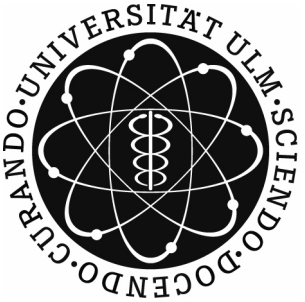
Universität Ulm
Fakultät für Mathematik und
Wirtschaftswissenschaften



Kerne und Kostenfunktionale

Statistische Lerntheorie und ihre Anwendungen

Seminararbeit
in dem Institut für Stochastik

Prüfer: Prof. Dr. U. Stadtmüller, Prof. Dr. E. Spodarev
Betreuer: Prof. Dr. U. Stadtmüller

vorgelegt von:
Name, Vorname: Häußler, Franziska

Abgabetermin: 16. Juli 2007

Inhaltsverzeichnis

1 Einleitung

In dieser Seminararbeit werde ich das Thema „Kerne und Kostenfunktionale“ im Rahmen der statisischen Lerntheorie behandeln. Ich werde mich dabei primär an dem zweiten und dritten Kapitel des Buches „Learning with Kernels“ von Schölkopf und Smola [1][3] orientieren.

Allgemein lässt sich meine Seminararbeit in zwei Themenbereiche untergliedern. Nach der Einleitung werden zunächst die „Kerne“, als Hilfsmittel zum Lernen aus Daten, und anschließend die „Kostenfunktionale“, zur Bewertung der Schätzer für die Klassifikation bzw. Regression, behandelt.

Im Kapitel „Kerne“ werde ich im ersten Abschnitt auf die Motivation dieses Themas, diese im Kontext der Statistischen Lerntheorie zu betrachten, eingehen. Dann werde ich näher auf den Begriff „Kern“ eingehen und dabei die zwei verschiedenen Arten erklären. Dazu werden im zweiten Abschnitt „Kerne als Ähnlichkeitsmaße“ definiert, ihre Eigenschaften genannt und auf die Konstruktion solcher Kerne eingegangen. Der dritte Abschnitt ist im wesentlichen analog zum zweiten aufgebaut, wobei nun auf „Kerne als Nicht-Ähnlichkeitsmaße“ eingegangen wird.

Im dritten Kapitel „Kostenfunktionale“ beginnen wir in dem ersten Abschnitt mit den Grundlagen. Hierbei stellt sich uns die Frage wie man beurteilen kann, ob eine Klassifikation oder Regression „gut“ ist bzw. welche „Kosten“ eine „schlechtere“ Schätzung verursacht. Dies ist nicht zwangsläufig trivial, da verschiedene Aspekte, wie z.B. asymmetrisch verteilte Kosten, oder die Wahrscheinlichkeit für eine Fehlklassifizierung mit berücksichtigt werden können. Das werden wir im ersten Abschnitt klären. Im zweiten Abschnitt werden wir verschiedene Ansätze zur Minimierung der Kosten bzw. des Fehlers kennen lernen. Im dritten Abschnitt, betrachten wir die statistischen Konzepte des Kostenfunktionals, bevor ich in Abschnitt vier Möglichkeiten zur Bewertung von Schätzern vorstellen werde.

Das vierte Kapitel wird die Schlussfolgerungen mit einer kurzen Zusammenfassung der wichtigsten Aspekte meiner Seminararbeit enthalten.

2 Kerne

2.1 Motivation

Das Grundproblem der statistischen Lerntheorie besteht darin, dass man aus beobachteten Daten bzw. Messungen Zusammenhänge erschließen will. Dabei unterscheidet man im Wesentlichen zwischen Klassifikation und Regression. Bei der Klassifikation wird einer Funktion mit Hilfe von Beobachtungen gelehrt, wie sie neue Beobachtungen in die richtige Klasse einteilen kann. Bei der Regression hingegen wird eine reellwertige Funktion gesucht, die die Beobachtungen möglichst gut interpoliert. Dazu haben wir im Laufe dieses Seminars bereits verschiedene Klassifikation- bzw. Regressionsvrfahren kennen gelernt. Zum einen gibt es den linearen Ansatz, z.B. die lineare Regression oder die lineare Diskriminanzanalyse, dabei werden relativ einfache, lineare Algorithmen verwendet. Hierbei kann man jedoch keine sehr gute Anpassung an das gegebene Problem erreichen, das bedeutet, dass z.B. es zu Falschklassifizierung kommen kann. Um diese unerwünschten Ereignisse zu beheben, könnte man dieses Probleme auch mit einem nicht-linearen Ansatz, z.B. die quadratische Diskriminanzanalyse, lösen. So könnte man die Entscheidungslinien bei der Klassifikation bzw. die Regressionskurven genauer an die Daten anpassen. Das führt jedoch zu einer hohen algorithmischen Komplexität, was mit einer hohen Rechenintensität verbunden ist. Das zeigt uns, dass die uns bisher bekannten Ansätze zwar Vorteile, aber auch unerwünschte Nachteile mit sich bringen. Das Buch „Learning with Kernels“ von Schölkopf und Smola bietet dafür eine Lösung aus diesem Dilema - die Kerne. Aber warum Kerne? Die Kerne verbinden die Vorteile des linearen und nicht-linearen Ansatzes in dem sie nicht-lineare Abbildungen als lineare Abbildungen in höher dimensionalen Räumen, sog. Merkmalräumen beschreiben. Zusätzlich bieten Kerne den Vorteil, dass man bei deren Verwendung einfache, exakt lösbare Optimierungsverfahren verwenden kann. Desweiteren sind sie auch eine Verbesserung gegenüber den bereits in einer früheren Seminararbeit vorgestellten Neuronalen Netzen, da sie, bis auf triviale Fälle, nur approximative Lösungen liefern. Im folgenden Kapitel werde ich nun genauer auf die zwei unterschiedlichen Arten von Kernen, „Kerne als Ähnlichkeitsmaße“ und „Kerne als Nicht-Ähnlichkeitsmaße“ eingehen.

2.2 Kerne als Ähnlichkeitsmaße

Nun kommen wir zu der Frage, wie denn ein Kern definiert ist.

Sei $\Phi : \mathcal{X} \to \mathcal{H}$ eine Merkmalabbildung, die von der nicht-leeren Eingabemenge $\mathcal{X} \subset \mathbb{R}^N$ in den Merkmalraum $\mathcal{H}$ abbildet. Der Merkmalraum hat im Allgemeinen eine viel größere Dimension als die Eingabemenge. Die nachfolgenden Konstanten seien aus $\mathbb{K} = \mathbb{R}$ oder $\mathbb{C}$.

Definition 1 *Eine Abbildung $k : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$ heißt **Kern als Ähnlichkeitsmaß**, falls sie die Symmetrie-Eigenschaft*

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

und die positive Definitheit

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

erfüllt.

Zur Veranschaulichung werden wir uns nun ein konkretes Beispiel [2] betrachten.

Beispiel 1 *Unter der Annahme, dass k das Skalarprodukt von Φ in dem Merkmalraum $\mathcal{H}$ ist, können wir bei gegebener Merkmalabbildung den dazugehörigen Kern berechnen.*

Berechnung des Kerns:

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

Wie aus der Bezeichung „Kerne als Ähnlichkeitsmaße“ schon hervor geht, spiegeln diese Kerne die Ähnlichkeit zweier Objekte wider. Hierbei gilt, dass je größer der Wert des Ähnlichkeitsmaßes ist, desto ähnlicher sind die zu vergleichenden Objekte. Dies folgt aus der Cauchy-Schwarzschen Ungleichung für positiv definite Kerne.

2.2.1 Konstruktion eines Merkmalraumes

Mit Hilfe von Kernen lassen sich nicht-lineare Abbildungen als lineare Abbildungen in sog. Merkmalräumen darstellen.

Als möglichen Merkmalraum könnte man z.B. den Hilbertraum $L_2(\mathcal{X})$ wählen. Dabei ergibt sich jedoch das Problem, dass dieser Raum auch viele nicht-glatte Funktionen enthält. Deshalb wird der $L_2(\mathcal{X})$ auf eine kleinere Menge von Funktionen eingeschränkt, was dann ein sog. „Reproducing Kernel Hilbert Space“ ist.

Definition 2 *Ein Hilbertraum $\mathcal{H}$ mit Funktionen $f : \mathcal{X} \to \mathbb{R}$, Skalarprodukt $\langle \cdot, \cdot \rangle$ (und Norm $\|f\| := \sqrt{\langle f, f \rangle}$) heißt* ***Reproducing Kernel Hilbert Space****, falls ein Kern $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ existiert mit den folgenden Eigenschaften:*

1. *Reproduzierungseigenschaft:*
   $\langle k(., x), f(.) \rangle = f(x)$ *und* $\langle k(., x), k(., x') \rangle = k(x, x') \quad \forall x, x' \in \mathcal{X}$

2. *Abgeschlossenheit des Raumes:*
   k spannt $\mathcal{H}$ auf: $\mathcal{H} = \overline{span\{k(., x) | x \in \mathcal{X}\}}$

Wenn man einen Kern $k(x, x')$ und eine reproduzierende Merkmalabbildung $\phi : \mathcal{X} \to \mathbb{R}^{\mathcal{X}} = \{f : \mathcal{X} \to \mathbb{R}\}$, $\phi(.) = k(., x)$ gegeben hat, sind wir in der Lage einen Vektorraum zu konstruieren mit k als Skalarprodukt. Dazu gehen wir wie folgt vor:

Zu Beginn konstruieren wir einen Vektorraum, der alle Linearkombinationen von $k(x, x')$ enthält. Dazu sei für beliebige $m \in \mathbb{N}$, $\alpha_i \in \mathbb{R}$ und $x_i \in \mathcal{X}$ die Linearkombination $f(.) = \sum_{i=1}^{m} \alpha_i k(., x_i)$ gegeben. Danach definieren wir auf diesem Vektorraum ein Skalarprodukt von $f(.)$ mit $g(.) = \sum_{j=1}^{k} \beta_j k(., x'_j)$ und erhalten $\langle f, g \rangle := \sum_{i=1}^{m} \sum_{j=1}^{k} \alpha_i \beta_j k(x_i, x'_j)$. Zuletzt müssen wir noch überprüfen, ob es sich hierbei wirklich um ein Skalarprodukt handelt.

Zu zeigen: $\langle f, g \rangle := \sum_{i=1}^{m} \sum_{j=1}^{k} \alpha_i \beta_j k(x_i, x'_j)$ erfüllt die Eigenschaften eines Skalarproduktes.

Beweis: Die Eigenschaften wie Symmetrie oder Linearität sind offensichtlich. Deshalb bleibt nur noch zu zeigen, dass aus $\langle f, f \rangle = 0$ folgt $f = 0$. Dazu berechnen wir das Skalarprodukt von k und f, in dem wir die Definition von f einsetzen, die Reproduzierungseigenschaft ausnutzen und die

Cauchy-Schwarzsche Ungleichung für Kerne anwenden. Somit erhalten wir $\langle k(.,x), f(.)\rangle = \sum \alpha_i k(x_i,x) = f(x) \Rightarrow \langle k(.,x), k(.,x')\rangle = k(x,x')$ und damit $|f(x)|^2 = |\langle k(.,x), f(x)\rangle|^2 \leq k(x,x) \cdot \langle f(x), f(x)\rangle \Rightarrow 0 \leq |f(x)|^2 \leq k(x,x) \cdot 0 \quad \Rightarrow f(x) = 0$ was zu zeigen war.

2.2.2 Darstellung in einem Merkmalraum

Doch welche Bedingungen müssen nun gelten, damit man zu einem Kern einen Hilbertraum finden kann, in dem der Kern als Skalarprodukt darstellbar ist?

Diese Frage beantwortet das Theorem von Mercer [4]. Es liefert uns eine Darstellung des Reproducing Kernel Hilbert Space mit einer Standardbasis. So kann jeder stetige, symmetrische, positiv definite Kern als Skalarprodukt in einem höher-dimensionalen Raum ausgedrückt werden. Das Theorem von Mercer lautet wie folgt:

Theorem 1 *Sei μ ein endliches Maß und $k \in L_\infty(\mathcal{X}^2)$ ein symmetrischer, stetiger und reellwertiger Kern, 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 für alle $f \in L_2(\mathcal{X})$ gilt:

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

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

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

Hierbei wurde die Merkmalabbildung als $\Phi(x) = (\sqrt{\lambda_1}\psi_1(x), \ldots, \sqrt{\lambda_J}\psi_J(x))$ gewählt.

Eine Anwendung findet das Theorem von Mercer in dem sog. „Kerntrick“, welcher eine Methode ist, um lineare Algorithmen in nicht-lineare umzuwandeln. Beispiele dafür wären u.a. der Perceptron Algorithmus, Support Vector Machines oder die Hauptkomponentenanalyse.

2.2.3 Beispiele für Kerne als Ähnlichkeitsmaße

Zuletzt noch einige Beispiele für Kerne als Ähnlichkeitsmaße:

- 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 $\kappa, \vartheta > 0$

2.3 Kerne als Nicht-Ähnlichkeitsmaße

Eine weitere Art von Kernen sind die „Kerne als Nicht-Ähnlichkeitsmaß“. Auch diese wollen wir unter der Annahme, dass $\Phi : \mathcal{X} \to \mathcal{H}$ eine Merkmalabbildung ist, die von der nicht-leeren Eingabemenge $\mathcal{X} \subset \mathbb{R}^N$ in den Merkmalraum $\mathcal{H}$ abbildet. Die Konstanten seien wieder aus $\mathbb{K} = \mathbb{R}$ oder $\mathbb{C}$.

Definition 3 *Eine Abbildung $k : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$ heißt **Kern als Nicht-Ähnlichkeitsmaß**, falls sie die Symmetrie-Eigenschaft*

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

und die bedingte positive Definitheit

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

erfüllt.

Diese Definition zeigt, dass auch „Kerne als Nicht-Ähnlichkeitsmaße“ die Ähnlichkeit zweier Objekte ermittelt. Hier gilt jedoch, dass je kleiner der Wert des Nicht-Ähnlichkeitsmaßes, desto ähnlicher sind die verglichenen Objekte. Deshalb werden sie auch Unähnlichkeitsmaße oder Distanzmaße genannt.

Doch wieso definiert man sich zusätzlich bedingt positiv definite Kerne? Zum einen arbeiten manche Kernalgorithmen mit ihnen. Zum anderen bilden positiv definite Kerne nur eine kleinere Klasse von möglichen Kernen. Die bedingt positiv definiten Kerne bilden eine größere Klasse, da sie keine so strengen Voraussetzungen erfüllen müssen.

2.3.1 Konstruktion positiv definiter Kerne aus bedingt positiv definiten Kernen

Wir haben bereits in Abschnitt 2.2.2 gesehen, dass man zum Teil (z.B. für den Satz von Mercer) die Voraussetzung, dass der Kern positiv definit ist, benötigt. Es besteht die Möglichkeit einen positiv definiten Kern aus einem bedingt positiven Kern zu konstruieren.

Lemma 1 *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

Dieses Lemma führt zu einer 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.

Satz 1 *Sei k ein reellwertiger, bedingt positiv definiter 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:

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

2.3.2 Beispiele für Kerne als Nicht-Ähnlichkeitsmaße

Beispiele für Kerne als Nicht-Änlichkeitsmaße wären:

- 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: $k(x, x') = -\sqrt{\|x - x'\|^2 + c^2}$

3 Kostenfunktionale

3.1 Grundlagen

Wie uns bereits bekannt ist, ist das Ziel der Statistischen Lerntheorie eine möglichst gute Klassifikation bzw. Regression zu erhalten.

Doch wie kann man beurteilen, was „gut“ ist? Und warum ist diese Beurteilung nicht trivial?

Um zu beurteilen, wie gut unsere berechnete Entscheidungsregel bei der Klassifikation, bzw. die bei der Regression erhaltene Kurve ist, bestrafen wir Fehlklassifikationen bzw. Abweichungen von der Regressionskurve. Die so entstandenen „Kosten“ spiegeln wider, wie gut unsere vorhergegangenen Berechnungen waren. Um diese zu „verbessern“ können wir ein Optimierungsprogramm aufstellen, mit dessen Hilfe wir die Kosten bzw. den Fehler minimieren. Dabei ist es teilweise sinnvoll, dass man beachtet, dass die Kosten z.B. auch asymmetrisch sein können. Das bedeutet, dass unterschiedliche Falschklassifikation unterschiedlich „schlimm“, also gravierender, sein können. Es ist z.B. weniger „schlimm“, eine gesunde Blutspende als unrein zu kategorisieren, als verunreinigtes Blut zu transfusieren. Ebenso könnte noch die Wahrscheinlichkeit für eine bestimmte Falschklassifizierung berücksichtigt werden.

Kommen wir nun zu den Kosten. Diese werden mathematisch mit Hilfe des sog. Kostenfunktionals beschrieben.

Definition 4 *Sei ein Tupel* $(x, y, f(x)) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Y}$ *gegeben, wobei x die Lerndaten, y die Beobachtungen und f(x) die Vorhersagen für eine neue Beobachtung seien. Dann heißt die Funktion* $c : \mathcal{X} \times \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ ***Kostenfunktional****, wenn sie die Eigenschaft*

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

besitzt.

Das Kostenfunktional wird auch Verlust- oder Risikofunktion genannt. Diese Definition zeigt, dass nur falsche Vorhersagen bestraft werden. „Bessere“ Vorhersagen gibt es nicht, deshalb können die Kosten nicht negativ werden und somit bekommt man keine Belohnung für eine „besonders gute“ Vorhersage.

Die Kosten unterscheiden sich bei der Klassifikation und der Regression.

Wie zu Beginn dieses Kapitels schon erwähnt, entstehen durch Falschklassifikation Kosten.

Wenn wir keinen Unterschied zwischen den Klassen machen, also annehmen, dass es nicht mehr „kostet“ die Beobachtung falsch in Klasse 1 statt in Klasse 2 einzuordnen oder umgekehrt. Wir sprechen dann von binärer Klassifikation und das Kostenfunktional ist $\tilde{c}(x) \equiv 1$.

Sonst ist $\tilde{c}(yf(x))$ eine beliebige, nicht-lineare Funktion, wie z.B. das Soft Margin Kostenfunktional $\tilde{c}(yf(x)) = \max(0, 1 - yf(x))$ oder das Logistische Kostenfunktional $\tilde{c}(yf(x)) = \ln(1 + \exp(-yf(x)))$.

Bei der Regression entspricht die Schätzung der Differenz von $y - f(x)$ der Quantität der Falsch-Vorhersage. Hier sind die bekanntesten Kostenfunktionale das Quadratisches Kostenfunktional $\tilde{c}(y - f(x)) = (y - f(x))^2$ und das ϵ - unempfindliche Kostenfunktional $\tilde{c}(y - f(x)) = \max(|y - f(x)| - \epsilon, 0) = |y - f(x)|_\epsilon$

3.2 Testfehler und erwartetes Risiko

Unser Ziel ist nun eine Methode zu finden, um die erkannten Fehler zu minimieren. Wenn wir annehmen, dass die Daten (x,y) identisch, unabhängig und mit P(x,y) verteilt sind, gibt es zwei mögliche Fälle.

Der erste Fall wäre, dass wir Wissen über die Testdaten zur Trainingszeit besitzen. Dann müssen wir den erwarteten Fehler auf dieser speziellen Testmenge minimieren.

Im zweiten Fall haben wir kein Vorwissen über die Testdaten während der Trainingszeit. Deshalb müssen wir nun den erwarteten Fehler auf allen möglichen Testmengen minimieren.
Bei den folgenden Definitionen seien $\{x_1, \ldots, x_m\}$ die Trainingsdaten mit den Zielwerten $\{y_1, \ldots, y_m\}$. Die Vorhersagen werden mit $f(x'_i),\ i = 1, \ldots, k$ bezeichnet.

Definition 5 *Die Testdaten $\{x'_1, \ldots, x'_k\}$, aus denen die $f(x'_i),\ i = 1, \ldots, k$ vorhergesagt werden sollen, seien bekannt. Dann ist*

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

*der **Testfehler**.*

Das Problem bei der Berechung dieses Ausdruckes ist, dass er sehr rechenintensiv und kompliziert ist, da sehr viel Vorwissen, z.B. in der bedingten Verteilungsfunktion, vorhanden ist und in den Ausduck mit einfließt.

Wenn wir, wie im zweiten Fall, weniger über die Testdaten wissen vereinfacht das die Minimierung. Hier werden die erwarteten Kosten bezüglich der gemeinsamen Verteilung P und dem Kostenfunktional c minimiert.

Definition 6 *Das **erwartete Risiko** sei als*

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

definiert.

Wenn man diesen Ausdruck minimieren will, ist er jedoch nur lösbar, wenn man die Verteilungsfunktion $P(x, y)$ explizit kennt. Das kann man umgehen, wenn man die empirische Verteilungsfunktion einsetzt.

Definition 7

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

*wird das **empirische Risiko** genannt.*

Der Vorteil bei der Minimierung des empirischen Risikos ist, dass bei gegebenen Testdaten einfach eine Lösung berechnet werden kann.

Wie man in diesem Abschnitt erkennen konnte, spielt die Verteilung der Daten eine elementare Rolle. Diese werden wir im nächsten Abschnitt aus statistischer Sicht genauer beleuchten.

3.3 Statistische Konzepte des Kostenfunktionals

Statt wie bisher nur über das erwartete Risiko, wollen wir nun auch ein genaueres Wissen über die bedingte Verteilungsfunktion, bzw. die bedingte Dichte erlangen. Als Hilfsmittel dient uns die Maximum-Likelihood-Schätzung. Diese liefert uns eine Funktion f, die höchstwahrscheinlich die Daten erzeugt hat, in dem man die bedingte Dichte $p(y|x, f(x))$ bestimmt und bezüglich f maximiert.

Dazu wiederhole ich zu Beginn die Hauptaspekte der Maximum-Likelihood-Schätzung [4]. Das Ziel dieser Methode ist, einen Schätzer zu finden, der eine

möglichst gute Anpassung der Modellverteilung bzw. Verteilungsfunktion an die beobachteten Daten erreicht. Hierzu definiert man sich die Maximum-Likelihood-Funktion als

$$L((x,y),f) := \prod p(x_i,y_i|f) = \prod p(y_i|x_i,f)p(x_i).$$

Da Summen leichter als Produkte zu minimieren sind, verwenden wir die Log-Likelihood-Funktion

$$logL((x,y),f) := \sum \ln(p(y_i|x_i,f)p(x_i)) = \sum \ln p(y_i|x_i,f) + \sum \ln p(x_i).$$

Bei der der Logarithmus auf die Likelihood-Funktion angewendet wurde. Aus Statistik II wissen wir, dass diese beiden Funktionen an der selben Stelle ihr Minimum haben. Eine weitere Tatsache ist uns aus Operations Research bekannt. Bei der Optimierung ist es egal, ob wir das Maximum einer Funktion oder das Minimum dieser Funktion multipliziert mit minus Eins bestimmen. Das führt uns zu

$$\max logL((x,y),f) = -\min(\sum \ln p(y_i|x_i,f) + \sum \ln p(x_i)).$$

Wobei die letzte Summe eine Konstante bezüglich der Maximierung von f ist und deshalb weggelassen werden kann.

Diese Vorüberlegungen führen uns zu dem Kostenfunktional der Regression. Somit kommen wir zu folgendem Minimierungsproblem

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

wenn $c(x,y,f(x)) = -\ln(p(y|x,f(x))$. Ein eindeutiges Minimum existiert hier jedoch selten und nur unter der Voraussetzung, dass $f(x)$ beliebig, aber fest gewählt ist. In der Praxis ist man meist mit Störungen der Funkionen f konfrontiert. In diesem Fall kann das Kostenfunktional auch mit Hilfe des folgenden Ausdrucks

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

angepasst werden, wobei p_ξ die gestörte Dichte ist.

Ebenso gibt es ein Kostenfunktional für die Klassifizierung. Bei der

binären Klassifizierung ist $P(y|f(x))$ direkt berechenbar mit

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

In diesem Fall ist $y \in \{-1, 1\}$ und es folgt, dass $P(-1|f(x)) = 1 - P(1|f(x))$ gilt. Zusätzlich besteht auch die Möglichkeit $P(y|f(x))$ nach Bedaft zu wählen, z.B. als logistische Linkfunktion $c(x, y, f(x)) = \ln(1 + \exp(-f(x)))$. Im Allgemeinen kann das Kostenfunktional jedoch nicht beliebig gewählt werden.

3.4 Bewertung von Schätzern

Aus der Statistik stehen uns zahlreiche Eigenschaften von Schätzern zur Verfügung, die uns die Bewertung der Schätzer ermöglichen. Die Erwartungstreue, die uns einen unverzerrten Schätzer liefert, oder die Varianz, die die Schätzgenauigkeit widerspiegelt und durch die Ungleichung von Craner-Rao nach unten beschränkt ist. Desweiteren spielt in unserem Fall die Effizienz eines Schätzers noch eine Rolle.

Definition 8 *Die **Effizienz** eines Schätzer ist*

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

wobei I die Fisher-Informationsmatrix und B die Kovarianzmatrix von $\hat{\theta}(Y)$ ist.

Ein Beispiel für einen effizienten Schätzer wäre der Maximum-Likelihood-Schätzer (ML-Schätzer). Dieser ist jedoch nur asymptotisch effizient, d.h. für einen Stichprobenumfang der gegen unendlich geht, ist dieser Schätzer effizient. In der Praxis ist der Stichprobenumfang meist „kleiner“, deshalb gibt es bessere Schätzer als den ML-Schätzer. Zusätzlich kann eventuell die wahre Dichte nicht bekannt sein. Dadurch besteht die Möglichkeit eines großen Fehlers. Robuste Schätzer bieten eine Lösung dieses Problemes.

Der Hintergedanke bei den robusten Schätzern ist, dass man denjenigen Anteil von „schlechten“ Beobachtungen, sog. Ausreißer, die die Qualität der Schätzung beeinträchtigen, entfernt. Diese Überlegung führt zu der Idee von Huber. Dort werden robuste Schätzer konstruiert, die so modifiziert sind, damit der Einfluss von jedem einzelnen Muster begrenzt ist. Diese Idee ist

auch im sog. robusten Kostenfunktional von Huber

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

erkennbar.

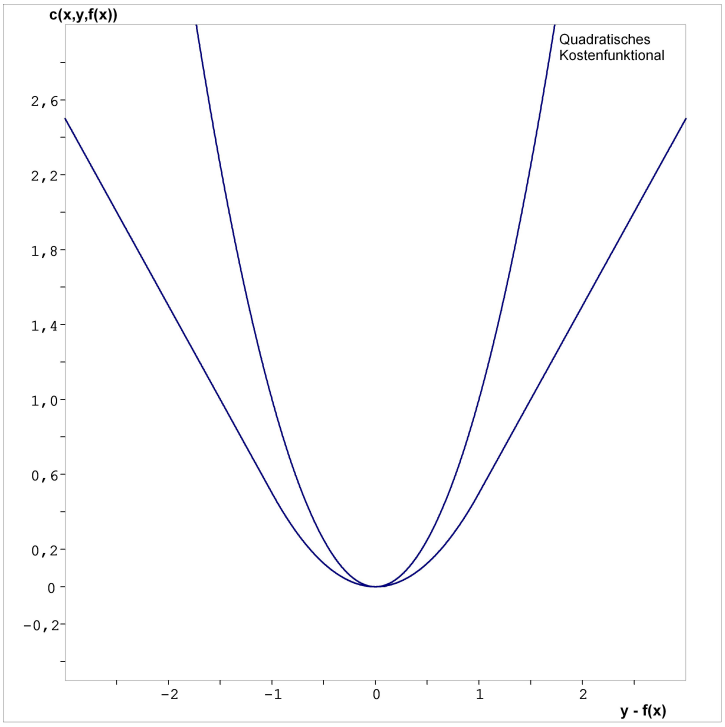


Abbildung 1: Robustes Kostenfunktional von Huber

Die Abbildung 1 vergleicht das quadratische Kostenfunktional mit dem Kostenfunktional von Huber. Zusätzlich zeigt sie, dass bei dem Kostenfunktionals von Huber innerhalb eines Toleranzbereiches $|y - f(x)| \leq \sigma$ die Kosten quadratisch und zusätzlich durch einen Faktor zwischen 0 und 1 geschwächt, von der Abweichung bzw. dem Fehler, abhängen. Außerhalb dieses Bereiches ist die Abhängigkeit und somit die Abweichung nur noch linear. Damit haben große, vereinzelte Ausreißer weniger Einfluss.

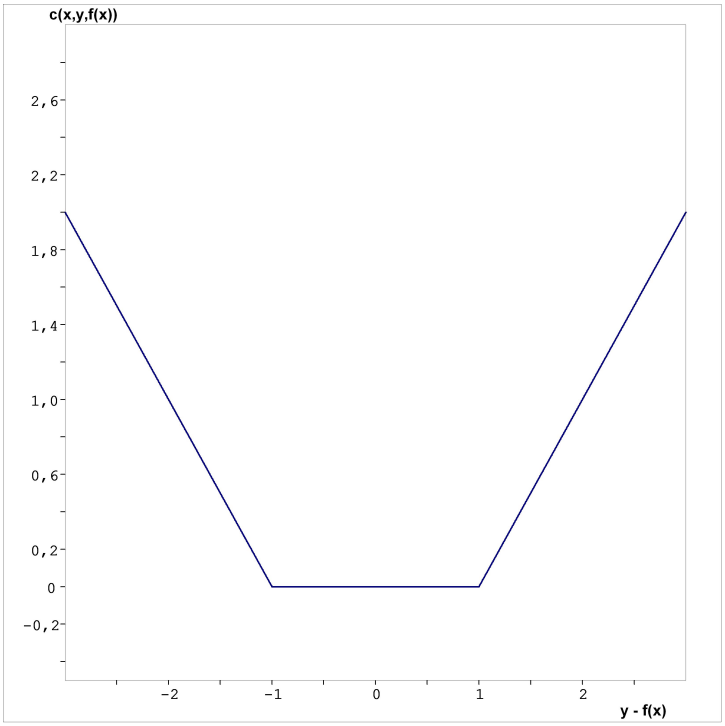


Abbildung 2: ϵ - unempfindliches Kostenfunktional

Ein weiteres robustes Kostenfunktional ist das ϵ - unempfindliche Kostenfunktional

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

Hier zeigt die Abbildung 2, dass „kleinere“ Abweichungen noch gar keine Kosten verursachen. „Klein“ wird hier über den Abstand zwischen Vorhersage und Zielwert, der kleiner als ϵ sein soll, definiert. Alle größeren Abweichungen sind dann linear zu den Kosten.

4 Schlussfolgerungen

Wir haben nun erfahren, dass Kerne, abhängig von ihrer Definition, ein Ähnlichkeitsmaß oder ein Nicht-Ähnlichkeitsmaße sein können. Somit ist durch diese Seminararbeit die Grundlage geschaffen, nun genauer darauf einzugehen, inwiefern man Kerne zur Klassifikation bzw. Regression verwenden kann. Hierzu gibt es in dem Buch „Learning with Kernels“ von Schölkopf und Smola im Abschnitt II „Support Vector Machines“, welche auf dem Prinzip der Kerne aufgebaut sind. Im Abschnitt III „Kernmethoden“ werden weitere Aspekte zur Konstruktion von Kernen und ihrer Verwendung, z.B. die Fisher-Diskriminanzanalyse mit Kernen oder die Bayesche Kernmethode, erörtert.

Der zweite Teil dieser Seminararbeit befasste sich mit den „Kostenfunktionalen“. Es wurden verschieden Ansätze zur Definition des Kostenfunktionals bzw. des Fehlers diskutiert. Dabei wurden verschiedene Kostenfunktionale vorgestellt und eine Anleitung zur Minimierung dieser Kostenfunktionale aufgezeigt. Hierbei konnte jedoch nur die Theorie umrissen werden, da ein Optimierungsproblem stark von der gegebenen Problemstellung abhängt.

Literatur

[1] B.Schölkopf, A. Smola, „Learning with Kernels“, MIT Press, 2002

[2] Vorlesungen von Prof. M. Pawlak, University of Manitoba, Kanada

[3] http://www.learning-with-kernels.org

[4] http://de.wikipedia.org/