Algorithms and Uncertainty, Winter 2018/19 Lecture 23 (4 pages)

Agnostic PAC Learning

Instructor: Thomas Kesselheim

Today, we will continue our study of learnability of hypothesis classes. Again, our task will be to classify points from a set X. There is a set of hypotheses $\mathcal{H}$, each being a function $X \to \{-1, 1\}$. There is a distribution $\mathcal{D}$ over pairs of points and labels $X \times \{-1, 1\}$, which we would like to predict correctly. In our results so far, we only considered the "realizable case". That is, there is a ground truth $f\colon X \to \{-1, 1\}$ and $f \in \mathcal{H}$. The distribution $\mathcal{D}$ is such that for all pairs (x, y) in the support $y = f(x)$.

Actually, in any machine learning setting, this is too strong an assumption. Usually, the features do not describe a concept entirely. Coming back to our original example of spam classification, typical features might be word counts, sender IP addresses, header data, and so on. Of course, based on only this information, it is impossible to fully correctly classify all e-mails. Even if it was possible, we might choose only a smaller hypothesis class $\mathcal{H}$ to allow efficient learning.

For example, we could have $X = [0, 1]^2$ and $\mathcal{H}$ as the set of linear classifiers. The ground truth might not be determined by a straight line but a more complex function $f\colon X \to \{-1, 1\}$ (left image in Figure 1). It might even be that there is no function f such that the label of x is always $f(x)$. We can capture this by the labels being random. The distribution $\mathcal{D}$ is over $X \times \{0, 1\}$: For a fixed x, it might return $(x, 0)$ with probability 70% and $(x, 1)$ with probability 30%.

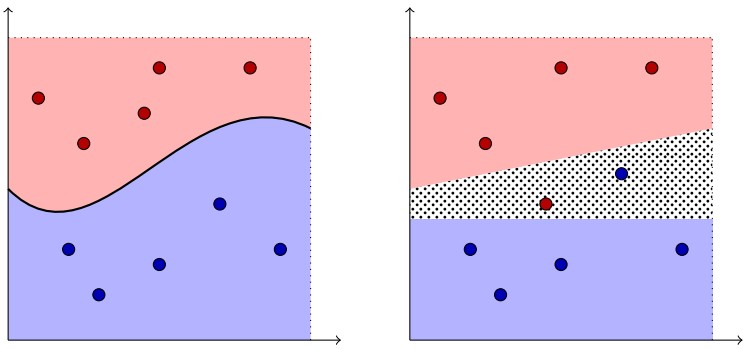

Figure 1: Examples of unrealizable cases: In the left image, no linear classifier matches the function f on all points. In the right image, in the dotted area, the labels are random; for example 0 or 1 with probability 50%. So there is no function $f\colon X \to \{0, 1\}$ that always returns the correct label.

1 PAC Learnability in the Unrealizable Case

Given a set of m samples $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$, the *training error* $\mathrm{err}_S(h)$ of a hypothesis h with respect to a training set S is

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

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

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

So far, we assumed that there was a hypothesis $f \in \mathcal{H}$ such that $\mathrm{err}_{\mathcal{D}}(f) = 0$. In this agnostic (or unrealizable) case, $\mathrm{err}_{\mathcal{D}}(f) > 0$ for all $f \in \mathcal{H}$. So, all we can hope for is to find a hypothesis h_S based on the traing set S such that $\mathrm{err}_{\mathcal{D}}(h_S)$ is as close as possible to $\min_{f\in\mathcal{H}} \mathrm{err}_{\mathcal{D}}(f)$.

Definition 23.1. *Hypothesis class $\mathcal{H}$ is PAC-learnable (in the agnostic sense) if there is a function $m_{\mathcal{H}}$ and a learning algorithm such that for any $\epsilon, \delta > 0$ and any distribution $\mathcal{D}$, given a set of random samples S of size at least $m_{\mathcal{H}}(\epsilon, \delta)$ drawn from $\mathcal{D}$, the algorithm produces a hypothesis $h_S \in \mathcal{H}$ such that* $\mathbf{Pr}\left[\mathrm{err}_{\mathcal{D}}(h_S) < \min_{f\in\mathcal{H}} \mathrm{err}_{\mathcal{D}}(f) + \epsilon\right] \geq 1 - \delta$.

Our focus will again be on *empirical risk minimizers*. In the realizable case, this means that given a training set S we choose a hypothesis h such that the training error is 0, that is, $\mathrm{err}_S(h) = 0$. Now, there might not be any such hypothesis. Instead, we will choose h that minimizes $\mathrm{err}_S(h)$.

2 Uniform Convergence for Finite Hypothesis Classes

Let us again start with finite hypothesis classes. For these, we can show the following theorem.

Theorem 23.2. *If $m \geq \frac{2}{\epsilon^2} \ln\left(\frac{2|\mathcal{H}|}{\delta}\right)$, then with probability at least $1-\delta$, all $h \in \mathcal{H}$ that minimize $\mathrm{err}_S(h)$ fulfill $\mathrm{err}_{\mathcal{D}}(h) < \min_{f\in\mathcal{H}} \mathrm{err}_{\mathcal{D}}(f) + \epsilon$.*

This theorem is a little complicated because it asks us to show a property of all $h \in \mathcal{H}$ that minimize $\mathrm{err}_S(h)$. Instead, we will show a sufficient and easier statement. Suppose that the set S fulfills that

$$|\mathrm{err}_{\mathcal{D}}(h) - \mathrm{err}_S(h)| < \frac{\epsilon}{2} \qquad \text{for all } h \in \mathcal{H}. \tag{1}$$

So, the true error is close to the training error for every possible hypothesis. Let h be a hypothesis that minimizes $\mathrm{err}_S(h)$; f be a hypothesis that minimizes $\mathrm{err}_{\mathcal{D}}(f)$. We then have

$$\mathrm{err}_{\mathcal{D}}(h) < \mathrm{err}_S(h) + \frac{\epsilon}{2} \leq \mathrm{err}_S(f) + \frac{\epsilon}{2} < \mathrm{err}_{\mathcal{D}}(f) + \epsilon \ .$$

Therefore, if Condition (1) is fulfilled, then, in particular, all $h \in \mathcal{H}$ that minimize $\mathrm{err}_S(h)$ fulfill $\mathrm{err}_{\mathcal{D}}(h) < \min_{f\in\mathcal{H}} \mathrm{err}_{\mathcal{D}}(f) + \epsilon$. Therefore, to prove Theorem 23.2 it suffices to show the following proposition.

Proposition 23.3. *If $m \geq \frac{2}{\epsilon^2} \ln\left(\frac{2|\mathcal{H}|}{\delta}\right)$,*

$$\mathbf{Pr}\left[\exists h \in \mathcal{H} : |\mathrm{err}_{\mathcal{D}}(h) - \mathrm{err}_S(h)| \geq \frac{\epsilon}{2}\right] < \delta \ .$$

We first show a statement about a single hypothesis.

Lemma 23.4. *Consider a fixed hypothesis $h \in \mathcal{H}$. Let S be a set of m samples drawn from $\mathcal{D}$. Then for all $\gamma > 0$*

$$\mathbf{Pr}\left[|\mathrm{err}_{\mathcal{D}}(h) - \mathrm{err}_S(h)| \geq \gamma\right] \leq 2\exp\left(-2m\gamma^2\right) \ .$$

Proof. This is a pretty straightforward application of Hoeffding's inequality.

Lemma 23.5 (Hoeffding's inequality)**.** *Let $Z_1, \ldots, Z_N$ be independent random variables such that $a_i \leq Z_i \leq b_i$ with probability 1. Let $\bar{Z} = \frac{1}{N}\sum_{i=1}^{N} Z_i$ be their average. Then for all $\gamma \geq 0$*

$$\mathbf{Pr}\left[|\bar{Z} - \mathbf{E}\left[\bar{Z}\right]| \geq \gamma\right] \leq 2\exp\left(-\frac{2N^2\gamma^2}{\sum_{i=1}^{N}(b_i - a_i)^2}\right) \ .$$

Let $Z_i = 1$ if $h(x_i) \neq y_i$ and 0 otherwise. Then $\bar{Z} = \text{err}_S(h)$ and $\mathbf{E}\left[\bar{Z}\right] = \text{err}_{\mathcal{D}}(h)$. Furthermore, $a_i = 0$, $b_i = 1$ and $N = m$. So the lemma follows. □

Now, proving Proposition 23.3 is straightforward too.

Proof of Proposition 23.3. By using the union bound and setting $\gamma = \frac{\epsilon}{2}$ in Lemma 23.4, we have

$$\mathbf{Pr}\left[\exists h \in \mathcal{H} : |\text{err}_{\mathcal{D}}(h) - \text{err}_S(h)| \geq \frac{\epsilon}{2}\right] \leq |\mathcal{H}| \cdot 2\exp\left(-m\frac{\epsilon}{2}\right) \leq \delta \ . \qquad \square$$

We have shown a stronger property: If the size of the training set is only large enough, with good probability the training error and the true error are close for every possible hypothesis. This is called *uniform convergence*.

Definition 23.6. *Hypothesis class $\mathcal{H}$ fulfills the* uniform-convergence property *if there is a function $m_{\mathcal{H}}$ such that for any $\epsilon, \delta > 0$ and any distribution $\mathcal{D}$, given a random training set S of size at least $m_{\mathcal{H}}(\epsilon, \delta)$ of correctly labeled data $|\text{err}_{\mathcal{D}}(h) - \text{err}_S(h)| < \epsilon$ for all $h \in \mathcal{H}$ with probability at least $1 - \delta$.*

3 Bounded Growth Function

We have already shown that every finite hypothesis class fulfills the uniform-convergence property and is therefore PAC learnable. As in the realizable case, we will now consider the case of hypothesis classes of bounded growth functions. We will show the following theorem.

Theorem 23.7. *For any hypothesis class $\mathcal{H}$, for all choices of $\epsilon > 0$, $\delta > 0$, if S is a training set of size m with*

$$m \geq \frac{8}{\epsilon^2}\ln\left(\frac{4\mathcal{H}[2m]}{\delta}\right) \ .$$

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

Note that we can again use Sauer's lemma to bound the growth function in terms of the VC dimension, just as we did for the realizable case. Therefore, we can derive that if the VC dimension is finite the class is also PAC learnable.

Proof of Theorem 23.7. We do a similar proof as in the realizable case. Once again, we let S' be another set of random samples of size m drawn from the distribution $\mathcal{D}$.

Let A denote the event that there is a $h \in \mathcal{H}$ such that $|\text{err}_{\mathcal{D}}(h) - \text{err}_S(h)| \geq \epsilon$; let B be the event that there is a $h \in \mathcal{H}$ such that $|\text{err}_{S'}(h) - \text{err}_S(h)| \geq \frac{\epsilon}{2}$.

We show that by our choice of m, we have

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

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

This then proves the claim exactly the same way as before.

We will apply Hoeffding's inequality to show both claims. Note that, for any fixed $h \in \mathcal{H}$, $\text{err}_{S'}(h)$, just as $\text{err}_S(h)$, can be considered an average of m independent random variables in $[0, 1]$ and $\mathbf{E}\left[\text{err}_{S'}(h)\right] = \text{err}_{\mathcal{D}}(h)$. So, Hoeffding's inequality tells us that for all $\gamma > 0$

$$\mathbf{Pr}\left[|\text{err}_{S'}(h) - \text{err}_{\mathcal{D}}(h)| \geq \gamma\right] \leq 2\exp\left(-2m\gamma^2\right) \ .$$

So, in particular

$$\mathbf{Pr}\left[|\mathrm{err}_{S'}(h) - \mathrm{err}_{\mathcal{D}}(h)| \geq \frac{\epsilon}{2}\right] \leq 2\exp\left(-2m\frac{\epsilon^2}{4}\right) \leq \frac{1}{2} \ .$$

Furthermore, if event A takes place and $|\mathrm{err}_{S'}(h) - \mathrm{err}_{\mathcal{D}}(h)| < \frac{\epsilon}{2}$ for this same h that caused A to happen, then event B takes place automatically. This happens with probability at least $\frac{1}{2}$.

For claim (b), we use a similar argument as in the realizable case but we have to be a little more careful. Think of S and S' being generated in a different way as follows. We draw $2m$ times from the distribution $\mathcal{D}$. Call the draws $(x_1, y_1), \ldots, (x_m, y_m), (x'_1, y'_1), \ldots (x'_m, y'_m)$. For every index i flip a coin: With probability $\frac{1}{2}$, (x_i, y_i) is added to S and (x'_i, y'_i) is added to S'. Otherwise (x'_i, y'_i) is added to S and (x_i, y_i) is added to S'.

Fix a hypothesis h. Let $Z_i = 0$ if h is correct or incorrect on both (x_i, y_i) and (x'_i, y'_i). Furthermore, let $Z_i = 1$ if h correctly classifies the one of (x_i, y_i) and (x'_i, y'_i) being added to S'; otherwise (so the one classified incorrectly is added to S), let $Z_i = -1$. By this definition

$$\mathrm{err}_{S'}(h) - \mathrm{err}_{S}(h) = \frac{1}{m}\sum_{i=1}^{m} Z_i \ .$$

Furthermore, $\mathbf{E}[Z_i] = 0$ for all i by symmetry reasons. Therefore, Hoeffding's inequality gives us

$$\mathbf{Pr}\left[|\mathrm{err}_{S'}(h) - \mathrm{err}_{S}(h)| \geq \frac{\epsilon}{2}\right] \leq 2\exp\left(-\frac{2m^2(\frac{\epsilon}{2})^2}{4m}\right) \leq \frac{\delta}{2\mathcal{H}[2m]}$$

The remainder works just as before: The hypotheses in $\mathcal{H}$ can label the $2m$ points (x_1, y_1), $\ldots, (x_m, y_m), (x'_1, y'_1), \ldots (x'_m, y'_m)$ in only $\mathcal{H}[2m]$ different ways. Therefore, we have to take a union bound over only this many hypotheses. □

4 Implications

Indeed, we have now shown that four properties of a hypothesis class $\mathcal{H}$ are equivalent:

(i) It is PAC-learnable in the realizable sense.

(ii) It is PAC-learnable in the agnostic sense.

(iii) It fulfills the uniform-convergence property.

(iv) Its VC-dimension is finite.

We showed the equivalence of (i) and (iv) in the previous lecture. Theorem 23.7 shows that (iv) implies (iii); generalizing the proof of Proposition 23.3 shows that (iii) implies (ii). It is clear that (ii) implies (i) because it is a stronger condition.

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.