Lecture 1: The PAC Model for Supervised Learning

Instructor: Aditya Bhaskara *Scribe: Josh McMillan*

CS 5966/6966: Theory of Machine Learning

January 13th, 2022

Abstract

Today we discussed the general set-up of supervised learning problems and gave definitions to terms like learn-ability with reference to Leslie Valiant's work. Other topics included: a brief history of machine learning, theory of the learnable, risk minimization, and data distributions.

1 Introduction

Until the 1980's there was generally no formal reasoning or definitions for learning problems. In 1983 Leslie Valiant's "The Theory of the Learnable" reasoned about learning a hypothesis which yielded low error with respect to the input distribution for supervised learning problems.

Supervised learning relies on input data which has a labelled corresponding output (label); it lends itself to problems like classification, prediction.

2 Theory of the Learnable

In this framework we have inputs with "features" and what we want to learn must be derived from these input features. That is, we build some hypothesis/model which is a function from the features of our input to the predicted label. We consider a hypothesis "good" if the predicted labels generally match the true labels (low error).

An important distinction is that this need not be true for any theoretically possible input, but only inputs of interest. One of Valiant's key assumptions is that there is some probability distribution on the space of all inputs which is unknown to the "learner". It is with respect to this distribution that we care about our error.

Low error on previously unseen inputs from the distribution is called generalization

3 Risk Minimization

The formula below shows the "risk" of a hypothesis with respect to the given distribution.

Risk is another word for error here

$\exists$ some distribution $\mathbf{D}$ on the space of all inputs and $\exists$ a true label for each input. Given a hypothesis h, a true label oracle function l, the risk of h with respect to the distribution $\mathbf{D}$ is:

$$R_{\mathbf{D}}(h) = \Pr_{x \in \mathbf{D}}[h(x) \neq l(x)]. \tag{1}$$

This risk takes into account that we must correctly predict for certain inputs frequently if $\mathbf{D}$ is biased

4 Definition of Learnability

We say a hypothesis l is learnable if for any $\epsilon > 0$, there exists an m such that given m *iid* examples from **D** along with labels we can produce a hypothesis such that: $R_{\mathbf{D}}(h) \leq \epsilon$, with probability $> 90\%$.

The term "we can produce" means that $\exists$ an efficient algorithm **A** (polynomial in m, number of training examples) that takes m inputs and labels $(x_1, l(x_1)...x_m, l(x_m))$ and outputs a hypothesis h.

Training and test samples come from the same probability distribution.

5 Hypothesis Classes

Assume that the label function l is in a certain hypothesis class H (which is known to algorithm A). Informally, any finite hypothesis class H is learnable with $\approx \log |H|$ training examples.

If the true label function l belongs to H, then $\forall \epsilon$, we can produce h such that risk $R_{\mathbf{D}}(h) \leq \epsilon$, with probability $\geq 90\%$ using $\frac{\log |H|}{\epsilon^2} samples$.