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

Scribe(s): [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 the *Hoeffding's Inequality* to establish a bound on the generalization gap for countable hypothesis classes $\mathcal{H}$.
In this lecture we continue our crash course on Statistical Learning Theory and learn about the *Occam's Razor Bound* and *PAC Bayes*, which build upon what we saw in the previous lecture.

2 PAC Learning

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 $S = \{z_1, z_2, ..., z_n\}$ with $z_i = (x_i, y_i)$ and z_i $\mathcal{D}$ i.i.d - where $\mathcal{D}$ is the data distribution. Assume also that we have defined a function $l : \mathcal{Y}x\mathcal{Y} \rightarrow [0, 1]$ (loss function) that quantifies the mismatch between two elements of $\mathcal{Y}$. In this PAC Learning setting, we define:

Definition 1 (Risk)**.**

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

Definition 2 (Empirical Risk (evaluated on the data samples))**.**

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

Definition 3 (Generalization Gap)**.**

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

We previously obtained the following bound for the generalization error:

Theorem 4. *Fix $h \in \mathcal{H}$, if*

$$n = O\left(\frac{log(\frac{1}{\delta})}{\epsilon^2}\right)$$

then

$$|R[h] - \hat{R}_s[h]| < \epsilon$$

with probability $\geq 1 - \delta$

Note: to obtain this bound we have assumed that $\mathcal{H}$ is countable and finite.
This bound is for a specific h, but how can we guarantee that all $h_i \in \mathcal{H}$ would have "good" generalization? Motivated by this, we find a bound for all h_i as follows.

We start by assuming we have a uniform probability distribution over $[H]$:

$$\delta(h) = \frac{\delta}{|\mathcal{H}|}$$

Observe that:

$$\sum_{h \in \mathcal{H}} \delta(h) = \delta \sum_{1}^{|\mathcal{H}|} \frac{1}{|\mathcal{H}|} = \delta$$

Theorem 5. *If*

$$n = O\left(\frac{log|H| + log\frac{1}{\delta}}{\epsilon^2}\right)$$

then

$$|R[h] - \hat{R}_s[h]| < \epsilon$$

with probability $\geq 1 - \delta$ *for all* $h \in \mathcal{H}$.

Now we show an example on how to compute $|\mathcal{H}|$. We define the *perceptron* class of models as:

$$l(w^T x + b), w \in \mathbb{R}^d, b \in \mathbb{R}$$

To represent this model, assuming we use 32 bit floats for each learnable parameter, we need $32(d+1)$ bits, so:

$$|H| = 2^{32(d+1)}$$

$$log|H| = log(2^{32(d+1)})$$

$$log|H| = (d+1)log(2^3 2)$$

3 Occam's (Razor) bound

Simply put, an *Occam's bound* states that if two hypotheses explain the data equally well, the one that makes less assumptions is generally preferable. In this context, we'll use it to state that the generalization loss is near training loss when the number of bits needed to write the rule is small compared to the sample size [1]. For the assumption to hold true, we consider that $\mathcal{H}$ is finite and we define the following:

Definition 6.

$$\epsilon(h) = \sqrt{\frac{ln\frac{1}{P(h)} + ln\frac{2}{\delta}}{2n}}$$

Definition 7.

$$\sum_{h \in \mathcal{H}} P(h) = 1$$

Theorem 8. *Given P on* $\mathcal{H}$, $\sum P(h) = 1$, *a sample size n, with probability at least* $\geq 1 - \delta$ *over S*

$$\forall h \in \mathcal{H}, R[h] \leq \hat{R}_s[h] + \sqrt{\frac{ln\frac{1}{P(h)} + ln\frac{2}{\delta}}{2n}}$$

Proof. We're gonna start by fixing a single hypothesis h and using the Chernoff Bound.

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

Where:

$$2e^{-2n\epsilon(h)^2} = \delta(h) = \delta\mathbb{P}(h)$$

For a fixed h, the following statement holds true with a probability of $\mathbb{P}(h)\delta$.

$$R[h] > \hat{R}_s[h] + \epsilon(h)$$

By union bound, we get that the probability it holds true for any h is

$$\sum_{h \in \mathcal{H}} P(h)\delta$$

$$\delta \sum_{h \in \mathcal{H}} P(h)$$

By the Definition 7, we get a probability of δ, which means that its opposite holds true with a probability of $1 - \delta$.

$$R[h] \leq \hat{R}_s[h] + \epsilon(h)$$

Substituting $\epsilon(h)$ by Definition 6 gives us Theorem 8.

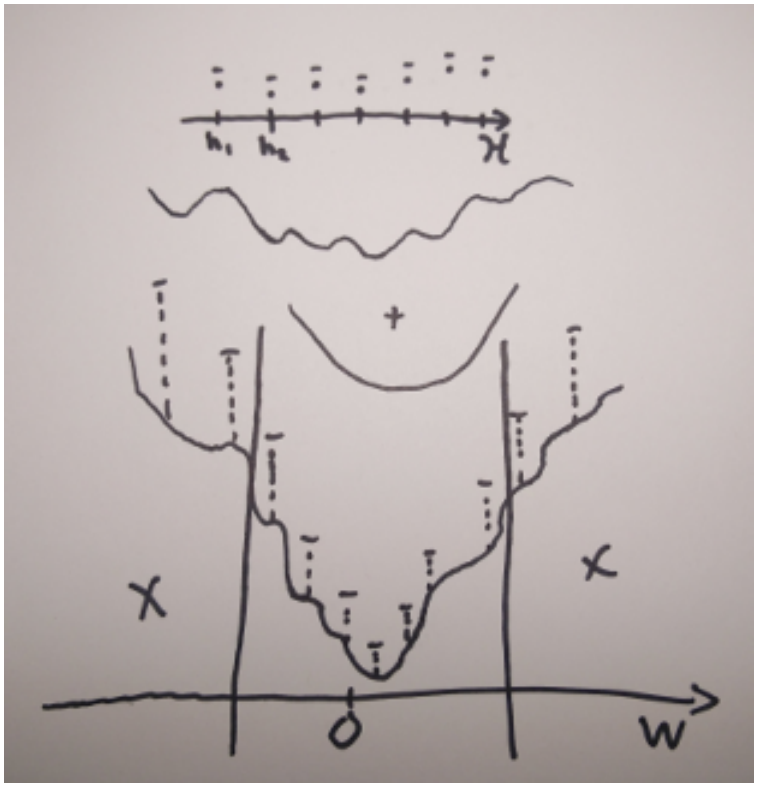

Figure 1: Visual representation of the bounds on all h's. The top part represents all different hypotheses with an uniform bound. The lower part represents the region of interest in which we want a stronger "bet".

4 PAC Bayes

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

Definition 9 (Kullback-Leibler Divergence)**.** *The Kullback-Leibler (KL) divergence between two distributions Q and P is defined as:*

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

The KL-divergence is a measure of how close two probability distributions are. Note that the KL-divergence is not necessarily symmetric - $D(Q||P) \neq D(P||Q)$ - and therefore it is not a metric. Also note that $P = Q \iff D(P||Q) = 0$.
In the following bound, 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.

Theorem 10 (PAC Bayes bound)**.** *Given a prior probability distribution P over a hypothesis class H and a posterior probability distribution Q over H. Then:*

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

with probability $\geq 1 - \delta$

References

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