Homework 1

Yunlin Zhang N17583629

2015/10/15

1 Problem A. PAC learning

1.1

Consider the following algorithm to learning the tightest interval that contains all the points that are labeled 1, given m total points in the sample:

Initiate the min of the interval as max double, and max of the interval as min double

Iterate through all points, if a point is labeled 1 and less than min or greater than the max, then update the min and max of the interval to the new value Return the resulting interval

Total time complexity is $O(m)$ to iterate through all points once.

Let this learned interval be $R_S = [a', b']$, whereas the target interval $R = [a, b]$.

R contains all points labeled 1, and R_S is the tightest interval that contains all points labeled 1 $\Rightarrow R_S \subseteq R$ and therefore, it is not possible to have false positives

Defining $Pr[\star]$ the same way as in textbook/class as the probability of drawing randomly from the distribution D and landing in $\star$

For a fixed ϵ, if $Pr[R] < \epsilon \Rightarrow Pr[R_S] \leq Pr[R] < \epsilon$

For the case of $Pr[R] \geq \epsilon$, define 2 intervals $r_1 = [a, c]$ and $r_2 = [d, b]$ s.t.

$Pr[r_1] = \epsilon/2$ and $Pr[r_2] = \epsilon/2$

Therefore if R_S intersect both r_1 and $r_2 \Rightarrow$

$$\begin{aligned}
\Re[R_S] = Pr[R] - Pr[R_S] &= Pr[r_1] + Pr[r_2] - Pr[R_S \cap r_1] - Pr[R_S \cap r_2] \\
&= Pr[r_1] + Pr[r_2] - Pr[R_S \cap r_1] - Pr[R_S \cap r_2] \\
&= \epsilon - Pr[R_S \cap r_1] - Pr[R_S \cap r_2] \leq \epsilon
\end{aligned}$$

$\Rightarrow R_S$ must miss either r_1 or r_2 if $\Re[R_S] > \epsilon$

$\Rightarrow Pr_{S \sim D^m}[\Re[R_S] > \epsilon] \leq Pr_{S \sim D^m}[\cup_{i=1}^{2}\{R_S \cap r_i = \emptyset\}]$

$\leq \sum_{i=1}^{2} Pr_{S \sim D^m}[\{R_S \cap r_i = \emptyset\}]$ (by union bound)

$\leq 2(1-\epsilon/2)^m$ (each region has $Pr[r_i] = \epsilon/2$)

$\leq 2exp(-m\epsilon/2)$ $((1-x)^y \leq exp(-xy))$

Setting r.h.s $\leq \delta$ gives

$2exp(-m\epsilon/2) \leq \delta$

$\Rightarrow m \geq (2/\epsilon)\log(2/\delta)$

Therefore, $\forall \epsilon > 0, \delta > 0$, if $m \geq (2/\epsilon)\log(2/\delta) \Rightarrow Pr_{S \sim D^m}[\Re[R_S] > \epsilon] \leq \delta$
And the algorithm of finding the tightest interval that contains all points labeled 1 is PAC learnable with data complexity of $m \geq (2/\epsilon)\log(2/\delta) \sim O((1/\epsilon)\log(1/\delta))$

1.2

Consider the following algorithm of finding the set of continuous intervals:

Initiate current region to 0, initiate previous point to null

Sort all points, and iterate through the sorted list

If the current region is 0, and current point is 1 then set region 1, current point is stored as the min of current interval

If the current region is 1, and current point is 1 then set store current point as previous point

If the current region is 1, and current point is 0, then set previous point as the max of the interval, and reset interval to 0

Return all intervals found this way.

The time complexity is $O(m\log m)$ to sort, and $O(m)$ to iterate, therefore the total time complexity is $O(m\log m)$

The proof of PAC learnability of this algorithm will be constructed from the different scenarios:

1. The two intervals are not disjoint
2. The two intervals are disjoint

Note that based on the construction of these cases, it is possible to admit false positives in case 2 when the sample does not contain points in between the two target intervals. Without loss of generality, let $a < d$

Case 1: $b \geq c$: the regions are not disjoint, and therefore this case reduces to a single interval. The algorithm will also give the tightest region containing all points labeled with 1. Using logic similar to the previous problem, proves that this case is PAC learnable with data complexity of $m \geq (2/\epsilon)\log(2/\delta) \sim O((1/\epsilon)\log(1/\delta))$

Case 2: $b < c$: the regions are disjoint and the algorithm can return the following two outputs depending on the sample

1. Two tightest intervals is there are points sampled from the interval (b, c)

2. The tightest interval that is the union of the two target intervals and (b, c) if no points are sampled from (b, c)

Again, the proof is constructed based on subregions (r_i) on the left and right sides of the intervals, but we need to find optimal size for these intervals.

Consider subregion on the left and right of R_1 and R_2, call them r_{ij} where subscript i corresponds to the region it is a subset of, and subscript j corresponds to $l = left$ or $r = right$. We would like to find a t such that $Pr[r_{ij}] = t\epsilon$ and for $\Re[R_S] > \epsilon$, R_S must either miss at least one of each of these such regions or it contains the false position region between the two target intervals (i.e. none of the sampled points labeled 0 is between the target intervals)

Let the region between the two target intervals be called R', and let $Pr[R'] = u\epsilon$ for some yet to be determined u

2 Problem B. Rademacher complexity, growth function

2.1

Consider the family of threshold functions $F = \{f : x \mapsto 1_{x>\theta} : \theta \in \mathcal{R}\}$

This function will map all $x \leq \theta$ to 0 and all $x > \theta$ to 1

For a set of m points, if $x_i/neqx_j$, $i \neq j$ then there are $m-1$ intervals between consecutive points when sorted, plus the two intervals to the left and right of all sampled points. For all values of θ in each such interval, the classification would be the same. If there are identical points in the sample, then the possible number of classifications is reduced by the number of such duplicated.

Therefore the maximum number of classification is $m+1$ for F

Similarly, for $G = \{g : x \mapsto 1_{x\leq\theta} : \theta \in \mathcal{R}\}$, there are $m-1$ intervals between sampled points if all of them are unique. And for the same value of θ, g would classify the points in the opposite way as f. Note, however, that the classification for θ greater or less than all the points have been already counted.

Therefore there are maximum of $m-1$ classifications for G that are unique from F

$\Rightarrow \Pi_H(m) = \Pi_{F\cup G}(m) \leq (m+1) + (m-1) = 2m$

The results defined in class is for hypotheses that map to $\{-1, +1\}$, whereas $H \mapsto \{0, +1\}$. However, note that in Massart's theorem, $R = max_{x\in H}||x||_2$ is still $\sqrt{m}$

$\Rightarrow \mathcal{R}_m[H] \leq \sqrt{2\log \Pi_H(m)/m} \leq \sqrt{2\log 2m/m}$

2.2