CS260: Machine Learning Theory
Lecture 4: The Agnostic Setting

October 5, 2011

Lecturer: Jennifer Wortman Vaughan

1 The Agnostic Setting

Until this point, we have been working under the assumption that there exists a target function c in a known concept class $\mathcal{C}$ that perfectly labels our data. This is not always a valid assumption to make. There could be random noise in the data, occasionally causing a label to be flipped. It could be the case that there is indeed a perfect target function, but it is not in the concept class $\mathcal{C}$ that we are considering. Or perhaps there is so much randomness in the labels that no function comes close to labeling all of our data correctly. In light of this, we would like to remove the assumption of a perfect target function. This is often referred to as the *agnostic* learning setting, since we make no assumptions about the origin of the labels. It is also referred to as the *unrealizable* setting, to contrast it with the *realizable* setting we have been studying so far.

We need to update all of our definitions and assumptions for this new setting. Before we assumed that each input point is drawn i.i.d. from a distribution $\mathcal{D}$ and labeled by the target function. We now instead assume that there exists a *joint* distribution over pairs of values $(\vec{x}, y)$ where $\vec{x}$ is the input point and y is the corresponding label. In light of this, we need to update our notion of error. Previously we defined error as

$$\mathrm{err}(h) = \mathrm{Pr}_{\vec{x}\sim\mathcal{D}}\left(h(\vec{x}) \neq c(\vec{x})\right).$$

We now define it in terms of the joint distribution over $(\vec{x}, y)$ pairs, with

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

Note that this is a strictly more general way to define the error. We can still model a perfect target function as a joint probability distribution for which the label y is deterministically equal to $c(\vec{x})$ conditioned on the input $\vec{x}$.

Next, we must update our goal for this new setting. Previously our goal was to output a function $h \in \mathcal{H}$ with $\mathrm{err}(h) \leq \epsilon$ for some small value ϵ. In the agnostic setting, such a function might not exist. We can only possibly hope to find a function as good as the best function in $\mathcal{H}$. Therefore, our new goal is instead to output a function $\hat{h}$ such that $\mathrm{err}(\hat{h})$ is close to $\min_{h\in\mathcal{H}} \mathrm{err}(h)$.

Finally, many of our previous results relied on the existence of algorithms that could return a hypothesis consistent with some data. It is no longer necessarily the case that there is any consistent hypothesis in $\mathcal{H}$, so we relax this notion as well. Instead, we consider algorithms that return the hypothesis $\hat{h} \in \mathcal{H}$ that is "most consistent" with the data. We measure the extent of the consistency of a hypothesis by its *empirical error* on a data set $\{(\vec{x}_1, y_1), \cdots, (\vec{x}_m, y_m)\}$, which is defined as

$$\widehat{\mathrm{err}}(h) = \frac{1}{m}\left|\{i : h(\vec{x_i}) \neq y_i\}\right| .$$

All CS260 lecture notes build on the scribes' notes written by UCLA students in the Fall 2010 offering of this course. Although they have been carefully reviewed, it is entirely possible that some of them contain errors. If you spot an error, please email Jenn.

The empirical error is an unbiased estimate of the true error, in that $\mathrm{E}[\widehat{\mathrm{err}}(h)] = \mathrm{err}(h)$. (Exercise: Convince yourself this is true.)

We consider algorithms that return a hypothesis in $\mathcal{H}$ with minimal empirical error, and denote this hypothesis $\hat{h}$.

2 Towards a General Learning Bound in the Agnostic Setting

We would like to derive a general learning bound that holds without the assumption of a perfect target function. Assume that we have an algorithm $\mathcal{A}$ that, given m independent samples from an arbitrary distribution $\mathcal{D}$ (joint over $\vec{x}$ and y), outputs a hypothesis $\hat{h} \in \mathcal{H}$ with minimal empirical error. (If there is more than one, $\mathcal{A}$ can choose in any arbitrary way.) We would like to derive a lower bound on m that guarantees that for any $\epsilon, \delta \in (0, 1/2)$,

$$\Pr\left(\mathrm{err}(\hat{h}) - \min_{h \in \mathcal{H}} \mathrm{err}(h) > \epsilon\right) < \delta\ .$$

How might we derive such a bound? Well, suppose that we were able to make the following assumption:

$$\forall h \in \mathcal{H},\ |\mathrm{err}(h) - \widehat{\mathrm{err}}(h)| \leq \frac{\epsilon}{2}.$$

If this assumption were true, then for any $h \in \mathcal{H}$,

$$\begin{aligned}
\mathrm{err}(\hat{h}) &\leq \widehat{\mathrm{err}}(\hat{h}) + \frac{\epsilon}{2} && \text{this follows from the assumption} \\
&\leq \widehat{\mathrm{err}}(h) + \frac{\epsilon}{2} && \text{this follows from the definition of } \hat{h} \\
&\leq \mathrm{err}(h) + \epsilon. && \text{this follows from the assumption}
\end{aligned}$$

Since the derivation above holds for any $h \in \mathcal{H}$, it must hold for the h that minimizes error, so we have that

$$\mathrm{err}(\hat{h}) \leq \min_{h \in \mathcal{H}}(\mathrm{err}(h)) + \epsilon.$$

This tells us that if we are able to make m large enough so that our assumption holds with high probability, we are effectively done. Below we consider how to do this.

3 Concentration Inequalities and Uniform Convergence

Above we argued that in order to prove an error bound for the agnostic setting, it will be sufficient to show that the empirical error of a hypothesis approaches its expectation (i.e., the true error of the hypothesis) as the number of samples grows. This idea is referred to as *uniform convergence* and can be proved using one of several concentration inequalities collectively referred to as the Chernoff bounds. The bound that is most useful for our purposes is Hoeffding's inequality.

Theorem 1. *(Hoeffding's inequality) Let $Z_1,\ Z_2,\ \dots,\ Z_m$ be independent random variables such that $Z_i \in [a_i,\ b_i]$ for all i. Let $\hat{p}$ be the empirical mean of these random variables,*

$$\hat{p} = \frac{1}{m}\sum_{i=1}^{m} Z_i$$

and let p *be its expectation,*

$$p = E[\hat{p}] = \frac{1}{m}\sum_{i=1}^{m} E[Z_i].$$

Then for any $\epsilon > 0$,

$$Pr(p - \hat{p} \geq \epsilon) \leq e^{-2\epsilon^2 m^2 / \sum_{i=1}^{m}(b_i - a_i)^2}$$

and

$$Pr(\hat{p} - p \geq \epsilon) \leq e^{-2\epsilon^2 m^2 / \sum_{i=1}^{m}(b_i - a_i)^2}.$$

Applying the union bound, we obtain the following corollary.

Corollary 1. *Under the same assumptions on* $Z_1,\ Z_2,\ \ldots,\ Z_m$, p, *and* $\hat{p}$ *as above,*

$$Pr\left(|p - \widehat{p}| \geq \epsilon\right) \leq 2e^{\frac{-2\epsilon^2 m^2}{\sum_{i=1}^{m}(b_i - a_i)^2}}.$$

4 Back to the General Learning Bound

We are now ready to prove a sample complexity bound for the agnostic setting. We start by showing how many samples we need to guarantee that the assumption above holds with high probability.

Lemma 1. *Let* $\{(\vec{x}_1, y_1), \cdots, (\vec{x}_m, y_m)\}$ *be a set of data points drawn i.i.d. from a distribution* $\mathcal{D}$. *If*

$$m \geq \frac{2}{\epsilon^2}\left(\ln|\mathcal{H}| + \ln\left(\frac{2}{\delta}\right)\right)$$

then with probability at least $1 - \delta$, *for all* $h \in \mathcal{H}$, $|\mathrm{err}(h) - \widehat{\mathrm{err}}(h)| \leq \epsilon/2$.

Proof: Consider any arbitrary $h \in \mathcal{H}$. For $i = 1, \ldots, m$, define

$$Z_i = \begin{cases} 1 & \text{if } h(x_i) \neq y_i \\ 0 & \text{otherwise.} \end{cases}$$

Since $\widehat{err}(h) = (1/m)\sum_{i=1}^{m} Z_i$ and $err(h) = \mathrm{E}[\widehat{err}(h)]$, we can apply (the corollary to) Hoeffding's inequality with $\hat{p} = \widehat{err}(h)$, $p = err(h)$, $Z_1, \ldots, Z_m$ defined as above, ϵ set to $\epsilon/2$, and $b_i - a_i = 1$ for all i. This gives

$$\Pr\left(|\mathrm{err}(h) - \widehat{\mathrm{err}}(h)| \geq \frac{\epsilon}{2}\right) \leq 2e^{-\epsilon^2 m/2}.$$

Applying the union bound, we get

$$\Pr\left(\exists h \in \mathcal{H} \,:\, |\mathrm{err}(h) - \widehat{\mathrm{err}}(h)| \geq \frac{\epsilon}{2}\right) \leq 2|\mathcal{H}|e^{-\epsilon^2 m/2}.$$

Since we want to show $|\mathrm{err}(h) - \widehat{\mathrm{err}}(h)| \leq \epsilon/2$, this is a bound on the probability of *failure.* We can compute the condition on the value of m that will make $2|\mathcal{H}|e^{-\epsilon^2 m/2} \leq \delta$. Rearranging terms, we find that this holds if

$$m \geq \frac{2}{\epsilon^2}\left(\ln|\mathcal{H}| + \ln\left(\frac{2}{\delta}\right)\right).$$

□

Combined with the argument that we gave earlier in class, this gives us the following theorem.

Theorem 2. *For any concept class $\mathcal{H}$, suppose that we have access to an algorithm that, given input $\{(\vec{x}_1, y_1), \cdots, (\vec{x}_m, y_m)\}$ outputs a function $\hat{h} \in \mathcal{H}$ with minimal empirical error. Then for any distribution $\mathcal{D}$ over input-label pairs, for any $\epsilon \in (0, 1/2)$ and $\delta \in (0, 1/2)$, if $\mathcal{A}$ is given access to m pairs drawn i.i.d. from $\mathcal{D}$ with*

$$m \geq \frac{2}{\epsilon^2}\left(\ln|\mathcal{H}| + \ln\left(\frac{2}{\delta}\right)\right),$$

$\mathcal{A}$ will output a function $\hat{h}$ such that with probability $\geq 1 - \delta$, $\operatorname{err}(\hat{h}) - \min_{h \in \mathcal{H}} \operatorname{err}(h) \leq \epsilon$.

4.1 Remarks about the General Learning Bound

When we compare the bound on the number of samples for the agnostic case with the bound for the realizable case, we see that they have a very similar form. The Occam's razor principle applies here as well. The fewer hypotheses we have the better because of the $\ln|\mathcal{H}|$ factor. Note that also we now have ϵ^2 where we used to have ϵ. This means that we are paying a penalty of a factor $1/\epsilon$ by not having a perfect target function.

By solving for ϵ in the bound we obtained for m and applying it to the inequality we obtained in an earlier result, we get

$$\operatorname{err}(\hat{h}) \leq \min_{h \in \mathcal{H}} (\operatorname{err}[h]) + \sqrt{\frac{2\ln|\mathcal{H}| + 2\ln(2/\delta)}{m}}.$$

This gives us an interesting perspective on the effect of the size of the hypothesis class. The first term in this equation tells us that we want a large hypothesis class because we increase our likelihood of minimizing the error. However, the second term tells us that a smaller hypothesis class is better because of the Occam's razor principle.

Note that way in which the Occam's razor principle shows up here can be described in terms of *overfitting*. When we have a large $\mathcal{H}$ and we try to fit our hypothesis to the data, the likelihood of picking the wrong hypothesis is larger, so overfitting becomes a bigger issue.