Algorithms and Uncertainty, Winter 2018/19 Lecture 21 (4 pages)

PAC Learning

Instructor: Thomas Kesselheim

Throughout this course, we have often seen optimization problems, in which we assumed to know a probability distribution. The degree of knowledge that is necessary certainly depends on what we want to use it for. Today, we will get to know a concept that allows us to describe how much we actually have to know. It originates in the theory of machine learning and therefore we will describe it in this context, but similar ideas spread well beyond this.

1 Learning Model: Realizable Setting

Our task will be to classify data points from some set X. For example, $X \subseteq \mathbb{R}$. The labels will be binary, i.e., -1 or 1. (The labels could also be 0 and 1 but -1 and 1 is often more convenient.) For example, the set X could be the set of all e-mails and the labels mean "not spam" or "spam". Eventually, for every data point x we are presented, we have to predict the correct label $y \in \{-1, 1\}$.

There is a class of hypotheses $\mathcal{H}$, each of the form $h\colon X \to \{-1, 1\}$. To keep things simple, we will first assume to be in the *realizable case*. This means that there is a ground truth $f \in \mathcal{H}$, which is one of our possible hypotheses, and the correct label to $x \in X$ is always $f(x)$. We want to find a function $h \in \mathcal{H}$ that is as close as possible to the correct f using only a bounded number of correctly labeled samples.

In more detail, there is a distribution $\mathcal{D}$ over $X \times \{-1, 1\}$ that we would like to understand. As we are in the realizable case, whenever $\mathcal{D}$ returns (x, y) then $y = f(x)$. We have access to m samples $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$. The (x_i, y_i) are drawn independently and identically distributed from distribution $\mathcal{D}$.

Definition 21.1. *The* training error *(or empirical risk)* $\mathrm{err}_S(h)$ *of a hypothesis h with respect to a sample set S is*

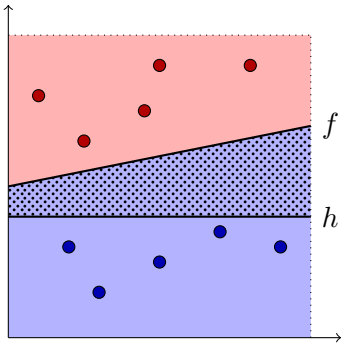
$$\mathrm{err}_S(h) := \frac{1}{m}|\{h(x_i) \neq y_i\}| \ .$$

Given any set S, the hypotheses $h \in \mathcal{H}$ that minimize $\mathrm{err}_S(h)$ are called *empirical risk minimizers*. Note that for the ground truth f clearly $\mathrm{err}_S(f) = 0$ for all S. So, for all empirical risk minimizers h we have $\mathrm{err}_S(h) = 0$. The difficulty is that there are usually multiple hypotheses h with no training error. Given only the samples, we have no idea which one is correct, meaning that it correctly predicts the labels for the rest of X. This is called the *true error*.

Definition 21.2. *The* true error *(or true risk)* $\mathrm{err}_{\mathcal{D}}(h)$ *of a hypothesis h with respect to a distribution $\mathcal{D}$ is*

$$\mathrm{err}_{\mathcal{D}}(h) := \mathbf{Pr}_{(x,y)\sim\mathcal{D}}\left[h(x) \neq y\right] \ .$$

Example 21.3. *Let $X = [0, 1]^2$ and $\mathcal{H}$ be the set of linear classifiers, that is, defined by a straight line. In this example, h correctly classifies all of S (indicated by the points) but incorrectly classifies the dotted area. If $\mathcal{D}$ is the uniform distribution, then* $\mathrm{err}_{\mathcal{D}}(h)$ *corresponds to the area of the dotted area.*



Naturally, $\mathrm{err}_{\mathcal{D}}(f) = 0$. But what about other empirical risk minimizers? How large does m have to be so that $\mathrm{err}_{\mathcal{D}}(h)$ becomes small for any empirical risk minimizer? This, of course, depends on the class of hypotheses $\mathcal{H}$ and the distribution $\mathcal{D}$. We will give bounds that hold for any distribution $\mathcal{D}$ and characterize the *sampling complexity* of $\mathcal{H}$.

As a start, let us consider the most basic case, namely that $\mathcal{H}$ is finite but arbitrary.

Theorem 21.4. *If* $m \geq \frac{1}{\epsilon} \ln\left(\frac{|\mathcal{H}|}{\delta}\right)$*, then with probability at least* $1-\delta$*, all* $h \in \mathcal{H}$ *with* $\mathrm{err}_S(h) = 0$ *have* $\mathrm{err}_{\mathcal{D}}(h) < \epsilon$*.*

Proof. Let us consider some $h \in \mathcal{H}$ with $\mathrm{err}_{\mathcal{D}}(h) \geq \epsilon$, so the true error of h is at least ϵ. Then

$$\begin{aligned}\mathbf{Pr}\left[\mathrm{err}_S(h) = 0\right] &= \mathbf{Pr}\left[h(x_1) = y_1, \ldots, h(x_m) = y_m\right] \\ &= \mathbf{Pr}\left[h(x_1) = y_1\right] \cdot \ldots \cdot \mathbf{Pr}\left[h(x_m) = y_m\right] \leq (1-\epsilon)^m \leq \mathrm{e}^{-\epsilon m} \quad .\end{aligned}$$

This means that it is a empirical risk minimizer with probability at most $\mathrm{e}^{-\epsilon m}$.

To get a bound that there exists such an h, we use the union bound.

Lemma 21.5 (Union Bound)**.** *Let* $\mathcal{E}_1, \ldots, \mathcal{E}_n$ *be (not necessarily disjoint) events. Then*

$$\mathbf{Pr}\left[\bigcup_{i=1}^{n} \mathcal{E}_i\right] \leq \sum_{i=1}^{n} \mathbf{Pr}\left[\mathcal{E}_i\right] \quad .$$

We apply the bound as follows. The event that exists an $h \in \mathcal{H}$ with $\mathrm{err}_{\mathcal{D}}(h) \geq \epsilon$ and $\mathrm{err}_S(h) = 0$ can be considered as the union of all events of the form $\mathrm{err}_S(h) = 0$ for a fixed h with $\mathrm{err}_{\mathcal{D}}(h) \geq \epsilon$, which we have bounded above. So

$$\mathbf{Pr}\left[\exists h \in \mathcal{H} : \mathrm{err}_{\mathcal{D}}(h) \geq \epsilon \text{ and } \mathrm{err}_S(h) = 0\right] \leq |\mathcal{H}|\mathrm{e}^{-\epsilon m} \leq \delta \quad . \qquad \square$$

This result shows that finite hypothesis classes are PAC-learnable, which requires that a similar statement as in the previous theorem is true.

Definition 21.6. *Hypothesis class* $\mathcal{H}$ *is PAC-learnable (in the realizable sense) if there is a function* $m_{\mathcal{H}}$ *and a learning algorithm*[1] *such that for any* $\epsilon, \delta > 0$ *and any distribution* $\mathcal{D}$ *admitting* $\mathrm{err}_{\mathcal{D}}(f) = 0$ *for some* $f \in \mathcal{H}$*, given a random sample* S *of size at least* $m_{\mathcal{H}}(\epsilon, \delta)$ *of correctly labeled data, the algorithm produces a hypothesis* $h_S \in \mathcal{H}$ *such that* $\mathbf{Pr}\left[\mathrm{err}_{\mathcal{D}}(h_S) < \epsilon\right] \geq 1 - \delta$*.*

PAC stands for "probably approximately correct". "Probably" means that the probability is at least $1 - \delta$, "approximately correct" means that $\mathrm{err}_{\mathcal{D}}(h_S) < \epsilon$.

[1] Sometimes, it is assumed that the algorithm has to run in polynomial time. We ignore computational issues in this lecture and only focus on the number of samples.

2 Growth Function

We will be interested in better guarantees than this one, particularly if $\mathcal{H}$ is infinite but has a nice structure. If $\mathcal{H}$ is the set of all functions, then a set of m data points could have 2^m different labelings. More structured classes of functions (like linear classifiers) only admit much few such labelings.

Definition 21.7. *Given $S \subseteq X$, let $\mathcal{H}[S]$ be the number of distinct ways to label S using a function in $\mathcal{H}$. That is, if $S = \{x_1, \ldots, x_m\}$ then*

$$\mathcal{H}[S] = |\{(h(x_1), \ldots, h(x_m)) \in \{-1, 1\}^m \mid h \in \mathcal{H}\}| \quad .$$

The growth function *of $\mathcal{H}$ is defined as $\mathcal{H}[m] = \max_{S \subseteq X, |S|=m} \mathcal{H}[S]$.*

Trivially, $\mathcal{H}[m] \leq 2^m$. If, for example, $X = \mathbb{R}$ and $\mathcal{H}$ is the class of functions of the form

$$h(x) = \begin{cases} -1 & \text{for } x \leq t \\ 1 & \text{otherwise} \end{cases}$$

then $\mathcal{H}[m] = m + 1$.

Theorem 21.8. *For all choices of $\epsilon > 0$, $\delta > 0$, if*

$$m \geq \max\left\{\frac{8}{\epsilon}, \frac{2}{\epsilon}\log_2\left(\frac{2\mathcal{H}[2m]}{\delta}\right)\right\} \quad ,$$

then with probability at least $1 - \delta$, all $h \in \mathcal{H}$ with $\mathrm{err}_S(h) = 0$ have $\mathrm{err}_{\mathcal{D}}(h) < \epsilon$.

Proof. Let A be the event that there is $h \in \mathcal{H}$ with $\mathrm{err}_{\mathcal{D}}(h) \geq \epsilon$ but $\mathrm{err}_S(h) = 0$. We would like to show $\mathbf{Pr}[A] \leq \delta$.

To bound the probability of A, we introduce an auxiliary event. To this end, let S' be another draw of m independent samples from $\mathcal{D}$. Let B be the event that there is $h \in \mathcal{H}$ with $\mathrm{err}_{S'}(h) \geq \frac{\epsilon}{2}$ but $\mathrm{err}_S(h) = 0$.

Below, we will show the following two claims:

(a) $\mathbf{Pr}[B \mid A] \geq \frac{1}{2}$

(b) $\mathbf{Pr}[B] \leq \frac{\delta}{2}$.

In combination, the theorem follows because $\mathbf{Pr}[A] \leq 2\mathbf{Pr}[B]$ by $\mathbf{Pr}[B] \geq \mathbf{Pr}[B \mid A]\,\mathbf{Pr}[A]$ and therefore $\mathbf{Pr}[A] \leq 2\frac{\delta}{2} = \delta$.

Let us first show Claim (a). Let us understand what this conditional probability expresses. Event A has already occured. This depends only on the set S and requires that there is $h \in \mathcal{H}$ with $\mathrm{err}_{\mathcal{D}}(h) \geq \epsilon$ but $\mathrm{err}_S(h) = 0$.

We now claim that with probability at least $\frac{1}{2}$ there is also $h \in \mathcal{H}$ with $\mathrm{err}_{S'}(h) \geq \frac{\epsilon}{2}$ but $\mathrm{err}_S(h) = 0$. More specifically, we claim that even for the same S that caused A to happen we have $\mathrm{err}_{S'}(h) \geq \frac{\epsilon}{2}$ with probability at least $\frac{1}{2}$. In other words, h has to classify at least $m\frac{\epsilon}{2}$ elements of S' incorrectly. We fill S' by independent draws from $\mathcal{D}$ and we know that $\mathrm{err}_{\mathcal{D}}(h) \geq \epsilon$. In other words, we perform m independent biased coin tosses. The probability the coin comes up heads is at least ϵ in each toss. We claim that with probability at least $\frac{1}{2}$ we see at least $\frac{\epsilon}{2}m$ times heads.

Let Z be the number of heads in the coin tosses. We have $\mathbf{E}[Z] \geq \epsilon m$ and $\mathrm{Var}[Z] \leq \epsilon(1-\epsilon)m$ if $\epsilon \leq \frac{1}{2}$. So, by Chebyshev's inequality

$$\mathbf{Pr}\left[Z \leq \frac{\epsilon}{2}m\right] \leq \mathbf{Pr}\left[|Z - \mathbf{E}[Z]| \geq \frac{\epsilon}{2}m\right] \leq \frac{\mathrm{Var}[Z]}{\left(\frac{\epsilon}{2}m\right)^2} \leq \frac{\epsilon(1-\epsilon)m}{\left(\frac{\epsilon}{2}m\right)^2} = \frac{4(1-\epsilon)}{\epsilon m} \leq \frac{1}{2} \ ,$$

where in the last step we use that $m \geq \frac{8}{\epsilon}$.

It remains to show Claim (b). To bound the probability of event B to happen, we devise a different but equivalent way of determining S and S': We draw $2m$ times from the distribution $\mathcal{D}$; let the outcome be called T. Now, draw m times *without replacement* from T, let the outcome be called S, and let $S' = T \setminus S$.

Fix the set T and consider a fixed $h \in \mathcal{H}$. Let $h(x) \neq f(x)$ for exactly k elements of T. The only way we can have $\mathrm{err}_{S'}(h) \geq \frac{\epsilon}{2}$ is that $k \geq \frac{\epsilon}{2}m$.

Furthermore, the probability that h is correct on all of S is given as

$$\begin{aligned}\mathbf{Pr}[\mathrm{err}_S(h) = 0 \mid T] &= \frac{2m-k}{2m} \cdot \frac{2m-k-1}{2m-1} \cdot \ldots \cdot \frac{m-k+1}{m+1} \\ &= \frac{m(m-1)\ldots(m-k+1)}{(2m)(2m-1)\ldots(2m-k+1)} \leq 2^{-k} \ .\end{aligned}$$

This means that for a fixed h and a fixed T

$$\mathbf{Pr}\left[\mathrm{err}_S(h) = 0 \text{ and } \mathrm{err}_{S'}(h) \geq \frac{\epsilon}{2} \;\middle|\; T\right] \leq \left\{ \begin{array}{ll} 0 & \text{if } k < \frac{\epsilon}{2}m \\ 2^{-k} & \text{otherwise} \end{array} \right\} \leq 2^{-\frac{\epsilon}{2}m} \ .$$

Now, the key difference to the previous, much simple proof happens: the set T has size only $2m$. This means, because only the function values on T matter, effectively there are at most $\mathcal{H}[2m]$ different choices for h. Therefore, the union bound now gives us

$$\mathbf{Pr}[B \mid T] = \mathbf{Pr}\left[\exists h \in \mathcal{H} : \mathrm{err}_S(h) = 0 \text{ and } \mathrm{err}_{S'}(h) \geq \frac{\epsilon}{2} \;\middle|\; T\right] \leq \mathcal{H}[2m]2^{-\frac{\epsilon}{2}m} \leq \frac{\delta}{2} \ .$$

This bound holds for all conditional probabilities, no matter which set T we use. Therefore also the unconditional probability is bounded this way. □

References and Further Reading

These notes are based on notes and lectures by Anna Karlin `https://courses.cs.washington.edu/courses/cse522/17sp/` and Avrim Blum `http://www.cs.cmu.edu/~avrim/ML14/`. Also see the references therein.