**COS 511: Theoretical Machine Learning**

Lecturer: Rob Schapire
Scribe: Kevin Hassani

Lecture #7
February 26, 2013

# 1 A Lower Bound on Sample Complexity

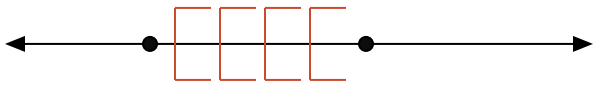
We have established upper bounds of $m$ involving training size, complexity, and the VC-dim. Last time, we discussed a lower-bound as a function of the VC-dim. Even if an adversarial agent, making the worst case scenario, chooses $\mathcal{C}$ and $D$, $A$ should still satisfy the conditions of PAC learning. We are trying to prove the opposite: that if $m$ is too small ($< \frac{d}{2}$), then the concept is not PAC learnable. We simply choose $\epsilon, \delta < 1/8$ to prove that it cannot be PAC learnable.

In the last lecture, we gave a false proof by choosing $\mathcal{C}$ after $A$ (and choosing $\mathcal{C}$ to disagree with over half of the training labels). We give a valid proof this time, paying particular attention to the order of steps.

**Theorem 1.** *Let $d = VC\text{-}dim(\mathcal{C})$. For any algorithm, $A$, there exists a concept, $c \in \mathcal{C}$, and distribution, $D$, such that if $A$ is given $m \leq \frac{d}{2}$ examples, then*

$$\mathbf{Pr}\left[err(h_A) > \frac{1}{8}\right] \geq \frac{1}{8}$$

*Proof.* Because $\mathcal{C}$ has VC-dim d, there exists $z_1, \ldots, z_d$ that are shattered by $\mathcal{C}$. Let $D$ be uniform random over the shattered set. We want to pick a $c$ such that no $A$ can be learnable on it. One approach might be to pick $c \in \mathcal{C}$ uniformly at random. The problem with this approach is that there could be many $c$ with the same labeling on the training set. In our go-to example of positive half-lines, there could be a concentration of $c$ between two points in the sample:



Instead, we choose uniformly at random over $c \in \mathcal{C}'$ where $\mathcal{C}' \subseteq \mathcal{C}$ with one representative for every labeling of $z_1, \ldots, z_d$.

It is important to keep track of the order in which steps occur. We will consider two experiments with different orderings:

Experiment 1:
$c$ chosen
training sample, $S$, chosen (according to $D$) and labeled by $c$
$h_A$ computed from $S$
test point, $x$, chosen
measure: $Pr[h_A(x) \neq c(x)]$

Experiment 2: $S$ $(x_1, \ldots, x_m)$ chosen without labels
random labels $(c(x_i))$ chosen for just $x_i \in S$

$h_A$ computed from labeled $S$
test point, $x$, chosen
if $x \notin S$ then label $c(x)$ is chosen at uniform random
measure: $Pr[h_A(x) \neq c(x)]$

The above two experiments produce the same probability measures since the choice of label $c$ is independent of the choice of $S$, and we choose the label for $x$ independently of the samples $S$. This probability is given over the randomness of concept $c$, the examples $S$, and the test point $x$. We denote it as $\mathbf{Pr}_{c,S,x}[h_A(x) \neq c(x)]$.

$$\begin{aligned}\mathbf{Pr}_{c,S,x}\left[h_A(x) \neq c(x)\right] &\geq \mathbf{Pr}_{c,S,x}\left[x \notin S \wedge h_A(x) \neq c(x)\right] \\ &= \underbrace{\mathbf{Pr}_{c,S,x}\left[x \notin S\right]}_{\geq \frac{1}{2} \text{ because it is chosen uniform and } m \leq \frac{d}{2}} \underbrace{\mathbf{Pr}_{c,S,x}\left[h_A(x) \neq c(x) | x \notin S\right]}_{=\frac{1}{2}} \\ &\geq \frac{1}{4}\end{aligned}$$

We are getting closer to the desired results, but we want to show that there exists a $c$ such that $\mathbf{Pr}_S\left[\text{err}_D(h_A) > \frac{1}{8}\right] \geq \frac{1}{8}$, where $\text{err}_D(h_A) \equiv \mathbf{Pr}_x\left[h_A(x) \neq c(x)\right]$. To do this, we will use marginalization:

$$\mathbf{Pr}\left[a\right] = \mathbf{E}_a\left[\mathbf{P}\left[a|x\right]\right]$$

So, we have:

$$\mathbf{E}_c\left[\mathbf{Pr}_{S,x}\left[h_A(x) \neq c(x)|c\right]\right] = \mathbf{P}_{c,S,x}\left[h_A(x) \neq c(x)\right]$$

Therefore there exists an $x \in \mathcal{C}' \subseteq \mathcal{C}$:

$$\mathbf{Pr}_{S,x}\left[h_A(x) \neq c(x)\right] \geq \frac{1}{4} \qquad \text{*adversary can pick } c \text{ that is at least the average}$$

$$\begin{aligned}\frac{1}{4} &\leq \mathbf{E}_S\left[\mathbf{Pr}_x\left[h_A(x) \neq c(x)\right]\right] = \mathbf{Pr}_{S,x}\left[h_A(x) \neq c(x)\right] \\ &= \mathbf{E}_S\left[\text{err}(h_A)\right] \\ &\leq \underbrace{\mathbf{Pr}_S\left[\text{err}(h_A) \leq \frac{1}{8}\right]}_{\leq 1} \cdot \frac{1}{8} + \mathbf{Pr}_S\left[\text{err}(h_A) > \frac{1}{8}\right] \\ &\leq \frac{1}{8} + \mathbf{Pr}_S\left[\text{err}(h_A) > \frac{1}{8}\right]\end{aligned}$$

The second to last line comes from the following for $X \in [0,1]$:

$$\begin{aligned}\mathbf{E}\left[x\right] &= \sum_{x: x \leq 1/8} \mathbf{P}\left[x\right] \cdot \underbrace{x}_{\leq 1/8} + \sum_{x: x > 1/8} \mathbf{P}\left[x\right] \cdot \underbrace{x}_{\leq 1} \\ &\leq \mathbf{P}\left[X \leq \frac{1}{8}\right] \cdot \frac{1}{8} + \mathbf{P}\left[X > \frac{1}{8}\right] \cdot 1\end{aligned}$$

□

## 2 Inconsistent Model Hypothesis

So far we have only dealt with the situation in which the hypothesis is consistent, but what to do if the hypothesis is not consistent? There are several reasons it may not be consistent:

- concept $c \notin \mathcal{H}$ ($\mathcal{H}$ not rich enough)
- computationally hard to find it
- $c$ may not exist. We've assumed that the label $c(x)$ is a function of $x$, but it may not be deterministic. For example, predicting weather is a random process (though not uniform).

In order to accommodate this new wrinkle, we must make a few modifications. We now have $(x, y) \sim D$, $D$ distribution over $X \times \{0, 1\}$ with the following chain rule:

$$\mathbf{Pr}[x, y] = \underbrace{\mathbf{Pr}[x]}_{\substack{\text{process of picking } x \\ \text{(e.g. current weather condition)}}} \underbrace{\mathbf{Pr}[y|x]}_{\substack{\text{process of labeling} \\ \text{(e.g. weather tomorrow)}}}$$

Before, we had $\mathbf{Pr}[y = 1|x] = 0$ or 1, but now it can be between 0 and 1.

We also have to modify our definition of the generalization error. Before, we had:

$$\text{err}_D(h) = \mathbf{Pr}_{x \sim D}[h(x) \neq c(x)]$$

Now we have:

$$\text{err}_D(h) = \mathbf{Pr}_{(x,y) \sim D}[h(x) \neq y]$$

The first question we must ask is, "What is the best this error can be?" Let's start with an easier case, where we are tossing a coin. It lands on heads with $p$ probability:

$$\begin{cases} \text{heads} & \text{with probability } p \\ \text{tails} & \text{else} \end{cases}$$

Of course, if we wanted to guess the outcome, the optimal prediction would be:

$$\begin{cases} \text{heads} & \text{if } p > \frac{1}{2} \\ \text{tails} & \text{if } p < \frac{1}{2} \end{cases}$$

If $p = \frac{1}{2}$, then we can choose heads or tails. We can do so deterministically since there is nothing to be gained in making randomized predictions.

We face a similar situation when assigning classifications through a hypothesis. The "Bayes optimal classifier" or "Bayes optimal decision rule" is as follows:

$$h_{\text{opt}}(x) = \begin{cases} 1 & \text{if } \mathbf{Pr}_D[y = 1|x] > \frac{1}{2} \\ 0 & \text{if } \mathbf{Pr}_D[y = 1|x] < \frac{1}{2} \end{cases}$$

The "Bayes error" is the theoretical minimum error we can achieve regardless of computational power involved:

$$\text{err}_D(h_{\text{opt}}) = \min_{\text{all } h} \text{err}_D(h)$$

For $h \in \mathcal{H}$, our goal is to find $\min_{h\in\mathcal{H}} \text{err}_D(h)$. Given $S = (x_1, y_1), \ldots, (x_m, y_m)$, we minimize training error using the empirical formula:

$$\widehat{\text{err}}(h) = \frac{1}{m}\sum_{i=1}^{m} \underbrace{\begin{cases} 1 & \text{if } h(x_i) \neq y_i \\ 0 & \text{else} \end{cases}}_{1\{h(x_i)\neq y_i\}}$$

In order for this formula to be useful, we want to show that minimizing $\widehat{\text{err}}(h)$ also minimizes $\text{err}_D(h)$. That is to say $\forall h \in \mathcal{H}$ :

$$|\text{err}(h) - \widehat{\text{err}}(h)| \leq \epsilon$$

Such a property implies that if $\hat{h}$ minimizes the empirical error, that is $\hat{h} = \arg\min_{h\in\mathcal{H}} \widehat{\text{err}}(h)$, then for all $h \in \mathcal{H}$:

$$\begin{aligned} \text{err}(\hat{h}) &\leq \widehat{\text{err}}(\hat{h}) + \epsilon \\ &\leq \widehat{\text{err}}(h) + \epsilon \\ &\leq \text{err}(h) + 2\epsilon \end{aligned}$$

In words, the above inequality states that given a hypothesis $\hat{h}$ with minimal empirical error, the true error of this hypothesis will be no bigger than the minimum true error over all hypotheses in $\mathcal{H}$ plus $2\epsilon$. Thus, this hypothesis will have a generalization error close to the lower bound of the error for all $\mathcal{H}$. To prove this, we must prove uniform convergence: $\forall h \in \mathcal{H}$:

$$|\text{err}(h) - \widehat{\text{err}}(h)| \leq \epsilon$$

In order to prove this, we use Chernoff bounds, and in particular, Hoeffdings' Inequality:

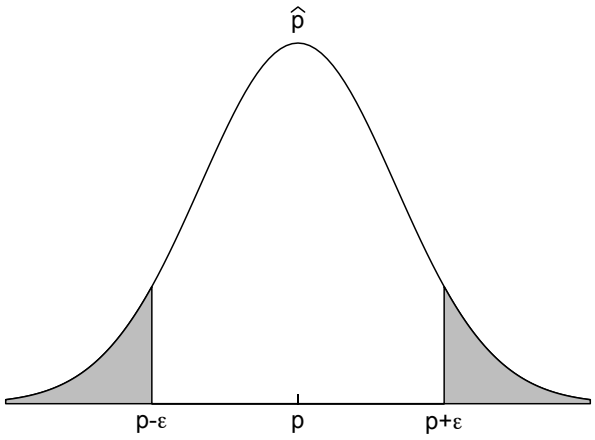
**Theorem 2.** *Let $X_1, \ldots, X_m$ be iid (independent and individually distributed) random variables with $X_i \in [0, 1]$, $p = \mathbf{E}[X_i]$, and $\hat{p} = \frac{1}{m}\sum_{i=1}^{m} X_i$. Then, we have:*

$$\begin{aligned} \mathbf{Pr}[\hat{p} \geq p + \epsilon] &\leq e^{-2\epsilon^2 m} \\ \mathbf{Pr}[\hat{p} \leq p - \epsilon] &\leq e^{-2\epsilon^2 m} \end{aligned}$$

Together, these imply:

$$\mathbf{Pr}[|\hat{p} - p| \geq \epsilon] \leq 2e^{-2\epsilon^2 m} = \delta$$

Such a "tail bound" or "concentration inequality" gives us a probability bound for $\hat{p}$:




From Hoeffdings' Inequality, we can derive that with probability $\geq 1-\delta$, $|\hat{p}-p| \leq \sqrt{\frac{\ln \frac{2}{\delta}}{2m}}$.

In the next lecture we will provide a proof of Chernoff bounds.