EECS 598: Statistical Learning Theory, Winter 2014 **Topic 1**

Probabilistic Setting

Lecturer: Clayton Scott *Scribe: Tianpei Xie*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

1 Introduction

This is a course on statistical learning theory. We will primarily focus on the theory of supervised classification, with some additional topics such as density estimation, online learning and/or weakly supervised learning. In the first part of the course, we focus on classification.

2 Classification

In classification we consider pairs (x, y), where x is a *feature vector* belonging to a *feature space* $\mathcal{X}$ (for example, $\mathcal{X} = \mathbb{R}^d$), and y is a label belonging to a *label space* $\mathcal{Y}$. In *binary classification*, for example, $\mathcal{Y} = \{0, 1\}$ (or, $\mathcal{Y} = \{-1, 1\}$, sometimes). A classifier is a function $h : \mathcal{X} \to \mathcal{Y}$. In Handwritten Digit Recognition [1], $\mathcal{X} = \{\text{digital images of a certain size}\}$ and $\mathcal{Y} = \{0, 1, \cdots, 9\}$. In classification we desire a classifier that accurately assigns labels to feature vectors.

The *key* assumption for statistical learning theory is that there exists a joint probability distribution on the feature-label space $\mathcal{X} \times \mathcal{Y}$, denoted as P_{XY}. Each observed feature-label pair (X, Y) is random and generated according to P_{XY}. We use (X, Y) to denote a random variable and (x, y) to denote its realization.

Remark. In this course we will assume that all events and functions are measurable, and will not concern ourselves with issues of measurability. Students aiming to do research in this area would do well to study measure theory.

To understand the meaning of the joint distribution P_{XY}, consider the following decomposition:

$$\begin{aligned} P_{XY} &= P_{X|Y}\, P_Y, \\ \text{where} \quad & P_Y \text{ is the } Y\text{-marginal of } P_{XY}, \text{ referred to as the \textit{prior} of label } Y, \\ & P_{X|Y=y} \text{ is the \textit{class-conditional} distribution of } X \text{ given } Y = y. \end{aligned}$$

There are two interpretations for the above decomposition: First, one can view it as a two-step random number generation procedure: first generate the label $Y = y$ via P_Y, then generate the feature vector according to $P_{X|Y=y}$. Second, one can interpret this decomposition via the total expectation theorem (aka. the disintegration Theorem), i.e., for any real-valued function $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$,

$$\mathbb{E}_{XY}[\phi(X, Y)] = \mathbb{E}_Y\, \mathbb{E}_{X|Y}[\phi(X, Y)]$$

There is an alternative decomposition, namely

$$\begin{aligned} P_{XY} &= P_{Y\,|X}\, P_X, \\ \text{where} \quad & P_X \text{ is the } X\text{-marginal of } P_{XY}, \\ & P_{Y|X=x} \text{ is the \textit{posterior} distribution of } Y \text{ given } X = x. \end{aligned}$$

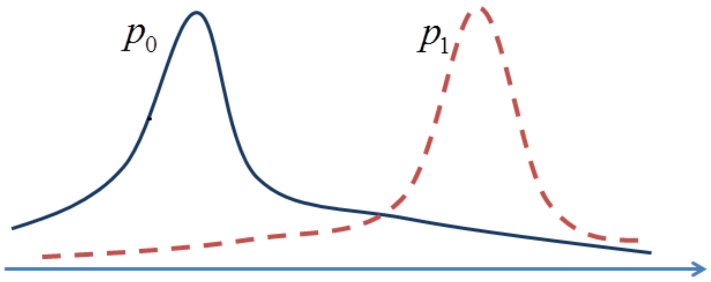

Figure 1: The illustration of a binary classification problem with $\mathcal{X} = \mathbb{R}$ and $\mathcal{Y} = \{0, 1\}$. Assume $P_{X|Y=y}$ is continuous, with density p_y, $y = 0, 1$. The red dashed (right) curve and blue solid curve (left) denote the two conditional densities. They overlap in the middle, which means that they cannot be completely separated.

The same comments apply here. Both decompositions are useful. Figure 1 shows an example for binary classification, with $\mathcal{X} = \mathbb{R}$ and $\mathcal{Y} = \{0, 1\}$. Note that in this case, the two class-conditional distributions cannot be completely separated: no classifier will be 100% accurate.

To learn a classifier $h : \mathcal{X} \to \mathcal{Y}$, assume we have access to a set of n labeled *training examples* or *training instances*, $(X_1, Y_1), \ldots, (X_n, Y_n)$, which are assumed to be i.i.d. according to P_{XY}. The collection of all training examples is called the *training data*. The domain of the training data is $(\mathcal{X} \times \mathcal{Y})^n$. Let $\mathcal{H}$ be a collection of classifiers of concern, e.g., the set of linear classifiers $h(x) = \mathrm{sign}\,(w^T x + b)$. A *learning algorithm/classification algorithm/discrimination rule* is a function

$$\mathcal{L}_n : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{H}$$

Denote $\widehat{h}_n := \mathcal{L}_n((X_1, Y_1), \ldots, (X_n, Y_n))$. Note that $\widehat{h}_n$ is a function of random variables, so it is a random variable as well.

3 The goal of classification

Define the *risk* of classifier h as

$$R(h) \quad := \quad P_{XY}(h(X) \neq Y) = \mathbb{E}_{XY}[\mathbf{1}_{\{h(X) \neq Y\}}], \tag{1}$$

where (X, Y) is independent of the training data. Also define the *Bayes risk* as

$$R^* \quad := \quad \inf_h R(h),$$

where the infimum is taken over *all classifiers* $h : \mathcal{X} \to \mathcal{Y}$, not just $h \in \mathcal{H}$. If $R(h^*) = R^*$, then h^* is called a *Bayes classifier.*

A learning algorithm $\mathcal{L}_n$ is called *weakly consistent* if

$$R(\widehat{h}_n) \xrightarrow{i.p.} R^*, \tag{2}$$

and *strongly consistent* if

$$R(\widehat{h}_n) \xrightarrow{a.s.} R^*. \tag{3}$$

Note that as $\widehat{h}_n$ is a random variable, $R(\widehat{h}_n)$ is a random variable as well. $\mathcal{L}_n$ is called *universally (weakly/strongly) consistent* if it is (weakly/strongly) consistent for $\forall P_{XY}$. That is, the consistency holds without any assumption on the distribution P_{XY}.

In the following lectures, we will study learning algorithms possessing asymptotic and/or finite sample performance guarantees. Asymptotic guarantees will include consistency and rates of convergence. Finite sample guarantees will include confidence intervals on the risk, and sample complexity bounds.

References

[1] Y. Le Cun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Handwritten digit recognition with a back-propagation network," *Advances in Neural Information Processing Systems*, 1990.