Advanced Machine Learning: Problem Set I Solutions

Problem 1: Prove that under realizability assumption the class of all axis parallel rectangles in two dimension is PAC learnable. An axis parallel rectangle labels all points lying on or inside it as positive class and all other points as negative class. Derive an expression for sample complexity for learning this class.

Proof:

Some notations and points:

- Let X is the set of all possible examples or instances.
- A concept $c : X \implies Y$ is a mapping from X to Y.
- A concept class is a set of concepts we may wish to learn and is denoted by C.
- We assume that examples are independently and identically distributed (i.i.d.) according to some fixed but unknown distribution D.
- Learner receives a sample $S = (x_1, ..., x_m)$ drawn i.i.d. according to D as well as the label $(c(x_1), ..., c(x_m))$,which are based on a specific target concept $c \in C$ to learn.
- Our task is to use the labeled sample S to select a hypothesis $h_S \in H$ that has a small generalization error with respect to the concept c. The generalization error of a hypothesis $h \in H$, also referred to as the true error or just error of h is denoted by $R(h)$.

A concept class C is said to be PAC-learnable if there exists an algorithm A and a polynomial function $poly(\cdot, \cdot, \cdot, \cdot)$ such that for any $\epsilon > 0$ and $\delta > 0$, for all distributions D on X and for any target concept $c \in C$, the following holds for any sample size $m \geq poly(1/\epsilon, 1/\delta, n, size(c))$:

$$\underset{S \sim D^m}{Pr}[R(h_S) > \epsilon] \leq 1 - \delta$$

As can be seen from the figure 1, R represents a target axis-aligned rectangle and R' a hypothesis. The error regions of R' are formed by the area within the rectangle R but outside the rectangle R' and the area within R' but outside the rectangle R. The first area corresponds to false negatives, that is, points that are labeled as 0 or negatively by R', which are in fact positive or labeled with 1. The second area corresponds to false positives, that is, points labeled positively by R' which are in fact negatively labeled.

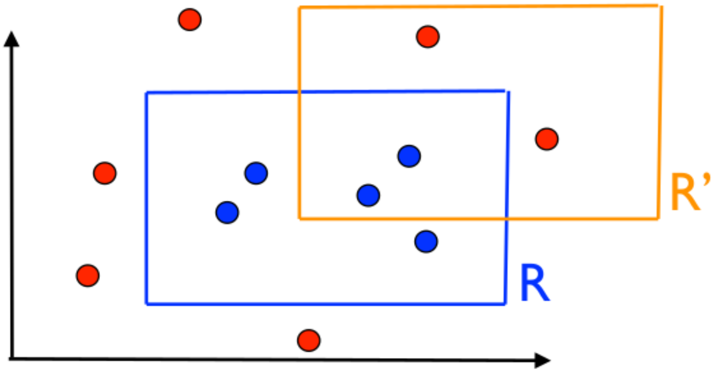

Figure 1: Target concept R and possible hypothesis R'. Circles represent training instances. A blue circle is a point labeled wit 1, since it falls within the rectangle R. Others are red and labeled with 0.

To show that the concept class is PAC-learnable, we describe a simple PAC-learning algorithm A. Given a labeled sample S, the algorithm consists of returning the tightest axis-aligned rectangle $R' = R_S$ containing the points labeled with 1. Figure 2 illustrates the hypothesis returned by the algorithm. By definition, R_S does not produce any false positive, since its points must be included in the target concept R. Thus, the error region of R_S is included in R.

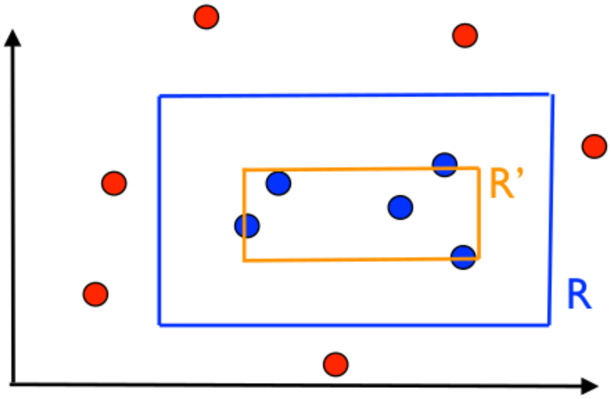

Figure 2: Illustration of the hypothesis $R' = R_S$ returned by the algorithm.

Let $R \in C$ be a target concept. Fix $\epsilon > 0$. Let $\Pr[R_S]$ denote the probability mass of the region defined by R_S, that is the probability that a point randomly drawn according to D falls with in R_S. Since errors made by our algorithm can be due only to points falling inside R_S, we can assume that $Pr[R_S] > \epsilon$; otherwise, the error of R_S is less than or equal to ϵ

regardless of the training sample S received.

Now, since $Pr[R_S] > \epsilon$, we can define four rectangular regions r_1, r_2, r_3, and r_4 along the sides of R_S, each with probability at least $\epsilon/4$. These regions can be constructed by starting with the empty rectangle along a side and increasing its size until its distribution mass is at least $\epsilon/4$. Figure 3 illustrates the definition of these regions.

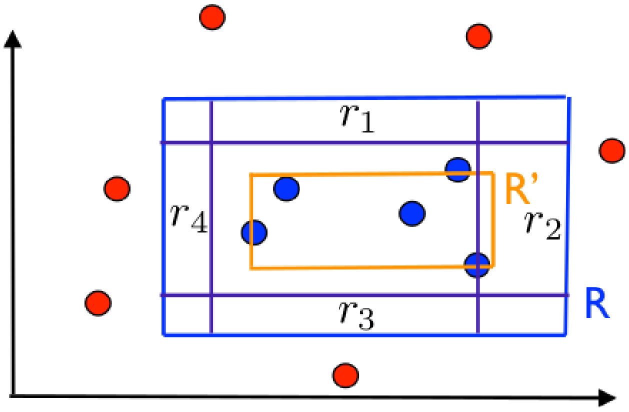

Figure 3: Illustration of the regions $r_1, ..., r_4$.

Observe that if R_S meets all of these four regions, then, because it is a rectangle,it will have one side in each of these four regions (geometric argument). Its error area, which is the part of R that it does not cover, is thus included in these regions and cannot have probability mass more than ϵ. By contraposition, if $R(R_S) > \epsilon$,then R_S must miss at least one of the regions $r_i, i \in [1, 4]$. As a result, we can write

$$\begin{aligned}
\underset{S\sim D^m}{Pr}[R(R_S) > \epsilon] &\leq \underset{S\sim D^m}{Pr}[\cup_{i=1}^{4}\{R_s \cap r_i = \phi\}] \\
&\leq \sum_{i=1}^{4} \underset{S\sim D^m}{Pr}[\{R_s \cap r_i = \phi\}] \\
&\leq 4(1 - \frac{\epsilon}{4})^m & \text{(since } Pr[r_i] > \frac{\epsilon}{4}) \\
&\leq 4exp(-m\epsilon/4) & \text{(since } 1 - x \leq e^{-x} \text{)}
\end{aligned}$$

For any $\delta > 0$, to ensure that $\underset{S\sim D^m}{Pr}[R(R_S) > \epsilon] \leq \delta$,we can impose

$$4exp(-m\epsilon/4) \leq \delta \iff m \geq \tfrac{4}{\epsilon}\log\tfrac{4}{\delta}$$

Thus, for any $\epsilon > 0$ and $\delta > 0$, if the sample size m is greater than $\frac{4}{\epsilon} \log \frac{4}{\delta}$, then $\underset{S \sim D^m}{Pr}[R(R_S) > \epsilon] \leq 1 - \delta$. Furthermore, the computational cost of the representation of points in R^2 and axis-aligned rectangles, which can be defined by their four corners, is constant. This proves that the concept class of axis-aligned rectangles is PAC-learnable and that the sample complexity of PAC-learning axis-aligned rectangles is in $\mathcal{O}(\frac{1}{\epsilon} log \frac{1}{\delta})$.

Problem 2: Prove that the class of conjunctions of Boolean literals of at most n variables is PAC learnable. A literal is either a Boolean variable x_i or its negation $\neg x_i$. For example, $x_1 \wedge \neg x_2 \wedge x_3$ is a conjunction of literals. All Boolean vectors which evaluates to TRUE by the conjunction belong to positive class and the rest belong to negative class.

Proof: Consider the class C of target concepts described by conjunctions of boolean literals. We need to show that any consistent learner will require only a polynomial number of training examples (say m) to learn any $c \in C$. For that, we will use theorem which states :

Let H be a finite set of functions mapping from $X \implies Y$. Let A be an algorithm that for any target concept $c \in H$ and i.i.d. sample S returns a consistent hypothesis $h_S : \widehat{R}(h_S) = 0$. Then, for any $\epsilon, \delta > 0$, the inequality $\underset{S \sim D^m}{Pr}[R(h_S) > \epsilon] \leq 1 - \delta$ holds if

$$m \geq \tfrac{1}{\epsilon}(\log |H| + \log \tfrac{1}{\delta})$$

Suppose H contains conjunctions of constraints on up to n boolean attributes (i.e., n boolean literals). Then $|H| = 3^n$, and

$$m \geq \frac{1}{\epsilon}(\log 3^n + \log \frac{1}{\delta})$$

$$m \geq \frac{1}{\epsilon}(n \log 3 + \log \frac{1}{\delta})$$

So, m grows linearly or below (i.e., polynomially). So it is PAC-learnable.