CS 560: Statistical Machine Learning

Instructor: Jie Shen Lecture 5

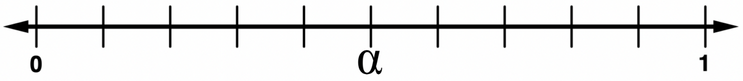
Scribe: Shay Dineen and Justin Depardieu October 1, 2021

Overview

In this lecture, we review and expand on PAC learning from the preceding week. We will also introduce the Hypothesis Class, the Representative Set, Radamacher Complexity, and the McDiarmid Inequality. Finally, we will prove three very important results that are derived from the previous concepts.

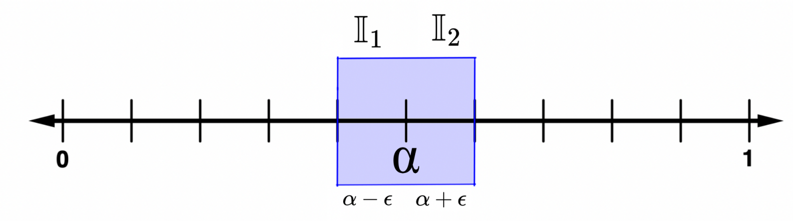
Review of PAC Learning

Let's first begin with an example to review PAC Learning. The below image shows the domain for the interval $[0, 1]$.



The Hypothesis Class (i.e. the prediction rules) are given by $H = \{x \geq \alpha \rightarrow y = +, x < \alpha \rightarrow y = -\}$. Even though we know the prediction rules, we still don't know the optimal α. Our goal is to to now estimate α by drawing samples from the domain. Let's assume that our underlying distribution D is uniform over the interval $[0, 1]$. It's trivial to see that sampling an infinite number of samples would give us optimal performance, however; we want to sample a discrete number of samples. The question now becomes how many samples do we need from $\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$ to approximate α ? For the sake of notation, let's denote our approximation to α as α'. We also want the difference between α' and α to be less than some error rate denoted ϵ. This is empirically defined as $\|\alpha' - \alpha\| \leq \epsilon$.

As you can see from the above diagram, if we get any point in $\mathbb{I}_1$ or $\mathbb{I}_2$ then we are done! Both $\mathbb{I}_1$ and $\mathbb{I}_2$ are less than ϵ. Let's now examine how we can get a point specifically from $\mathbb{I}_1$. Let's recall that D is uniform over the interval $[0, 1]$. So the size of $\mathbb{I}_1$ is ϵ. Let's let $P(x \in \mathbb{I}_1) = \epsilon$ be the probability that a randomly drawn point is from $\mathbb{I}_1$. And the



probability that an example is not from $\mathbb{I}_1$ is given by $P(x \notin \mathbb{I}_1) = 1 - \epsilon$. Then number of samples that we should draw (denoted k) is given by $O(\frac{1}{e})$. Moreover, if $k = O(\frac{1}{e}\log(\frac{1}{\delta}))$ then we can say with probability $1 - \delta$ that a randomly drawn sample is from $\mathbb{I}_1$. Since we are drawing random samples, we needed to define all of this in a probabilistic framework.

The above equations also hold for $\mathbb{I}_2$. If we draw another k samples then with probability $1 - \delta$ a random sample x_i will be in $\mathbb{I}_2$. If x_i is in either $\mathbb{I}_1$ or $\mathbb{I}_2$ then we can set $\alpha' = \alpha$. It's worth noting that this is a guaranteed ϵ approximation. A key takeaway from the above example is that when the sample size n is large enough, then with probability $1 - \delta$ we can do α' approximation to α. Moreover, any x from domain $[0, 1]$ has probability $\leq \epsilon$ that we miss classify the example.

So in the above example, there are two different types of probabilities. The first probability is related to the random draw of the training data given by $1-\delta$. The other probability is related to the draw of the unseen data (i.e. the future data). And this is the probability that we miss classify an example x. And as we talked about in last weeks lecture, X_{train} and X_{future} must come from the same distribution D or learning becomes impossible.

To review, PAC Learning stands for "probably approximately correct learning." And the name comes from the fact that it's impossible to exactly predict the optimal threshold α so we hope to use some approximation given by the error rate ϵ. But unfortunately, we can't always be approximately correct so we'd like to be probably approximately correct.

PAC Learning Assumptions

There are also three major empirical assumptions to the PAC Learning Model that we will now go over. These assumptions are related to the domain, the hypothesis class, and the loss function. We've already extensively covered domains in last week's lecture but we will again review the assumptions associated with it. Examples (x, y) come from some domain X and some label space Y. In the above example $X = [0, 1]$ and $Y = \{+, -\}$. And we also have an underlying distribution D over the domain X. The next assumption in the PAC Learning Model is in relation to the hypothesis class H. The hypothesis class defines our prediction rules and is known to the algorithm ahead of time. In the example above our hypothesis class was $H = \{x \geq \alpha \rightarrow y = +, x < \alpha \rightarrow y = -\}$. In the next section, we will go over several types of popular hypothesis classes.

The last assumption in PAC learning is related to the loss function L. L is dependent on D so we write L as L_D. In addition, L_D is also dependent on the current model $h \in H$ so our final notation for the loss is written as $L_D(h)$. And since our loss is over the distribution D, $L_D(h)$ is defined by expectation. $L_D(h)$ is known as the expected loss. In practice, we can't evaluate the expected loss because we don't know anything about D. We instead consider the loss over the training examples called the empirical loss (denoted $L_S(h)$). The bulk of today's lecture is dedicated to going over sufficient conditions such that a small $L_S(h)$ implies a small $L_D(h)$. We will cover these conditions in the coming sections.

Types of Hypothesis Classes

One of the most famous hypothesis spaces is called the "half-space." The half-space is simply defined as a linear weighted sum of some variables (i.e. $(w_1x_1 + \cdots + w_nx_n)$). Below is an example of the half space in $\mathbb{R}^2$. The line W_{True} indicates the ground truth decision boundary. As you can see, W_{True} perfectly separates the green circles from the red minuses.

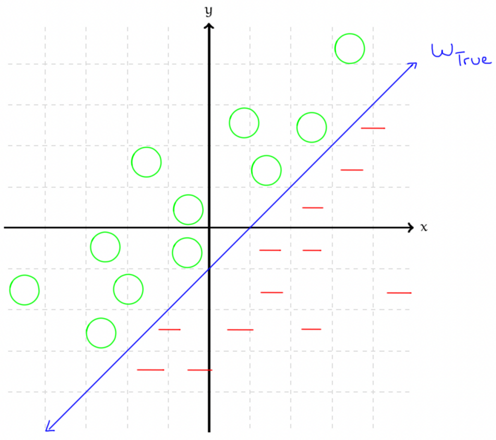


Figure 1: Half-Space Diagram

There are many other types of hypothesis classes like the polynomial class and the majority vote class. The majority vote hypothesis class with K classifiers can be empirically defined as:

$$y = \text{sign}\left[\sum_{k=1}^{K} \text{sign}(w_k \cdot x)\right]$$

If the output of three classifiers is $\{+, -, +\}$ then the final output will be $y = +$.

The Representative Set

To reiterate, our goal is to $\min_{h \in H} L_D(h)$. It is worth noting that minimizing $L_D(h)$ is the equivalent to minimizing the error. In practice, it is impossible to directly evaluate the expected loss $L_D(h)$ so we instead evaluate the empirical loss $L_S(h)$. The question remains- if $L_S(h)$ is small, does that imply that $L_D(h)$ is also small? In fact, the expected loss is bounded by the empirical loss plus some small constant:

$$L_D(h) \leq L_S(h) + O(1)$$

And if our sample size grows to infinity ($n \to \infty$) then $O(1)$ will go to 0. However, to assert the above we need several strong conditions. The first strong condition is called the Epsilon Representative Set. What we mean by the Epsilon Representative Set is that our training set S is representative of the ground truth distribution D. This is empirically defined as for $\forall h \in H$: $\|L_D(h) - L_S(h)\| \leq \epsilon$. If S has infinite samples from D then this condition will always hold.

Given S, we need to run gradient descent to find h such that $L_S(h) \approx \epsilon$. $\hat{h}$ is our global minimizer for the predicted function on S and h^* is our perfect, ground truth model. Given that S is representative:

$$L_D(\hat{h}) - L_D(h^*) \leq L_S(\hat{h}) + \epsilon - L_D(h^*)$$

And since $L_S(\hat{h})$ is the global minimizer on S then $L_S(\hat{h}) \leq L_S(h^*)$. We can now write:

$$L_D(\hat{h}) - L_D(h^*) \leq L_S(\hat{h}) + \epsilon - L_D(h^*) \leq L_S(h^*) + \epsilon - L_D(h^*)$$

Now we want to find the difference between $L_S(\hat{h}) - L_D(h^*)$. We know that $L_S(\hat{h}) - L_D(h^*) \leq 2\epsilon$ and we can rewrite this inequality as $L_S(\hat{h}) \leq L_D(h^*) + 2\epsilon$. And if ϵ is small, this implies that $L_S(\hat{h}) \approx L_D(h^*)$ (i.e. they have the same performance). The question now becomes, how do we build the representative set such that these conditions hold.

How Do We Build the Representative Set?

We are going to denote our representative set as $Rep(H, S)$ because the representative set is a function of hypothesis space H and the training data S.

$$Rep(H, S) = \forall h \in H : \; \|L_D(h) - L_S(h)\|$$

Our goal is to find a S such that $Rep(H, S)$ is small. However, it's not trivial to directly evaluate $Rep(H, S)$. It is easy to evaluate $L_S(h)$ but it is hard to evalaute $L_D(h)$. Since $L_D(h)$ is an expectation, we could potentially use Markov, Chebyshev, or Hoeffding's inequalites. However, in practice we break S into two subsets S_1 and S_2 such that $S_1 \cup S_2 = S$ and $S_1 \cap S_2 = \emptyset$. S_1 and S_2 both contain half of the training examples. By breaking S into two subsets, we can use the empirical loss on S_1 to evaluate the expected loss (i.e. $L_{S_1}(h) \approx L_D(h)$). Moreover, if we have sufficent data then the empirical loss on S_2 should be close to the loss on all of S (i.e. $L_{S_2}(h) \approx L_S(h)$). And if $L_{S_1}(h) \approx L_D(h)$ and $L_{S_2}(h) \approx L_S(h)$ then we can say:

$$\|L_D(h) - L_S(h)\| \approx \|L_{S_1}(h) - L_{S_2}(h)\|$$

As we stated above, each of the subsets of S contain half the training examples n. Therefore, $P(x_i \in S_1) = .5$ and $P(x_i \in S_2) = .5$. We will now write the above as:

$$\|L_D(h) - L_S(h)\| = \frac{2}{n}\sum_{i=1}^{n/2} f(h, x_i, y_i) - \frac{2}{n}\sum_{i=1}^{n/2} f(h, x_i, y_i)$$

And if we add the condition:

$$\sigma_i = \{i \in S_1 \rightarrow \sigma_i = 1, \; i \in S_2 \rightarrow \sigma_i = -1\}$$

Then we can write the above as:

$$\|L_D(h) - L_S(h)\| = \sup_h \frac{2}{n}\sum_{i=1}^{n} \sigma_i f(h, x_i, y_i)$$

And since σ_i is a random variable, we don't have to define a strategy to split the data if we take the expectation over σ_i. And if we take the expectation we get:

$$\mathbb{E}_\sigma[\sup_h \frac{2}{n}\sum_{i=1}^{n} \sigma_i f(h, x_i, y_i)]$$

The above equation is really good approximation for $Rep(H, S)$ and is called the Radamacher Complexity of H and S.

Radamacher Complexity

We denote Radamacher complexity of H on S by R(F, H, S), where F is the Loss Function, H is our Hypothesis Class and S is the training data . The great thing about Radamacher Complexity is that it can be evaluated because it does not depend on the distribution D. Another important consequence of Radamacher Complexity is that if we take its expectation over the random draw of training samples S we can write:

$$\mathbb{E}_S[Rep(H,S)] \leq 2 \cdot \mathbb{E}_S[R(F,H,S)]$$

Now let's establish another very important property: the generalization ability of any function given S. But first, we must understand the McDiarmid Inequality to understand this property.

The McDiarmid Inequality

Let's first begin by defining some notation. Let f be a function that maps our training data $x^n \to \mathbb{R}$. We also have a constant c such that $c > 0$ and $\forall i \leq n$. We also construct two set S and S' that only differ by one sample located at index i.

$$S = \{x_1, x_2, \cdots, x_i, \cdots, x_n\}$$

$$S' = \{x_1, x_2, \cdots, x_i', \cdots, x_n\}$$

And if $\|f(S) - f(S')\| \leq c$ then we can say with probability $1 - \delta$, the function evaluated on a random draw of training data concentrates around its expectation. This is written as:

$$\|f(x_1, x_2, \cdots, x_n)\| - \mathbb{E}[f(x_1, x_2, \cdots, x_n)]$$

The above equation is bounded by $c\sqrt[2]{\frac{n}{2}\log(\frac{2}{\delta})}$. We can then rewrite the above as:

$$\|f(x_1, x_2, \cdots, x_n)\| - \mathbb{E}[f(x_1, x_2, \cdots, x_n)] \leq c\sqrt[2]{\frac{n}{2}\log(\frac{2}{\delta})}$$

The above is important because if $\|f(S) - f(S')\| \leq c$ then we can say that the loss function is always centered around its mean.

Important Results

For the previous section, we can derive some important results that we will prove in the next section:

1. With probability $1 - \delta$, we have

$$\sup_{h \in H} L_D(h) - L_S(h) \leq 2 \cdot \mathbb{E}_{\mathbb{S}}[R(f, H, S)] + c \cdot \sqrt{\frac{2 \log \frac{1}{\delta}}{n}}$$

where S is randomly drawn from D: $\mathrm{E}_{S \sim D}$
However, although we know $R(f, H, S)$ because we know the loss function f, the Hypothesis Class H and the training data S, we cannot use this result in practice because we do not know $\mathbb{E}_{\mathbb{S}}[R(f, H, S)]$. The following result is a more practical one.

2. With probability $1 - \delta$, we have

$$\sup_{h \in H} L_D(h) - L_S(h) \leq 2 \cdot R(f, H, S) + 4 \cdot c \cdot \sqrt{\frac{2 \log \frac{1}{\delta}}{n}}$$

Here, we are not using the expectation of the Radamacher $\mathbb{E}_{\mathbb{S}}[R(f, H, S)]$ but the Radamacher itself, which we know, making this result a much better one in practice. Finally, from the result 2, we can derive the following result:

3. With probability $1 - \delta$, we have

$$\sup_{h \in H} L_D(\hat{h}) - L_D(h^*) \leq 2 \cdot R(f, H, S) + 4 \cdot c \cdot \sqrt{\frac{2 \log \frac{1}{\delta}}{n}}$$

Proving These Results

Let's now prove these results:

1. Remember that from the Representative set that:

$$\mathit{Rep}(H, S) = \sup_{h \in H}(L_D(h) - L_S(h)) \quad (1)$$

that is given by definition.

Now, let's consider the fact that $\forall f, g$ loss functions:

$$|\sup_{h\in H} f(h) - \sup_{h\in H} g(h)| \leq \sup_{h\in H} |f(h) - g(h)|$$

The following figure clearly illustrate that fact:

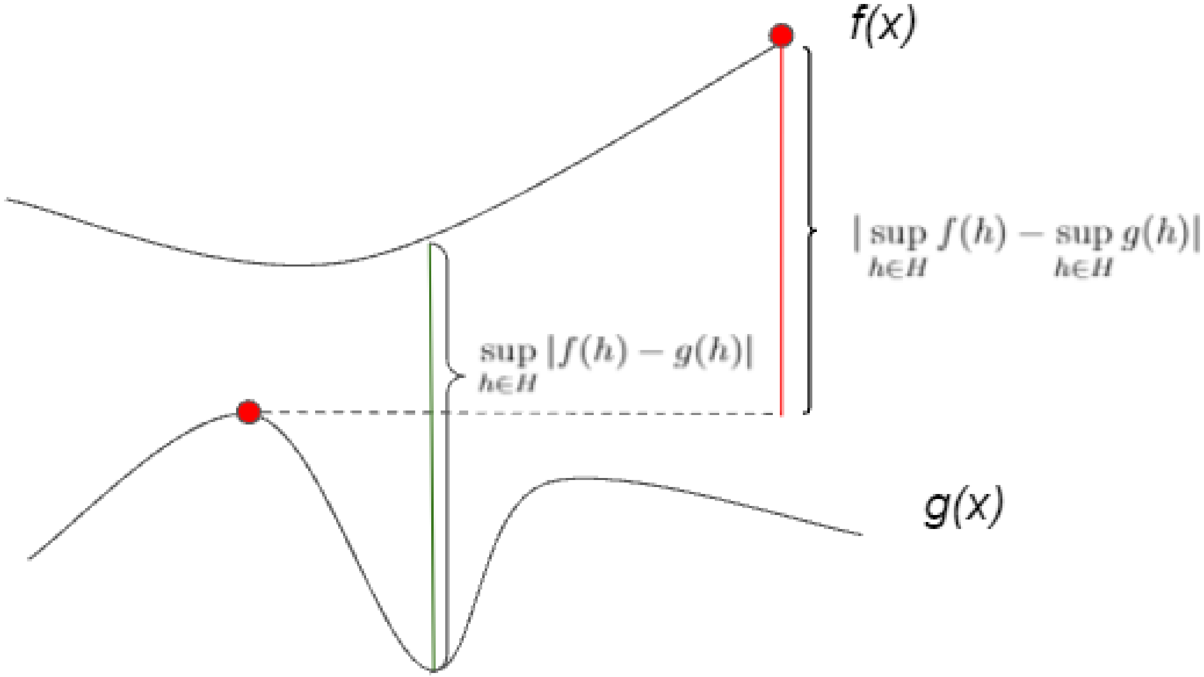

f(x)
$|\sup_{h\in H} f(h) - \sup_{h\in H} g(h)|$
$\sup_{h\in H} |f(h) - g(h)|$
g(x)

Figure 2: $|\sup f - \sup g| \leq \sup |f - g|$

Knowing this fact, let's have

$f = L_D(h) - L_S(h)$,

$g = L_D(h) - L_{S'}(h)$

where S' is different from S at one Sample.

$$|f - g| = \frac{1}{n}[l(z_i) - l(z_{i'})]$$

where l : loss function and $(x, y) \rightarrow z$

This is because looking at the expression of f and g, the leading terms $L_D(h)$ are the same, and the data set S and S' only differ at one point so subtracting $L_{S'}$ from L_S, it just remains the difference of the loss at that particular point. And then we take the average loss, that is why we have $\frac{1}{n}$.

Now, on the left-hand-side of (1) we have:

$$\sup_{h\in H} |f(h) - g(h)| \leq \frac{2}{n} \cdot c$$

where we assume that c is the bound of all the loss function : $|l(h, z)| \leq c$

Let's now look at the right-hand-side of (1), we now have :

$$|Rep(H, S) - Rep(H, S')| \leq \frac{2 \cdot c}{n} \ \ (2)$$

Now, if we apply McDiarmid Inequality, (2) implied that:

with probability $1 - \delta$:

$$|Rep(H, S) - \mathbb{E}_{\mathbb{S}}[Rep(H, S)]| \leq \frac{2 \cdot c}{n} \cdot \sqrt{\frac{n}{2} \cdot \log\frac{2}{\delta}} = c \cdot \sqrt{\frac{2}{n} \cdot \log\frac{2}{\delta}}$$

$$\Leftrightarrow |Rep(H, S)| \leq \mathbb{E}_{\mathbb{S}}[Rep(H, S)] + c \cdot \sqrt{\frac{2}{n} \cdot \log\frac{2}{\delta}} \leq 2 \cdot \mathbb{E}_{\mathbb{S}}[R(l, H, S)] + c \cdot \sqrt{\frac{2}{n} \cdot \log\frac{2}{\delta}}$$

(derived from Radamacher's section) Note that $c \cdot \sqrt{\frac{2}{n} \cdot \log\frac{2}{\delta}}$ is a deviation part which is very close to 0 for large enough n. That proves the first result.

2. Let's consider Radamacher, by definition:

$$R(l, H, S) = \frac{1}{n} \cdot \mathbb{E}_{\sigma}\left[\sup_{h\in H} \sum \sigma_i l(h, z_i)\right]$$

for a matter of consistency, we will keep l as our loss function from now on.

$$|R(l, H, S) - R(l, H, S')| \leq \frac{2 \cdot c}{n}$$

Now, by applying McDiarmid again, this time on $R(l, H, S)$, we obtain that with probability $1 - \delta$:

$$|R(l, H, S) - \mathbb{E}_{\mathbb{S}}[R(l, H, S)]| \leq \frac{2 \cdot c}{n} \cdot \sqrt{\frac{n}{2} \cdot \log\frac{2}{\delta}} = c \cdot \sqrt{\frac{2}{n} \cdot \log\frac{2}{\delta}}$$

This time we have

$$\Leftrightarrow \mathbb{E}_{\mathbb{S}}[R(l,H,S)] \leq R(l,H,S) + c \cdot \sqrt{\frac{2}{n} \cdot \log \frac{2}{\delta}} \ \ (3)$$

Now, back to the result from our first proof, we had

$$|Rep(H,S)| \leq 2 \cdot \mathbb{E}_{\mathbb{S}}[R(l,H,S)] + c \cdot \sqrt{\frac{2}{n} \cdot \log \frac{2}{\delta}}$$

Finally, substituting $\mathbb{E}_{\mathbb{S}}[R(l,H,S)]$ with $R(l,H,S) + c \cdot \sqrt{\frac{2}{n} \cdot \log \frac{2}{\delta}}$ from (3), we obtain

$$|Rep(H,S)| \leq 2 \cdot R(l,H,S) + 4c \cdot \sqrt{\frac{2}{n} \cdot \log \frac{2}{\delta}}$$

3. For this third and final proof, we have $L_D(\hat{h}) - L_D(h^*)$, and this is equal to

$$[L_D(\hat{h}) - L_S(\hat{h})] + L_S(\hat{h}) + [-L_D(h^*) + L_S(h^*)] - L_S(h^*) \ \ (3)$$

Notice that we just added and subtracted $L_S(\hat{h})$ from the first term, and we added and subtracted $L_S(h^*)$ from the second term. Take the two terms outside the brackets, we have $L_S(\hat{h}) - L_S(h^*) \leq 0$, and recognize that $L_S(\hat{h}) \leq L_S(h^*)$ because $\hat{h}$ is the minimized $L_S(h)$, and therefore $L_S(\hat{h}) - L_S(h^*) \leq 0$, so we can see that (3) is bounded by

$$\leq [L_D(\hat{h}) - L_S(\hat{h})] + [L_S(h^*) - L_D(h^*)]$$

Now, by using the second proof we have

$$L_D(\hat{h}) - L_S(\hat{h}) \leq 2 \cdot R(l,H,S) + 4 \cdot c \cdot \sqrt{\frac{2 \log \frac{1}{\delta}}{n}}$$

and

$$L_S(h^*) - L_D(h^*) \leq \epsilon$$

so combining the two we have

$$L_D(\hat{h}) - L_D(h^*) \leq [L_D(\hat{h}) - L_S(\hat{h})] + [L_S(h^*) - L_D(h^*)] \leq 2 \cdot R(l,H,S) + 4 \cdot c \cdot \sqrt{\frac{2 \log \frac{1}{\delta}}{n}} + \epsilon$$

or

$$L_D(\hat{h}) - L_D(h^*) \leq 2 \cdot R(l,H,S) + 4 \cdot c \cdot \sqrt{\frac{2 \log \frac{1}{\delta}}{n}}$$