Game Theory

Manik Bhandari
Department of Computational Data Science
Indian Institute of Science
Bangalore, India
mbbhandarimanik2@gmail.com

Abstract

Notes of Theory of Deep Learning mainly from lectures at IISC.

1 Introduction

Define error of a classifier (aka *true error*) as probability of it making a mistake given a random data point.

$$L_{D,f}(h) = \underset{x \sim D}{P}[h(x) \neq f(x)] = D(\{x : h(x) \neq f(x)\})$$

where D is the distribution from where a data point x is drawn, f is a known *correct* function which always gives the correct labels to a data point. By this definition $D(A)$ is the probability of observing a random point x from A.

Define *training error* or *emperical risk* as

$$L_S(h) = \frac{|i \in [m] : h(x_i) \neq y_i|}{m}$$

where S is the training *set* (it is actually a sequence since points can repeat and classifiers often take into account their order) of the form $\{(x_i, y_i)\}$. If you naively minimize this emperical risk then you are likely to overfit. To avoid it, you use some prior knowledge about the *kind of classifier* that can possibly fit to the data and restrict your hypothesis search space to those types of classifiers.

This kind of restriction induces a *bias* in the model (aka *inductive bias*). In this setting, define

$$h_S = ERM_h(S) \in \underset{h \in \mathcal{H}}{argmin}\, L_S(h).$$

This is a tradeoff – choosing a restricted H can add too much bias but choosing a large H may lead to overfitting.

Finite hypothesis class If we restrict H to have an upper bound on its size then ERM_h will not overfit if we have *large* training data (how large will depend on size of H).

Realizability Assumption There exists $h^* \in \mathcal{H}$ such that $L_{D,f}(h^*) = 0$ i.e. it never makes a mistake which means that $L_S(h^*) = 0$. Since this is the least possible error, this means that for every ERM hypothesis $L_S(h_S) = 0$. We are however interested in true error of h_S i.e. $L_{D,f}(h_S)$.

iid assumption Assume that elements of S are identically and independently distributed according to D denoted by $S \sim D^m$.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.