Assignment 2 Solutions

COMP 599 Autumn 2020 McGill University

Question 1.[15 points] An axis-aligned hyper-rectangle in $\mathbb{R}^n$ is a set of the form $[a_1, b_1] \times \ldots \times [a_n, b_n] \subseteq \mathbb{R}^n$. Show that the hypothesis class of hyper-rectangles is efficiently PAC-learnable (in the realizable case) by generalizing the reasoning in the class notes for the case $n = 2$.

Solution:

The algorithm selects the smallest axis-aligned hyper-rectangle R' containing all the positive sample points. We bound the error as follows.

Let $R = [a_1, b_1] \times \ldots \times [a_n, b_n]$ be a target concept and let D be a distribution on $\mathbb{R}^n$. Assume that $\Pr_D(R) \geq \varepsilon$, otherwise the error is trivially at most ε. For $i \in \{1, \ldots, n\}$, define

$$E_{2i} = [a_1, b_1] \times \ldots \times [a_{i-1}, b_{i-1}] \times [a_i, b'] \times [a_{i+1}, b_{i+1}] \times \ldots \times [a_n, b_n]$$

and

$$E_{2i+1} = [a_1, b_1] \times \ldots \times [a_{i-1}, b_{i-1}] \times [a', b_i] \times [a_{i+1}, b_{i+1}] \times \ldots \times [a_n, b_n]$$

such that $\Pr_D(E_i) = \varepsilon/2n$ for $i = 1, \ldots 2n$.

If $err(R') = \Pr_D(R \setminus R') > \varepsilon$ then R' must miss one of the regions E_i. The probability that none of the m sample points falls in the region E_i is at most $(1 - \varepsilon/2n)^m$. By a union bound, this shows that

$$\Pr[err(R') > \varepsilon] \leq 2n(1 - \varepsilon/2n)^m \leq 2n \exp\left(-\frac{\varepsilon m}{2n}\right).$$

It follows that if

$$m \geq \frac{2n}{\varepsilon} \ln\left(\frac{2n}{\delta}\right)$$

then $err(R') \leq \varepsilon$ with probability at least $1 - \delta$.

Question 2.[25 points]

(a) Let D be a distribution on $\mathbb{R}$ and (b, c) an interval with $\Pr_D((b, c)) > \varepsilon$ for some $\varepsilon > 0$. Show that the probability that m points drawn i.i.d. from D all fall outside (b, c) is at most $e^{-m\varepsilon}$.

(b) Show that the hypothesis class formed by the collection of unions of two closed bounded intervals of reals of the form $[a, b] \cup [c, d]$, with $a \leq b \leq c \leq d$, is efficiently PAC learnable in the realizable case.

Solution:

1. Consider an open interval (c, d) such that $\Pr_D((c, d) > \varepsilon$. Then the probability for m points chosen according to distribution D to avoid (c, d) is a most $(1 - \varepsilon)^m \leq e^{-m\varepsilon}$.

2. If the positive examples form a single contiguous sequence, then return the smallest interval containing the positive examples. Otherwise, the positive examples form two blocks, with some negative examples inbetween. In this case return the union of the two smallest intervals containing each block of positive examples.

   Let $[a, b] \cup [c, d]$ be the target concept and let $\varepsilon > 0$. We consider the case that $\Pr_D([a, b]) > \varepsilon/3$, $\Pr_D([c, d]) > \varepsilon/3$, and $\Pr_D((b, c) > \varepsilon/3$; the other cases can handled equally straightforwardly.

   Define "error regions" $E_1 := [a, a']$, $E_2 := [b', b]$, $E_3 := (b, c)$, $E_4 := [c, c']$, $E_5 := [d', d]$, where $E_1, E_2 \subseteq [a, b]$ and $E_3, E_4 \subseteq [c, d]$, such that $\Pr_D(E_i) = \varepsilon/5$ for $i = 1, \ldots, 5$. Observe that if the sample S contains points in each E_i then the hypothesis has error at most ε.

   The probability for a sample of size m to miss E_1 is a most $e^{-m\varepsilon/5}$. By a union bound, the probabiity to miss some E_i is at most $5e^{-m\varepsilon/5}$. This is a most δ if $m \geq \frac{5}{\varepsilon} \ln\left(\frac{5}{\delta}\right)$.

Question 3.[20 points]

The two-distribution model is defined as follows. Let H be a hypothesis class on domain X and and let $c \in H$ be the target concept. The learning algorithm $\mathcal{A}$ requests *separately* a set of m_+ positive examples and a set of m_- negative examples. The positive examples are drawn according to some distribution D_+ on $\{x \in X : c(x) = 1\}$ and the negative examples are drawn according to some fixed distribution D_- on the set $\{x \in X : c(x) = 0\}$. If either of the above sets is empty the learning algorithm receives the empty set in response to its request.

Fixing the accuracy ε and confidence δ the learning algorithm must output a hypothesis h such that with probability at least $1 - \delta$:

$$\Pr_{x \sim D_+}[h(x) = 0] \leq \varepsilon \text{ and } \Pr_{x \sim D_-}[h(x) = 1] \leq \varepsilon.$$

We say that $\mathcal{A}$ PAC-learns H in the two-distribution model if the required sample sizes are bounded by a polynomial in $1/\delta$ and $1/\varepsilon$.

Show that if the hypothesis class H is efficiently PAC-learnable in the standard (one-distribution) model, then it is also efficiently PAC learnable in the two-distribution model.

Solution:

We must show that learnability in the standard model implies learnability in the two-distribution model. Suppose $\mathcal{A}$ learns H in the standard model. We construct an algorithm $\mathcal{B}$ that learns H in the two-distribution model.

Let $c \in H$ be a target concept and assume that $\mathcal{B}$ has access to positive examples according to distribution D_+ and negative examples according to distribution D_-. Let $\varepsilon, \delta > 0$ be the desired accuracy and confidence.

We supply examples to $\mathcal{A}$ according to the distribution $D := \frac{1}{2}(D_- + D_+)$ and labelled by c. Given sufficiently many examples, with probability at least $1-\delta$ $\mathcal{A}$ outputs a hypothesis h such that $\Pr_{x\sim D}(h(x) \neq c(x)) \leq \varepsilon/2$. We take this h to be the hypothesis output by $\mathcal{B}$. Then

$$\Pr_{x\sim D}(h(x) \neq c(x)) = \frac{1}{2}\Pr_{x\sim D_-}(h(x)=1) + \frac{1}{2}\Pr_{x\sim D_+}(h(x)=0) \leq \varepsilon/2\,.$$

In this case $\Pr_{x\sim D_-}(h(x)=1)$ and $\Pr_{x\sim D_+}(h(x)=0)$ are both at most ε.

Question 4.[20 points]

1. For each fixed k, what is the VC dimension of the class of subsets of the real line expressible as the union of k or fewer closed intervals? Justify your answer.
2. Prove that the class of hyper-rectangles in $\mathbb{R}^n$, of the form $[a_1, b_1] \times \ldots \times [a_n, b_n]$, has VC dimension $2n$.

Solution.

1. The VC dimension is $2k$. In fact, any set S of size $2k$ can be shattered since any dichotomy on such a set, seen as a sequence in $\{-,+\}^{2k}$, comprises at most k contiguous blocks of positively labelled points, and so can be realised by a most k intervals. However no subset of size $2k+1$ can be shattered since the dichotomy $+-+-\ldots+-+$ is not realisable.
2. Let $\boldsymbol{e}_i \in \mathbb{R}^n$ denote the i-the coordinate vector. Then the set $S = \{\boldsymbol{e}_i, -\boldsymbol{e}_i : i = 1, \ldots, n\}$ can be shattered by the class of hyper-rectangles. An arbitrary subset $T \subseteq S$ can be realised by the rectangle $[a_1, b_1] \times \ldots \times [a_n, b_n]$, where

$$a_i = \begin{cases} -1 \text{ if } -\boldsymbol{e}_i \in S \\ 0 \text{ otherwise} \end{cases} \qquad b_i = \begin{cases} 1 \text{ if } \boldsymbol{e}_i \in S \\ 0 \text{ otherwise} \end{cases}$$

   However no set S of cardinality $2n+1$ can be shattered. Given such a set S, there is a subset $T = \{\boldsymbol{a}^{(i)}, \boldsymbol{b}^{(i)} : i = 1, \ldots, n\} \subseteq S$ of cardinality at most $2n$, such that $a_i^{(i)} \leq x_i \leq b_i^{(i)}$ for all $\boldsymbol{x} \in S$. Then any hyper-rectangle that contains all the elements of must include all of S.