ECE901 Spring 2004 Statistical Regularization and Learning Theory **Lecture: 1**

Statistical Regularization and Learning Theory

Lecturer: Rob Nowak *Scribe: Rob Nowak*

1 Three Elements of Statistical Data Analysis

1. **Probabilistic Formulation** of learning from data and prediction problems.
2. **Performance Characterization:**

   concentration inequalities
   uniform deviation bounds
   approximation theory
   rates of convergence
3. **Practical Algorithms** that run in polynomial time (e.g., decision trees, wavelet methods, support vector machines).

2 Learning from Data

To formulate the basic learning from data problem, we must specify several basic elements: data spaces, probability measures, loss functions, and statistical risk.

2.1 Data Spaces

Learning from data begins with a specification of two spaces:

$$\mathcal{X} \equiv \text{Input Space}$$

$$\mathcal{Y} \equiv \text{Output Space}$$

The input space is also sometimes called the "feature space" or "signal domain." The output space is also called the "class label space," "outcome space," "response space," or "signal range."

Example 1

$$\mathcal{X} = \mathbf{R}^d \quad \textit{d-dimensional Euclidean space of "feature vectors"}$$
$$\mathcal{Y} = \{0, 1\} \quad \textit{two classes or "class labels"}$$

Example 2

$$\mathcal{X} = \mathbf{R} \quad \textit{one-dimensional signal domain (e.g., time-domain)}$$
$$\mathcal{Y} = \mathbf{R} \quad \textit{real-valued signal}$$

A classic example is estimating a signal f in noise:

$$Y = f(X) + W$$

where X is a random sample point on the real line and W is a noise independent of X.

2.2 Probability Measure and Expectation

Define a joint probability distribution on $\mathcal{X} \times \mathcal{Y}$ denoted $P_{X,Y}$. Let (X, Y) denote a pair of random variables distributed according to $P_{X,Y}$. We will also have use for marginal and conditional distributions. Let P_X denote the marginal distribution on X, and let $P_{Y|X}$ denote the conditional distribution of Y given X. For any distribution P, let p denote its density function with respect to the corresponding dominating measure; e.g., *Lebesgue measure* for continuous random variables or *counting measure* for discrete random variables.

Define the expectation operator:

$$E_{X,Y}[f(X,Y)] \equiv \int f(x,y) dP_{X,Y}(x,y) = \int f(x,y) p_{X,Y}(x,y) dx dy.$$

We will also make use of corresponding marginal and conditional expectations such as E_X and $E_{Y|X}$.

Wherever convenient and obvious based on context, we may drop the subscripts (e.g., E instead of $E_{X,Y}$) for notational ease.

2.3 Loss Functions

A loss function is a mapping

$$\ell : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbf{R}$$

Example 3 *In binary classification problems,* $\mathcal{Y} = \{0, 1\}$. *The* $0/1$ *loss function is usually used:* $\ell(y_1, y_2) = 1_{y_1 \neq y_2}$, *where* 1_A *is the indicator function which takes a value of 1 if condition* A *is true and zero otherwise. We typically will compare a true label* y *with a prediction* $\hat{y}$, *in which case the* $0/1$ *loss simply counts misclassifications.*

Example 4 *In regression or estimation problems,* $\mathcal{Y} = R$. *The squared error loss function is often employed:* $\ell(y_1, y_2) = (y_1 - y_2)2$, *the square of the difference between* y_1 *and* y_2. *In application, we are interested in a true value* y *in comparison to an estimate* $\hat{y}$.

2.4 Statistical Risk

The basic problem in learning is to determine a mapping $f : \mathcal{X} \mapsto \mathcal{Y}$ that takes an input $x \in \mathcal{X}$ and predicts the corresponding output $y \in \mathcal{Y}$. The performance of a given map f is measured by its expected loss or *risk*:

$$R(f) \equiv E_{X,Y}\left[\ell(f(X), Y)\right]$$

The risk tells us how well, on average, the predictor f performs with respect to the chosen loss function. A key quantity of interest is the mininum risk value, defined as

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

where the infinum is taking over all measurable functions.

2.5 The Learning Problem

Suppose that (X, Y) are distributed according to $P_{X,Y}$ ($(X, Y) \sim P_{X,Y}$ for short). Our goal is to find a map so that $f(X) \approx Y$ with high probability. Ideally, we would chose f to minimize the risk $R(f) = E[\ell(f(X), Y)]$. However, in order to compute the risk (and hence optimize it) we need to know the joint distribution $P_{X,Y}$. In many problems of practical interest, the joint distribution is unknown, and minimizing the risk is not possible.

Suppose that we have some exemplary samples from the distribution. Specifically, consider n samples $X_i, Y_i{}_{i=1}^n$ distributed independently and identically (iid) according to the otherwise unknown $P_{X,Y}$. Let us call these samples *training data*, and denote the collection by $D_n \equiv X_i, Y_i{}_{i=1}^n$. Let's also define a collection

of candidate mappings $\mathcal{F}$. We will use the training data D_n to pick a mapping $f_n \in \mathcal{F}$ that we hope will be a good predictor. This is sometimes called the *Model Selection* problem. Note that the selected model f_n is a function of the training data:

$$f_n(X) = f(X; D_n),$$

which is what the subscript n in f_n refers to. The risk of f_n is given by

$$R(f_n) = E_{X,Y}[\ell(f_n(X), Y)]$$

Note that since f_n depends on D_n in addition to a new random pair (X, Y), the risk is a random variable (i.e., a function of the training data D_n). Therefore, we are interested in the *expected risk*, computed over random realizations of the training data:

$$E_{D_n}[R(f_n)]$$

We hope that f_n produces a small expected risk.

The notion of expected risk can be interpreted as follows. We would like to define an algorithm (a model selection process) that performs well on average, over any random sample of n training data. The expected risk is a measure of the expected performance of the algorithm with respect to the chosen loss function. That is, we are not gauging the risk of a particular map $f \in \mathcal{F}$, but rather we are measuring the performance of the algorithm that takes any realization of training data and selects an appropriate model in $\mathcal{F}$.

This course is concerned with determining "good" model spaces $\mathcal{F}$ and useful and effective model selection algorithms.