EECS 598-005: Theoretical Foundations of Machine Learning **Fall 2015**

Lecture 6: Risk and PAC-Learnable Class

Lecturer: Jacob Abernethy *Scribes: Sentao Miao*
Editors: Luke Brandl, Nghia Nguyen, and Shuang Qiu

6.1 Error Bound and Sample Size Requirement

Example 1. Say you receive a sequence random variables $X_1, ..., X_n$ known to be sampled i.i.d. from one of the two distributions:

- $\mathcal{D}_0$: $X_i = 0$ always.
- $\mathcal{D}_1$: $X_i = 1$ with probability ϵ and $X_i = 0$ with probability $1 - \epsilon$.

and our goal is to determine the true underlying distribution. Consider the algorithm $\mathcal{A}$ whose output is:

$$\mathcal{A}(X_1, \ldots, X_n) = \begin{cases} \mathcal{D}_0 & \text{if } \sum_{i=1}^n X_i = 0 \\ \mathcal{D}_1 & \text{if } \sum_{i=1}^n X_i > 0 \end{cases}$$

If the underlying distribution is $\mathcal{D}_0$, this algorithm always returns the correct result; however, if the underlying distribution is $\mathcal{D}_1$, the algorithm may get unlucky and not recieve any samples of value 1. We can bound this probability:

$$\mathbb{P}\left[\sum_{i=1}^n X_i = 0 \middle| X_i \text{ sampled from } \mathcal{D}_1\right] = (1-\epsilon)^n \leq e^{-\epsilon n}.$$

Therefore, in order to guarantee that the mistake happens with probability at most δ, we need the sample size n to satisfy:

$$n \geq \frac{1}{\epsilon} \log(\frac{1}{\delta}).$$

Example 2. Now consider the case where we know that the i.i.d. samples follow one of the two distributions:

- $\mathcal{D}_f$: $X_i = 1$ w.p. $\frac{1}{2}$ and $X_i = 0$ w.p. $\frac{1}{2}$ (fair coin).
- $\mathcal{D}_b$: $X_i = 1$ w.p. $\frac{1}{2} + \epsilon$ and $X_i = 0$ w.p. $\frac{1}{2} - \epsilon$ (biased coin).

Similarly to the previous example, consider the following simple algorithm:

$$\mathcal{A}(X_1, \ldots, X_n) = \begin{cases} \mathcal{D}_b & \text{if } \sum_{i=1}^n X_i \geq \left(\frac{1+\epsilon}{2}\right) n \\ \mathcal{D}_f & \text{otherwise.} \end{cases}$$

We can apply Hoeffding's concentration inequality on the probability that the algorithm predicts $\mathcal{D}_b$ while the true distribution is $\mathcal{D}_f$. Similarly, we upper bound the other type of error. It is an easy exercise to show that the sample size has to be $n \geq \Theta(\frac{\log \frac{1}{\delta}}{\epsilon^2})$ in order to bound the error probability with δ.

Remark. In both of the examples above, we want to identify the ϵ gap in the distribution. Interestingly, the sample size requirement scales in $\frac{1}{\epsilon}$ for Example 1, while it scales in $\frac{1}{\epsilon^2}$ for Example 2.

The above two examples give us a question: what class of problems can we learn? How big should the sample size should be? These questions lead us to the so-called **PAC-Learnable class**, to be defined later. Let us first introduce the concept of **risk** and **empirical risk**.

6.2 Risk and Empirical Risk

Consider that we have a input space X, output space Y, and concept class C which is a set of functions $h : X \longrightarrow Y$. A learning instance includes a distribution $D \in \Delta(X)$ (a distribution on X) and a target concept $c \in C$. We then have the following definitions.

Definition 6.1 (True Risk)**.** *True risk (or generalized error) of a concept $h : X \longrightarrow Y$ is defined as:*

$$R(h) := E_{x\sim D}[1(h(x) \neq c(x))] = \mathbb{P}_{x\sim D}(h(x) \neq c(x)).$$

Definition 6.2 (Empirical Risk)**.** *The empirical risk of h for a sample S drawn from X with respect to distribution D is defined as:*

$$\hat{R}_n(h) := \frac{1}{n}\sum_{x\in S} 1[h(x) \neq c(x)]$$

where $n = |S|$.

Remark 6.3. *If S is drawn from X with respect to to D, an easy check gives us $E[\hat{R}_n(h)] = R(h)$. Hence, the empirical risk $\hat{R}_n(h)$ is indeed an unbiased estimator.*

6.3 PAC-Learnable Class

Now we can define the so-called PAC-Learnable Class.

Definition 6.4 (PAC-Learnable)**.** *A class C is said to be PAC-Learnable if there exists an algorithm $\mathcal{A}$ and a polynomial of two variables (denoted as* $\text{poly}(a, b)$*) such that $\forall \epsilon, \delta > 0$, $\forall$ distributions $D \in \Delta(X)$, and for any target concept $c \in C$, we have:*

$$\mathbb{P}_{S\sim D,|S|=n}(R(h_S) \leq \epsilon) \geq 1 - \delta$$

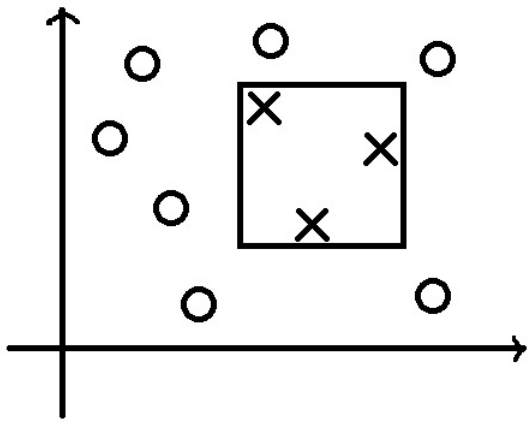
as long as the sample size n is at least $\text{poly}(\frac{1}{\epsilon}, \frac{1}{\delta})$*. Here, h_S is the output of the algorithm $\mathcal{A}$ on sample S.*

We shall have another example which is PAC-Learnable to illustrate this definition.

6.3.1 Learning Axis-Aligned Rectangles

The concept class $\mathcal{C}$ is defined as $\mathcal{C} := \{[a_1, a_2] \times [b_1, b_2] \subset \mathbb{R}^2 : \forall a_1 \leq a_2, b_1 \leq b_2 \in \mathbb{R}\}$. Note that concepts are now being viewed as rectangular subsets of $\mathbb{R}^2$; that is, points within the subset have label 1, and those outside have label 0.

Consider the following figure.

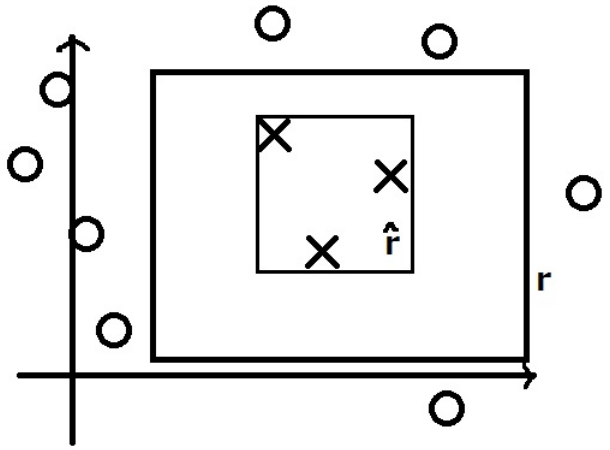


Here we have points on the x/y plane, where points drawn with a $\times$ have a label of 1, and points drawn with a $\circ$ have label 0.

Consequently our goal is to decide which area is C. Let us analyze an algorithm $\mathcal{A}(S)$ which we will call the "smallest enclosed rectangle algorithm" which given sample S drawn drawn randomly from D, $\mathcal{A}(S)$ outputs the rectangle $\hat{r} = r_S := [\hat{a_1}, \hat{a_2}] \times [\hat{b_1}, \hat{b_2}]$ such that:

- $\hat{a_1} = \min\{x$ **coordinate of input labelled** $1\}$.
- $\hat{a_2} = \max\{x$ **coordinate of input labelled** $1\}$.
- $\hat{b_1} = \min\{y$ **coordinate of input labelled** $1\}$.
- $\hat{b_2} = \max\{y$ **coordinate of input labelled** $1\}$.

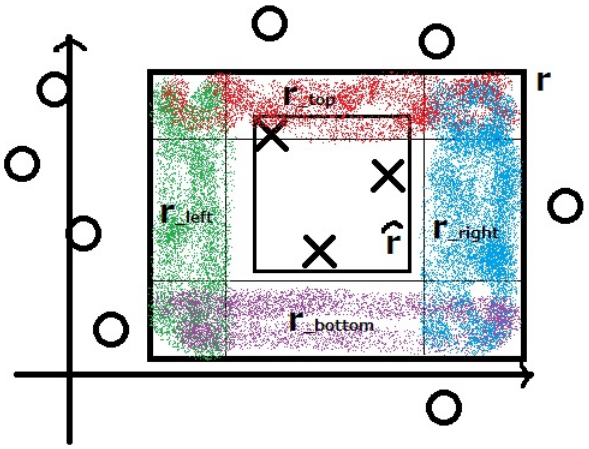
How likely is this algorithm to produce an error larger than epsilon? To work toward a bound on this problem, take a look at the following figure.



We see that the real C (defined here as the rectangle r) is larger than our estimate $\hat{r}$. We are interested in the chance of making mistakes, i.e. $\mathbb{P}_{S\sim D}(R(\hat{r}) > \epsilon)$ for some ϵ.

Note that if $\mathbb{P}_{x\sim D}(x \in r) < \epsilon$, i.e. the probability of x falling to rectangle r is smaller than ϵ, $\mathbb{P}_{S\sim D}(R(\hat{r}) > \epsilon)$ is definitely 0 because $\hat{r}$ cannot be bigger than r.

We shall consider the case that $\mathbb{P}_{x\sim D}(x \in r) \geq \epsilon$. To compute $\mathbb{P}_{S\sim D}(R(\hat{r}) > \epsilon)$, we define the following four regions $r_{bottom}, r_{top}, r_{left}, r_{right}$. r_{bottom} is defined as starting from bottom of r with height chosen such that $\mathbb{P}(x \in r_{bottom}) = \frac{\epsilon}{4}$. The others are similarly defined. The following figure illustrates this construction.



Claim: If $R(\hat{r}) > \epsilon$ then there must be some $i \in \{top, bottom, left, right\}$ for which $r_i \cap \hat{r} = \emptyset$.

Proof: We will prove the contrapositive statement, that if $\hat{r}$ intersects with all four regions (top,bottom,left,right), then the error rate of $\hat{r}$ cannot exceed ϵ. Notice that if the output of our algorithm, the smallest enclosed rectangle is incorrect for some example, i.e. the smallest rectangle $\hat{r}$ disagrees with the true r, then this example must lie in the "error region" $r \setminus \hat{r}$. Indeed, the true risk of $\hat{r}$ is equal to the probability mass in this region; that is, $R(\hat{r}) = \mathbb{P}_{x \sim D}(x \in r \setminus \hat{r})$. However, each region has probability $\frac{\epsilon}{4}$ so the union of the regions has probability no more than ϵ. On the other hand, if each of the regions has a non-empty intersection with $\hat{r}$, then it is clear from the picture that the union of the regions covers $r \setminus \hat{r}$. So if all four regions has a non-empty intersection with $\hat{r}$ we can conclude that $R(\hat{r}) = \mathbb{P}_{x \sim D}(x \in r \setminus \hat{r}) \leq \mathbb{P}_{x \sim D}(x \in r_i$ for some $i =$top,bottom,left, or right$) \leq \epsilon$ as desired. ■

Therefore, we can compute $\mathbb{P}_{S \sim D}(R(\hat{r}) > \epsilon)$ as the following:

$$\begin{aligned}
\mathbb{P}_{S \sim D}(R(\hat{h}) > \epsilon) &\leq \mathbb{P}(\exists i \in \{top, bottom, left, right\} : \hat{r} \cap r_i = \emptyset) \\
&\leq \sum_i \mathbb{P}(\hat{r} \cap r_i = \emptyset) \\
&\leq \sum_i (1 - \frac{\epsilon}{4})^n \\
&= 4(1 - \frac{\epsilon}{4})^n \\
&\leq 4e^{-\frac{\epsilon n}{4}}
\end{aligned}$$

From this result, we see that if we want this probability to be at most δ, then choosing $n \geq \frac{4}{\epsilon} \log(\frac{4}{\delta})$ suffices. Consequently, this shows it is PAC-Learnable via the smallest enclosed rectangle algorithm.