COS 511: Theoretical Machine Learning

Lecturer: Rob Schapire
Scribe: Chee Wei Tan

Lecture #7
February 25, 2008

1 A Lower Bound on Sample Complexity

In the last lecture, we stopped at the lower bound on sample complexity. We discussed intuitively why lower bounds must be determined by the target concept class $\mathcal{C}$, rather than the hypothesis class $\mathcal{H}$. We stated the theorem of lower bound and gave a bogus proof. In the following, we give a formal and correct proof on the lower bound:

Theorem 1 *Let d =VC-dim($\mathcal{C}$). $\forall$ algorithm A, $\exists c \in \mathcal{C}$ and $\exists D$ such that if A gets $m \leq d/2$ examples from D labeled by c, then*

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

In other words, if $\epsilon \leq 1/8$ and $\delta \leq 1/8$, then PAC learning is not possible with fewer than $d/2$ examples.

The outline of the proof is: To prove that there exists a concept $c \in \mathcal{C}$ and a distribution D, we are going to construct a fixed distribution D, but we do not know the exact target concept c used. Instead, we will choose c at random. If we get an expected probability of error over c, then there must exist some $c \in \mathcal{C}$ that satisfy some criteria and there is thus no need to construct c explicitly.

Proof:

As d =VC-dim($\mathcal{C}$), we can have $\overline{x}_1$,...,$\overline{x}_d$ shattered by $\mathcal{C}$. Let $\mathcal{C}' \subseteq \mathcal{C}$ with one representative from $\mathcal{C}$ for every dichotomy of $\overline{x}_1$,...,$\overline{x}_d$. Then, $|\mathcal{C}'| = 2^d$. Let $c \in \mathcal{C}'$ be chosen uniformly at random. Let distribution D be uniform over $\overline{x}_1$,...,$\overline{x}_d$.

Let us look at the following two experiments:

Experiment 1:
c is chosen at random.
S is chosen at random and labeled by c.
(c and S are chosen independent of each other.)
h_A is computed from S.
x is the test point chosen.
Consider: what is the probability of $h_A(x) \neq c(x)$?

Experiment 2:
S is chosen at random (without labels).
Random labels $c(x_i)$ assigned to $x_i \in S$.
h_A is computed from S.
x is the test point chosen.
If $x \notin S$, then label $c(x)$ at random.

Consider: what is the probability of $h_A(x) \neq c(x)$?

The above two experiments produce the same probability of $h_A(x) \neq c(x)$, 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 $\Pr_{c,S,x}[h_A(x) \neq c(x)]$.

Considering experiment 2, we have:

$$\begin{aligned}
\Pr_{c,S,x}[h_A(x) \neq c(x)] &\geq \Pr[x \notin S \wedge h_A(x) \neq c(x)] \\
&= \Pr[x \notin S] \Pr[h_A(x) \neq c(x) | x \notin S] \text{ (Definition of conditional probability)} \\
&\geq \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4},
\end{aligned}$$

where the last inequality comes from the fact that $\Pr[x \notin S] \geq \frac{1}{2}$ since there are only $m \leq d/2$ examples in set S and the distribution of x is uniform over d points; and the fact that $\Pr[h_A(x) \neq c(x) | x \notin S] = \frac{1}{2}$ since we label $c(x)$ by random guess (cf. Experiment 2).

Next, we consider marginalizing $\Pr_{c,S,x}[h_A(x) \neq c(x)]$ over c. Then $\Pr_{c,S,x}[h_A(x) \neq c(x)]$ can be written as:

$$\frac{1}{4} \leq \Pr_{c,S,x}[h_A(x) \neq c(x)] = \mathbb{E}_c[\Pr_{S,x}[h_A(x) \neq c(x) | c]] .$$

From the fact that $\mathbb{E}[x] \geq k$ implies $\exists x$ such that $x \geq k$, we know $\exists c \in \mathcal{C}' \subseteq \mathcal{C}$ such that:

$$\Pr_{S,x}[h_A(x) \neq c(x)] \geq \frac{1}{4} .$$

Observe the following chain of inequalities:

$$\begin{aligned}
&\Pr_{S,x}[h_A(x) \neq c(x)] = \mathbb{E}_S[\underbrace{\Pr_x[h_A(x) \neq c(x)]}_{err_D(h_A)}] \quad \text{(Marginalizing over } S\text{)} \\
= \; &\Pr[err(h_A) > 1/8] \underbrace{\mathbb{E}[err(h_A) | err(h_A) > 1/8]}_{\leq 1} + \underbrace{\Pr[err(h_A) \leq 1/8]}_{\leq 1} \underbrace{\mathbb{E}[err(h_A) | err(h_A) \leq 1/8]}_{\leq 1/8} \\
\leq \; &\Pr[err(h_A) > 1/8] + 1/8 .
\end{aligned}$$

Thus, combining with equation above, we establish the result:

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

The above proof illustrates an example of the randomization proof technique.

2 Inconsistent Hypothesis Model

Recall that the PAC learning model requires the hypothesis to be consistent with the training data. However, in reality, it is often difficult or impossible or unwise to choose a hypothesis that is consistent with the training data. In fact, inconsistent hypotheses are commonly seen in machine learning problems. This occurs when there is no functional relationship between the sample and the label, or when there might be a functional relationship between the two, but it may be NP-complete to find a consistent hypothesis. We summarize the above observations as follows. In practice, i.e., in more realistic learning situations, we cannot possibly require hypotheses to be consistent because of the following reasons:

- concept $c \notin \mathcal{H}$
- there might not exist the target concept
- concept $c \in \mathcal{H}$ but intractable to find.

When target concept c is not in the hypothesis space, $\forall h \in \mathcal{H}$, there exists at least one $x \in \mathbb{X}$ such that $c(x) \neq h(x)$. In case of the training set including all points of the possible instance space, there will be no consistent hypothesis for this training set. There also might not exist a target concept related with some set of data in the case that there are both ("+") and ("−") labels for the same example because of noise. In this case, there will be no consistent model. Even if there exists such a consistent model, sometimes it may be too difficult to find. Instead of looking for the complex consistent model, we try to find an inconsistent but simple one.

In short, we are going to modify the PAC learning model by not making the requirement $y = c(x)$. In the following, we state this in a more rigorous way.

Let (x, y) denote one example and its label: $x \in \mathbb{X}$ where $\mathbb{X}$ is the instance space and $y \in \{0, 1\}$. The example (x, y) is random according to some joint distribution D on $\mathbb{X} \times \{0, 1\}$. (Unlike our earlier model, the label y is also random.)

According to definition of conditional probability:

$$\Pr[x, y] = \Pr[x] \Pr[y|x].$$

Thus, we can think of x being generated according to its marginal distribution $\Pr[x]$ and then y being generated according to its conditional distribution $\Pr[y|x]$. This form is like the PAC model where the example is random with some distribution and its label is deterministic, i.e. $\Pr[y|x]$ is either 0 or 1. In this inconsistency model, we can generate x according to its marginal distribution and then generate y according to $0 \leq \Pr[y|x] \leq 1$.

The m examples from distribution D are denoted as: $S = \langle (x_1, y_1), ..., (x_m, y_m) \rangle$.
The hypothesis $h : \mathbb{X} \to \{0, 1\}$. Then the generalization error, which measures the probability of h misclassifying a random example, is defined to be:

$$err_D(h) = \Pr_{(x,y)\sim D}[h(x) \neq y] .$$

Note the error definition subsumes the one of consistency model. In consistency model, the error is defined to be:

$$err_D(h) = \Pr_{D}[h(x) \neq c(x)] .$$

The distribution D here is only over x instead of over (x, y) and there is a true label $y = c(x)$, which is related with x deterministically.

It is imperative to ask what is the optimal $h(x)$ that minimizes $err_D(h)$. If we have known the distribution D, it is easy to construct an optimal hypothesis with minimal error, i.e.

$$h_{opt}(x) = \begin{cases} 1 & \text{if } \Pr_{y|x}[y = 1|x] > 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

This hypothesis is called the *Bayes Optimal Classifier* and $err_D(h_{opt})$ is termed the *Bayes error*. Note that $err_D(h_{opt})$ provides a lower bound on the error over all hypotheses, because it captures the intrinsic uncertainties regardless of the computation power.

But, in reality, we usually do not know the conditional distribution ahead of time. Indeed, it is the goal of machine learning to approximate the true conditional distribution. In summary, our goal is to find the best hypothesis that minimizes the $err_D(h_{opt})$ over some hypothesis space, e.g., the hypothesis space can be constrained with a bounded VC dimension. The idea is to minimize this error on a sample $S = \langle (x_1, y_1), ..., (x_m, y_m) \rangle$ as illustrated in the following section.

3 Empirical Error and Expected Error

Given m examples $S = \langle (x_1, y_1), ..., (x_m, y_m) \rangle$, the empirical or training error of $h \in \mathcal{H}$ is defined as:

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

Note the empirical error can also be called empirical risk or training error. The expected empirical error is just the true error of h: $err(h)$.

Later, we will discuss results showing that

$$\forall h \in \mathcal{H}, \quad |err(h) - \hat{err}(h)| \leq \epsilon \tag{1}$$

for suitable sample size. Such a statement of the relation between empirical error and expected error must be true for all $h \in \mathcal{H}$, and thus is known as a "uniform convergence" theorem, because it is concerned about the (uniform) convergence of training error to the true error.

Such a theorem implies a nice property of empirical error. To be more precise, let $\hat{h} = \arg\min_{h \in \mathcal{H}} \hat{err}(h)$. Note $\hat{h}$ needs not be unique, but all $\hat{h}$ have the same complexity in minimizing $\hat{err}(h)$. Then, observe the following chain of inequalities, assuming (1):

$$\begin{aligned} err(\hat{h}) &\leq \hat{err}(\hat{h}) + \epsilon \quad \text{(by (1))} \\ &\leq \hat{err}(h) + \epsilon \quad \text{(since } \hat{h} \text{ minimizes } \hat{err}(h)) \\ &\leq err(h) + 2\epsilon \quad \text{(again by (1))} \end{aligned}$$

This holds for all h. 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 2ϵ. Therefore, the hypothesis that minimizes empirical error will be very close in generalization error to the best hypothesis in the class.

To prove uniform convergence, we need the following theorem. This theorem is called *Hoeffding's Inequality*, and is an example of a Chernoff bound.

Theorem 2 *Assume random variables $X_1, ..., X_m$ are i.i.d. (independent identically distributed) with bounded range. Let*

$$p = \mathbb{E}X_i \qquad X_i \in [0, 1] \qquad \hat{p} = \frac{1}{m} \sum_{i=1}^{m} X_i .$$

Then,

$$\Pr[\hat{p} \geq p + \epsilon] \leq e^{-2\epsilon^2 m} \qquad \textit{and} \qquad \Pr[\hat{p} \leq p - \epsilon] \leq e^{-2\epsilon^2 m} .$$

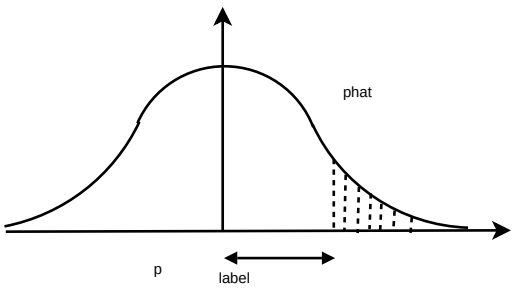

Figure 1: Illustration of concentration inequality or tail bound on $\hat{p}$.

Now, both $\Pr[\hat{p} \geq p + \epsilon]$ and $\Pr[\hat{p} \leq p - \epsilon]$ state how close the random variable $\hat{p}$ is with respect to the constant p. We will prove it in the next lecture, but we make the following remarks:

1. From *Hoeffding's Inequality*, we can derive an error ϵ, with probability $> 1 - \delta$, $|\hat{p} - p| < \epsilon$:

$$\Pr[|\hat{p}-p| \geq \epsilon] \leq 2e^{-2\epsilon^2 m} = \delta \qquad \Rightarrow \qquad |\hat{p}-p| < \epsilon \leq \sqrt{\tfrac{\ln 2/\delta}{2m}} \text{ with probability } 1-\delta.$$

2. In our learning setting, we can use Hoeffding's Inequality to prove a statment such as (1) for a single $h \in \mathcal{H}$: For fixed h, we draw m examples (x_i, y_i) independently from D. Denote

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

In other words, we get m i.i.d. random variables $X_1, ..., X_m$ and:

$$p = \mathbb{E}X_i = err(h) \qquad \hat{p} = \frac{1}{m}\sum_{i=1}^{m} X_i = \hat{err}(h)\,.$$

Thus, Hoeffding's inequality will imply that $\hat{err}(h)$ rapidly approaches $err_D(h)$.

In the next lecture, we prove a Chernoff bound to determine how fast $\hat{err}(h) \to err_D(h)$.