CS281B/Stat241B (Spring 2003) Statistical Learning Theory **Lecture: 1. Jan 21, 2003**

Probabilistic Formulation of Prediction Problems

Lecturer: Peter L. Bartlett *Scribe: Yuan Yao*

1 Overview

In this course we will mainly focus on three theoretical aspects in learning: probabilistic formulation of prediction problems, such as pattern classification; algorithms, such as kernel methods and ensemble methods (eg. boosting), etc.; and performance estimation, which is based on concentration inequalities and uniform convergence results.

In this lecture we will discuss the probabilistic formulation of general prediction problems and as a special case, the binary classification problem.

2 Prediction Problems

Let $\mathcal{X}$ be the *observation space* (input space) and $\mathcal{Y}$ be the *outcome space* (output space). Define P as a joint probability distribution on $X \times Y$. Let P_X and $P_{Y|X}$ be the marginal distribution and conditional distribution, respectively. Let $\mathbb{E}_{X,Y}$, $\mathbb{E}_X$ and $\mathbb{E}_{Y|X}$ be the expectation w.r.t. P, P_X and $P_{Y|X}$, respectively. When it is clear what random variables are associated with the expectations, we will omit the subscripts.

Denote the *loss function* by $l : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}$. For example, in classification, we wish to classify observations into a finite number of distinct classes, $|\mathcal{Y}| < \infty$, and we incur a loss of 1 if we assign the class label incorrectly, $l(\hat{y}, y) := 1[\hat{y} \neq y]$. (We denote the indicator function for a predicate Φ by $1[\Phi]$; it takes value 1 when Φ is true and 0 otherwise.) In regression, we have $\mathcal{Y} = \mathbb{R}$, and we might use $l(\hat{y}, y) = (\hat{y} - y)^2$.

Now we are ready to give the probabilistic formulation for general prediction problems.

Problem. Let (X_i, Y_i) $(i = 1, \ldots, n)$ and (X, Y) be the i.i.d. (independent identically distributed) with probability distribution P. The basic prediction problem is to find a map $f : \mathcal{X} \to \mathcal{Y}$ such that the *risk* $R(f) := \mathbb{E}l(f(X), Y)$ is as small as possible, *i.e.* close to the optimal value,

$$R^* = \inf_f R(f),$$

where the infimum is over all measurable functions f. Since an algorithm uses the training data (X_i, Y_i) to choose a suitable map, we write it as

$$f_n : x \mapsto f_n(x) = f_n(x; X_1, Y_1, \ldots X_n, Y_n)$$

which is random, since it depends on $(X_i, Y_i)_i^n$ and hence

$$R(f_n) := \mathbb{E}l(f_n(X), Y) = \mathbb{E}_{X,Y}\left[l\left(f_n(X; (X_i, Y_i)_1^n), Y\right) |(X_i, Y_i)_1^n\right]$$

is a random variable too. Thus we might wish to ensure that f_n performs well on average, in the sense that $\mathbb{E}R(f_n)$ is small, or that it performs well for most training sets, in the sense that $R(f_n)$ is small with high probability.

Remark. The key issues in the problem above lie in the so called *approximation problem* (how good is f_n with $n \to \infty$?), *estimation problem* (what is the effect of finite sample of size n?) and *computational or algorithmic problem* (what is an efficient algorithm to find f_n?)

3 Pattern Classification

For simplicity we just consider the binary classification case where $\mathcal{Y} = \{\pm 1\}$. The distribution P on $\mathcal{X} \times \mathcal{Y}$ be represented by (μ, η), where μ is the marginal distribution on $\mathcal{X}$ and η is the conditional distribution defined as

$$\eta(X) := P(Y = 1|X).$$

The loss function is $l(\hat{y}, y) = 1[\hat{y} \neq y]$, so the risk of a function f is $R(f) = \mathbb{E}(l(f(X), Y)) = P(f(X) \neq Y)$. Define the *Bayes risk* R^* as the smallest risk achievable by any measurable function,

$$R^* = \inf_f R(f).$$

Define

$$f^*(x) := \begin{cases} 1, & \eta(x) \geq 1/2 \\ -1, & \text{otherwise.} \end{cases}$$

For any x, $f^*(x)$ is a value of $y \in \{\pm 1\}$ that maximizes $P(Y = y|X = x)$. The following theorem shows that this function has risk equal to the Bayes risk, $R^* = R(f^*)$. For this reason, we refer to f^* as the *Bayes decision function*. The theorem also shows how the performance of an arbitrary f degrades as it moves away from f^*.

Theorem 3.1. For any $f : \mathcal{X} \to \mathcal{Y}$,

$$R(f) - R(f^*) = \mathbb{E}(1[f(X) \neq f^*(X)]|2\eta(X) - 1|).$$

PROOF.

$$\begin{aligned} R(f) - R(f^*) &= \mathbb{E}_{X,Y}(1[f(X) \neq Y] - 1[f^*(X) \neq Y]) \\ &= \mathbb{E}_{X,Y}(1[f(X) \neq f^*(X)](1[f(X) \neq Y] - 1[f^*(X) \neq Y])) \\ &= \mathbb{E}_{X,Y}(1[f(X) \neq f^*(X)](2 \cdot 1[f^*(X) = Y] - 1)) \\ &= \mathbb{E}_X\{(1[f(X) \neq f^*(X)](2\mathbb{E}_{Y|X}(1[f^*(X) = Y]) - 1)\} \\ &= \mathbb{E}_X\{1[f(X) \neq f^*(X)] \cdot [2(1/2 + |\eta(X) - 1/2|) - 1)]\} \\ &= \mathbb{E}_X(1[f(X) \neq f^*(X)]|2\eta(X) - 1|) \end{aligned}$$

□

Notice that the increase in risk depends on how far the conditional probability, η, is from 1/2 on the set where f and f^* disagree. In particular, any function that disagrees with f^* only on the set $\{s : \eta(x) = 1/2\}$ also has $R(f) = R^*$.

One approach to minimizing risk is to estimate the regression function η with some $\hat{\eta} : \mathcal{X} \to [0, 1]$ and use the following classification function.

$$f_{\hat{\eta}}(X) = \begin{cases} 1, & \text{if } \hat{\eta}(X) \geq 1/2, \\ -1, & \text{otherwise.} \end{cases}$$

This is called the *plug-in decision function*. The following theorem gives an upper bound on $R(f_{\hat{\eta}}) - R(f^*)$ in terms of the difference $\hat{\eta} - \eta$.

Theorem 3.2.

$$R(f) - R^* \leq 2\mathbb{E}|\eta(X) - \hat{\eta}(X)|$$

Proof.

$$\begin{aligned} R(f_{\hat{\eta}}) - R^* &= 2\mathbb{E}(1[f_{\hat{\eta}}(X) \neq f^*(X)]|\eta(X) - 1/2|) \\ &\leq 2\mathbb{E}|\hat{\eta}(X) - \eta(X)| \end{aligned}$$

where the last step is due to

$$f_{\hat{\eta}}(X) \neq f^*(X) \quad \Rightarrow \quad |\hat{\eta}(X) - \eta(X)| = |\hat{\eta}(X) - 1/2| + |\eta(X) - 1/2| \geq |\eta(X) - 1/2|$$

since in that case η and $\hat{\eta}$ must lie to either side of $1/2$. □

Remark. Thus, estimating η accurately (in $L_1(P)$) gives near-minimal risk. But estimating η is not necessary, and it might be more difficult than the problem of finding an f with small risk.

An alternative approach is to first fix F, a set of functions $f : \mathcal{X} \to \{\pm 1\}$; then choose

$$f_n(X; X_1, Y_1, \ldots, X_n, Y_n) = f(X)$$

for $f \in F$. For instance, in *empirical risk minimization*, we choose $f \in F$ to minimize empirical risk,

$$\hat{R}(f_n) = \hat{\mathbb{E}} l(f(X), Y) = \frac{1}{n} \sum_{i=1}^{n} l(f(X_i), Y_i).$$

For example if $\mathcal{X} = \mathbb{R}^d$, consider $F = \{x \mapsto \text{sign}(w^T x) : w \in \mathbb{R}^d\}$ which are linear threshold functions with decision regions as the half spaces.