IFT 6085 - Lecture 8
Statistical learning theory: PAC-Bayes bounds

Winter 2020: Howard Huang
Winter 2019: Zafarali Ahmed, Nadeem Ward
Winter 2018: Jérémy Trudel, Lluís Castrejón **Instructor:** Ioannis Mitliagkas

1 Summary

In the previous lecture we introduced the basics of Statistical Learning Theory. We established the setting for PAC Learning and defined the concepts of *risk*, *empirical risk* and *generalization gap*. We then used *Hoeffding's Inequality* to establish a bound on the generalization gap for finite hypothesis classes $\mathcal{H}$.

In this lecture we continue our crash course on Statistical Learning Theory by introducing new concepts in order to get tighter bounds on the generalization gap, namely *Occam's (Razor) Bound* and *PAC Bayesian learning*.

2 PAC Learning

In this section we recap our notation from last time. Probably Approximate Correct (PAC) Learning is a framework for analyzing machine learning algorithms. Assume that we have a hypothesis class $\mathcal{H}$ - the set of all possible model configurations - and a set of samples that form our dataset $S = \{z_1, z_2, ..., z_n\}$ with $z_i = (x_i, y_i)$ and $z_i \sim \mathcal{D}$ i.i.d - where $\mathcal{D}$ is the data distribution. Assume also that we have defined a bounded function $l : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ (loss function) that quantifies the mismatch between two elements of $\mathcal{Y}$. In this PAC Learning setting, we define the performance of a hypothesis using:

Definition 1 (Risk)**.**

$$R[h] = \mathbb{E}_{(x,y)\sim\mathcal{D}}[l(h(x), y)]$$

Recall that we do not have access to the data generating procedure, $\mathcal{D}$, so we resort to using the empirical risk evaluated on our data set, S:

Definition 2 (Empirical Risk)**.**

$$\hat{R}_s[h] = \frac{1}{n}\sum_{i=1}^{n} l(h(x_i), y_i) = \frac{1}{n}\sum_{i=1}^{n} l(h, z_i)$$

Since we can only use S to discover hypotheses, h_s, we define a generalization gap based on how well h_s does on the true risk (Equation 1):

Definition 3 (Generalization Gap)**.**

$$\epsilon_{gen}(h_s) = |R[h_s] - \hat{R}_s[h_s]|$$

Our main result gave us a bound on the sample size:

Theorem 4. *For all $h \in \mathcal{H}$, If*

$$n = \Omega\left(\frac{log(\frac{|\mathcal{H}|}{\delta})}{\epsilon^2}\right)$$

then with probability at least $1-\delta$:

$$\left|R[h] - \hat{R}_S[h]\right| < \epsilon$$

Recall that to get this result we bounded $P\big(\left|R[h_s] - \hat{R}_S[h_s]\right| > \epsilon\big)$ by $P\big(\bigcup_{h\in\mathcal{H}} \left|R[h] - \hat{R}_s[h]\right| > \epsilon\big)$. The reason why we couldn't just apply *Hoeffding's inequality* directly on $P\big(\left|R[h_s] - \hat{R}_S[h_s]\right| > \epsilon\big)$ is because the term $\hat{R}_S[h_s]$ is **not** the average of i.i.d random variables. If we look closely at the term

$$\hat{R}_S[h_s] = \frac{1}{n}\sum_{\forall z\in S} l(h_s, z)$$

h_s is constructed based on the data $z \in S$. These same $z \in S$ are used in the selection of h_s and used again in the losses $l(h_s, z)\ \forall z \in S$ making them dependent. When we consider h to be arbitrarly chosen, then $\hat{R}_S[h]$ doesn't suffer from the dependence problem which makes it an average of independent random variables.

3 Occam's (Razor) bound

Simply put, *Occam's bound* tells us to put a distribution over the countably infinite hypothesis class $\mathcal{H}$ that is independent of dataset S we will receive (The PAC bound can be treated as Occam's bound with a uniform prior). This can be thought of as "placing our bets" on the different hypotheses $h \in \mathcal{H}$ prior to seeing the actual data. We call this distribution the *prior* P on the hypothesis class $\mathcal{H}$. We will see that in doing so we will get bounds on the generalization gap that no longer depend on the size of the hypothesis class, $|\mathcal{H}|$. These bounds now become variable depending on how we weigh each individual hypothesis h, i.e. $P(h)$.
This analysis will use the assumption that $\mathcal{H}$ is discrete (countable) in order to use the union bound. Also, we will assume that the loss function $l(h, z)$ is bounded in order to use Hoeffding's inequality. Without loss of generality assume it's bounded on the interval $[0, 1]$.

Theorem 5. *(McAllester [1]) Given a prior distribution P on $\mathcal{H}$, $\sum_{\forall h} P(h) = 1$, with probability at least $\geq 1-\delta$ over $S \sim \mathcal{D}^n$ we have that the following holds true for all h $\in \mathcal{H}$:*

$$R[h] \leq \hat{R}_S[h] + \sqrt{\frac{\log\frac{1}{P(h)} + \log\frac{2}{\delta}}{2n}}$$

Proof. This proof follows the same outline as the one seen for PAC learning. For an arbitrary hypothesis h we demand that:

$$\mathbb{P}_{S\sim D}(|\hat{R}_s[h] - R[h]| > \epsilon) \leq \delta P(h)$$

Recall that, by Hoeffding's inequality, we have that:

$$\mathbb{P}_{S\sim D}(|\hat{R}_s[h] - R[h]| > \epsilon) \leq 2e^{-n\epsilon^2}$$

we can solve for ϵ by solving $2e^{-n\epsilon^2} = \delta P(h)$ and obtain $\epsilon(h) = \sqrt{\frac{\log\frac{1}{P(h)} + \log\frac{2}{\delta}}{2n}}$.

Given this, we can proceed to bounding our hypothesis of interest, h_S:

$$\begin{aligned}
\mathbb{P}_{S\sim D}(|\hat{R}_s[h_S] - R[h_S]| > \epsilon(h)) &\leq \mathbb{P}_{S\sim D}\big(\bigcup_{\forall h} |\hat{R}_s[h] - R[h]| > \epsilon\big) \\
&\leq \sum_{\forall h} \mathbb{P}_{S\sim D}(|\hat{R}_s[h] - R[h]| > \epsilon) \\
&\leq \sum_{\forall h} \delta P(h) = \delta
\end{aligned}$$

We used the union of all the hypothesis h to bound our hypothesis of interest h_S. line 2 follows from the union bound. What this tells us is that with probability at least $1-\delta$ we have:

$$|\hat{R}_s[h_S] - R[h_S]| \leq \sqrt{\frac{\log \frac{1}{P(h_s)} + \log \frac{2}{\delta}}{2n}}$$

□

Notice that this bound is no longer dependent on $|\mathcal{H}|$ and instead has the term $\log \frac{1}{P(h)}$. If our prior distribution gives more probability to h than this term will decrease therefore giving a tighter bound and vise versa. If, however, we don't give any probability to a hypothesis h (i.e $P(h) = 0$) then $\log \frac{1}{P(h)}$ will be undefined which provides a vacuous bound.

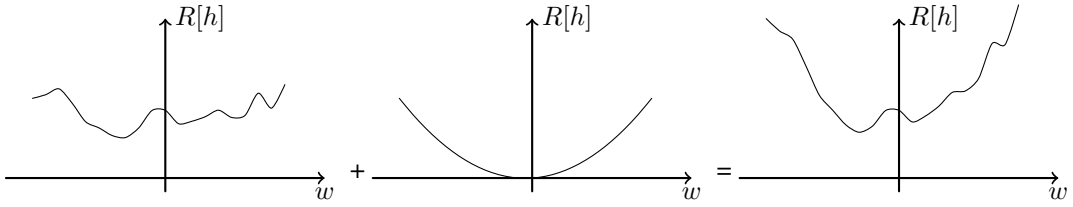

Figure 1: When an ℓ_2 regularization term is added to the learning algorithm, it adds concavity to the loss function, leading to a higher likelihood of choosing a set of weights near the origin. In such a situation, using a prior which places heavier "bets" near the origin makes sense. This figure is a visual representation of what this loss function looks like after the addition of a regularization term.

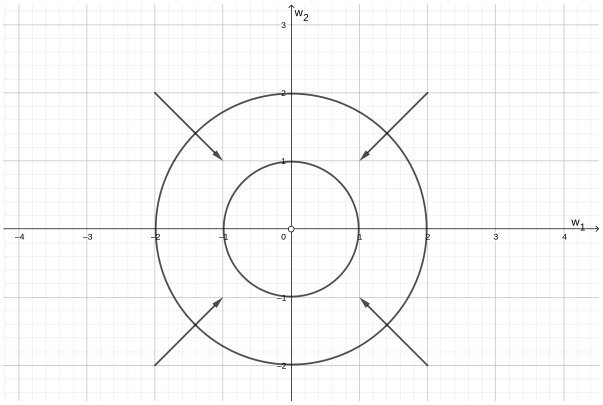

Figure 2: **Example:** Consider a simple linear classifier with 2 weights $\vec{w} = (w_1, w_2)$, which are stored using a 32 bit floats. This implies that the hypothesis class is finite with $|H| = 2^{32\times 2}$. Consider also that the loss function is bounded and has a *regularization* term of the form $\lambda||\vec{w}||$. The regularization term forces the weights to be as small as possible and corresponds to a prior P with higher probability assigned to $\vec{w}$ near the origin. In doing so we will get tight theoretical bounds for desirable hypotheses. Notice that this says nothing about the algorithm itself, we are just affecting how good our bounds are.

4 PAC Bayes

In PAC Bayes the basic idea is that we add a "posterior" Q on $\mathcal{H}$, in addition to the prior P we already had in the Occam's Bound. The basic recipe we follow is:

1. Set our "bets" using the prior, P, independent of the data.
2. We then collect some finite dataset $S \sim D$.
3. Select a posterior, Q, based on the data.

If we select the $Q = P$ as the posterior, then we get the same trade-off as in Occam's bound, wherein we get tight bounds on certain hypotheses, but not necessarily those we care about. With the addition of this posterior, we can derive a new bound on the generalization gap that depends on the KL-divergence between the prior and the posterior:

Theorem 6 (PAC Bayes bound)**.** *Given a prior probability distribution P over a hypothesis class H and a posterior probability distribution Q over H. Then:*

$$\mathbb{E}_{h\sim Q}[R[h]] \leq \mathbb{E}_{h\sim Q}[\hat{R}_S[h]] + \sqrt{\frac{D(Q||P) + \log(\frac{n}{\delta})}{2(n-1)}}$$

with probability $\geq 1 - \delta$.

where

Definition 7 (Kullback-Leibler Divergence)**.** *The Kullback-Leibler (KL) divergence between two distributions Q and P is defined as:*

$$D(Q||P) = \mathbb{E}_{h\sim Q}\left[\log \frac{Q(h)}{P(h)}\right]$$

is a measure of how far two probability distributions are[1]. In Theorem 6, the KL-divergence serves as a complexity measure.

Proof. (See Theorem 31.1 of [2]) Recall the Markov inequality states that $P(X \geq a) \leq \frac{\mathbb{E}(X)}{a}$, and we can apply this to a function $f(S)$ to get

$$\mathbb{P}_S(f(S) \geq \epsilon) = \mathbb{P}_S(e^{f(S)} \geq e^{\epsilon}) \leq \frac{\mathbb{E}_S[e^{f(s)}]}{e^{\epsilon}} \tag{1}$$

Let $\Delta(h) = \mathbb{E}_{h\sim Q}[R[h]] - \mathbb{E}_{h\sim Q}[\hat{R}_S[h]]$. We will apply equation (1) with the function

$$f(S) = \sup_Q \left(2(n-1)\mathbb{E}_{h\sim Q}(\Delta(h))^2 - D(Q||P)\right)$$

We now bound $\mathbb{E}_S[e^{f(S)}]$.

$$2(n-1)\mathbb{E}_{h\sim Q}(\Delta(h))^2 - D(Q||P) = \mathbb{E}_{h\sim Q}\left[\log(e^{2(n-1)\Delta(h)^2}P(h)/Q(h))\right]$$

Apply Jensen's inequality, which tells us that $\mathbb{E}[\log(\cdot)] \leq \log \mathbb{E}(\cdot)$

$$\leq \log \mathbb{E}_{h\sim Q}\left[e^{2(n-1)\Delta(h)^2}P(h)/Q(h)\right]$$

Use a "twist of measures" trick: $\mathbb{E}_{h\sim Q}[(\cdots)P(h)/Q(h)] = \sum_{h\in Q} Q(h)[(\cdots)P(h)/Q(h)] = \sum_{h\in Q}[(\cdots)P(h)] = \sum_{h\in P}[(\cdots)P(h)] = \mathbb{E}_{h\in P}[\cdots]$.

$$= \log \mathbb{E}_{h\sim P}\left[e^{2(n-1)\Delta(h)^2}\right]$$

Therefore

$$\mathbb{E}_S[e^{f(S)}] \leq \mathbb{E}_S\mathbb{E}_{h\sim P}[e^{2(n-1)\Delta(h)^2}]$$

Since P does not depend on S, we can swap the expectations

$$\mathbb{E}_S[e^{f(S)}] \leq \mathbb{E}_{h\sim P}\mathbb{E}_S[e^{2(n-1)\Delta(h)^2}] \tag{2}$$

[1]The KL-divergence is not symmetric - $D(Q||P) \neq D(P||Q)$ - and therefore is not a metric. Also note that $P = Q \iff D(P||Q) = 0$.

Next, we claim that for all h, we have $\mathbb{E}_S[e^{2(n-1)\Delta(h)^2}] \leq n$. Recall that Hoeffding's inequality tells us that

$$\mathbb{P}_S[\Delta(h) \geq \epsilon] \leq e^{-2n\epsilon^2}$$

This implies that $\mathbb{E}_S[e^{2(n-1)\Delta(h)^2} \leq n]$. Combining this with equation (2) and plugging into equation (1), we get

$$\mathbb{P}_S[f(S) \geq \epsilon] \leq \frac{n}{e^\epsilon}$$

With $\delta = n/e^\epsilon$, we get $\epsilon = \log(n/\delta)$, and we obtain that with probability at least $1-\delta$ and for all Q,

$$2(n-1)\mathbb{E}_{h\sim Q}(\Delta(h))^2 - D(Q||P) \leq \epsilon = \log(n/\delta)$$

Rearranging the inequality and using Jensen's inequality again, we conclude that

$$(\mathbb{E}_{h\sim Q}\Delta(h))^2 \leq \mathbb{E}_{h\sim Q}(\Delta(h))^2 \leq \frac{\log(n/\delta) + D(Q||P)}{2(n-1)}$$

□

If we update our distribution over hypotheses using the posterior, Q, so that h performs well on the empirical risk, it ensures that we can reduce $\mathbb{E}_{h\sim Q}[\hat{R}_S[h]]$ more than if we just sampled from the prior. However, we need to be careful of the complexity term that bounds our generalization on the true data distribution. If we choose a "good" posterior that is close to the prior, then the KL-divergence will become smaller and our bound will be tighter.
Here are some examples of posteriors:

1. Q assigns h_S a probability of 1, i.e. $\mathbb{E}_{h\sim Q}[h] = h_S$. This means that $D(Q||P) \to \infty$ and the bound explodes.
2. The posterior and the prior are the same, $Q = P$. This means that $D(Q||P) = 0$ and the bound becomes tight. Even though this bound is tight, it is tight for hypotheses we may not care about, i.e. $\mathbb{E}_{h\sim P}[\hat{R}_S[h]]$ can be large.

References

[1] D. McAllester. A PAC-Bayesian Tutorial with A Dropout Bound. *ArXiv e-prints*, July 2013.

[2] B.-D. S. Shalev-Shwartz S. Understanding machine learning: From theory to algorithms. *Cambridge*, 2014.