TTIC 31120: Statistical and Computational Learning Theory

Instructor: Nati Srebro

Autumn 2018

Contents

TTIC 31120: Learning Theory **October 1, 2018**

Lecture 1: What is Learning?

Lecturer: Nati Srebro *Scribes: Akilesh Tangella*

TTIC 31120: Learning Theory **October 3, 2018**

Lecture 2: Statistical Learning and PAC Learning

Lecturer: Nati Srebro *Scribes: Owen Melia*

2.1 Introduction

We have moved from the online learning setting to a "batch learning" one. Now we have an unknown Source Distribution , a probability distribution over $(, Y)$. It may be convenient to think of as a distribution over and $|Y$, and at other times we will think of as a distribution over Y and $Y|$. We will be attempting to construct a predictor h given $S \sim D^m$, an independent and identically distributed sample of m training points. We make two assumptions on the training set S:

1. S is an i.i.d sample.
2. S is drawm from the same source distribution which we will later use the predictor on.

These assumptions are often left unsatisfied in the real world (Prof. Srebro's example of a sample of dash-cam pictures to create a pedestrian recognition tool) Our goal is to construct a predictor $h :\to Y$ with expected error [1] $L_(h)$ is minimized.

$$L_(h) =_{(,Y)\sim} [h(x) \neq y] \tag{2.1}$$

This type of learning takes the following process:

1. Sample $S \sim^m$
2. Use some learning rule to construct $h = A(S)$ where $A : S \to Y$
3. Ship h and use it to predict on some future examples drawn from

2.2 From Expected Error to Empirical Error

Because we cannot observe $L_(h)$ at the time of learning, we must turn to an estimate. One natural estimate is the empirical error, $L_S(h)$

$$L_S(h) = \frac{1}{m}\sum_{t=1}^{m} [[h(x_t) \neq y_t]]$$

A natural question arises: how good is this estimate? What is the distribution of $|L_S(h) - L_(h)|$? The first way to approach this question is to observe that $L_S(h)$ is a Binomially-distributed random variable (scaled by $\frac{1}{m}$). It is a success/failure experiment with m trials and probability of success $L_(h)$. This leads to some observations:

$$L_S(h) \sim \frac{1}{m} binom(m, L_(h))$$

The mean of a Binomial random variable is the product of its parameters:

$$\mathbb{E}[L_S(h)] = \frac{1}{m} m L_(h) = L_(h)$$

We can also apply the Normal approximation:

$$L_S(h) \approx \mathcal{N}\left(L_(h), \sqrt{\frac{L_(h)(1 - L_(h))}{m}}\right)$$

[1] Expected error is sometimes called 'risk' and 'loss'

To consider tail probabilities, we need a bound which is easier to work with analytically. Hoeffding's Inequality states that for a random variable $X \sim binom(n, p)$,

$$[|X - \mathbb{E}X| \geq t] \leq 2e^{-2nt^2}$$

So if we want to bound the probability of a tail event as δ, we set $\delta = 2e^{-2nt^2}$ and solve for t, which gives us a closeness guarantee for each value of δ.

$$t = \sqrt{\frac{\ln\left(\frac{\delta}{2}\right)}{2m}}$$

So we can say that with probability $\geq 1 - \delta$,

$$|L_S(h) - L_(h)| \leq \sqrt{\frac{\ln\left(\frac{\delta}{2}\right)}{2m}}$$

2.3 Empirical Risk Minimization

We now have empirical risk, an estimate for the actual risk (expected error), which is both computable for every $h \in$ and gives a bound on the actual risk. This leads us to a learning rule called Empirical Risk Minimization.

$$\text{ERM}_(S) = \hat{h} = \arg\min_{h\in} L_S(h)$$

The learning rule chooses the hypothesis h which minimizes empirical error.

But is the following inequality valid (with probability $\geq 1 - \delta$)?

$$|L_(\hat{h}) - L_S(\hat{h})| \leq \sqrt{\frac{\ln\left(\frac{\delta}{2}\right)}{2m}}$$

The answer is no, and to illustrate why, we turn to the Monkey Pollster example. Suppose there are 8 elections occurring next Tuesday, and a bunch of pollsters are attempting to predict the outcomes. Since these pollsters are actually monkeys flipping 8 fair coins, [all correct] $= 2^{-8} = 0.004$. But there are 1,000 such pollsters, and a newspaper will break the story about the best pollster on Wednesday morning. The best pollster of the 1,000 is very likely completely correct:

$$[\text{best pollster all correct}] \approx 1 - e^{-\frac{1000}{256}} = 0.98$$

We see that the probability of a randomly selected pollster being very far from correct half the time (far from the expected error) is relatively low, but when we choose the best pollster after observing a sample, the pollster's performance on the sample deviates from its mean.

Consider a slight formalization of the above example. Let y be random and unpredictable. Then $L_(h) = \frac{1}{2}$. For each h we construct, we have $L_S(h) = 0$ with probability $\frac{1}{2^{|S|}}$. Thus for any hypothesis, there is a low probability that the empirical error will deviate far from the expected error. But if we have a large hypothesis class, $|| > 2^{|S|}$, then it becomes likely that the hypothesis chosen by ERM will have $L_S(\hat{h}) = 0$

The following bound on empirical error distance holds true before we see the sample, and it still leaves some tiny probability that the deviance is huge:

$$\forall h, \left[|L_S(h) - L_(h)| \geq t\right] \leq 2e^{-2mt^2} \tag{2.2}$$

What we want is a guarantee about the probability that every hypothesis comes with a bounded difference between empirical and expected losses- we want to switch the quantifiers. So we look to bond the following probability:

$$\Big(\forall h|L_S(h) - L_(h)| \geq t\Big) \geq \dots$$

We can use equation (2.2) and a union bound:

$$\Big(\exists h \in |L_(h) - L_S(h)| \geq t\Big) \leq \sum_{h\in} \Big(|L_S(h) - L_(h)| \geq t\Big) \leq ||2e^{-2mt^2}$$

By setting $\delta = ||2e^{-2mt^2}$, we can bound the probability of a tail event. For any hypothesis class and any source distribution ,

$$_{S\sim^m}\left(\forall_{h\in H}, |L_(h) - L_S(h)| \leq \sqrt{\frac{\ln|| + \ln\frac{2}{\delta}}{2m}}\right) \geq 1 - \delta \tag{2.3}$$

2.4 ERM Learning Guarantees

So we can now begin making guarantees about the difference between the loss and the empirical loss. The first guarantee is quantified over any hypothesis class and any source distribution , any sample $S \sim^m$, with probability $1 - \delta$.

$$L_(\hat{h}) \leq L_S(\hat{h}) + \sqrt{\frac{\ln|| + \ln\frac{2}{\delta}}{2m}} \tag{2.4}$$

The bound (2.4) is a post-hoc guarantee which can only be calculated after is sampled and $\hat{h}$ is constructed. This is relatively powerful: there is no assumption made on , and the bound grows relatively slowly with the size of . But this bound can be improved in two different ways: first by using a sharper approximation of binomial tail probabilities. We rely on Hoeffding's bound, but sharper ones exist. The second method of approximating $L_(\hat{h})$ is evaluating performance on a set of testing observations S'. From a similar Hoeffding approximation and an arbitrary learning rule, with probability $\geq 1 - \delta$

$$L_(A(S)) \leq L_{S'}(A(S)) + \sqrt{\frac{\ln\frac{1}{\delta}}{2|S'|}} \tag{2.5}$$

The second guarantee is an extention of (2.4). If we let $h^* = \arg\min_{h\in} L_(h)$. Then $L_S(\hat{h}) \leq L_S(h^*)$. From (2.3) we see $L_(h) \leq L_S(h) + \sqrt{\frac{\ln||+\ln\frac{2}{\delta}}{2m}}$. So we have

$$L_(\hat{h}) \leq L_S(h^*) + \sqrt{\frac{\ln|| + \ln\frac{2}{\delta}}{2m}} \leq L_(h^*) + \sqrt{\frac{\ln|| + \ln\frac{2}{\delta}}{2m}} + \sqrt{\frac{\ln|| + \ln\frac{2}{\delta}}{2m}}$$

$$L_(\hat{h}) \leq \inf_{h\in} L_(h) + 2\sqrt{\frac{\ln|| + \ln\frac{2}{\delta}}{2m}} \tag{2.6}$$

The bound in (2.6) is an a-priori guarantee which can be calculated before any sampling or estimation. It can be useful to think of the first term of (2.6) as the approximation error, and the second term as the estimation error. If we know from expert knowledge that there exists some good predictor $h^* \in$, then we know that we can create a good approximation, and our only problem lies in estimating h^* So then, in the ERM setting, we can state that with probabilty ≥ 1 - δ, if we have a sample size of

$$m = 2\frac{\ln|| + \ln\frac{2}{\delta}}{\epsilon^2} = O(\frac{\log||}{\epsilon^2}) \tag{2.7}$$

then we can ensure $L_(\hat{h}) \leq L_(h^*) + \epsilon$, and we can call m the sample complexity bound, which gives us a minimum sample size to ensure our expected error is within ϵ of an optimum. This can point toward a general notion of a hypothesis class' learnability: with a large enough sample, we can bound the risk arbitrarily close to the optimum. The second equality in (2.7) gives us important asymptotics for m with respect to and ϵ.

2.5 PAC Learning

Now for some definitions of learnability under the Probably Approximately Correct learning model, first introduced by Leslie Valiant.

1. A hypothesis class is **PAC-Learnable** (in the realizable case) if $\exists$ some learning rule A such that $\forall \epsilon, \delta > 0$, $\exists m(\epsilon, \delta)$ such that $\forall$ ($\exists h \in s.t. L_(h) = 0$ i.e. realizable by), $\forall^{\delta}_{S \sim m(\epsilon,\delta)}, L_(A(S)) \leq \epsilon$

2. A hypothesis class is **Agnostically PAC-Learnable** if $\exists$ some learning rule A such that $\forall \epsilon, \delta > 0$, $\exists m(\epsilon, \delta)$ such that $\forall$, $\forall^{\delta}_{S \sim m(\epsilon,\delta)}, L_(A(S)) \leq \inf_{h \in} L_(h) + \epsilon$

From above, we have seen that all finite hypothesis classes are PAC-Learnable, and that

$$m_(\epsilon, \delta) \leq m_{ERM,}(\epsilon, \delta) \leq O\left(\frac{\ln || + \ln \frac{1}{\delta}}{\epsilon^2}\right)$$