = \min_{\xi \geqslant 0} -L(w^*, t^*, \xi).$$

$$D = \min_{\xi \geqslant 0} \frac{1}{2} \sum_{i,j=1}^{s} f(x_i)\xi_i \cdot \langle \phi(x_i), \phi(x_j) \rangle \cdot f(x_j)\xi_j - \sum_{i=1}^{s} \xi_i$$

$$\text{so dass } \xi \geqslant 0 \text{ und } \underbrace{\sum_{i=1}^{s} \xi_i \cdot f(x_i) = 0}_{\text{KKT-Bedingung}}.$$

Hard Margin: Das duale Problem (3/3)

Es ist $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$. Das duale Problem hat also die Form

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

$$\text{so dass } \xi \geqslant 0 \quad \text{und} \quad \sum_{i=1}^{s} \xi_i \cdot f(x_i) = 0.$$

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

Support-Vektoren (1/2)

Eine der KKT Bedingungen: $\xi_i^* \cdot (f(x_i) \cdot (\langle w^*, \phi(x_i)\rangle + t^*) - 1) = 0$.

(1) Setze $\mathrm{SV} = \{\, i \mid f(x_i) \cdot (\, \langle w^*, \phi(x_i)\rangle + t^*) = 1 \,\}$.

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

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

$$w^* = \sum_{i=1}^{s} f(x_i)\xi_i^* \cdot \phi(x_i) = \sum_{i \in \mathrm{SV}} f(x_i)\xi_i^* \cdot \phi(x_i).$$

Nur die Support-Vektoren $\phi(x_i)$, also

die am nächsten an der trennenden Hyperebene liegenden Beispiele,

bestimmen den optimalen Gewichtsvektor w^*.

Support-Vektoren (2/2)

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

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

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

Soft Margin

Im Feature-Raum zu lösen:

$$\min_{w,\beta,\eta} \quad ||w||^2 + C \cdot \sum_{i=1}^{m} \eta_i$$

so dass für jedes i: $\quad f(x_i) \cdot (\langle w, \phi(x_i) \rangle + \beta) \geqslant 1 - \eta_i$ und $\eta_i \geqslant 0$.

Nach Analyse des dualen Programms löse im Beispielraum:

$$\min_{\xi \geqslant 0} \quad \frac{1}{2} \sum_{i,j=1}^{s} f(x_i)\xi_i \cdot K(x_i, x_j) \cdot f(x_j)\xi_j - \sum_{i=1}^{s} \xi_i$$

$$\text{sodass } 0 \leqslant \xi_j \leqslant C \text{ für alle } j \text{ und } \sum_{i=1}^{s} \xi_i \cdot f(x_i) = 0.$$

Zusammenfassung

(1) Der Anwender wählt eine Feature-Funktion ϕ und einen Kern ψ aus dem Werkzeugkasten (Polynome, Gauß-Funktion etc.)

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

(2) Die Bestimmung einer trennenden Hyperebene kann im „*niedrig-dimensionalen*“ Beispielraum durchgeführt werden.

(3) Bei s Beispielen der Länge höchstens R und Margin ρ gilt

$$\underbrace{\mathrm{Loss}_D(h)}_{\text{erwarteter Hinge Loss}} \leqslant \underbrace{\mathrm{Loss}^{S,\rho}(h)}_{\text{empirischer Margin-Loss}} + \frac{1}{\sqrt{s}} \cdot \left(2\sqrt{\frac{R^2}{\rho^2}} + \sqrt{\frac{\ln \frac{1}{\delta}}{2}} \right)$$

mit Wahrscheinlichkeit mindestens $1 - \delta$.

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