COS 511: Foundations of Machine Learning

Rob Schapire
Scribe: Lan Dong

Lecture #7
February 28, 2006

1 Lower Bound on Sample Complexity

Last time in class, we began to look at the lower bound on sample complexity. We discussed the intuition why lower bounds must be in terms of the target concept class $\mathcal{C}$, rather than the hypothesis class $\mathcal{H}$. We stated the theorem of lower bound and gave a wrong proof. Let's restate the theorem first and we are going to prove it this time:

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*

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

Note here the hypothesis h_A is not required to be consistent with the examples. Later in this class we will discuss why we don't require consistency model and how to deal with it.

The outline of the proof is: In order to prove there exists such a concept c and a distribution D, we are going to construct a fixed distribution D but we don't know the exact target concept c. If we get an expected probability of error over c, we know there must exist some c to satisfy some criteria and there is no need to construct the 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.

Let distribution D be uniform over $\overline{x}_1$,...,$\overline{x}_d$.

Let's look at two experiments:

Experiment1:
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 $h_A(x) \neq c(x)$?

Experiment2:
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 $h_A(x) \neq c(x)$?

It is not difficult to see after some consideration that these two experiments produce the same probability of $h_A(x) \neq c(x)$. This probability is over the randomness of concept c, the examples S and the test point x. We denote it as $Pr_{c,S,x}[h_A(x) \neq c(x)]$.

¿From the definition of experiment 2, we have:

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

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.

If we write the probability in the form of expected value over c, we have:

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

¿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:

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

We also know that:

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

Thus, combining with equation above, we have proved the result:

$$Pr[err > \frac{1}{8}] \geq \frac{1}{8} .$$

2 Introduction to Inconsistent Hypothesis Model

In the last section, we have seen a hypothesis h_A which is not required to be consistent with the training data. In fact, the inconsistent hypotheses are commonly seen in machine learning problems. In practice, we cannot generally 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 ("−x") 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 bothering to look for the complex consistent model, we try to find an inconsistent but simple one.

Now, we are going to state the problem 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 $y \in \{0, 1\}$.

(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 error is defined to be:

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

Note the error definition is different from 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 $c(x)$ related with x which is deterministic.

If we have known the distribution D, it is easy to construct an optimal hypothesis with minimal error, i.e.

$$h^*(x) = \begin{cases} 1 & \text{if } Pr_{y|x}[y = 1|x] > 1/2 \\ 0 & \text{else} \end{cases}$$

This hypothesis is called the *Bayes Optimal Classifier* and the error is called the *Bayes error*. But in real life, we usually don't know the conditional distribution and actually it is the goal of machine learning to approximate the true conditional distribution. How to find a best hypothesis that generates minimal error is the topic to be discussed in the next section.

3 Empirical Error and Expected Error

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

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

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

There is a nice theorem of the relation between empirical error and expected error. We state it here and will prove it in the next lecture.

Theorem 2 *Given m examples, assume $\mathcal{H}$ is finite, with probability $\geq 1-\delta$,*

$$\forall h \quad |err(h) - e\hat{r}r(h)| \leq \epsilon \qquad \textit{if} \quad m = O\left(\frac{\ln|\mathcal{H}| + \ln\frac{1}{\delta}}{\epsilon^2}\right)$$

The statement of the relation between empirical error and expected error is true $\forall h \in \mathcal{H}$. Therefore, we call this theorem a "uniform convergence" theorem.

This theorem implies a nice property of empirical error. To be more precise, let $\hat{h} = \arg\min_{h\in\mathcal{H}} e\hat{r}r(h)$, then:

$$\begin{aligned} err(\hat{h}) &\leq e\hat{r}r(\hat{h}) + \epsilon \\ &\leq e\hat{r}r(h) + \epsilon \\ &\leq err(h) + 2\epsilon \qquad \forall h \in \mathcal{H} \end{aligned}$$

The inequality says: choose a hypothesis $\hat{h}$ with minimal empirical error, then the true error of this hypothesis will be no bigger than the true error of any hypothesis (including *Bayes Optimal Classifier*) plus 2ϵ. Therefore, the minimal empirical error can be very close to the true minimal error.

Before we prove Theorem 2, we prove another theorem which will help to prove Theorem 2. This theorem is called *Hoeffding's Inequality*.

Theorem 3 *Assume random variables $X_1, ..., X_m$ are i.i.d. (independent identically distributed). 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,

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

We will prove it in the next lecture but here are two notes for this theorem:

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

$$Pr[|\hat{p} - p| \geq \epsilon] \leq 2e^{-2\epsilon^2 m} = \delta \qquad \Rightarrow \qquad \epsilon = \sqrt{\frac{\ln 2/\delta}{2m}}\,.$$

2. For all $h \in \mathcal{H}$, if we draw m examples (x, y) independently from D, and denote

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

I.e. we get m i.i.d. random variables $X_1, ..., X_m$ and:

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

Now we can see the intuition of using Theorem 3 to prove Theorem 2.

We can also think of the process like flipping a coin, i.e. X is a random variable with distribution:

$$X = \begin{cases} 1 & \text{w/ prob} \quad err(h) \\ 0 & \text{w/ prob} \quad 1 - err(h) \end{cases}$$

In the next lecture, we are going to prove the above two Theorems.