Lecture #5: Topic

Instructor: Aditya Bhaskara *Scribe: Anthony Chyr*

CS 5966/6966: Theory of Machine Learning

January 25^{th}, 2022

Abstract

In this lecture we look into Empirical Risk Minimization and drawing Representative samples from distribution, and the consequences resulting from this including the Chernoff bound (Hoeffding).

1 Recap

- Goal of statistical learning: given some independent identically distributed (iid) samples from unknown distribution D with some labels, to learn a good h. $D; h; \text{risk}_D(h) = Pr[h(x) \neq \text{label}(x)]$.
- No free lunch theorem: we need to restrict the hypothesis/concept class before we start learning. Otherwise the problem becomes too hard.
- Probably Approximately Correct (PAC): learnability of a concept class H over domain X. Informally, for all $f \in H$ for any distribution D over X, given examples of the form $(x, f(x))$, we can learn a hypothesis h such that $\text{risk}_D(h)$ is $< \epsilon$, with high probability $(1-\delta)$ for some parameter δ.
- Agnostic learning: means f need not belong to H.

2 Generic Algorithm

Empirical Risk Minimization (ERM) is given samples $(x, f(x))$ find hypothesis $h \in H$ that minimizes the total error on the samples.

An example of doing empirical risk minimization efficiently is learning half spaces in the realignable case.

If the samples are representative of the distribution, for every hypothesis in a class, the error on samples is approximately equal to the error on the distribution (the risk). In other words it outputs h with small generalization error.

Example: let H be the set of all linear threshold functions in 1-D space. Let D be between -1 and 1. The threshold θ defines H. The goal is to learn with error (risk bound) ≤ 0.1

3 Representative Sample

Let H be a hypothesis class and X be an input space with a distribution D on it, and let f be a target function. Samples $S \subseteq X$ is said to be ϵ — "representative" if for all h in H, we have:

$$\left| \frac{1}{|X|} \text{error}(S, h) - \text{risk}_D(h, f) \right| < \epsilon$$

This applies to the realignable and agnostic cases.

4 Bounding Risk

If we happen to get a representative sample, the ERM gives the desired bound on risk!

A question we can ask: is a sample representative "with high probability"?

The claim we're trying to prove is: If we perform ERM on a representative sample S, we obtain a hypothesis with risk $\leq$ (best-risk-in-H) ϵ.

So let's suppose you have some sample S and through ERM we get h, with optimal hypothesis $h*$. By $h*$ we mean this is the hypothesis with the least risk with respect to D.

Another way to say "Empirical Risk Minization" is to call it "Sample Risk Minimization".

$$\text{error}(S, h) := \sum_{x \in S} \mathbb{1}[h(x) \neq \text{label}(x)]$$

NOTE: $\mathbb{1}$ is the indicator function, which is 1 when true and 0 otherwise.

$$risk_D(h) \leq \frac{1}{|S|} \text{error}(S, h) + \epsilon \leq \frac{1}{|S|} \text{error}(S, h*) + \epsilon$$

The reason why you can have a lower risk for the empirical case than the optimal case is because you can have overfitting!

The question we're trying to solve is: what is the smallest sample we can acquire that will still generalize well when doing empirical risk minimization. This results in the **Chernoff bound (Hoeffding)**. Suppose $X_1, X_2, ...X_n$ are n iid samples from a distributions with mean μ and support $[a, b]$. Then we have

$$Pr\left[\left| \frac{1}{n}(X_1 + \cdots + X_n) - \mu \right| > \epsilon \right] \leq 2 \exp\left(-\frac{\epsilon^2 n}{(a-b)^2} \right)$$

In other words:

$$Pr[|\text{sample avg} - \mu| > \epsilon] \leq \text{exponentially small in number of samples}$$

This also occurs in the central limit theorem in probability and statistics.

The support $[a, b]$ defines where the probability distribution function is not zero.

Furthermore, this implies that $n \approx \frac{1}{\epsilon^2} \log\left(\frac{2}{\delta}\right)$.

This further implies that finite classes are learnable scaling on the order of $O\left(\frac{1}{\epsilon^2} \log \frac{|H|}{\delta}\right)$.

NOTE: these bounds are very useful to know. It says samples capture the space we are trying to learn.