CS281B/Stat241B (Spring 2008) Statistical Learning Theory **Lecture: 1**

Overview plus probabilistic formulations of prediction problems

Lecturer: Peter Bartlett *Scribe: Peter Bartlett*

1 Organizational issues

The course web page is at http://www.cs.berkeley.edu/∼bartlett/courses/281b-sp08/
See the web page for details of office hours, the syllabus, assignments, readings, lecture notes, and announcements.

1.1 Assignments

There will be roughly five homework assignments, approximately one every two weeks. The first has been posted on the web site. It is due at the lecture on Thursday, January 31. You will also need to act as scribe for a small number of lectures, preparing a latex version of lecture notes. There is a template on the web site, and the latex file of the lecture notes for this lecture. (Please email the GSI, David, to choose the lecture that you'd like to prepare lecture notes for.) Also, there will be a final project, in an area related to the topics of the course.

2 Overview

The course will focus on the theoretical analysis of prediction methods.

1. Probabilistic formulation of prediction problems
2. Algorithms:
   (a) Kernel methods
   (b) Boosting algorithms
3. Risk bounds
4. Game theoretic formulation of prediction problems
5. Model selection

3 Probabilistic Formulations of Prediction Problems

In a prediction problem, we wish to predict an outcome y from some set $\mathcal{Y}$ of possible outcomes, on the basis of some observation x from a feature space $\mathcal{X}$. Some examples:

x	y
phylogenetic profile of a gene (i.e., relationship to genomes of other species)	gene function
gene expression levels of a tissue sample	patient disease state
image of a signature on a check	identity of the writer
email message	spam or ham

For such problems, we might have access to a data set of n pairs, $(x_1, y_1), \ldots, (x_n, y_n)$, and we would like to use the data to produce a function $f : \mathcal{X} \to \mathcal{Y}$ so that, for subsequent (x, y) pairs, $f(x)$ is a good prediction of y.

To define the notion of a 'good prediction,' we can define a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, so that $\ell(\hat{y}, y)$ quantifies the cost of predicting $\hat{y}$ when the true outcome is y. Then the aim is to ensure that $\ell(f(x), y)$ is small. For instance, in *pattern classification* problems, the aim is to classify an x into one of a finite number of classes (that is, the label space $\mathcal{Y}$ is finite). If all mistakes are equally bad, we could define

$$\ell(\hat{y}, y) = 1[\hat{y} \neq y] = \begin{cases} 1 & \text{if } \hat{y} \neq y, \\ 0 & \text{otherwise.} \end{cases}$$

As another example, in a *regression* problem, with $\mathcal{Y} = \mathbb{R}$, we might choose the quadratic loss function, $\ell(\hat{y}, y) = (\hat{y} - y)^2$.

We can formulate such problems using probabilistic assumptions: we assume that there is a probability distribution P on $\mathcal{X} \times \mathcal{Y}$, and that the pairs $(X_1, Y_1), \ldots, (X_n, Y_n), (X, Y)$ are chosen independently according to P. The aim is to choose f so that the *risk* of f,

$$R(f) = \mathbb{E}\ell(f(X), Y),$$

is small. For instance, in the pattern classification example, this is the misclassification probability.

$$R(f) = \mathbb{E}1[f(X) \neq Y] = \Pr(f(X) \neq Y).$$

Some things to notice:

1. We are using capital letters to denote random variables.
2. The distribution P can be viewed as modelling both the relative frequency of different features or covariates X, together with the conditional distribution of the outcome Y given X.
3. The assumption that the data is i.i.d. is a strong one.
4. The function $x \mapsto f_n(x) = f_n(x; X_1, Y_1, \ldots, X_n, Y_n)$ is random, since it depends on the random (X_i, Y_i). Thus, the risk

$$\begin{aligned} R(f_n) &= \mathbb{E}\left[\ell(f_n(X), Y) | X_1, Y_1, \ldots, X_n, Y_n\right] \\ &= \mathbb{E}\left[\ell(f_n(X; X_1, Y_1, \ldots, X_n, Y_n), Y) | X_1, Y_1, \ldots, X_n, Y_n\right] \end{aligned}$$

   is a random variable. We might aim for $\mathbb{E}R(f_n)$ small, or $R(f_n)$ small with high probability (over the training data).

We might choose f_n from some class F of functions, for instance, by choosing the structure and parameters of a decision tree, or by choosing the parameters of a neural net or a kernel machine.

There are several questions that we are interested in:

1. Can we design algorithms for which f_n is close to the best that we could hope for, given that it was chosen from F? (that is, is $R(f_n) - \inf_{f \in F} R(f)$ small?)
2. How does the performance of f_n depend on n? On other parameters of the problem?
3. Can we ensure that $R(f_n)$ approaches the best possible performance (that is, the infimum over all f of $R(f)$)?
4. What do we need to assume about P? About F?

In this course, we are concerned with results that apply to large classes of distributions P, such as the set of *all* joint distributions on $\mathcal{X} \times \mathcal{Y}$. In contrast to parametric problems, we will not (often) assume that P comes from a small (e.g., finite-dimensional) space, $P \in \{P_\theta : \theta \in \Theta\}$.

Several key issues arise in designing a prediction method for these problems:

Approximation How good is the best f in the class F that we are using? That is, how close to $\inf_f R(f)$ is $\inf_{f \in F} R(f)$?

Estimation Since we only have access to the distribution P through observing a finite data set, how close is our performance to that of the best f in F?

Computation We need to use the data to choose f_n, typically through solving some kind of optimization problem. How can we do that efficiently?

In this course, we will not spend much time on the approximation properties, beyond observing some universality results (that particular classes can achieve zero approximation error). We will focus on the estimation issue. We will take the approach that efficiency of computation is a constraint. Indeed, the methods that we spend most of our time studying involve convex optimization problems. (For example, kernel methods involve solving a quadratic program, and boosting algorithms involve minimizing a convex criterion in a convex set.)

4 The Probabilistic Formulation of Pattern Classification Problems

Assume, for simplicity, that $\mathcal{Y} = \{\pm 1\}$ (We'll consider extensions of the results of this lecture to the multiclass case in a homework problem.) Let's fix some notation: We'll represent the joint distribution P on $\mathcal{X} \times \mathcal{Y}$ as the pair (μ, η), where μ is the marginal distribution on $\mathcal{X}$ and η is the conditional expectation of Y given X,

$$\eta(x) = \mathbb{E}(Y|X = x) = P(Y = 1|X = x).$$

If we knew η, we could use it to find a decision function that minimized risk. To see this, notice that we can write the expected loss as an expectation of a conditional expectation,

$$\begin{aligned}
R(f) &= \mathbb{E}\ell(f(X), Y) \\
&= \mathbb{E}\mathbb{E}[\ell(f(X), Y)|X] \\
&= \mathbb{E}\left(\ell(f(X), 1)P(Y = 1|X) + \ell(f(X), -1)P(Y = -1|X)\right) \\
&= \mathbb{E}\left(1[f(X) \neq 1)\eta(X) + 1[f(X) \neq -1)(1 - \eta(X))\right) \\
&= \mathbb{E}\left(1[f(X) \neq 1)\eta(X) + (1 - 1[f(X) \neq 1])(1 - \eta(X))\right) \\
&= \mathbb{E}\left(1[f(X) \neq 1)(2\eta(X) - 1) + 1 - \eta(X)\right). \qquad (1)
\end{aligned}$$

Clearly, this expectation is minimized by choosing $f(x) = 1$ when $\eta(x) > 1/2$ and $f(x) = -1$ when $\eta(x) < 1/2$. Obviously, if $\eta(x) = 1/2$, the choice does not affect the risk. Let's define f^* as a function of this kind:

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

Denote the optimal risk (the *Bayes risk*), by $R^* = \inf_f R(f)$. We have shown that f^* achieves the Bayes risk. It is called the *Bayes decision function*.

Notice that any choice for $f^*(x)$ is equally good when $\eta(x) = 1/2$, so there can be several Bayes decision functions.

The following theorem shows something a little stronger: that the amount by which the risk of any other decision function exceeds the Bayes risk can be quantified in terms of a certain distance from f^*. (Actually, it's not quite a distance, since differences between functions at an x with $\eta(x) = 1/2$ have no influence on the risk.)

Theorem 4.1. For any $f : \mathcal{X} \to \mathcal{Y}$,

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

Proof. Using the identity (1), we have

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

But

$$\begin{aligned} &\left(1[f(X) \neq 1] - 1[f^*(X) \neq 1]\right)(2\eta(X) - 1) \\ &= 1[f(X) \neq f^*(X)]\left(1[f(X) \neq 1] - 1[f^*(X) \neq 1]\right)(2\eta(X) - 1) \\ &= \begin{cases} 1[f(X) \neq f^*(X)](2\eta(X) - 1) & \text{if } 2\eta(X) - 1 \geq 0, \\ 1[f(X) \neq f^*(X)](-1)(2\eta(X) - 1) & \text{if } 2\eta(X) - 1 < 0. \end{cases} \\ &= 1[f(X) \neq f^*(X)]|2\eta(X) - 1|, \end{aligned}$$

where the second inequality used the definition of f^*. □

This suggests one family of approaches to the pattern classification problem, known as *plug-in* methods. The idea is to use the data to come up with an estimate $\hat{\eta}$ of η, and then use

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

In estimating η, what criterion should we aim to minimize? We can use the earlier result to show that if the $L_1(\mu)$ distance between $\hat{\eta}$ and η is small, that suffices to ensure that the risk of $f_{\hat{\eta}}$ is close to the Bayes risk.

Theorem 4.2. For any $\hat{\eta} : \mathcal{X} \to \mathbb{R}$,

$$R(f_{\hat{\eta}}) - R^* \leq 2\mathbb{E}\left|\eta(X) - \hat{\eta}(X)\right|.$$

Proof. The previous theorem shows that the excess risk of $f_{\hat{\eta}}$ can be written as

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

Now, if $f_{\hat{\eta}}(X) \neq f^*(X)$, then $\hat{\eta}(X)$ and $\eta(X)$ must lie on opposite sides of $1/2$, and so we can write

$$|\eta(X) - \hat{\eta}(X)| = |\eta(X) - 1/2| + |\hat{\eta}(X) - 1/2| \geq |\eta(X) - 1/2|.$$

Thus, when the indicator inside the random variable in (2) is 1, we have

$$1[f_{\hat{\eta}}(X) \neq f^*(X)]2|\eta(X) - 1/2| \leq 2|\eta(X) - \hat{\eta}(X)|$$

And this inequality is also true when the indicator is zero, since the right hand side is non-negative. Plugging this inequality into (2) gives the result. □

Another family of approaches to pattern classification problems is to fix a class F of functions that map from $\mathcal{X}$ to $\mathcal{Y}$ and choose f_n from F. Next lecture, as an introduction to kernel methods, we'll consider the class of linear threshold functions on $\mathcal{X} = \mathbb{R}^d$,

$$F = \left\{ x \mapsto \text{sign}(\theta' x) : \theta \in \mathbb{R}^d \right\}.$$

The decision boundaries are hyperplanes through the origin ($d-1$-dimensional subspaces), and the decision regions are half-spaces.