Kerne und Kostenfunktionale

Franziska Häußler

Universität Ulm
Fakultät für Mathematik und Wirtschaftswissenschaften

22. Mai 2007

Inhaltsverzeichnis

Inhaltsverzeichnis

Inhaltsverzeichnis

Einleitung

- Grundproblem der statistischen Lerntheorie: aus Daten lernen
- Linearer Ansatz
  - Vorteil: einfache Algorithmen
  - Nachteil: geringe Flexibilität
- Nicht-linearer Ansatz
  - Vorteil: bessere Anpassung
  - Nachteil: zu hohe algorithmische Komplexität
- Lösung dieses Dilemmas: **Kerne**

Warum Kerne?

1. Verbindung der Vorteile:
   Beschreibung von nicht-linearen als lineare Abbildungen in sog. Merkmalräumen
2. Verwendung von einfachen, exakt lösbaren Optimierungsverfahren
3. Auch Verbesserung gegenüber Neuronalen Netzen

Was sind Kerne?

1. **Ähnlichkeitsmaße**
2. **Nicht-Ähnlichkeitsmaße**

Konventionen

- Auf den folgenden Folien gilt:
- Merkmalabbildung: $\Phi : \mathcal{X} \to \mathcal{H}$,
- Eingabemenge: $\mathcal{X} \subset \mathbb{R}^N$, nicht leer
- Merkmalraum: $\mathcal{H}$, i.A. mit $dim(\mathcal{X}) << dim(\mathcal{H})$
- $\mathbb{K} = \mathbb{R}$ oder $\mathbb{C}$

Inhaltsverzeichnis

Kern als Ähnlichkeitsmaß

- Definition: Abbildung

$$k : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$$

- mit **Symmetrie-Eigenschaft**

$$\forall x, x' \in \mathcal{X} : \quad k(x, x') = \overline{k(x', x)}$$

- und **positiver Definitheit**

$$\sum_{i,j=1}^{m} (c_i \overline{c}_j k(x_i, x_j)) \geq 0, \qquad \forall c_i \in \mathbb{K}$$

Rechenbeispiel

- Geg.: $\Phi : \mathbb{R}^2 \to \mathbb{R}^3 = \mathcal{H},\ (x_1, x_2) \mapsto (x_1^2, \sqrt{2}x_1x_2, x_2^2)$
- Wähle k als Skalarprodukt von Φ im Merkmalraum $\mathcal{H}$
- Berechnung des Kerns:

$$\begin{aligned}
k(x, x') &= \langle \Phi(x), \Phi(x') \rangle \\
&= (x_1^2, \sqrt{2}x_1x_2, x_2^2)^T (x_1'^2, \sqrt{2}x_1'x_2', x_2'^2) \\
&= x_1^2 x_1'^2 + 2x_1x_2x_1'x_2' + x_2^2 x_2'^2 \\
&= (x_1x_1' + x_2x_2')^2 \\
&= \langle x, x' \rangle^2
\end{aligned}$$

Interpretation

- Kerne als Ähnlichkeitsmaße spiegeln Ähnlichkeit zweier Objekte wider
- **Je größer** der Wert des **Ähnlichkeitsmaß**es, **desto ähnlicher** sind die **Objekt**e (folgt aus der Cauchy-Schwarzschen-Ungleichung für Kerne)
- Eigenschaft: Positive Definitheit, Symmetrie

Konstruktion eines Merkmalraumes (1)

- Wähle z.B. Hilbertraum $L_2(\mathcal{X})$
- Problem: enthält viele nicht-glatte Funktionen
- Lösung: Hilbertraum auf kleinere Menge von Funktionen einschränken
  $\hat{=}$ **Reproducing Kernel Hilbert Space**

Reproducing Kernel Hilbert Space (RKHS)

- Definition: Sei $\mathcal{H}$ Hilbertraum mit Funktionen $f : \mathcal{X} \to \mathbb{R}$, Skalarprodukt $\langle \cdot, \cdot \rangle$ (und Norm $\|f\| := \sqrt{\langle f, f \rangle}$), falls $\exists\ k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ mit folgenden Eigenschaften:

1. **Reproduzierungseigenschaft**:

$$\langle k(.,x), f(.) \rangle = f(x) \text{ und } \langle k(.,x), k(.,x') \rangle = k(x,x') \ \forall x, x' \in \mathcal{X}$$

2. **Abgeschlossenheit des Raumes**:

$$\text{k spannt } \mathcal{H} \text{ auf: } \mathcal{H} = \overline{span\{k(.,x) | x \in \mathcal{X}\}}$$

Konstruktion eines Merkmalraumes (2)

- Geg.: Kern k(x,x') und reproduzierende Merkmalabbildung $\phi : \mathcal{X} \to \mathbb{R}^{\mathcal{X}} = \{f : \mathcal{X} \to \mathbb{R}\}$, $\phi(.) = k(., x)$
- Ziel: Vektorraum konstruieren mit k als Skalarprodukt
  1. Konstruiere Vektorraum aller Linearkombinationen von k(.,x)
  2. Definiere darauf ein Skalarprodukt
  3. Prüfe, ob wirklich Skalarprodukt

Konstruktion eines Merkmalraumes (3)

- zu 1.: Für $m \in \mathbb{N}$, $\alpha_i \in \mathbb{R}$ und $x_i \in \mathcal{X}$ beliebig

$$f(.) = \sum_{i=1}^{m} \alpha_i k(., x_i)$$

- zu 2.: Skalarprodukt von f und $g(.) = \sum_{j=1}^{k} \beta_j k(., x_j')$:

$$\langle f, g \rangle := \sum_{i=1}^{m} \sum_{j=1}^{k} \alpha_i \beta_j k(x_i, x_j')$$

Konstruktion eines Merkmalraumes (4)

- zu 3.: fast alle Eigenschaften sind offensichtlich
- noch zu zeigen: $\langle f, f \rangle = 0 \Rightarrow f = 0$
- Beweis: $f(.) = \sum \alpha_i k(., x_i)$:

$$\langle k(., x), f(.) \rangle = \sum \alpha_i k(x_i, x) = f(x)$$

- $\Rightarrow \langle k(., x), k(., x') \rangle = k(x, x')$
- $|f(x)|^2 = |\langle k(., x), f(x) \rangle|^2 \leq k(x, x) \cdot \langle f(x), f(x) \rangle$
- $\Rightarrow 0 \leq |f(x)|^2 \leq k(x, x) \cdot 0 \quad \Rightarrow f(x) = 0$

Darstellung im Merkmalraum

- **Theorem von Mercer** gibt Bedingungen an, wann zu einem Kern ein Hilbertraum gefunden werden kann, in dem der Kern ein Skalarprodukt darstellt
- Liefert Darstellung des RKHS mit einer Standardbasis
- Interpretation des Theorems:
  Jeder stetige, symmetrische, positiv definite Kern kann als Skalarprodukt in einem höher-dimensionalen Raum ausgedrückt werden
- Anwendung: **„Kerntrick“**

Theorem von Mercer (1)

- Ann.: Sei μ ein endliches Maß und $k \in L_\infty(\mathcal{X}^2)$ symmetrisch, stetig und reellwertig, so dass der Integraloperator

$$T_k : L_2(\mathcal{X}) \to L_2(\mathcal{X})$$

$$(T_k f)(x) := \int_{\mathcal{X}} k(x, x') f(x') d\mu(x')$$

positiv definit ist und so dass $\forall\ f \in L_2(\mathcal{X})$ gilt:

$$\int_{\mathcal{X} \times \mathcal{X}} k(x, x') f(x) f(x') d\mu(x) d\mu(x') \geq 0$$

Theorem von Mercer (2)

- In dieser Situation gibt es für den Operator T_k höchstens abzählbar viele nicht-negative Eigenwerten $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_J \geq 0$ und dazugehörige, orthogonale Eigenfunktionen $\psi_j \in L_2(\mathcal{X})$ ($J \in \mathbb{N}$ oder $J = \infty$)
- Und es gilt:

$$k(x, x') = \sum \lambda_j \psi_j(x) \psi_j(x') \text{ für fast alle } (x, x')$$

$\Rightarrow$ Wahl von $\Phi(x) = (\sqrt{\lambda_1}\psi_1(x), \ldots, \sqrt{\lambda_J}\psi_J(x))$

„Kerntrick“

- Methode um lineare Algorithmen in nicht-lineare umzuwandeln
- Dazu Verwendung von nicht-lineare Funktionen, die die originalen Beobachtungen in höher-dimensionale Merkmalräume abbilden
- Anwendungen:
  - Perceptron Algorithmus
  - Support Vector Machines
  - Hauptkomponentenanalyse
  - ...

Beispiele

- Homogener Polynom-Kern: $k(x, x') = \langle x, x' \rangle^d$
- Inhomogener Polynom-Kern: $k(x, x') = (\langle x, x' \rangle + c)^d$
- S-förmige Kerne: $k(x, x') = \tanh(\kappa \langle x, x' \rangle + \vartheta)$ mit $\kappa, \vartheta > 0$

Inhaltsverzeichnis

Kern als Nicht-Ähnlichkeitsmaß

- Definition: Abbildung

$$k : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$$

- mit **Symmetrie-Eigenschaft**

$$\forall x, x' \in \mathcal{X} : \quad k(x, x') = \overline{k(x', x)}$$

- und **bedingt positiver Definitheit**

$$\sum_{i,j=1}^{m} (c_i \overline{c}_j k(x_i, x_j)) \geq 0, \forall c_i \in \mathbb{K} \quad \mathit{mit} \quad \sum_{i=1}^{m} c_i = 0 \quad \mathit{und} \quad m \geq 2$$

Interpretation

- Kerne als Nicht-Ähnlichkeitsmaße ermitteln auch Ähnlichkeit zweier Objekte
- **Je kleiner** der Wert des **Nicht-Ähnlichkeitsmaß**es, **desto ähnlicher** sind die verglichenen **Objekt**e
- Auch Unähnlichkeitsmaße oder Distanzmaße genannt
- Eigenschaft: Bedingte positive Definitheit, Symmetrie

Warum bedingt positiv definite Kerne?

- größere Klasse als positiv definite Kerne
- aber es gibt Verbindung zu positiv definiten Kernen
- manche Kernalgorithmen arbeiten mit ihnen

Konstruktion von pd Kerne aus bedingt pd Kernen

- Sei $x_0 \in \mathcal{X}$ und k ein bedingt pd Kern auf $\mathcal{X} \times \mathcal{X}$

  $$\Leftrightarrow \tilde{k}(x, x') := \frac{1}{2}(k(x, x') - k(x, x_0) - k(x_0, x') + k(x_0, x_0))$$

  ist ein pd Kern

- Führt zur Hilbertraum-Darstellung für $\tilde{k}(x, x')$ mit $\tilde{k}(x, x') = \langle \Phi(x), \Phi(x') \rangle$, indem man die obige Schreibweise in $\|\Phi(x) - \Phi(x')\|^2 = \tilde{k}(x, x) - 2\tilde{k}(x, x') + \tilde{k}(x', x')$ einsetzt

Darstellung von bedingt pd Kerne im Hilbertraum

- Sei k ein reellwertiger, bedingt pd Kern auf $\mathcal{X} \times \mathcal{X}$ dann existiert ein Hilbertraum $\mathcal{H}$ und eine Abbildung $\Phi : \mathcal{X} \to \mathcal{H}$, so dass

$$\|\Phi(x) - \Phi(x')\|^2 = -k(x, x') + \frac{1}{2}(k(x, x) + k(x', x'))$$

- Für $k(x, x) = 0 \quad \forall x \in \mathcal{X}$ gilt dann:

$$\|\Phi(x) - \Phi(x')\|^2 = -k(x, x')$$

Beispiel

- Exponential Kern: $e^{-c\|x-x'\|^{\beta}},\ 0 \leq \beta \leq 2$
- Inverser Multiquadratischer Kern: $k(x, x') = \frac{1}{\sqrt{\|x-x'\|^2+c^2}}$
- Multiquadratischer Kern: $k(x, x') = -\sqrt{\|x - x'\|^2 + c^2}$

Inhaltsverzeichnis

Fragestellung

- Ziel der Lerntheorie: Möglichst gute Klassifikation bzw. Regression
- Wie kann man dies beurteilen?
- Warum ist das nicht trivial?
    - Kosten müssen nicht symmetrisch sein
    - Kosten können von den Eingabedaten abhängen
    - Wahrscheinlichkeit für eine Fehlklassifizierung kann berücksichtigt werden
    - usw.

Definition des Kostenfunktionals

- Geg.: $(x, y, f(x)) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Y}$
- x: Lerndaten, y: Beobachtung, f(x): Vorhersage
- $c : \mathcal{X} \times \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$
- mit der Eigenschaft

$$c(x, y, f(x)) = 0 \quad \text{für} \quad y = f(x)$$

- wird **auch Verlust- oder Risikofunktion** genannt

Klassifizierung

- Durch Falschklassifizierung entstehen Kosten
- Kein Unterschied zw. Klassen und Fehlertypen
  - Kostenfunktional bei binärer Klassifikation: $\tilde{c}(x) \equiv 1$
- Sonst ist $\tilde{c}(yf(x))$ eine beliebige, nicht-negative Funktion (z.B. bei asymmetrischen Kosten)
  - Soft Margin Kostenfunktional: $\tilde{c}(yf(x)) = \max(0, 1 - yf(x))$
  - Logistisches Kostenfunktional: $\tilde{c}(yf(x)) = \ln(1 + \exp(-yf(x)))$

Regression

- Schätzung der Differenz von $y - f(x)$ entspricht der Quantität der Falsch-Vorhersage
- Bekannteste Kostenfunktionale:
  - Quadratisches Kostenfunktional: $\tilde{c}(y - f(x)) = (y - f(x))^2$
  - ϵ - unempfindliches Kostenfunktional: $\tilde{c}(y - f(x)) = \max(|y - f(x)| - \epsilon, 0) = |y - f(x)|_\epsilon$

Inhaltsverzeichnis

Testfehler und erwartetes Risiko

- Ziel: **Finde Methode um** erkannte **Fehler zu minimieren**
- Ann.: Daten (x,y) iid mit P(x,y)
- 1.Fall: Wissen über Testdaten während der Trainingszeit
  - Minimierung des erwarteten Fehlers auf dieser speziellen Testmenge
- 2.Fall: kein Wissen über Testdaten während der Trainingszeit
  - Minimierung des erwarteten Fehlers auf allen möglichen Testmengen

Definition des Testfehlers

- Trainingsdaten $\{x_1, \ldots, x_m\}$ mit Zielwerten $\{y_1, \ldots, y_m\}$
- bekannte Testdaten $\{x'_1, \ldots, x'_k\}$ aus denen $y'_i, \; i = 1, \ldots, k$, vorhergesagt werden sollen
- Minimiere erwarteten Fehler auf Testmenge

$$R_{test}[f] := \frac{1}{k} \sum_{i=1}^{k} \int_{\mathcal{Y}} c(x'_i, y, f(x'_i)) dP(y|x'_i)$$

- Problem: rechenintensiv und kompliziert

Definition des erwarteten Risikos

- kein Wissen über das Testdaten vereinfacht die Minimierung
- Minimiere erwarteten Fehler auf allen möglichen Testmengen
- Minimiere dazu erwartete Kosten bezüglich P und c

$$R[f] = E[R_{test}[f]] := \int_{\mathcal{X} \times \mathcal{Y}} c(x, y, f(x)) dP(x, y)$$

- Problem: nur lösbar, wenn $P(x, y)$ explizit bekannt; alternativ durch empirische Verteilung ersetzen

Definition des empirischen Risikos

- Minimiere das empirische Risiko

$$R_{emp}[f] = \frac{1}{m} \sum_{i=1}^{m} c(x_i, y_i, f(x_i))$$

- Vorteil: bei gegebenen Testdaten einfach zu berechnen und minimieren

Inhaltsverzeichnis

Statistische Sichtweise

- Statt (oder zusätzlich zum) erw. Risiko für festes Muster: Wunsch nach Wissen über $P(y|\tilde{x})$
- Ziel: Berechnung der bedingten Dichte $p(y|\tilde{x})$
- Hilfsmittel: **Maximum-Likelihood-Schätzung**
  Liefert Funktion f, die höchstwahrscheinlich die Daten erzeugt hat, in dem man $p(y|x, f(x))$ bestimmt und bzgl. f maximiert

Wiederholung: Maximum-Likelihood

- **Maximum-Likelihood-Funktion:**
  $L((x,y),f) := \prod p(x_i, y_i|f) = \prod p(y_i|x_i, f)p(x_i)$
- **Log-Likelihood-Funktion:**
  $logL((x,y),f) := \sum \ln(p(y_i|x_i, f)p(x_i))$
  $= \sum \ln p(y_i|x_i, f) + \sum \ln p(x_i)$
- Aus OR bekannt: $\max f(x) = -\min f(x)$
- Somit:
  $\max logL((x,y),f) = -\min(\sum \ln p(y_i|x_i, f) + \sum \ln p(x_i))$
- Wobei die letzte Summe eine Konstante bzgl. der Maximierung f ist und weggelassen werden kann

Kostenfunktional für die Regression

- Es gilt:

$$\min \sum_{i=1}^{m} (-\ln(p(y_i|x_i, f))) = \min(R_{emp}[f])$$

wenn $c(x, y, f(x)) = -\ln(p(y|x, f(x))$

- Eindeutiges Minimum existiert selten und nur bei festem f(x)
- Bei Störung der Funktion f:

$$c(x, y, f(x)) = -\ln(p_\xi(y - f(x)))$$

mit gestörter Dichte p_ξ

Kostenfunktional für die Klassifizierung

- Bei binäre Klassifizierung ist $P(y|f(x))$ direkt berechenbar mit

$$c(x, y, f(x)) = -\ln(P(y|f(x)))$$

- Hier: $y \in \{-1, 1\} \;\Rightarrow\; P(-1|f(x)) = 1 - P(1|f(x))$
- Möglich: $P(y|f(x))$ nach Bedarf zu wählen, z.B. als logistische Linkfunktion $c(x, y, f(x)) = \ln(1 + \exp(-f(x)))$
- Kostenfunktionale jedoch nicht frei wählbar

Inhaltsverzeichnis

Eigenschaften von Schätzern

- Ann.: **Erwartungstreue** (wobei Verzerrung nicht unbedingt schlechte Eigenschaft)
- **Varianz** als Größe für die Schätzgenauigkeit (mit Cramer-Rao als untere Schranke)
- **Effizienz** als Indikatior dafür, wie „gestört“ ein Schätzer ist

Effizienz

- Definition:

$$e := \frac{1}{\det(I\ B)}$$

- I : Fisher-Informationsmatrix
- B : Kovarianzmatrix von $\hat{\theta}(Y)$
- Interpretation: je näher e an 1 ist, desto kleiner ist die Varianz

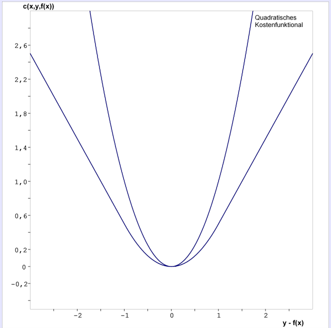
Beispiel für effizienten Schätzer

- ML-Schätzer ist asymptotisch effizent
- Aber
  - bei „kleinem“ Stichprobenumfang gibt es bessere Schätzer
  - eventuell wahre Dichte nicht bekannt, also großer Fehler möglich
- Lösung: robuste Schätzer

Robuste Schätzer

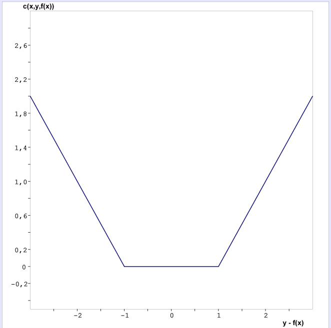
- Ziel: Entfernung des Anteils von „schlechten“ Beobachtungen (sog. Ausreißer), die die Qualität der Schätzung beeinträchtigen
- **Idee von Huber:** Robuste Schätzer konstruieren, die so modifiziert sind, damit der Einfluss von jedem einzelnen Muster begrenzt ist

Robustes Kostenfunktional von Huber



$$c(x, y, f(x)) = \begin{cases} \frac{1}{2\sigma}(y - f(x))^2 & |y - f(x)| \leq \sigma \\ |y - f(x)| - \frac{\sigma}{2} & sonst \end{cases}$$

ϵ - unempfindliches Kostenfunktional



$$c(x, y, f(x)) = |y - f(x)|_\epsilon$$

Fragen

Habt Ihr noch Fragen?

Fragen

Vielen Dank für eure Aufmerksamkeit

Quellen

- B.Schölkopf, A. Smola, „Learning with Kernels“, MIT Press, 2002
- Vorlesungen von Prof. M. Pawlak, University of Manitoba, Kanada
- http://www.learning-with-kernels.org
- http://www.wikipedia.org