Lernen mit Kernen

Alexander Johannes Smola

GMD FIRST
Rudower Chaussee 5
12489 Berlin
`smola@first.gmd.de`

Lernen mit Kernen stellt ein neues Paradigma im Bereich der statistischen Lerntheorie und des maschinellen Lernens dar. Während lineare Modelle oft nicht die erwünschte Flexibilität und Komplexität bieten, die zur korrekten Klassifikation bzw. Regression bei schwierigen Lernproblemen erforderlich sind, haben nichtlineare Modelle oft den Nachteil zu hoher algorithmischer Komplexität. Kernmethoden kombinieren die Vorteile der beiden Ansätze, indem sie *nichtlineare* Methoden als *linear* in bestimmten Hilberträumen, sogenannten *Merkmalsräumen* beschreiben.
Weiterhin ermöglicht die Verwendung von Verfahren der mathematischen Programmierung, vormals komplexe nichtlineare Minimierungsprobleme als sogenannte "quadratische" oder sogar "lineare Programme" zu formulieren, die sich exakt lösen lassen. Dies ist eine bedeutende Verbesserung gegenüber Neuronalen Netzen, die (mit Ausnahme trivialer Sonderfälle) nur approximative Lösungen erlauben. Dies stellt eine theoretisch fundierte Alternative zum de facto Standard gegenwärtig verfügbarer nichtlinearer Schätzer dar.
Ferner läßt sich die statistische Komplexität von Kernalgorithmen einfacher bestimmen, da es sich um mathematisch wohldefinierte Objekte handelt. Insbesondere die Verwendung moderner Methoden der Funktionalanalysis wie die metrische Entropie von Teilmengen von Banachräumen ermöglicht es, genaue Fehlerschranken zur uniformen Konvergenz statistischer Schätzer anzugeben.
Die gegenwärtige Arbeit soll einen Überblick über die beim Lernen mit Kernen verwendeten Konzepte und Algorithmen geben. Ziel ist es weniger, die genauen mathematischen Formalismen zu erläutern, als die zugrundeliegenden Ideen zu vermitteln.

1 Einleitung

Das Grundproblem statistischer Lerntheorie ist, aus Daten zu lernen, d.h. aus Beobachtungen und Messungen Zusammenhänge zu erschließen. Oder einfach gesagt, aus Beobachtungen X, möglicherweise gepaart mit Reaktionen Y des zu betrachtenden Systems, eine Aussage über neue Beobachtungen X' treffen zu können. Dies beinhaltet die Vorhersage darauffolgender Reaktionen Y' des Systems, oder möglicherweise auch nur die Bestimmung charakteristischer Merkmale von X' auf Basis von X.
Doch nun zu zwei wichtgen Bespielen: Klassifikation und Regression (siehe Abbildung 1). Ziel der Klassifikation ist es, aus Beobachtungen $(x_i, 1)$ oder $(x_i, -1)$ eine Funktion $f : x \to \{-1, 1\}$ zu lernen, die dann für (unbekannte) neue Beobachtungen x'_i das korrekte Ergebnis 1 (oder -1) liefert. Es könnte sich beispielsweise um die Erkennung handgeschriebener Zeichen [Sch97] oder auch die Klassifikation physikalischer Prozesse in einem Elementarteilchendetektor [VMSSR99] handeln.

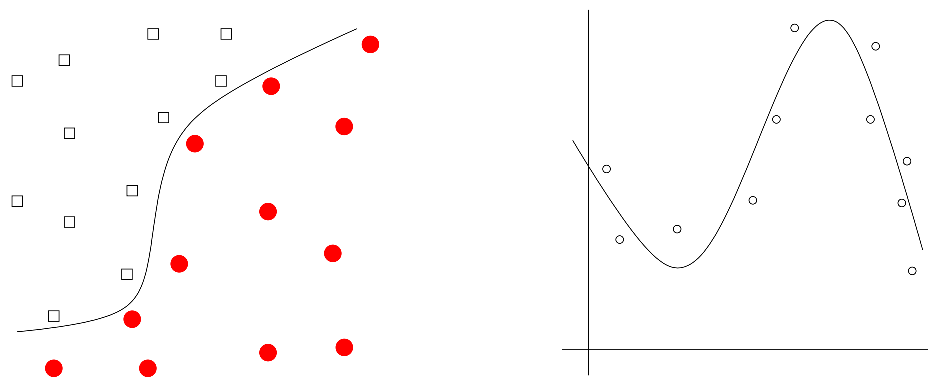

Abbildung 1: Links: Finde eine einfache Trennlinie, die Quadrate von kreisen trennt. Rechts: Finde eine einfache Kurve, die die Beobachtungen (x_i, y_i) gut interpoliert, gleichzeitig aber dem Rechnung trägt, daß die Beobachtungen fehlerbehaftet sein können.

Regression hingegen befaßt sich mit der Modellierung reellwertiger Funktionen. So könnte es sich bei x_i um diverse Marktindikatoren und bei y_i um den Börsenkurs einer bestimmten Aktie handeln.

Risikominimierung All diese Probleme lassen sich in einen gemeinsamen Rahmen fassen, den der Minimierung eines Risikofunktionals. Gegeben eine Kostenfunktion $c(f(x), y)$, die bestimmt, wie sehr Abweichungen zwischen den Vorhersagen $f(x)$ und dem wirklichen Reaktionen y des Systems zu 'bestrafen' sind, sowie eine Wahrscheinlichkeitsverteilung $p(x,y)$, die allen Beobachtungen zugrunde liegt, soll folgendes Funktional minimiert werden.

$$R[f] := \int_{X \times Y} c(f(x), y) p(x,y) dx dy \tag{1}$$

Bei Klassifikation ist $c(f(x), y) = 0$ für den Fall einer korrekten Vorhersage, d.h. falls $f(x) = \text{class}(x) = y$, und ansonsten 1. Bei Regression könnte es sich be $c(f(x), y)$ um den quadratischen Fehler $(f(x) - y)^2$, den absoluten Fehler $|f(x) - y|$ oder den absoluten Fehler mit ε Toleranz (also $\max(0, |f(x) - y| - \varepsilon)$ [Vap95]) handeln.

Gleichung (1) ist wohldefiniert, jedoch leider meist nicht explizit berechenbar. Dies liegt daran, daß fast immer $p(x,y)$ unbekannt ist und nur die Beobachtungen $\{(x_1, y_1), \ldots, (x_m, y_m)\}$ zur Verfügung stehen, die unabhängig und gleichverteilt von $p(x,y)$ erzeugt wurden. Gewöhnlich besteht der Ausweg dann darin, p durch die empirische Dichte $1/m \sum_i \delta_{x_i, y_i}(x,y)$ zu ersetzen, d.h. das Integral (1) nur an den Beobachtungen (x_i, y_i) auszuwerten und das sich daraus ergebende empirische Risikofunktional zu minimieren.

$$R_{\text{emp}}[f] := \frac{1}{m} \sum_{i=1}^{m} c(f(x_i), y_i) \tag{2}$$

Kapazitätskontrolle und Regularisierung Je nach Definition der Funktionenklasse $\mathcal{F}$, aus denen $f \in \mathcal{F}$ ausgewählt wird, um $R_{\text{emp}}[f]$ zu minimieren, handelt es sich um ein schlecht gestelltes inverses Problem [TA77]. Ferner garantiert eine Lösung $\hat{f}$ mit $R_{\text{emp}}[\hat{f}] = 0$ nicht, daß auch $R[\hat{f}]$ klein ist. Dies hängt in großem Maße von der Reichhaltigkeit von $\mathcal{F}$ ab (siehe Abbildung 2).

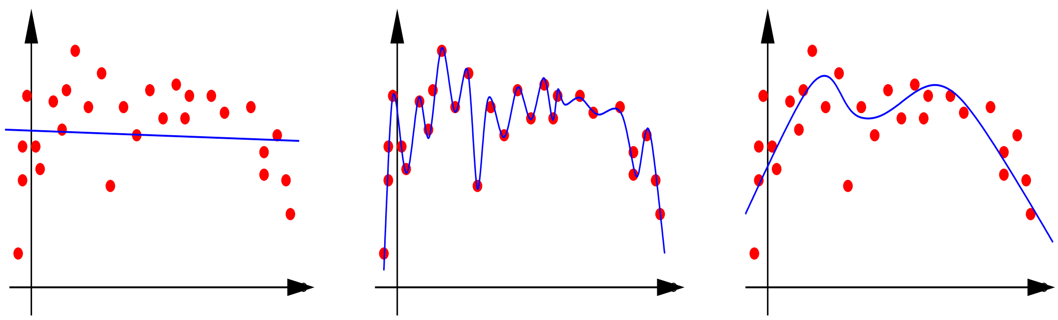

Abbildung 2: Links: zu einfaches Modell; Mitte: zu komplexes Modell; Rechts: optimale Komplexität

Die praktische Lösung des Dilemmas besteht darin, eine Variante von (2) zu minimieren, die bereits der Komplexität der Lösung Rechnung trägt. Dies führt zum regularisierten Risikofunktional

$$R_{\mathrm{reg}}[f] := R_{\mathrm{emp}}[f] + \lambda\Omega[f] = \frac{1}{m}\sum_{i=1}^{m} c(f(x_i), y_i) + \lambda\Omega[f]. \tag{3}$$

$\Omega[f]$ ist Stabilisierungsterm [TA77] und $\lambda > 0$ die Regularisierungskonstante. Je nachdem, wie groß λ gewählt wird, werden mehr oder weniger 'einfache' Funktionen favorisiert. Kapitel 3 zeigt Lösungen für dieses Problem auf.
Die Wahl von $\Omega[f]$ selbst beschreibt den Typ der Funktionenklasse. Die zwei wichtigsten Funktionale sind linear bzw. quadratisch. Erstere führen zu spärlichen Entwicklungen in ihren Basisfunktionen. Für Details in puncto lineare Regularisierer sei wegen der Kürze der Darstellung auf [Man65, CDS99, WGS+99, Smo98, Ben99] verwiesen. Positive quadratische Funktionale entsprechen Skalarprodukten in Hilberträumen. Hierbei kann es sich beispielsweise um Sobolevräume [SS98a], Hilberträume mit reproduzierendem Kern [KW71, GJP95, SSM98b, Gir98], oder allgemeine Merkmalsräume [ABR64, BGV92] handeln. Es hat sich herausgestellt, daß alle quadratischen Ansätze äquivalent [SSM98b, Gir98], und daß die statistischen Eigenschaften linearer und quadratischer Regularisierungsterme gleichwertig sind [Smo98, SWS98].

2 Algorithmen

Wie bereits eingangs erwähnt, besteht der Vorteil von Kernmethoden darin, daß sich Probleme als *linear* in bestimmten Merkmalsräumen beschreiben lassen. Aus diesem Grund werden die entsprechenden Algorithmen zuerst für den linearen Fall eingeführt.

Klassifikation Ziel ist es, die Beobachtungen (x_i, y_i) mit möglichst großer Konfidenz zu klassifizieren. Betrachtet man nun lineare Funktionen

$$f(x) = \langle w, x\rangle + b \text{ wobei } w, x \in X, b \in \mathbb{R} \tag{4}$$

und die zugehörigen Hyperebenen $\langle w, x\rangle + b = 0$, so entspricht hohe Konfidenz einem großen Abstand der Hyperebene von beiden Klassen. Dies kann man erreichen [VL63, CV95], indem man folgendes Optimierungsproblem löst.

$$\begin{array}{ll} \text{minimiere} & \frac{\lambda}{2}\|w\|^2 + \frac{1}{m}\sum_{i=1}^{m}\xi_i \\ \text{unter der Bedingung} & y_i(\langle w, x_i\rangle + b) \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{array} \tag{5}$$

Abbildung 3 erläutert diesen Sachverhalt. Der Term $\frac{1}{2}\|w\|^2$ entspricht aber nicht nur dem Kriterium eines möglichst großen Randes, sondern ist auch gleichzeitig ein quadratischer Regularisierungsterm. Dies bedeutet, daß hier die *flachste* lineare Funktion gesucht wird, die die Beobachtungen mit Genauigkeit ± 1 klassifiziert. Es ist nicht empfehlenswert, das Optimierungsproblem

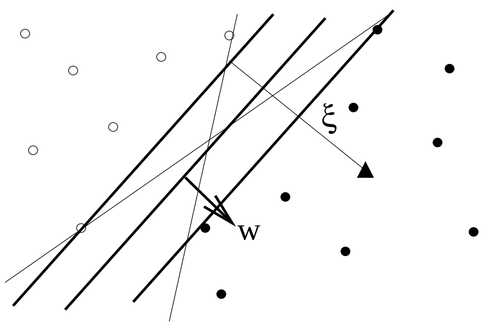

Abbildung 3: Kreise werden von Kugeln durch einen Hyperebene getrennt. Dabei ist die Breite der Region, die keine Punkte beinhaltet invers zur Länge von w, wenn man $y_i f(x_i) \geq 1$ annimmt. Punkte, die nicht korrekt klassifiziert werden können (Dreieck) tragen linear zum Gesamtfehler bei.

(5) direkt zu lösen, da es eine große Zahl linearer Randbedingungen enthält. Ferner erfordert es die *explizite* Berechnung von w, was, wie sich im folgenden herausstellen wird, bei nichtlinearen Funktionen teuer oder sogar unmöglich ist. Daher berechnet man das duale Optimierungsproblem [Fle89, CV95]

$$\begin{array}{ll} \text{minimiere} & \frac{1}{2}\sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle - \sum_{i=1}^{m} \alpha_i \\ \text{unter der Bedingung} & \alpha_i \in \left[0, \frac{1}{m\lambda}\right] \\ & \sum_{i=1}^{m} y_i \alpha_i = 0 \\ \text{wobei} & w = \sum_{i=1}^{m} \alpha_i y_i x_i \end{array} \tag{6}$$

Dies kann mit Standardmethoden des Operations Research [Van94] bzw. leichten Modifikationen [Joa99, Pla99, SS98a] gelöst werden.

Regression Eine ähnliche Situation ergibt sich bei Regressionsproblemen. Während eine allgemeine Formulierung in Form beliebiger konvexer Kostenfunktionen sowohl theoretisch als auch algorithmisch möglich ist [SSM98a] soll hier der Einfachheit halber nur auf die ε unempfindliche Kostenfunktion [Vap95] eingegangen werden. Wieder verwenden wir einen quadratischen Regularisierungsterm $Q[f] = \frac{1}{2}\|w\|^2$. Es ergibt sich folgendes Optimierungsproblem.

$$\begin{array}{ll} \text{minimiere} & \frac{\lambda}{2}\|w\|^2 + \frac{1}{m}\sum_{i=1}^{m}(\xi_i + \xi_i^*) \\ \text{unter der Bedingung} & y_i - \langle w, x_i \rangle - b \leq \varepsilon + \xi_i \\ & \langle w, x_i \rangle + b - y_i \leq \varepsilon + \xi_i^* \\ & \xi_i, \xi_i^* \geq 0 \end{array} \tag{7}$$

Das entsprechende duale Problem lautet

$$
\begin{array}{lll}
\text{minimiere} & \frac{1}{2}\sum_{i,j=1}^{m}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)\langle x_i,x_j\rangle & \\
& +\sum_{i=1}^{m}\alpha_i+\alpha_i^*+\sum_{i=1}^{m}y_i(\alpha_i-\alpha_i^*) & \\
\text{unter der Bedingung} & \alpha_i,\alpha_i^*\in\left[0,\frac{1}{m\lambda}\right] & (8)\\
& \sum_{i=1}^{m}(\alpha_i-\alpha_i^*)=0 & \\
\text{wobei} & w=\sum_{i=1}^{m}\alpha_i y_i x_i &
\end{array}
$$

Auch in diesem Fall sind Standardmethoden der quadratischen Optimierung ausreichend, um (8) zu lösen [SS98a].
Wichtig bei der Betrachtung der Lösung ist, daß die Beobachtungen x_i nur in Form von Skalarprodukten mit anderen Beobachtungen auftauchen. Ferner kann w als Linearkombination von x_i dargestellt werden. Zudem ist das Problem unabhängig von der Dimensionalität des Datenraums und besitzt als konvexes Minimierungsproblem ein *globales* Minimum.

Kerne Die bisher vorgestellten Verfahren haben jedoch den Nachteil, daß sie nur lineare Funktionen bestimmen können. Dies läßt sich dadurch umgehen, daß man die Beobachtungen x_i in einen Merkmalsraum F abbildet und dann lineare Funktionen in diesem Merkmalsraum konstruiert (siehe Abbildung 4). Dies ist möglich, da, wie bereits zuvor erwähnt, die Formulierung des Algo-

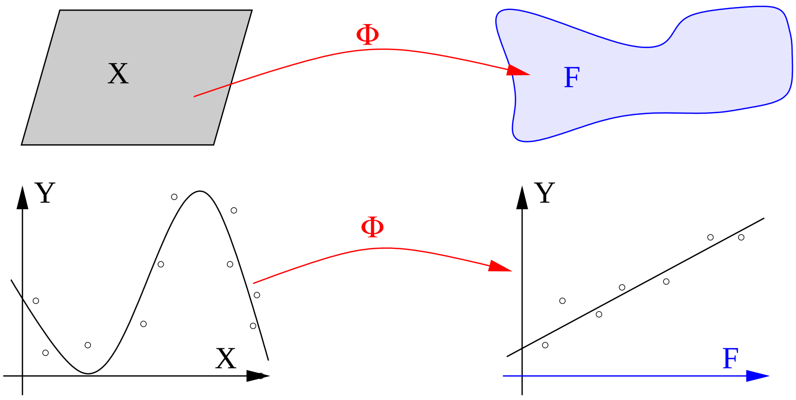

Abbildung 4: Oben: Die Beobachtungen x_i werden durch die Abbildung Φ in einen Merkmalsraum F abgebildet. Nichtlineare Funktionen im Eingaberaum werden dadurch zu linearen Funktionen in F.

rithmus nur von Skalarprodukten $\langle x_i,x_j\rangle$ abhängt. Die 'Nichtlinearisierung' erfolgt somit durch Ersetzen obiger Terme durch $\langle\Phi(x_i),\Phi(x_j)\rangle$.
In vielen Fällen ist es allerdings numerisch zu teuer, die Abbildung Φ in den Merkmalsraum explizit zu berechnen. Beispielsweise gibt es ca. $2.5\cdot 10^{12}$ verschiedene Monome fünften Grades bei einem $28\cdot 28 = 784$ dimensionalen

Eingaberaum. Letzteres ist eine realistische Annahme — die zur Handschrifterkennung eingesetzten Bilder besitzen eine Auflösung von 28×28 Pixels. Der Ausweg besteht darin, die Skalarprodukte $\langle \Phi(x_i), \Phi(x_j) \rangle$ nur *implizit* zu berechnen [ABR64, BGV92] — man verwendet Funktionen

$$k(x, x') := \langle \Phi(x), \Phi(x') \rangle. \tag{9}$$

Es zeigt sich beispielsweise, daß Skalarprodukte zwischen *sämtlichen* Monomen der Ordnung p sich durch $k(x, x') := \langle x, x' \rangle^p$ darstellen lassen. Inhomogene Polynome erzeugt $k(x, x') = (\langle x, x' \rangle + 1)^p$. Nachdem auch w als Linearkombination der abgebildeten Daten im Merkmalsraum dargestellt werden kann erhält man (im Falle der Regression)

$$f(x) = \langle w, \Phi(x) \rangle + b = \sum_{i=1}^{m} (\alpha_i - \alpha_i^*) k(x_i, x) + b. \tag{10}$$

Analog können auch die Gleichungen (6) und (8) unter Verwendung von Kernfunktionen umformuliert werden.

Regularisierung Es stellt sich die Frage, inwieweit beliebige symmetrische Funktionen $k(x, x')$ als Kerne herangezogen werden können. Boser et al. [BGV92] haben gezeigt, daß jede Funktion k, die Mercers Bedingung erfüllt, als Skalarprodukt in einem Merkmalsraum F dargestellt werden kann.

Theorem 1 (Mercer, 1909) *Sei X kompakt, $k : X \times X \to \mathbb{R}$ mit $k \in (L_2(X) \times L_2(X)) \cap (\mathcal{C}^0(X) \times \mathcal{C}^0(X))$. Sofern für alle $f \in L_2(X)$ gilt, daß $\int_{X \times X} k(x, x') f(x) f(x') dx dx' \geq 0$, gibt es eine Zerlegung von $k(x, x')$ in ein Eigensystem mit positiven Eigenwerten λ_i und Eigenfunktionen $\phi_i(\cdot)$, d.h. $k(x, x') = \sum_i \lambda_i \phi_i(x) \phi_i(x')$. Ferner konvergiert diese Reihe gleichmäßig, die Funktionen $\phi_i(x)$ sind durch eine Konstante C beschränkt und $(\lambda_i) \in \ell_1$.*

Daraus läßt sich explizit die Abbildung in den Merkmalsraum konstruieren. Funktionen wie $k(x, x') = e^{-\|x - x'\|^2}$ oder $k(x, x') = e^{-\|x - x'\|}$ sind zulässige Kerne, da sie Mercers Bedingung erfüllen. Die Berechnung des Eigensystems von k in der Praxis gestaltet sich jedoch oft schwierig, ebenso auch der Beweis, daß die Bedingung gilt.

Die Einführung von Regularisierungsoperatoren erlaubt es dann, den Zusammenhang zwischen Merkmalsräumen und Kernen explizit herzustellen. Es zeigt sich [SSM98b], daß die Minimierung von $\frac{1}{2}\|w\|^2$ im Merkmalsraum der Minimierung von $\frac{1}{2}\|Pf\|^2$ im Eingaberaum entspricht, sofern

$$k(x, x') = \langle Pk(x, \cdot), Pk(x', \cdot) \rangle. \tag{11}$$

Hierbei ist P ein Operator, der f in einen Skalarproduktraum abbildet. Beispielsweise nimmt der zu $\exp(-\|x-x'\|)$ gehörige Operator folgende Form an: $\|Pf\|^2 = \int |f(x)|^2 + \|\nabla f(x)\|^2 dx$. Damit erzwingt P Glattheit der Funktion f im Eingaberaum.

3 Fehlerschranken und Modellselektion

Die Vielzahl verfügbarer Kerne, unterschiedlicher Modellparameter, Regularisierungskonstanten, etc. läßt die Frage nach Methoden aufkommen, diese Parameter optimal einstellen zu können. Naiverweise könnte man die Trainingsdaten dazu heranziehen.
Damit liefe man aber Gefahr, schlechte Generalisierungsfähigkeit durch "Overfitting" zu erhalten (indem man nicht nur die eigentlichen Trainings-, sondern auch die Modellparameter lernt). Beispielsweise würde ein zu kleiner Wert für λ zu $R_{\text{emp}}[f] = 0$ führen. Der Versuch, die Modellparameter auf einer separaten Validierungsmenge einzustellen führt zu ähnlichen Problemen: wird diese Menge zu klein gewählt, oder handelt es sich um zu viele Variablen, so kann dies ebenfalls die Wahl eines zu komplexen Modells zur Folge haben (man trainiert auf Trainings- und Validierungsmenge).

Uniforme Konvergenz Intuitiv würde man allerdings erwarten, daß mit ansteigender Zahl der Trainingsbeispiele sich der Trainingsfehler dem wahren Fehler nähert. Dies gilt, sofern die Komplexität der Funktionsklasse $\mathcal{F}$, aus denen f gewählt werden darf, beschränkt ist. Hoeffdings Theorem [Hoe63] ist der Ausgangspunkt solcher Abschätzungen. Es besagt, daß die empirischen Mittelwerte von Zufallsvariablen exponentiell gegen ihren Erwartungswert konvergieren.

Theorem 2 (Hoeffding, 1963) *Seien $\xi_1, \ldots, \xi_m$ unabhängige Zufallsvariablen, denen dieselbe Wahrscheinlichkeitsverteilung zugrundeliegt und die mit Wahrscheinlichkeit 1 ins Intervall $[0,1]$ fallen. Dann gilt für beliebige $\varepsilon > 0$*

$$\Pr\{|S_m - E[S_m]| \geq \varepsilon\} \leq 2e^{-2m\varepsilon^2} \tag{12}$$

wobei $S_m := \frac{1}{m}\sum_{i=1}^m \xi_i$ das empirische Mittel und $E[\cdot]$ den Erwartungswert einer Zufallsvariable darstellen.

Vapnik und Chervonenkis [VC71] zeigten, daß eine ähnliche Aussage auch für die Konvergenz des empirischen zum erwarteten Risiko gilt. Allerdings muß der Tatsache Rechnung getragen werden, daß es sich um beliebige Funktionen aus einer ganzen Funktionenklasse $\mathcal{F}$ handeln kann (aus denen man dann typischerweise das f mit minimalem $R_{\text{emp}}[f]$ auswählt). In der Folgezeit wurden diese Absätzungen verfeinert und an erweitert. Daher soll nur die allgemeine Form einer solchen Fehlerschranke angegeben werden. Für Details siehe [DGL96, AB99].

$$\Pr\left\{\sup_{f\in\mathcal{F}} |R_{\text{emp}}[f] - R[f]| > \varepsilon\right\} \leq c_1(m) E\left[\mathcal{N}(\varepsilon, \mathcal{F}, \ell_p^m)\right] e^{-\varepsilon^\beta m/c_2}. \qquad (13)$$

Der Term $c_1(m)$ ist linear oder konstant in m, $c_2 > 0$ und $\beta \in \{1,2\}$ je nach der betrachteten Situation. $\mathcal{N}(\varepsilon, \mathcal{F}, \ell_p^m)$ schließlich ist die sogenannte Überdeckungszahl der Funktionenklasse. Sie ist ein Maß dafür, wieviele Funktionen man in der ℓ_p^m Metrik benötigt, um $\mathcal{F}$ mit ε Genauigkeit zu approximieren. Mit andern Worten — $\mathcal{F}$ wird durch eine endliche Menge ersetzt, für deren einzelne Elemente dann Theorem 2 zur Anwendung kommt.
Aus (13) ersieht man, daß die Konfidenzaussage wesentlich von der Wahl der Funktionenklasse $\mathcal{F}$ *und* der Güte der Abschätzung der Überdeckungszahl abhängt.

Bestimmung der Modellkomplexität Der von Vapnik und Chervonenkis beschrittene Weg zur Ermittlung einer Obergrenze von $\mathcal{N}$ führt über die Berechnung der sogenannten *VC-Dimension*. Dies ist die maximale Zahl von Elementen, die durch Funktionen aus $\mathcal{F}$ in beliebiger Weise getrennt werden können (d.h. $f \in \mathcal{F}$ beliebige Vorzeichen annehmen kann). Während dieser Ansatz theoretisch vielversprechend ist und zu sehr eleganten Beweisen führt, ist es in der Praxis oft sehr schwierig, die VC-Dimension explizit abzuschätzen. Ferner ist die VC-Dimension nicht skalenabhängig. Das bedeutet, daß Funktionenklassen, die nur auf sehr feiner Skala komplex erscheinen (während sie auf grober Skala sehr einfach sein mögen), insgesamt als sehr komplex beurteilt werden, ohne zu betrachten, ob die Komplexität überhaupt relevant ist. Die *direkte* Abschätzung der Überdeckungszahlen wurde als zu schwierig angesehen.
Erst durch den Einsatz funktionalanalytischer Methoden [Mau81, CS90, WSS98, STW99, SSTSW99] gelang es, Obergrenzen für $\mathcal{N}$ direkt anzugeben. Dies trifft zumindest auf Kernfunktionen zu, da diese als linear in Merkmalsräumen dargestellt, und somit durch *lineare* Operatoren beschrieben werden können. Insbesondere gelang es, die Abschätzungen in vielen Fällen dramatisch zu verbessern, indem dem Effekt der Kerne Rechnung getragen wurde. Bei bestimmten Kernen wurde anstelle einer polynomialen eine logarithmische Skala erreicht [WSS98, Smo98].

Eine detaillierte Analyse der dabei verwendeten Techniken würde zuweit führen, da moderne Methoden der Theorie der Banachräume und zugehöriger Operatoren benötigt werden. Es stellt sich jedoch heraus, daß sich die entsprechenden Theoreme leicht auf allgemeine lineare Modelle und damit auf eine große Zahl wichtiger statistischer Schätzverfahren ausdehnen lassen können.

4 Ausblick

Dieser Artikel konnte nur einen kleinen Einblick in moderne Methoden der statistischen Lerntheorie bieten. Als weitergehende Literatur sei auf [Bur98, SS98b, Vap95] verwiesen. Des weiteren bietet die Support Vector Homepage

`http://svm.first.gmd.de/`

Information und Links zu Veröffentlichungen auf dem Gebiet der Kernmethoden. Es stellt sich natürlich die Frage nach weiteren Entwicklungen und aktuellen Problemen. Wichtig erscheinen mir die folgenden Punkte:

– Anwendung der bei Kernmethoden entwickelten Techniken auf einen weiteren Bereich statistischer Verfahren (unüberwachtes Lernen, Reinforcement Learning, etc.).

– Verbesserung der Abschätzungen (teilweise sind nur die Konvergenz*raten* optimal, die Konstanten allerdings zu groß).

– Einfache Abschätzungen, die auch von Praktikern ohne großes Vorwissen verwendet werden können, als Ersatz für bisherige Verfahren.

– Erweiterung der funktionalanalytischen Methoden durch Resultate der statistischen Lerntheorie.

Literatur

[AB99] M. Anthony and P. Bartlett. *A Theory of Learning in Artificial Neural Networks.* Cambridge University Press, 1999.

[ABR64] M. A. Aizerman, E. M. Braverman, and L. I. Rozonoér. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–837, 1964.

[Ben99] K. Bennett. Combining support vector and mathematical programming methods for induction. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods - SV Learning*, pages 307–326, Cambridge, MA, 1999. MIT Press.

[BGV92] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, *5th Annual ACM Workshop on COLT*, pages 144–152, Pittsburgh, PA, 1992. ACM Press.

[Bur98] C. J. C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Knowledge Discovery and Data Mining*, 1998. in press.

[CDS99] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *Siam Journal of Scientific Computing*, 20(1):33–61, 1999.

[CS90] B. Carl and I. Stephani. *Entropy, compactness, and the approximation of operators.* Cambridge University Press, Cambridge, UK, 1990.

[CV95] C. Cortes and V. Vapnik. Support vector networks. *M. Learning*, 20:273 – 297, 1995.

[DGL96] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Number 31 in Applications of mathematics. Springer, New York, 1996.

[Fle89] R. Fletcher. *Practical Methods of Optimization.* John Wiley & Sons, New York, 1989.

[Gir98] F. Girosi. An equivalence between sparse approximation and support vector machines. *Neural Computation*, 10(6):1455–1480, 1998.

[GJP95] F. Girosi, M. Jones, and T. Poggio. Regularization theory and neural networks architectures. *Neural Computation*, 7:219–269, 1995.

[Hoe63] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[Joa99] T. Joachims. Making large–scale SVM learning practical. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods — Support Vector Learning*, pages 169–184, Cambridge, MA, 1999. MIT Press.

[KW71] G. S. Kimeldorf and G. Wahba. A correspondence between Bayesan estimation on stochastic processes and smoothing by splines. *Ann. Math. Statist.*, 2:495–502, 1971.

[Man65] O. L. Mangasarian. Linear and nonlinear separation of patterns by linear programming. *Operations Research*, 13:444–452, 1965.

[Mau81] B. Maurey. In: "Remarques sur un resultat non publiè de B. Maurey" by G. Pisier. In Centre de Mathematique, editor, *Seminarie d'analyse fonctionelle 1980–1981*, Palaiseau, 1981.

[Pla99] J. Platt. Fast training of support vector machines using sequential minimal optimization. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods — Support Vector Learning*, pages 185–208, Cambridge, MA, 1999. MIT Press.

[Sch97] B. Schölkopf. *Support Vector Learning.* R. Oldenbourg Verlag, Munich, 1997.

[Smo98] A. J. Smola. *Learning with Kernels.* PhD thesis, Technische Universität Berlin, 1998.

[SS98a] A. J. Smola and B. Schölkopf. From regularization operators to support vector kernels. In *Advances in Neural information processings systems 10*, pages 343–349, San Mateo, CA, 1998.

[SS98b] A. J. Smola and B. Schölkopf. A tutorial on support vector regression. NeuroCOLT Technical Report NC-TR-98-030, Royal Holloway College, University of London, UK, 1998.

[SSM98a] A. Smola, B. Schölkopf, and K.-R. Müller. Convex cost functions for support vector regression. In L. Niklasson, M. Bodén, and T. Ziemke, editors, *Proceedings of the 8th International Conference on Artificial Neural Networks*, Perspectives in Neural Computing, Berlin, 1998. Springer Verlag.

[SSM98b] A. J. Smola, B. Schölkopf, and K.-R. Müller. The connection between regularization operators and support vector kernels. *Neural Networks*, 11:637–649, 1998.

[SSTSW99] B. Schölkopf, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Generalization bounds via eigenvalues of the gram matrix. Technical Report NC-TR-99-035, NeuroColt2, University of London, UK, 1999.

[STW99] J. Shawe-Taylor and R. C. Williamson. Generalization performance of classifiers in terms of observed covering numbers. In *Proc. EUROCOLT'99*, 1999.

[SWS98] A. J. Smola, R. C. Williamson, and B. Schölkopf. Generalization bounds for convex combinations of kernel functions. NeuroCOLT Technical Report NC-TR-98-022, Royal Holloway College, University of London, UK, 1998.

[TA77] A. N. Tikhonov and V. Y. Arsenin. *Solution of Ill–Posed Problems.* Winston, Washington, DC, 1977.

[Van94] R. J. Vanderbei. LOQO: An interior point code for quadratic programming. TR SOR-94-15, Statistics and Operations Research, Princeton Univ., NJ, 1994.

[Vap95] V. Vapnik. *The Nature of Statistical Learning Theory.* Springer, N.Y., 1995.

[VC71] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probab. and its Applications*, 16(2):264–280, 1971.

[VL63] V. Vapnik and A. Lerner. Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24, 1963.

[VMSSR99] P. Vannerem, K.-R. Müller, A.J. Smola, and S. Söldner-Rembold. Classifying LEP data with support vector algorithms. In *Proceedings of AIHENP'99.* Elsevier, 1999.

[WGS+99] J. Weston, A. Gammerman, M. Stitson, V. Vapnik, V. Vovk, and C. Watkins. Support vector density estimation. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods — Support Vector Learning*, pages 293–306, Cambridge, MA, 1999. MIT Press.

[WSS98] R. C. Williamson, A. J. Smola, and B. Schölkopf. Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. NeuroCOLT NC-TR-98-019, Royal Holloway College, 1998.

Alexander Johannes Smola, geboren am 7. Mai 1971, erhielt 1996 das Diplom in Physik (TU München). Während des Studiums verbrachte er jeweils ein Jahr bei AT&T Bell Laboratories und am Collegio Ghislieri (Pavia). Mit einer bei GMD FIRST sowie während Gastaufenthalten an der Australian National University angefertigten Arbeit über Algorithmen und Generalisierungsschranken für das Lernen mit Kernen promovierte er 1998 in Informatik (TU Berlin). Er ist Stipendiat der Stiftung Maximilianeum und der Studienstiftung des Deutschen Volkes.