CS 536 : Statistical Learning Theory - PAC Learning

16:198:536

What does it mean to learn from data? How can we quantify and study the act of learning itself? If we are to understand what is possible and impossible to learn, we need to formalize the notion.

In the usual way, we assume that data is coming to us in an i.i.d. fashion, feature vectors $\underline{X}$ drawn from some underlying distribution D, and for each vector $\underline{X}$, there is some corresponding value Y that we would like to predict. Typically, we have a class of hypotheses or models H, and we want to choose $f \in H$ such that $f(\underline{X})$ matches Y as close and as frequently as possible. In the case of classification problems, when $Y = 0$ or $Y = 1$, we can define the 'true' error of a hypothesis as the probability that it mis-identifies a new sample:

$$\text{err}(f) = \mathbb{P}_D \left(f(\underline{X}) \neq Y \right). \tag{1}$$

If we could compute the true error of a hypothesis, it would be easy enough to select the $f \in H$ with minimal error. However, we frequently can't, for lack of information about D. If we do not know the underlying distribution of the data, we may need to *approximate* it from a set of data sampled from it. Let S_m be a set of i.i.d. samples $\{(\underline{X}_i, Y_i)\}_{i=1,\ldots,m}$ sampled from the underlying distribution D. Given the training data set S_m, we can compute the sample or training error of a given hypothesis as

$$\text{err}_{S_m}(f) = \sum_{i=1}^{m} \mathbb{I}\left\{ f(\underline{X}^i) \neq Y^i \right\}, \tag{2}$$

i.e., the average number of mistakes made over the data set.

There are a number of questions we can ask at this point, but the central one is this:

Given a data set S_m, how can we select an $f \in H$ with small true error?

At this point, we can ask a couple of questions:

- Given a data set S_m, how can we select an $f \in H$ with small error?
- Given an algorithm for selecting $f \in H$, how can we analyze how good this algorithm is?

For most of the models we've seen so far, the core idea for selecting a hypothesis to fit to our training data is **Empirical Risk Minimization**: choose $f \in H$ such that f minimizes $\text{err}_{S_m}(f)$.

Given a data set S_m, the **Empirical Risk Minimizer** is any $f_{\text{ERM}} \in H$ such that

$$f_{\text{ERM}} \in \text{argmin}_{f \in H} \, \text{err}_{S_m}(f). \tag{3}$$

This effectively maximizes the performance of the hypothesis on the training data, in the hopes that the training data is 'representative' and thus the hypothesis generalizes well.

To study the effectiveness of this algorithm, we introduce the notion of Probably Almost Correct (PAC) Learning (Valiant 1984):

An algorithm A **PAC-learns** a hypothesis class H if for any $\epsilon > 0, \delta > 0$, given enough data A will find a hypothesis with error at most ϵ, with probability at least $1 - \delta$.

Typical questions studied in PAC learning are:

- For a given ϵ, δ, how much data is needed to PAC learn a given hypothesis class?
- For a given δ and number of samples, how accurate can we assume a hypothesis is, with probability at least $1 - \delta$?
- How does the hypothesis space effect learnability?

We can answer these questions roughly, in the following way:

- The more accurate the hypothesis we want, the more data we need to achieve high confidence.
- Accuracy should fall off as $\delta \to 0$, but should improve as the number of samples increases $m \to \infty$.
- The more hypotheses there are to differentiate between, the more data will be needed.

But within the framework of PAC learning, we can provide more accurate and informative answers to these questions. For the moment, we restrict H to be **finite**, though generalizations will be discussed.

$f_{\mathbf{ERM}}$ when $Y = f^*(\underline{X})$ and $f^* \in H$

The importance of this assumption is that if there is a 'correct' hypothesis in the hypothesis space, then that hypothesis will always have zero training error on every set - this implies that the training error of f_{ERM} will always be zero, even if f_{ERM} fails to recover the correct f^*. We generalize to the case where f_{ERM} might have non-zero training error in the next section.

We would like to try to analyze $\text{err}(f_{\text{ERM}})$. Note that because f_{ERM} depends on the training data set, which is random, f_{ERM} is a random variable as well. So we can't control it perfectly - but in the spirit of PAC Learning, what we are really interested in is

$$\mathbb{P}\left(\text{err}(f_{\text{ERM}}) > \epsilon\right) < \delta. \tag{4}$$

It's convenient to define a set of 'bad' hypotheses: $\text{BAD} = \{f \in H : \text{err}(f) > \epsilon\}$. Note, since H is finite, BAD must also be finite. We can then bound the above probability in the following way:

$$\mathbb{P}\left(\text{err}(f_{\text{ERM}}) > \epsilon\right) \leq \mathbb{P}\left(\text{err}_{S_m}(f) = 0 \text{ for some } f \in \text{BAD}\right). \tag{5}$$

The intuition here is that for f_{ERM} to be in the bad set, there must be *some* function in the bad set with zero training error. Since BAD is a finite set, we can take a union bound over all the elements of BAD:

$$\begin{aligned} \mathbb{P}\left(\text{err}(f_{\text{ERM}}) > \epsilon\right) &\leq \mathbb{P}\left(\text{err}_{S_m}(f) = 0 \text{ for some } f \in \text{BAD}\right) \\ &\leq \sum_{f \in \text{BAD}} \mathbb{P}\left(\text{err}_{S_m}(f) = 0\right) \\ &\leq \sum_{f \in \text{BAD}} \mathbb{P}\left(f(\underline{X}) = Y\right)^m. \end{aligned} \tag{6}$$

The last step above is the observation that in order for f to have 0 training error, it must have made the correct prediction m times in a row on independent samples. Noting that $\mathbb{P}\left(f(\underline{X}) = Y\right) = 1 - \text{err}(f)$, the above gives us

$$\mathbb{P}\left(\text{err}(f_{\text{ERM}}) > \epsilon\right) \leq \sum_{f \in \text{BAD}} \left[1 - \text{err}(f)\right]^m \leq \sum_{f \in \text{BAD}} \left[1 - \epsilon\right]^m. \tag{7}$$

The last step above comes from the definition of the BAD set, as those with $\text{err}(f) > \epsilon$. Observing first the bound that $(1-\epsilon) \leq e^{-\epsilon}$, and the fact that $|\text{BAD}| \leq |H|$, we get the final bound

$$\mathbb{P}\left(\text{err}(f_{\text{ERM}}) > \epsilon\right) \leq |H| e^{-\epsilon m}. \tag{8}$$

If we want to ensure that the probability f_{ERM} is bad is sufficiently small, we want $|H|e^{-\epsilon m} \leq \delta$, or, rearranging, we get our first sample complexity result:

If there exist hypotheses that will *always* have zero training error, then if

$$m \geq \frac{1}{\epsilon} \ln\left(\frac{|H|}{\delta}\right), \tag{9}$$

we have that $\mathbb{P}\left(\text{err}(f_{\text{ERM}}) \leq \epsilon\right) > 1 - \delta$.

This result agrees with our intuition - the more accurate we want a model to be, the more data we need; the higher the confidence we want in our model, the more data we need.

f_{ERM} when perfect hypotheses might not exist

If no perfect hypotheses exist, then every hypothesis in H has some error, and we have no guarantee at all that we'll be able to achieve zero training error. In this case, $\text{err}(f_{\text{ERM}}) > \epsilon$ might not be bad at all, if the smallest possible true error is greater than ϵ.

Hence in this case, we are more concerned with whether a given hypothesis generalizes well - that is, is the training error reflective of the true over all error? Under what conditions could we be certain that

$$\mathbb{P}\left(\text{err}(f_{\text{ERM}}) > \text{err}_{S_m}(f_{\text{ERM}}) + \epsilon\right) < \delta, \tag{10}$$

i.e., the true error of the empirical risk minimizer is not too far away from the training error of the empirical risk minimizer.

We can begin with a similar bound as to the previous section, though the notion of 'bad' hypotheses becomes more implicit, since the badness of a hypothesis depends on the random training data set.

$$\begin{aligned} &\mathbb{P}\left(\text{err}(f_{\text{ERM}}) > \text{err}_{S_m}(f_{\text{ERM}}) + \epsilon\right) \leq \mathbb{P}\left(\text{err}(f) > \text{err}_{S_m}(f) + \epsilon \text{ for some } f \in H\right) \\ &\leq \sum_{f \in H} \mathbb{P}\left(\text{err}(f) > \text{err}_{S_m}(f) + \epsilon\right) \end{aligned} \tag{11}$$

Here we observe that $\text{err}_{S_m}(f)$ is the empirical estimate for the probability of an error, while $\text{err}(f)$ is the probability being estimated. Previous concentration inequalities for boolean event probabilities (the coin flip example we began with) tell us that $\mathbb{P}\left(p > \hat{p} + \epsilon\right) < e^{-2\epsilon^2 m}$ - and this is regardless of the true value of p. Applying this here, we get

$$\mathbb{P}\left(\text{err}(f_{\text{ERM}}) > \text{err}_{S_m}(f_{\text{ERM}}) + \epsilon\right) \leq |H| e^{-2\epsilon^2 m}. \tag{12}$$

In order to achieve a confidence level of δ, we therefore would want $|H|e^{-2\epsilon^2 m} < \delta$. This gives us our second sample complexity result:

Given a finite hypothesis space H, if

$$m \geq \frac{1}{2\epsilon^2} \ln\left(\frac{|H|}{\delta}\right), \tag{13}$$

we have that

$$\mathbb{P}\left(\text{err}(f_{\text{ERM}}) \leq \text{err}_{S_m}(f_{\text{ERM}}) + \epsilon\right) > 1 - \delta. \tag{14}$$

An Important Observation Here: Note that this bound is explicitly worse than the bound of the previous section. There we had a dependence like $O(1/\epsilon)$, here we have $O(1/\epsilon^2)$ - more data is needed. What gives? The problem is that if there are no perfect hypotheses, if a hypothesis f makes a mistake on the training data you must ask - is this mistake because f doesn't really describe the data, or is this mistake because there are no perfect hypotheses and f is doing the best it can? Is the mistake due to f being bad, or due to H being too small? Being able to differentiate between 'good mistakes' and 'bad mistakes' in this way requires more data.

We can also flip this around in the following way: for a data set of size m and confidence level δ, what can we say about how the training error relates to the true error? Note that if $m \geq 1/(2\epsilon^2)\ln(|H|/\delta)$, we have that $\epsilon \geq \sqrt{\ln(|H|/\delta)/(2m)}$. This gives us our third PAC-Learning result, and an answer to the second initial question:

If H is finite, then given m, δ, we have that with probability at least $1 - \delta$:

$$\text{err}(f_{\text{ERM}}) \leq \text{err}_{S_m}(f_{\text{ERM}}) + \sqrt{\frac{\ln|H| + \ln\frac{1}{\delta}}{2m}}. \tag{15}$$