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.

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. Additionally, it is convenient to consider two cases: let's define the 'best' hypothesis to be the one that minimizes the true error,

$$\text{err}(f_{\text{best}}) = \min_{f \in H} \text{err}(f). \tag{4}$$

With this, consider the following cases:

- $\text{err}(f_{\text{best}}) = 0$: In this case, there is a classifier that is 'perfect' and makes no errors. An important aspect of this is that the *training* error of f_{best} will also be zero - so you know immediately that any function that makes errors on the training data cannot be f_{best}.
- $\text{err}(f_{\text{best}}) > 0$: In this case, no classifier is perfect, which means that even the best classifier may make mistakes on the training error. Just because a hypothesis makes mistakes on the training data then, we cannot dismiss it as not the 'best'.

The main conclusions in the sections to follow is that as long as there is enough data, f_{ERM} approximates f_{best} in the sense that, with high probability,

$$\text{err}(f_{\text{best}}) \approx \text{err}(f_{\text{ERM}}). \tag{5}$$

This will not only illustrate a number of PAC Learning concepts, but also justify the idea that minimizing the training error is a good method for generating good models.

Learning when $\text{err}(f_{\text{best}}) = 0$

The importance of this assumption is that if there is a 'perfect' classifier, then no matter what there is always at least one function with zero training error: for data set S_m, we therefore know that

$$\text{err}_{S_m}(f_{\text{ERM}}) = 0. \tag{6}$$

But how can we be confident that we have selected a hypothesis that generalizes well - has small true error, and is not just overfit to the data? Since f_{ERM} is random and depends on the data, this could potentially happen, but we can ask what is the probability of this happening: can we ensure that

$$\mathbb{P}\left(\text{err}(f_{\text{ERM}}) > \epsilon\right) < \delta. \tag{7}$$

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:

$$\mathbb{P}\left(\text{err}(f_{\text{ERM}}) > \epsilon\right) \leq \mathbb{P}\left(\text{err}_{S_m}(f) = 0 \text{ for some } f \in \text{BAD}\right). \tag{8}$$

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:

$$\begin{aligned}\mathbb{P}\left(\text{err}(f_{\text{ERM}}) > \epsilon\right) &\leq \mathbb{P}\left(\text{err}_{S_m}(f) = 0 \text{ for some } f \in \text{BAD}\right) \\ &\leq \sum_{f \in \text{BAD}} \mathbb{P}\left(\text{err}_{S_m}(f) = 0\right) \\ &\leq \sum_{f \in \text{BAD}} \mathbb{P}\left(f(\underline{X}) = Y\right)^m.\end{aligned} \tag{9}$$

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

$$\mathbb{P}\left(\text{err}(f_{\text{ERM}}) > \epsilon\right) \leq \sum_{f \in \text{BAD}} \left[1 - \text{err}(f)\right]^m \leq \sum_{f \in \text{BAD}} \left[1 - \epsilon\right]^m. \tag{10}$$

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

$$\mathbb{P}\left(\text{err}(f_{\text{ERM}}) > \epsilon\right) \leq |H| e^{-\epsilon m}. \tag{11}$$

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 $\text{err}(f_{\text{best}}) = 0$, then if

$$m \geq \frac{1}{\epsilon} \ln\left(\frac{|H|}{\delta}\right), \tag{12}$$

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

$$\mathbb{P}\left(\text{err}(f_{\text{ERM}}) - \text{err}(f_{\text{best}}) \geq \epsilon\right) < \delta. \tag{13}$$

Hence, if we have enough data, then the empirical risk minimizer has a true error close to the true error minimum.

Learning when $\text{err}(f_{\text{best}}) > 0$

The issue that we run into here in terms of learning is that if there is no perfect hypothesis that makes no mistakes, any mistake on the training data could either be do to the model being 'bad' and having large true error, or simply because any model will make some mistakes on the data. However, even if the training error is bad, with any luck the training error will be representative of the true error - thus minimizing the training error will approximate minimizing the true error.

To complete the result in this case, I want to start with a broader property, that of uniform convergence:

If there is enough data, then we have

$$\mathbb{P}\left(\forall f \in H : |\text{err}(f) - \text{err}_{S_m}(f)| \leq \epsilon\right) > 1 - \delta, \tag{14}$$

i.e., if we have enough data, no function in f over fits (with high probability). The training error is indicative of the true error for all functions (up to ϵ).

This is a very powerful property, but not necessarily surprising. If we have enough data, then the performance of the function on that data should (with high probability) be indicative of the performance on the whole - for any function f! To see this, consider bounding the opposite probability:

$$\mathbb{P}\left(\exists f \in H : |\text{err}(f) - \text{err}_{S_m}(f)| > \epsilon\right) \leq \sum_{f \in H} \mathbb{P}\left(|\text{err}(f) - \text{err}_{S_m}(f)| > \epsilon\right). \tag{15}$$

For any of these individual probabilities, observe that $\text{err}(f)$ is a constant, representing the probability on some random instance of misclassification. $\text{err}_{S_m}(f)$ represents the estimate of that probability based on m i.i.d. samples. We can draw on an early result here, the Hoeffding inequality, which tells us that for a probability p and an estimate $\hat{p}$ based on m samples each done with probability p, we have $\mathbb{P}\left(|p - \hat{p}| > \epsilon\right) \leq 2e^{-2\epsilon^2 m}$. Applying that here,

$$\begin{aligned} \mathbb{P}\left(\exists f \in H : |\text{err}(f) - \text{err}_{S_m}(f)| > \epsilon\right) &\leq \sum_{f \in H} 2e^{-2\epsilon^2 m} \\ &= 2|H|e^{-2\epsilon^2 m}. \end{aligned} \tag{16}$$

For any given threshold δ, we can choose m large enough to make the above probability less than δ. Thus we can formalize the uniform convergence result above:

Uniform Convergence: If H is finite, and we have

$$m \geq \frac{1}{2\epsilon^2} \log\left(\frac{2|H|}{\delta}\right), \tag{17}$$

then with high probability we have uniform convergence for all f in H:

$$\mathbb{P}\left(\forall f \in H : |\text{err}(f) - \text{err}_{S_m}(f)| \leq \epsilon\right) < \delta. \tag{18}$$

Why is this useful? If we have good approximation for any f, we have good approximation for f_{ERM} as well as f_{best}:

$$\begin{aligned} |\text{err}(f_{\text{best}}) - \text{err}_{S_m}(f_{\text{best}})| &\leq \epsilon \\ |\text{err}(f_{\text{ERM}}) - \text{err}_{S_m}(f_{\text{ERM}})| &\leq \epsilon \end{aligned} \tag{19}$$

We can link these two with the fact that since f_{ERM} is the empirical risk minimizer, we have

$$\text{err}_{S_m}(f_{\text{ERM}}) \leq \text{err}_{S_m}(f_{\text{best}}). \tag{20}$$

Linking the three results above, we have (under the uniform convergence event),

$$\text{err}(f_{\text{ERM}}) - \epsilon \leq \text{err}_{S_m}(f_{\text{ERM}}) \leq \text{err}_{S_m}(f_{\text{best}}) \leq \text{err}(f_{\text{best}}) + \epsilon, \tag{21}$$

or

$$\text{err}(f_{\text{ERM}}) \leq \text{err}(f_{\text{best}}) + 2\epsilon. \tag{22}$$

Hence,

$$\mathbb{P}\left(\forall f \in H : |\text{err}(f) - \text{err}_{S_m}(f)| \leq \epsilon\right) \leq \mathbb{P}\left(\text{err}(f_{\text{ERM}}) \leq \text{err}(f_{\text{best}}) + 2\epsilon\right). \tag{23}$$

This combines the the probability bound to give us a PAC learning result on the ERM (rescaling the ϵ as appropriate):

PAC Learning for ERM: If H is finite, and we have

$$m \geq \frac{2}{\epsilon^2} \log\left(\frac{2|H|}{\delta}\right), \tag{24}$$

then with high probability we have uniform convergence for all f in H:

$$\mathbb{P}\left(\text{err}(f_{\text{ERM}}) - \text{err}(f_{\text{best}}) > \epsilon\right) < \delta. \tag{25}$$

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 $m \geq 2/(\epsilon^2)\ln(|H|/\delta)$, we have that $\epsilon \geq \sqrt{2\ln(|H|/\delta)/m}$. This gives us our fourth 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 $1-\delta$:

$$\text{err}(f_{\text{best}}) \leq \text{err}(f_{\text{ERM}}) \leq \text{err}_{S_m}(f_{\text{ERM}}) + \sqrt{\frac{2\ln|H| + 2\ln\frac{1}{\delta}}{m}}. \tag{26}$$

Note additionally that uniform convergence allows us to make a statement about *any* function (with high probability) if we have enough data. For instance, suppose that instead of computing the true empirical risk minimizer, we were to instead approximate it to some degree of accuracy. What uniform convergence would tell us is that if

$$\text{err}_{S_m}(f_{\text{approx-ERM}}) \leq \text{err}_{S_m}(f_{\text{ERM}}) + \epsilon, \tag{27}$$

then we would have

$$\text{err}(f_{\text{approx-ERM}}) - \epsilon \leq \text{err}_{S_m}(f_{\text{approx-ERM}}) \leq \text{err}_{S_m}(f_{\text{ERM}}) + \epsilon \leq \text{err}_{S_m}(f_{\text{best}}) + \epsilon \leq \text{err}(f_{\text{best}}) + 2\epsilon, \tag{28}$$

or

$$\text{err}(f_{\text{approx-ERM}}) \leq \text{err}(f_{\text{best}}) + 3\epsilon. \tag{29}$$

So approximating the empirical risk minimizer, in the situation of uniform convergence, is enough to get your true error within spitting distance of the best error.