Math 547: Mathematical Foundations of Statistical Learning Theory Fall 2017

Lecture 1 — August 21

Instructor: Stas Minsker *Scribe: Mose Wintner*

1.1 Binary Classification

Binary classification will be performed using collections of

$$(X, Y) \in \mathbb{S} \times \{\pm 1\},$$

where $(\mathbb{S}, \mathcal{B})$ is a measure space.

For example, a 100×100 grayscale image can be represented as a function

$$\{1, \dots, 100\}^2 \to [0, 1],$$

where in practice the meaning of $[0, 1]$ is dependent upon the data type chosen.

We may have a set of small grayscale portraits which we wish to classify by sex; this corresponds to $\mathbb{S} = \{f : \{1, \dots, 100\}^2 \to [0, 1]\}$, where $\{\pm 1\} \cong \{\text{male}, \text{female}\}$.

We say (X, Y) has distribution $\mathbb{P}$ when

$$\mathbb{P}(A) = Prob((X, Y) \in A)$$

and we will denote the marginal probability distribution of X by Π, i.e.

$$\Pi(x) = Prob(x = X).$$

The goal of binary classification is to predict the label $Y = \pm 1$ ***based on the observation*** $X \in \mathbb{S}$***.***

Prediction is carried out via a **binary classifier**, which is a $\mathcal{B}$-measurable function

$$g : \mathbb{S} \to \{\pm 1\}.$$

The **generalization error** of a classifier g is given by

$$L(g) := \mathbb{P}(Y \neq g(X)).$$

We have that

$$L(g) = \mathbb{E}_p(I\{Y \neq g(X)\}) = \int_{\mathbb{S} \times \{\pm 1\}} I\{y \neq g(x)\} \, d\mathbb{P}(x, y).$$

The "best" binary classifier is

$$g_* = \underset{\substack{g : \mathbb{S} \to \{\pm 1\} \\ g \text{ measurable}}}{\operatorname{argmin}} L(g).$$