COS 511: Theoretical Machine Learning

Lecturer: Rob Schapire
Scribe: Di Qi

Lecture #7
February 26, 2018

1 Overview

For a hypothesis class H with $\text{VCdim}(H) = d$, what we showed in previous lectures implies that $m = O\left(\frac{1}{\epsilon}\left(\ln(1/\delta) + d\ln(1/\epsilon)\right)\right)$ is sufficient for PAC learning. In essence, this provides an upper bound on sample complexity that is linear in the VC dimension. In this lecture, we wish to prove that the VC dimension is also a lower bound on sample complexity, which provides an insight into how much data is necessary for learning under a particular model. To do this, we will consider the VC dimension of the concept class, as opposed to the hypothesis class. The intuition is that if we don't see enough points, then we have no insight into the labeling of the remaining points. We will also talk about how to generalize the PAC model to be more realistic.

2 A Lower Bound on Sample Complexity

In prior lectures we discussed the amount of data that was sufficient for learning. We introduced the concept of VC-dimension, which gives insight into the expressiveness of a hypothesis class. In the following section, let d denote the VC-dimension of the concept class C. A given concept class C has VC dimension d if the size of the largest set that the concept class can shatter is d. This means that all possible labellings can be realized on a set of size d. This suggests that even if the learner sees part of a shattered set, the labels on the remaining examples are still unpredictable. We will formally show that this is the case.

2.1 A Plausible False Argument

We begin with a plausible attempt proving a lower bound based on the VC dimension. However, the argument used turns out to be incorrect.

Theorem 1. ***(False)*** *Any algorithm that sees less than or equal to $d/2$ examples has high error under the PAC model.*

We take the role of an adversary. Let D be a uniform distribution on d points $z_1, z_2, \ldots, z_d$ that form a shattered set. Train the algorithm A on a sample S with $m \leq d/2$ examples, labeled arbitrarily, and suppose it returns the hypothesis h_A. As the adversary, choose $c \in C$ to be any concept which is consistent with the labels in S and such that $c(x) \neq h_A(x)$ for points $x \notin S$. Observe that since the algorithm predicts incorrectly on the unseen examples, and it sees at most half of the examples, we have

$$P_D(c(x) \neq h_A(x)) \geq 1/2$$

Thus, we have shown that any algorithm A has error at least $1/2$, if we only give it half of the examples during training.

The problem with this argument is that if we pick the concept after determining h_A, then the concept depends on the sample and the algorithm. This is not allowed under the PAC-learning model. Rather, the concept c must be chosen before the training set is randomly chosen.

2.2 Theorem

We now give a correct argument proving a lower bound on sample complexity.

Theorem 2. *For any algorithm A, there exists a $c \in C$ and a distribution D such that if A gets a sample of size $\leq d/2$, then*

$$P_S(err_D(h_A) > 1/8) \geq 1/8$$

This is equivalent to saying that if we want $\epsilon < 1/8$ and $\delta \leq 1/8$ then we need $m > d/2$. To show this, we choose c in a uniform way such that all possibilities in a shattered set are equally likely.

Proof. Let $C' \subseteq C$ consist of one "representative" for each labelling of the shattered set. Let $c \in C'$ be chosen uniformly at random.
Let D be uniform on the sampled set. Consider the following setups.

Experiment 1

- c is chosen at random, as above
- S is chosen at random according to D and is labeled by c
- A computes h_A from the sample S
- x, a test point, is chosen at random
- measure $P(h_A(x) \neq c(x))$

Experiment 2

- unlabeled parts of S, $x_1, x_2, \ldots, x_m$, are chosen according to D
- random labels $c(x_i)$ are computed for $x_i \in S$
- h_A is computed from the labeled S
- a test point x is chosen.
- if $x \notin S$ then the label $c(x)$ is chosen at random
- measure $P(h_A(x) \neq c(x))$

These are experiments in a different order. This is because the relevant variables are c, S, and x, and in both cases they are chosen independently of each other. Experiment 1 satisfies the requirements of the PAC model, with c chosen before the sample S is generated. But under Experiment 2, it is perhaps easier to see why the test label of x is hard to guess. As such, we prove the theorem using the setup from Experiment 2.

Since what we measure in Experiment 2 is dependent on the variables c, S, x, we denote it $P_{c,S,x}(h_A(x) \neq c(x))$. Observe that

$$\begin{aligned} P_{c,S,x}(h_A(x) \neq c(x)) &\geq P(x \notin S \wedge h_A(x) \neq c(x)) && (1) \\ &= P(x \notin S)\, P(h_A(x) \neq c(x) \,|\, x \notin S) && (2) \\ &\geq 1/2 \cdot 1/2 = 1/4 && (3) \end{aligned}$$

Step (1) follows from the fact that $x \notin S \wedge h_A(x) \neq c(x)$ logically implies $h_A(x) \neq c(x)$. Step (2) follows from the definition of conditional probability. Step (3) follows from the fact that the first term is $\geq 1/2$ since S is chosen uniformly and $m \leq d/2$, and the second term is $1/2$ since the labels are chosen at random.

We wish to show that there exists some $c \in C'$ such that $P_S(err_D(h_A) > 1/8) \geq 1/8$ where $err_D(h_A) = P_x(h_A(x) \neq c(x))$. To do this, we use marginalization, defined as $P(a) = E_x(P(a|x))$. Observe that

$$E_c(P_{S,x}(h_A(x) \neq c(x)) = P_{c,S,x}(h_A(x) \neq c(x)) \geq 1/4$$

This implies that there exists some c such that $P_{S,x}(h_A(x) \neq c(x)) \geq 1/4$. We can apply marginalization again to see

$$E_S(P_x(h_A(x) \neq c(x))) \geq 1/4$$

$P_x(h_A(x) \neq c(x)$ is the generalization error, $\text{err}(h_A)$. We have

$$\begin{aligned}
&1/4 \leq E_S(\text{err}(h_A)) \\
&= P_S(\text{err}(h_A) > 1/8) \cdot E(\text{err}(h_A) \,|\, \text{err}(h_A) > 1/8) + P_S(\text{err}(h_A) \leq 1/8) \cdot E(\text{err}(h_A) \,|\, \text{err}(h_A) \leq 1/8) \\
&\leq P_S(\text{err}(h_A) > 1/8) \cdot 1 + P_S(\text{err}(h_A) \leq 1/8) \cdot (1/8) \\
&\leq P_S(\text{err}(h_A) > 1/8) + 1/8
\end{aligned}$$

Thus, we have shown that $P_S(\text{err}(h_A) > 1/8) \geq 1/8$, i.e., that the generalization error is at least $1/8$ with probability at least $1/8$.

□

3 Generalizing the PAC Model

Often it is not possible to find hypotheses that are consistent with the training set. It may be that the true concept is not in the hypothesis class, that it is computationally difficult to find a consistent relationship, or that no functional relationship exists between the instances and their labels. In the last case the relationship may be fundamentally probabilistic, as in the case of the weather. This motivates adjusting the PAC learning framework.

Old Framework	New Framework
• observe $x, c(x)$ where $x \in X$ and $c \in C$	• observe x, y where $(x, y) \sim D$
• D on X	• D on $\mathbb{X} \times \{0, 1\}$
• $\text{err}_D(h) = P_{x \sim D}(h(x) \neq c(x))$	• $\text{err}_D(h) = P_{(x,y) \sim D}(h(x) \neq y)$

In the new framework, we now have $(x, y) \sim D$, where D is a distribution over $X \times \{0, 1\}$. To contrast the frameworks, observe that

$$P(x, y) = P(x) \cdot P(y \,|\, x)$$

We note here that the probability distribution is with respect to D. We can think of (x, y) being generated as a pair. On the right hand-side, we can think of x as being generated first, and then the label y being generated probabilistically, dependent on x. Before, we had $P(y = 1 \,|\, x) = 0$ or 1. In the new framework, it can take any value in $[0, 1]$. To handle this new distribution of data, we needed to generalize the notion of error.

3.1 Bayes Optimal Hypothesis

To better understand the error under the new framework, consider the question: if we don't restrict h, how small can the generalization error be?

As a toy example, consider flipping a coin that lands heads with probability p and tails with probability $1-p$. If we wanted to guess the outcome, the optimal prediction would be

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

If $p = 1/2$, choosing either heads or tails would be optimal. Choosing deterministically is optimal because there is nothing to be gained from choosing randomly on our part. This suggests that in general, when assigning classifications through a hypothesis, the optimal hypothesis h_{opt} would be

$$h_{\text{opt}}(x) = \begin{cases} 1 & \text{if } P(y = 1 \mid x) > 1/2 \\ 0 & \text{if } P(y = 1 \mid x) < 1/2 \end{cases}$$

This is known as the "Bayes optimal classifier" or "Bayes optimal decision rule." The "Bayes error" is the theoretical minimum error we can achieve

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

3.2 PAC Learning

The Bayes error is helpful when it comes to understanding the new model. However, usually people are not aiming to find the Bayes error, but rather simply the best hypothesis in a given hypothesis class H, $\min_{h \in H} \text{err}_D(h)$.

Consider a sample $S = \langle (x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m) \rangle$ where $(x_i, y_i) \sim D$. We can define the empirical error of a hypothesis $h \in H$ as

$$\hat{\text{err}}(h) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\{h(x_i) \neq y_i\}$$

and the optimal hypothesis over the sample as

$$\hat{h} = \arg\min_{h \in H} \hat{\text{err}}(h)$$

Suppose we could prove that the empirical error of h is approximately equal to its true error, namely that

$$\forall h \in H : \qquad |\hat{\text{err}}(h) - \text{err}(h)| \leq \epsilon$$

This would be sufficient to prove that $\hat{h}$ has low generalization error:

$$\begin{aligned} \text{err}(\hat{h}) &\leq \hat{\text{err}}(\hat{h}) + \epsilon && \text{by the assumption} \\ &\leq \hat{\text{err}}(h) + \epsilon && \forall h, \text{ since } \hat{h} = \arg\min_{h \in H} \hat{\text{err}}(h) \\ &\leq \text{err}(h) + 2\epsilon && \text{by the assumption} \end{aligned}$$

Since this is true for all $h \in H$, it is also true for the one with minimum generalization error. Therefore, $\hat{h}$ will have generalization error that is within 2ϵ of the best hypothesis in H.