IFT 6085 - Lecture 9
Stability and generalization - Part 1

This version of the notes has not yet been thoroughly checked. Please report any bugs to the scribes or instructor.

Scribe(s): Seyedarian Hosseini and João Monteiro **Instructor:** Ioannis Mitliagkas

1 Summary

In the previous lecture we discussed about the Occam's Razor Bound and PAC Bayes.

In this lecture we introduce the concept of a learning algorithm $\mathcal{A}$. We also present the first step in order to define bounds dependent on $\mathcal{A}$. We will first go over the PAC Learning framework and restate the definitions that are also important for stability and generalization.

2 Recap: PAC Learning

Before we go on, we need to make a couple of assumptions and definitions for our PAC learning framework, which we will also use when we discuss stability.

Definition 1 (Hypothesis)**.**

$$h \in \mathcal{H}$$

we assume that we get our hypothesis h from a hypothesis class $\mathcal{H}$.

It is important to note that this $h \in \mathcal{H}$ is a specific model, not a specific architecture.

Definition 2 (Training dataset)**.**

$$S = \{z_1, z_2, \cdots, z_n\}$$

such that each z_i is i.i.d. sampled from $\mathcal{D}$, where $\mathcal{D}$ is the true data distribution.

Definition 3 (Chosen hypothesis)**.**

$$h_S$$

This is the hypothesis obtained given a particular dataset S.

We also need to make a couple of definitions for risk minimization.

Definition 4 (Empirical risk)**.**

$$\hat{R}_S[h_S] = \frac{1}{n}\sum_{i=1}^{n} \ell[h_S(x_i), y_i] = \frac{1}{n}\sum_{i=1}^{n} \ell[h_S, z_i]$$

where ℓ is the loss function: $\ell : \mathcal{Y} \times \mathcal{Y} \to R_+$

Definition 5 (Population risk)**.**

$$R[h_S] = \mathbb{E}_{z \sim \mathcal{D}} \ell[h_S(x), y]$$

This is the true risk which we cannot compute since we do not have access to the distribution D.

The population risk measures how well our model performs on unseen data. Although we do not have the population risk, we can find a bound for it. We start by defining the generalization error (generalization gap):

$$\epsilon =\mid \hat{R}_S[h] - R[h] \mid$$

What we care about in this framework is the generalization error for our chosen hypothesis h_s, which we denote as $\epsilon = \epsilon(h_S)$. In the previous lectures we obtained the following bound for the generalization error.
We showed that for a fixed h if $n = O\left(\frac{\log(\frac{1}{\delta})}{\epsilon^2}\right)$, the difference between the population risk and the empirical risk will be smaller than ϵ with probability $\geq 1-\delta$:

$$R[h_S] \leq \hat{R}_S[h_S] + \epsilon$$

where ϵ is shown to be equal to$\sqrt{\frac{\log|\mathcal{H}|+\log\frac{2}{\delta}}{2n}}$, so

$$R[h_S] \leq \hat{R}_S[h_S] + \sqrt{\frac{\log\mid\mathcal{H}\mid+\log\frac{2}{\delta}}{2n}}$$

Note that in finding this bound we assumed that $\mathcal{H}$ is countable and finite.
Next step was to find a bound for all $h_i \in \mathcal{H}$. We used Occam's bound and union bound to ensure the former.
We defined the prior P:

Definition 6.

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

And proved:

Theorem 7 (Given prior P on set of hypothesis $\mathcal{H}$, with probability $\geq 1-\delta$ over training dataset S)**.**

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

For the case of uncountable $\mathcal{H}$, we defined the PAC-Bayes bound. For this, besides the prior P defined for Occam's bound, we introduced the "Posterior" Q over $\mathcal{H}$. We then defined another bound dependent on the KL-divergence between Q and P.

Theorem 8 (PAC-Bayes bound)**.** *With probability* $1-\delta$*:*

$$\forall h \in \mathcal{H}, \qquad \mathbb{E}_{h\sim Q}[R[h]] - \mathbb{E}_{h\sim Q}[\hat{R}_S[h]] \leq \sqrt{\frac{KL(Q||P) + \ln\frac{2}{\delta}}{2(n-1)}}$$

where $KL(Q||P) = \mathbb{E}_{h\sim Q}\ln\left[\frac{Q(h)}{P(h)}\right]$.

Note that the bound will be as tight as Q is close to P. Note also that P has to be chosen in advance, before data is seen and any hypothesis is evaluated.

Possible choices of Q:

Different choices of Q result in:

Choice a

- $Q = h_S$ with probability 1
  - $P(h_S) \to \infty$ and the bound explodes

Choice b

- $Q = P$
  - $KL(Q||P) = 0$
  - bound becomes $\mathbb{E}_{h\sim Q}[R[h]] - \mathbb{E}_{h\sim Q}[\hat{R}_S[h]] \leq \sqrt{\frac{\ln \frac{2}{\delta}}{2(n-1)}}$
  - bound is tighter due to $KL(Q||P) = 0$ but there is no way of bounding $R_S[h]$ with $\mathbb{E}_{h\sim Q}[R[h]]$

Choice c - common "works"

- $Q = \mathcal{N}(h_S, I)$
  - $KL(Q||P)$ is finite
  - We can find bounds for $R_S[h]$, for instance, if $R_S[h]$ is convex around h, Jensen's inequality will give $R_S[h] \leq \mathbb{E}_{h\sim Q}[R[h]]$

3 Stability

In past lectures, we defined generalization bounds regardless of the learning algorithm. Now we introduce the notion of stability along with its effect on the previous bounds seen in class.
In order to do so, we define a "new variable" $\mathcal{A}$ such that:

Definition 9 (Learning Algorithm)**.**

$$\mathcal{A} : \mathcal{Z}^n \to \mathcal{H}$$

where $\mathcal{Z}^n = \mathcal{X} \times \mathcal{Y}$.

Above definition implies that $\forall S \in \mathcal{Z}^n$:

$$h_S = \mathcal{A}(S). \tag{1}$$

Moreover, before proceeding to the definition of stability, we define the perturbed dataset $S^{i,z}$ and the defect $D[h_s]$ of hypothesis h_s by the following:

Definition 10 (Perturbed dataset)**.** *Consider* $S = \{z_1, ..., z_i, ..., z_n\}$, *thus* $\forall z \in \mathcal{Z}^n$ *and* $\forall i \in \{1, ..., n\}$*:*

$$S^{i,z} = \{z_1, ..., z_{i-1}, z, z_{i-1}, ..., z_n\}$$

Definition 11 (Defect)**.**

$$D[h_S] = \hat{R}[h_S] - R_s[h_S]$$

As per the notion of stability, we define:

Definition 12 (Stability)**.** $\mathcal{A}$ *is* β*-uniformly stable* $\forall (S, z) \in \mathcal{Z}^n$ *and* $\forall i \in \{1, ..., n\}$ *if:*

$$\sup_{z' \in \mathcal{Z}} |\ell[h_S, z'] - \ell[h_{S^{i,z}}, z']| \leq \beta$$

where $h_{S^{i,z}} = \mathcal{A}(S^{i,z})$.

Note that $h_{S^{i,z}}$ is the resulting h_S after a perturbation in S.
Now we state the McDiarmid's inequality without proof. It will come useful in future steps:

Theorem 13 (McDiarmid's inequality)**.** *Let $V_1, V_2, \ldots V_n$ be a set of independent random variables; and $v_1, v_2, \ldots v_n$ be samples drawn from these random variables. Let also F be a function that has the following property:*

$$\sup_{v_1, v_2, \ldots, v_n, v_i'} |F(v_1, v_2, \ldots, v_n) - F(v_1, v_2, \ldots, v_{i-1}, v_i', v_{i+1} \ldots v_n)| \leq c_i$$

then

$$P(|F(V_1, V_2, \ldots, V_n) - \mathbb{E}[F(V_1, V_2, \ldots, V_n)]|) > \epsilon) \leq 2 \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right)$$

4 Bounding the expectation of the defect

We need two steps to define $\mathcal{A}$ dependent bounds for $D[h_S]$:

1. Bounding $\mathbb{E}_s D[h_S]$;
2. Bounding $D[h_S]$ when $S \to S^{i,z}$.

We proceed to the first step and leave the remaining results for the next class.

Theorem 14 (Bound for the defect's expectation)**.** *if $\mathcal{A}$ is β-stable, thus:*

$$\mathbb{E}_S D[h_S] \leq \beta$$

Proof.

$$\begin{aligned} \mathbb{E}_S D[h_S] &= \mathbb{E}_S \left[\hat{R}[h_S] - R_S[h_S]\right], \\ &= \mathbb{E}_S \left[\frac{1}{n} \sum_i \ell[h_S, z_i] - \mathbb{E}_z \ell[h_S, z]\right], \end{aligned} \tag{2}$$

which can be rewritten as (see Fubini's theorem[1] [1] for details on changing order of integration of double integrals):

$$\begin{aligned} &= \mathbb{E}_{S,z} \left[\frac{1}{n} \sum_i \ell[h_S, z_i] - \ell[h_S, z]\right], \\ &= \mathbb{E}_{S,z} \left[\frac{1}{n} \sum_i (\ell[h_{S^{i,z}}, z] - \ell[h_S, z])\right], \end{aligned} \tag{3}$$

now we use the β-stability of $\mathcal{A}$ assumption, which upper-bounds $\ell[h_{S^{i,z}}, z] - \ell[h_S, z]$, and write:

$$\begin{aligned} &\leq \mathbb{E}_{S,z} \left[\frac{1}{n} \sum_i \beta\right], \\ &\leq \beta. \end{aligned} \tag{4}$$

□

References

[1] G. Fubini. Sugli integrali multipli. *Rom. Acc. L. Rend. (5)*, 16(1):608–614, 1907. ISSN 0001-4435.

[1] https://en.wikipedia.org/wiki/Fubini's_theorem