ON THRESHOLD-BASED CLASSIFICATION RULES

LEILA MOHAMMADI AND SARA VAN DE GEER
Leiden University

Suppose we have n i.i.d. copies $\{(X_i, Y_i); i = 1, \ldots, n\}$ of an example (X, Y), where $X \in \mathcal{X}$ is an instance and $Y \in \{-1, 1\}$ is a label. A decision function (or classifier) f is a function $f\colon \mathcal{X} \to [-1, 1]$. Based on f, the example (X, Y) is misclassified if $Yf(X) \leq 0$. In this paper, we first study the case $\mathcal{X} = \mathbb{R}$, and the simple decision functions $h_a(x) = 2\mathbb{1}\{x \geq a\} - 1$ based on a threshold $a \in \mathbb{R}$. We choose the threshold $\hat{a}_n$ that minimizes the classification error in the sample, and derive its asymptotic distribution. We also show that, under monotonicity assumptions, $\hat{a}_n$ is a nonparametric maximum likelihood estimator. Next, we consider more complicated classification rules based on averaging over a class of base classifiers. We allow that certain examples are not classified due to lack of evidence, and provide a uniform bound for the margin. Moreover, we illustrate that when using averaged classification rules, maximizing the number of examples with margin above a given value, can overcome the problem of overfitting. In our illustration, the classification problem then boils down to optimizing over certain threshold-based classifiers.

1. Introduction

Suppose we have n i.i.d. realizations $\{(X_i, Y_i); i = 1, \ldots, n\}$ of an example (X, Y), where $X \in \mathcal{X}$ is an instance and $Y \in \{-1, 1\}$ is a label. A decision function f is a function $f\colon \mathcal{X} \to [-1, 1]$. We will also refer to f as a classifier. The decision rule based on f is to attach to an instance $x \in \mathcal{X}$ the label $y = 1$ if $f(x) > 0$, and otherwise to attach the label $y = -1$. Using this rule, the example (X, Y) is misclassified if $Yf(X) \leq 0$. A base classifier h is a function $h\colon \mathcal{X} \to \{-1, 1\}$, attaching the label $h(x)$ to the instance x.

Given a class of classifiers $\mathcal{F}$, the problem is to choose the "best" one. Let $L(f) = P(Yf(X) \leq 0)$ be the theoretical loss, or prediction error, of the classifier $f \in \mathcal{F}$. Thus, if X_{n+1} is a new instance for which we want to predict the label, $L(f)$ is the mean error of the prediction

$$\hat{Y}_{n+1} = \begin{cases} 1, & \text{if } f(X_{n+1}) > 0 \\ -1, & \text{if } f(X_{n+1}) \leq 0. \end{cases}$$

The smallest possible prediction error over $\mathcal{F}$ is $\min_{f \in \mathcal{F}} L(f)$. Consider now the empirical loss $L_n(f)$ of a particular classifier f, which is defined as the fraction of examples in the sample that have been misclassified by f. We will study, for some choices of $\mathcal{F}$, the classifier $\hat{f}_n$ that minimizes $L_n(f)$ over

AMS subject classifications: 62G05, 62G20.
Keywords and phrases: bounded variation; classification; classification error; convex hull; cube root asymptotics; entropy; threshold; VC class.