Introduction to Data Science

EMPIRICAL RISK MINIMIZATION

BRIAN D’ALESSANDRO

<u>Fine Print</u>: *these slides are, and always will be a work in progress. The material presented herein is original, inspired, or borrowed from others’ work. Where possible, attribution and acknowledgement will be made to content’s original source. Do not distribute, except for as needed as a pedagogical tool in the subject of Data Science.*

STATISTICAL LEARNING THEORY

Let's first set up some notation and ideas:

- Let $X \in \mathbb{R}^p$ be a p-dimensional real valued vector of predictor variables

- Let Y be a target variable, where
  - $Y \in \mathbb{R}$ is real valued
  - $Y \in C$ is an element in some set of classes $C = \{c_1, c_2, \dots c_k\}$

- X and Y are governed by a joint distribution $P(Y, X)$ (that we likely don't know)

- We seek a function $f(X)$ for predicting Y, given X, whose output can be
  - Real valued, i.e. $f(X) = E[Y|X]$
  - Discrete valued, i.e. $f(x) \in \{c_1, c_2, \dots c_k\}$

STATISTICAL LEARNING THEORY

Second, let's define two more things

- $\mathbb{F}$ is a family of functions, such that $f(x) \in \mathbb{F}$, examples are:
  - All linear hyper-planes, such that $f(x) = \alpha + \beta x$
  - All quadratic polynomials, such that $f(x) = \alpha + \beta_1 x + \beta_2 x^2$
  - All decision trees with max(depth)=k

- A loss function $\mathbb{L}(f(X), Y)$ that measures how well $f(X)$ approximates Y.
  - Squared Loss: $\mathbb{L}(f(x), y) = (f(x) - y)^2$
  - 0-1 Loss: $\mathbb{L}(f(x), y) = \mathbb{I}(f(x) == y)$
  - Logistic Loss: $\mathbb{L}(f(x), y) = -[y * Ln(f(x)) + (1 - y) * Ln(1 - f(x))]$
  - Hinge Loss: $\mathbb{L}(f(x), y) = max(0, 1 - f(x) * y)$

STATISTICAL LEARNING THEORY

The main goal of Supervised Learning can be stated using the Empirical Risk Minimization framework of Statistical Learning.

We are looking for a function $f \in \mathbb{F}$ that minimizes the expected loss:

$$E[\mathbb{L}(f(x), y)] = \int \mathbb{L}(f(x), y)\, P(x, y)\ \mathrm{d}x\mathrm{d}y$$

Because we don't know the distribution $P(X, Y)$, we can't minimize the expected loss. However, we can minimize the empirical loss, or risk, by computing the average loss over our training data.

Thus, in Supervised Learning, we choose the function $f(X)$ that minimizes the loss over training data:

$$f^{opt} = \underset{f \in \mathbb{F}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{L}(f(x_i), y_i)$$

This concept of Empirical Risk Minimization will the basis for understanding the linear models and the general notion of fitting mathematical models to data, as well as for understanding the very important principle of bias-variance tradeoffs.