COS597A: Machine Learning Seminar 2017
Generalization bounds

Lecture 2 notes – 09/21

Lecturer: *Sanjeev Arora* Scribe: *Hrishikesh Khandeparkar*

1 Theories, explanations, beliefs and why they work

1.1 Occam and his Razor

As humans, we seek to explain phenomenon that occur around us by trying to explain occurrences of events by simple underlying theories. Although this intuition of using "simple" theories to explain complex phenomenon may seem quite natural to us, it wasn't as obvious to mankind several times in the past. Nicholas Copernicus was called a heretic when he first proposed his theory that the sun was at the center of the solar system, despite his theory being vastly simpler than the one that had the earth at the center.

This intuition of "simpler" theories being "better" was first made well known by William of Ockham/Occam who said "Entia non sunt multiplicanda praeter necessitatem" which roughly translates to

Occams Razor (hand-wavy form): "More things should not be used than are necessary".

Famously known as Occam Razor (referring to the "razor" shaving away bad explanations), this principle is philosophically widespread and heavily used in science. It says that simpler hypothesis (and theories) are usually better at explaining phenomenon we see. However, this principle is very vague and hand wavy. It may work in practice, but this is a theory class so we need to better understand why this principle seems to work. What does "simpler" even mean? And what makes "simpler" theories "better"?

1.2 Empirical Risk Minimization (ERM)

If we were to critically think of the way we seek to explain things, we would see that all of science falls into a simple general framework. (Caution: Philosophers of science may find this framework simplistic.)

- Start with a set of initial candidate theories.
- Collect experimental data that the theories claim to explain
- Check which theories are consistent with the data, and throw away the rest. Amongst those theories that don't make errors, we usually settle on the simplest theories.

Machine learning can be seen as a search for theories that explain data. Unlike in mathematics or physics, where theories are incorrect if they even make one wrong prediction, our framework allows errors.

- Start with a set of initial candidate classifiers.
- Collect experimental data that the classifiers claim to explain
- Check which classifiers are **largely** consistent with the data, and throw away the rest. Amongst classifiers that don't make **many** errors, we usually settle on the **"simplest"** ones.

Let us formalize this framework. We use the usual notation of data points as ordered pairs of inputs and labels $\mathbf{z_i} = (\mathbf{x_i}, \mathbf{y_i})$. We further define

$$\begin{aligned}
\mathbf{D} &: \text{Distribution on labelled data} \\
\mathbf{S_m} &: \text{m i.i.d samples of data points } z_i \text{ drawn from } \mathbf{D} \\
\mathcal{H} &: \text{The set of classifiers } h \text{ which we consider to explain the data} \\
l(h,z) &: \text{Loss of classifier } h \in \mathcal{H} \text{ on point } z = (x,y) \\
L_{S_m}(h) &: \text{Average loss of classifier } h \text{ on the sample } S_m
\end{aligned}$$

Then, for a hypothesis $h \in \mathcal{H}$ we have that $L_{S_m}(h) = \mathbb{E}_{z \sim S_m}[l(h,z)]$. Here, $\mathbb{E}_{z \sim S_m}$ means expectation over z drawn uniformly at random from $\mathbf{S_m}$. We drop the m and say S from now on, with the m being implied. This loss $L_S(h)$ models the steps of measuring how well our theory (here, hypothesis h) can explain the data. A simple loss function is binary —loss 0 for correct classification and loss 1 for incorrect. In addition it may have a term *complexity*(h) that measures the complexity of h. One simple such measure is the sum of the absolute values of the numbers that describe h.

Now we seek to find a

$$h_S = \arg\min_h L_S(h)$$

which best explains the dataset. We denote $L_S = L_S(h_S)$ and call it the empirical loss. Thus, we call this method **Empircal Risk Minimization** as it minimizes the empirical error.

However the empirical error isn't the true error/loss that we are actually interested in.

In order to say that a classifier h is actually good, we need it to perform well on real world data. Specifically, we are seeking to minimize $L_D(h) = \mathbb{E}_{z \sim D}[l(h,z)]$. We call this $L_D(h)$ the true loss of the hypothesis h. How is $L_D(h)$ related to $L_S(h)$ for a given h?

We define the generalization error as

$$\Delta_S(h) = L_D(h) - L_S(h)$$

This measures how well hypothesis h generalizes to a distribution D when its performance on a sample S drawn from D is known.

Intuitively, if the generalization error is large then the hypothesis's performance on sample S does not accurately reflect the performance on the full distribution of examples, so we say it *overfitted* to the sample S.

A trivial example of this is the hypothesis class that assigns the known label to all seen examples, and the label 0 to all unseen examples. Clearly, this hypothesis class can achieve 0 loss on any dataset but won't perform well in the real world. Another example from folklore is how conspiracy theorist can join seemingly random facts to explain an outcome with a theory but they are clearly not to be relied on for making good predictions about future.

2 Generalization Theory

Generalization theory tries to upper bound this error $\Delta_S(h)$. Sanjeev has a different way of phrasing this: *If overfitting occured and $\Delta_S(h)$ was high, then the hypothesis class was complex in some way.* Generalization theory formalizes what it means for classes to be complex. Sanjeev emphasizes that it is primarily a *descriptive* theory that gives a name for the type of complexity. But it is not a *prescriptive* theory, in that it gives no insight into how estimate this complexity.

2.1 Rademacher Complexity OR "ability to correlate with random labels"

Now we turn to Rademacher complexity. Sanjeev cautions that often this topic confuses students, or falsely impresses them. Possibly because standard accounts use the wrong definition and don't clarify that the basic point is rather trivial.

We first formalize the idea of the "complexity" of a hypothesis class. To this this we use the notion of Rademacher complexity inspired by our intuition of classifying random labels. First, let

$$\begin{aligned}
\mathcal{H} &: \text{Hypothesis class} \\
S &: 2m \text{ i.i.d samples from D} \\
\sigma_i &= \begin{cases} 1 & i \in \{1, ..m\} \\ -1 & i \in \{m+1, ..2m\} \end{cases}
\end{aligned}$$

Now

$$\mathcal{R}_{m,D}(\mathcal{H}) = \mathop{\mathbb{E}}_{S \sim D^m} \left[\frac{1}{2m} \sup_{h \in \mathcal{H}} \left| \sum \sigma_i h(z_i) \right| \right]$$

We call $\mathcal{R}_{m,D}(H)$ the *Rademacher Complexity* of H on a distribution D.

Note that flipping the sign in front of the loss function turns high loss into low and vice versa, so it is effectively like flipping the label of the underlying datapoint. Thus effectively we are flipping the labels of half the datapoints randomly and retaining the labels of the other half. The definition requires finding classifier h in the class that correlates well with this random relabeling; this is the usual interpretation of Rademacher complexity.

(Sanjeev's definition is different from the one used in literature where σ_i is picked randomly for each i, but you can convince yourself that picking exactly half -1s and half +1s isn't too different.)

Claim 1. For a given loss function,$\forall \delta > 0$, with probability $> 1-\delta$, we have that the generalization error of all hypothesis $h \in \mathcal{H}$, on a sample S of m i.i.d. samples drawn from a distribution D, is

$$\Delta_S(h) \leq 2\mathcal{R}_{m,D}(\mathcal{H}) \quad \left(+\, O\left(\frac{1}{m} ln\left(\frac{1}{\delta}\right)\right)\right)$$

The main takeaway of this claim is that generalization error can be upper bounded by the Rademacher complexity.

(The part in the square brackets comes from concentration bounds from the sample S being "representative" of the distribution and having the generalization bound hold for all $h \in \mathcal{H}$)

Proof Sketch 1. Suppose for a random sample S the generalization error is high.consider the following thought experiment

- Split S_m into sets S_1 and S_2 randomly, with the sets being of equal size.
- For a given h (**picked independently of S**), consider $L_{S_1}(h)$ and $L_{S_2}(h)$
- For large enough m, we have that $L_{S_2}(h) \approx L_D(h)$ and thus $L_D(h) - L_{S_1}(h) \approx L_{S_2}(h) - L_{S_1}(h)$ Here S_2 is like the "test set" and S_1 is like the "training set". Thus,

$$\Delta_S(h) \approx L_{S_1}(h) - L_{S_2}(h)$$

- But since S_1 and S_2 are randomly picked, we can just consider S_1 as the first half of the sample S and then the difference reduces to

$$\underset{S\sim D^m}{\mathbb{E}}\left[\underset{z\sim S_2}{\mathbb{E}}[L(h,z)] - \underset{z\sim S_1}{\mathbb{E}}[L(h,z)]\right] \leq \underset{S\sim D^m}{\mathbb{E}}\left[\frac{1}{m}\left|\sum \sigma_i h(z_i)\right|\right] \leq \sup_{h\in\mathcal{H}} \underset{S\sim D^m}{\mathbb{E}}\left[\frac{1}{m}\left|\sum \sigma_i h(z_i)\right|\right]$$

- Thus we have

$$\Delta_S(h) \leq \sup_{h\in\mathcal{H}} \underset{S\sim D^m}{\mathbb{E}}\left[\frac{1}{m}\left|\sum \sigma_i h(z_i)\right|\right] \leq \underset{S\sim D^m}{\mathbb{E}}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\left|\sum \sigma_i h(z_i)\right|\right] = 2\mathcal{R}_{m,D}(\mathcal{H})$$

  (The 2 in the end is simply because we defined Rademacher complexity with a set of size 2m. We also leave out the concentration term that arrives due to the approximation of the generalization error using a training and test set. For a more formal treatment of this topic refer to the chapter in Understanding Machine Learning: From Theory to Algorithms, Shalev-Shwartz, Shai and Ben-David, Shai)

Example: We can show that the Rademacher complexity of the set of linear classifiers (unit norm vectors $U = \{w | w \in \mathbb{R}^d, |w|_2 = 1\}$), on a given sample $S = (x_1, x_2, x_3, ..x_m)$ (each $x_i \in \mathbb{R}^d$) is $\leq \frac{\max_i |x_i|_2}{\sqrt{m}}$

2.2 PAC-Bayesian generalization bounds

Now we consider the Bayesian approach of having a prior on the hypothesis class, and (possibly) using data to arrive at a posterior distribution.

In this setting, we define a prior as a distribution P on the hypothesis class $\mathcal{H}$. This in a way represents a hierarchy over hypotheses in $\mathcal{H}$. Now, instead of picking a single prediction hypothesis, an algorithm outputs a distribution Q called the posterior over the hypothesis class $\mathcal{H}$. This Q represents the algorithms "belief" in hypothesis of $\mathcal{H}$. The prediction rule for a new data point x picks a random hypothesis $h \in \mathcal{H}$ according to Q and predicts $h(x)$. Thus, the error of this classifier is

$$\mathop{\mathbb{E}}_{h \sim Q}[L_D(h)]$$

Now, the main result in PAC-Bayesian learning is that

Claim 2. Consider a distribution D on the data. Let P be a prior distribution over hypothesis class $\mathcal{H}$ and $\delta > 0$. Then with probabilty $\geq 1 - \delta$, on a i.i.d. sample $S = (z_1, z_2.., z_m)$ of size m from D, for all distributions Q over $\mathcal{H}$ (which could possibly depend on S), we have that

$$\Delta_S(Q(\mathcal{H})) = \mathop{\mathbb{E}}_{h \sim Q}[L_D(h)] - \mathop{\mathbb{E}}_{h \sim Q}[L_S(h)] \leq \sqrt{\frac{D(Q||P) + \ln m/\delta}{2(m-1)}}$$

What this claim is saying is that the generalization error is upper bounded by the square root of the KL-divergence of the distributions (plus some terms that arise from concentration bounds).

Thus, in order to minimize the error on the real distribution, we should try to simultaneously minimize the empirical error as well as the KL-divergence between the posterior and the prior.

Proof Sketch 2. First, lets observe that for a fixed h, using a standard Hoeffdings inequality, we have that

$$\Pr_S[\Delta(h) > \epsilon] \leq e^{-2m\epsilon^2}$$

which means the generalization error of a given hypothesis exceeds a given constant is exponentially small. This means that with very high probability, the generalization error is bounded by a small constant. Roughly, this says that $\sqrt{m}\Delta_S(h)$ behaves like a univariate gaussian. Using concentration bounds, this further implies that

$$\mathop{\mathbb{E}}_S[e^{2(m-1)\Delta(h)^2}] \leq m$$

and therefore, with high probability over S,

$$e^{2(m-1)\Delta(h)^2} = O(m) \tag{1}$$

Now consider the expression (derived by working backwards from statement of the claim)

$$2(m-1)\mathop{\mathbb{E}}_{h\sim Q}[\Delta(h)]^2 - D(Q||P) \leq 2(m-1)\mathop{\mathbb{E}}_{h\sim Q}[\Delta(h)^2] - D(Q||P)$$

where the inequality is by convexity of squares. This in turn is now

$$\begin{aligned} 2(m-1)\mathop{\mathbb{E}}_{h\sim Q}[\Delta(h)^2] - D(Q||P) &= \mathop{\mathbb{E}}_{h\sim Q}[2(m-1)\Delta(h)^2 - \ln\frac{Q(h)}{P(h)}] \\ &= \mathop{\mathbb{E}}_{h\sim Q}\left[\ln\left(e^{2(m-1)\Delta(h)^2}\frac{P(h)}{Q(h)}\right)\right] \\ &\leq \ln\mathop{\mathbb{E}}_{h\sim Q}\left[\left(e^{2(m-1)\Delta(h)^2}\frac{P(h)}{Q(h)}\right)\right] \end{aligned}$$

where the last inequality uses Jensens inequality along with the concavity of ln. Also, we have

$$\ln\mathop{\mathbb{E}}_{h\sim Q}\left[\left(e^{2(m-1)\Delta(h)^2}\frac{P(h)}{Q(h)}\right)\right] = \ln\mathop{\mathbb{E}}_{h\sim P}\left[\left(e^{2(m-1)\Delta(h)^2}\right)\right]$$

This last step is a standard trick – using the KL-Divergence term to switch the distribution over which expectation is taken!

Recapping, we thus have that

$$2(m-1)\mathop{\mathbb{E}}_{h\sim Q}[\Delta(h)]^2 - D(Q||P) \leq \ln\left(\mathop{\mathbb{E}}_{h\sim P}\left[e^{2(m-1)\Delta(h)^2}\right]\right) \tag{2}$$

Now using (1),

$$\mathop{\mathbb{E}}_{S}\left[\mathop{\mathbb{E}}_{h\sim P}\left[\ e^{2(m-1)\Delta(h)^2}\right]\right] = \mathop{\mathbb{E}}_{h\sim P}\left[\mathop{\mathbb{E}}_{S}\left[e^{2(m-1)\Delta(h)^2}\right]\right] \leq m$$

as P is independent of S. Thus, (1) implies that with high probability over S,

$$\mathop{\mathbb{E}}_{h\sim P}\left[e^{2(m-1)\Delta(h)^2}\right] = O(m) \tag{3}$$

Thus, combining (2),(3) we get

$$\begin{aligned} 2(m-1)\mathop{\mathbb{E}}_{h\sim Q}[\Delta(h)]^2 - D(Q||P) &\leq O(\ln(m)) \\ \mathop{\mathbb{E}}_{h\sim Q}[\Delta(h)]^2 &\leq \frac{O(\ln(m)) + D(Q||P)}{2(m-1)} \\ \mathop{\mathbb{E}}_{h\sim Q}[\Delta(h)] &\leq \sqrt{\frac{O(\ln(m)) + D(Q||P)}{2(m-1)}} \end{aligned}$$

which completes our proof sketch.