DISS. ETH NO. 17690

Fast rates of convergence for adaptive classification

A dissertation submitted to
ETH ZURICH

for the degree of
Doctor of Sciences

presented by
BERNADETTA TARIGAN
MSc., Institut Teknologi Bandung
born September 26, 1972
citizen of Indonesia

accepted on the recomendation of
Prof. Dr. Sara A. van de Geer, examiner
Prof. Dr. Peter Bühlmann, co-examiner
Dr. Gilles Blanchard, co-examiner

2008

Abstract

Classification refers to the problem of predicting the category of an observation, based on the categories of previously observed examples and with as small an error as possible. A predictor is a function that assigns a data value to one out of a fixed number of mutually exclusive categories, called classes. Determining a suitable predictor is a statistical learning problem, since the properties of the data source are not known explicitly but have to be inferred from examples.

Let $x \in \mathcal{X}$ denote an observation and $y \in \mathcal{Y} = \{1, \ldots, m\}$ denote a category. The input-output pair (x, y) is a realization of a random pair (X, Y) governed by a joint probability distribution P on the space $\mathcal{X} \times \mathcal{Y}$. The assumption in the statistical learning is that the examples $\{(x_i, y_i)\}_{i=1}^n$ are drawn independently from the same distribution as (x, y). A predictor $g : \mathcal{X} \to \mathcal{Y}$ makes an error when $g(x) \neq y$. Under the 0–1 loss, the prediction error of g is $\widetilde{R}(g) := P(g(X) \neq Y)$. This prediction error is also called the standard risk or the true risk. Based on a given set of examples $D_n = (X_i, Y_i)_{i=1}^n$, the empirical risk minimization (ERM) method looks for the predictor $\hat{g}_n(X) = \hat{g}_n(X, D_n)$ that minimizes the empirical prediction error $(1/n) \sum_{i=1}^n \mathbb{1}(g(X_i) \neq Y_i)$ over a class $\mathcal{G}$ of candidate predictors. There exists a best theoretical predictor, the so-called Bayes predictor g^*, that has the smallest achievable risk, $\widetilde{R}(g^*)$. The performance of (a sequence of) predictors is now measured by the rates of convergence to zero of the excess prediction errors $\widetilde{R}(\hat{g}_n) - \widetilde{R}(g^*)$.

Recent results of statistical learning apply empirical process theory and concentration inequalities to show that convergence rates depend on the complexity condition of the class of candidate predictors $\mathcal{G}$ and the so-called *margin condition* or *noise condition*. To handle the complexity, we can apply either a regularized ERM-based method by defining the appropriate penalty for the class under consideration, or a complexity constraint on the class in terms of its entropy, parameterized by a constant $\rho \in (0, 1)$ (smaller ρ means simpler class).

The margin condition quantifies the identifiability of the Bayes predictor, and is parameterized by a constant $\kappa \geq 1$. When the underlying distributions behave well (low noise level), it places a small probability around the Bayes decision boundaries (κ small). The convergence rates obtained are functions of the unknown parameters ρ and κ.

Under the conditions above and for the binary case ($m = 2$), the rates of convergence to zero of $\widetilde{R}(\hat{g}_n) - \widetilde{R}(g^*)$ are adaptive to the unknown parameters and faster than the classical non-parametric rate $n^{-1/2}$. However, since the true 0–1 loss is a non-convex function, the ERM method is computationally infeasible. We replace the 0–1 loss with a convex upper bound surrogate loss $l(g(X), Y)$ so that the method can be implemented. To obtain fast convergence rates, two questions of concern are: (a) whether the Bayes predictor g^* also minimizes the surrogate l-risk $R(g) := \mathbb{E}[l(g(X), Y)]$ over all possible predictors (Bayes consistency of l), and (b) whether the minimization of the excess l-risk $R(\hat{g}_n) - R(g^*)$ of the predictors obtained implies to the minimization of the excess prediction error $\widetilde{R}(\hat{g}_n) - \widetilde{R}(g^*)$.

This thesis obtains fast convergence rates of the excess risk with respect to the hinge loss as the surrogate that adapt to the unknown margin and complexity parameters. These results help to statistically explain the practical efficiency of the support vector machine (SVM) algorithms that use the hinge loss. Furthermore, this thesis analyses the so-called classification with reject option, where we allow a predictor to reject taking a decision on the categories if the observation is too hard to classify. Under suitable margin condition and complexity constraint, we obtain fast convergence rates of the excess true risk.

In this work we do not assume that the model class contains the Bayes predictor. The results we obtain are in terms of the approximation error $\inf_{g \in \mathcal{G}} R(g) - R(g^*)$.

For the binary problem with hinge loss, we consider classifiers that are linear combinations of base functions. Instead of an ℓ_2-penalty, which is used by the SVM, we put an ℓ_1-penalty on the coefficients. Under certain conditions on the base functions, hinge loss with this complexity penalty is shown to lead to an oracle inequality involving both model complexity and margin.

While statistical properties of binary classifiers are quite well understood, their extensions to multicategory cases ($m > 2$) are not trivial and certainly more involved. However, the so-called multi-hinge loss—an extension of binary hinge loss that considers all of the categories at once—has been shown to be Bayes consistent. Furthermore, the convergence to zero (in probability) of the

excess multi-hinge risk implies the convergence to zero with the same rate (in probability) of the excess prediction error. In this thesis we show a moment bound for the so-called multi-hinge loss minimizers based on two kinds of complexity constraints: entropy with bracketing and empirical entropy. Obtaining such a result based on the latter is harder than finding one based on the former. We obtain fast rates of convergence that adapt to the unknown margin and complexity parameters, that is, $n^{-\kappa/(2\kappa-1+\rho)}$.

The reject option can improve performance in applications for which the cost of rejecting certain samples, and handling them with different procedures (for example, manual classification), is not larger than the cost of misclassifying. We can think of embedding the reject option as adding the rejection category into the output space. We consider the case in which acceptable misclassification is given as a parameter α (α-*reject loss*). Based on this reject loss, we investigate the margin condition and impose a complexity constraint on the class of predictors that lead the fast rates of convergence for the excess true risk.

Zusammenfassung

Der Begriff Klassifikation beschreibt das Problem, die Kategorie einer Beobachtung vorherzusagen, basierend auf den Kategorien gegebener Beobachtungsbeispiele. Dabei soll der Erwartungswert des mit einem Fehler verknüpften Verlusts so klein wie möglich sein. Ein Prädiktor ist eine Funktion, welche einen Datenpunkt einer von mehreren möglichen, paarweise disjunkten Kategorien zuordnet. Diese Kategorien werden als Klassen bezeichnet.

Sei $x \in \mathcal{X}$ eine Beobachtung und $y \in \mathcal{Y} = \{1, \ldots, m\}$ ein Kategorie-Bezeichner. Das Paar (x, y) ist eine Realisierung des Paares (X, Y) von Zufallsvariablen mit gemeinsamer Verteilung P auf dem Raum $\mathcal{X} \times \mathcal{Y}$. In der statistischen Lerntheorie werden die Beispiele $\{(x_i, y_i)\}_{i=1}^n$ als unabhängige Züge von der gemeinsamen Verteilung P vorausgesetzt. Ein Prädiktor $g : \mathcal{X} \to \mathcal{Y}$ erzeugt einen Fehler falls $g(x) \neq y$. Unter 0–1-Verlust ist der Vorhersagefehler von g gegeben durch $\widetilde{R}(g) := P(g(X) \neq Y)$. Dieser Vorhersagefehler wird auch als Standardrisiko oder wahres Risiko bezeichnet. Die Methode der empirischen Risiko-Minimierung (ERM) bestimmt einen Prädiktor $\hat{g}_n(X) = \hat{g}_n(X, D_n)$ welcher, gegeben eine Menge $D_n = (X_i, Y_i)_{i=1}^n$ von Beispielen, den empirischen Vorhersagefehler $(1/n) \sum_{i=1}^n \mathbb{1}(g(X_i) \neq Y_i)$ über eine Klasse $\mathcal{G}$ von möglichen Prädiktoren minimiert. Es existiert ein bester theoretischer Pädiktor, der sogenannte Bayes-Prädiktor g^*, welcher durch kleinstmögliches Risiko $\widetilde{R}(g^*)$ gekennzeichnet ist. Um die Qualität eines Prädiktors (bzw. einer Folge von Prädiktoren) zu messen, wird die Konvergenz der Risikodifferenz $\widetilde{R}(\hat{g}_n) - \widetilde{R}(g^*)$ gegen Null betrachtet.

Neuere Resultate der statistischen Lerntheorie verwenden die Theorie empirischer Prozesse und Konzentrations-Ungleichungen für Wahrscheinlichkeitsmasse, um die Abhängigkeit der Konvergenzraten von der Komplexität der Hypothesenklasse $\mathcal{G}$ und der sogenannten Margin-Bedingung (*margin condition*) oder Rausch-Bedingung (*noise condition*) zu zeigen. Um die Komplexität der

Klasse kann entweder mit Hilfe regularisierter ERM durch Formulierung eines passenden Straftermes kontrolliert werden, oder durch eine auf der Entropie der Klasse basierende Komplexitätsbedingung. Letztere wird durch einen Konstante $\rho \in (0, 1)$ parametrisiert, wobei ein kleiner Wert von ρ einer einfachen Klassenstruktur entspricht. Die Margin-Bedingung misst die Identifizierbarkeit des Bayes-Prädiktors, und wird durch einen Parameter $\kappa \geq 1$ kontrolliert. Für kleine Werte des Parameters treten entlang der Bayes-Entscheidungsgrenze nur kleine Wahrscheinlichkeiten auf, sofern das Rausch-Niveau der zugrundeliegenden Verteilungen hinreichend niedrig ist. Die resultierenden Konvergenzraten sind Funktionen der Parameter ρ und κ.

Unter den oben genannten Bedingungen und für den Fall binärer Klassenzuordnungen ($m = 2$) sind die Konvergenzraten von $\widetilde{R}(\hat{g}_n) - \widetilde{R}(g^*)$ gegen Null abhängig von den unbekannten Parametern ρ und κ, und höher als die klassische nicht-parametrische Rate von $n^{-1/2}$. Allerdings ist die ERM-Methode nicht rechnerisch durchführbar, da die 0–1-Verlustfunktion nicht konvex ist. Wir ersetzen den 0–1-Verlust durch eine konvexe Ersatzfunktion $l(g(X), Y)$, welche eine obere Schranke an den wahren Verlust darstellt, und die Methode praktisch durchführbar macht. Um hohe Konvergenzraten zu erhalten, sind zwei Fragen zu beantworten: (a) Minimiert der Bayes-Prädiktor g^* auch das dem Ersatzverlust l entsprechende l-Risiko $R(g) := \mathbb{E}[l(g(X), Y)]$, und (b) ob ein durch Minimierung des l-Differenzrisikos $R(\hat{g}_n) - R(g^*)$ bestimmter Prädiktor tatsächlich den Prädiktionsfehler $\widetilde{R}(\hat{g}_n) - \widetilde{R}(g^*)$ minimiert.

Die vorliegende Arbeit bestimmt schnelle Konvergenzraten für das Differenzrisiko bei Ersatz des 0–1-Verlusts durch eine Hinge-Verlustfunktion (*hinge loss*). Die Konvergenzraten passen sich den unbekannten Margin- und Komplexitäts-Parametern an. Diese Resultate helfen, die praktische Effizienz von Support Vector Machine-Algorithmen mit Hinge-Verlust statistisch zu erklären. Desweiteren wird die sogenannte Klassifikation mit Zurückweisung (*reject option*) untersucht, bei welcher der Prädiktor für schwer klassifizierbare Beobachtungen eine Entscheidung verweigern kann. Unter geeigneten Margin- und Komplexitäts-Bedingungen erhalten wir schnelle Konvergenzraten für das wahre Differenzrisiko. Wir nehmen dabei nicht an, dass der Bayes-Prädiktor in der Modellklasse enthalten ist. Die vorgestellten Resultate werden in Bezug auf den Approximationsfehler $\inf_{g \in \mathcal{G}} R(g) - R(g^*)$ formuliert.

Für das Zwei-Klassen-Problem mit Hinge-Verlust betrachten wir Klassifikatoren, die als Linearkombinationen von Basisfunktionen definiert sind. Anstelle des von der Support Vector Machine verwendeten ℓ_2-Strafterms untersuchen wir eine ℓ_1-Bedingung an die Koeffizienten. Bei Verwendung des Hinge-Verlusts

können wir unter gewissen Anforderungen an die Basisfunktionen die Gültigkeit einer
Oracle-Ungleichung (*oracle inequality*) nachweisen, die von Modellkomplexität und Margin abhängt.

Während die statistischen Eigenschaften des Zwei-Klassen-Falls relativ gut verstanden sind, erweist sich deren Verallgemeinerung auf mehr Kategorien ($m > 2$) als nicht-trivial und oft schwerer handhabbar. Für den sogenannten multiplen Hinge-Verlust—eine Verallgemeinerung des Hinge-Verlusts, welche sämtliche Kategorien gleichzeitig berücksichtigt—konnte allerdings Bayes-Konsistenz nachgewiesen werden. Desweiteren folgt aus der Konvergenz des auf dem multiplen Hinge-Verlusts basierenden Differenzrisikos gegen Null die entsprechende Konvergenz des wahren Differenzrisikos, und zwar mit derselben Konvergenzrate. In der hier vorgestellten Arbeit zeigen wir Momenten-Schranken für die Minimierer des multiplen Hinge-Verlusts. Die Schranken basieren auf zwei Typen von Komplexitäts-Bedingungen, nämlich auf Entropie mit sogenanntem *bracketing*, und auf empirischer Entropie. Die Herleitung des letzteren Resultats erweist als deutlich aufwendiger. Wir erhalten schnelle Konvergenzraten in Abhängigkeit von den unbekannten Margin- und Komplexitäts-Parametern, nämlich $n^{-\kappa/(2\kappa-1+\rho)}$.

Klassifikation mit Zurückweisung kann die Leistung eines Prädiktors verbessern, sofern in einer gegebenen Anwendung die Kosten für die Klassifikation eines zurückgewiesenen Beispiels mit Hilfe eines Ausweichverfahrens (z.B. Klassifikation von Hand) kleiner sind als die Kosten einer Fehlklassifikation. Zurückweisung durch den Prädiktor lässt sich als zusätzliche Kategorie im Ausgaberaum beschreiben. Wir betrachten den Fall, in welchem akzeptablen Kosten einer Fehlklassifikation in Form eines Parameters α gegeben sind (α-*reject loss*). Wir untersuchen die Margin-Bedingung und geben eine Komplexitäts-Bedingung für die Hypothesenklasse an, die zu schneller Konvergenz des wahren Differnzrisikos führt.