# 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:

$min = double.max,\ max = double.min$

foreach $x_i \in X$:

if $y_i = 1 \& x_i > max$ then $max = x_i$

if $y_i = 1 \& x_i < min$ then $min = x_i$

return $[min, max]$

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 $\epsilon$, 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 $r_2 \Rightarrow$

$\Re[R_S] = Pr[R] - Pr[R_S] = Pr[r_1] + Pr[r_2] - Pr[R_S \cap r_1] - Pr[R_S \cap r_2]$

$= Pr[r_1] + Pr[r_2] - Pr[R_S \cap r_1] - Pr[R_S \cap r_2]$

$= \epsilon - Pr[R_S \cap r_1] - Pr[R_S \cap r_2] \leq \epsilon$

$\Rightarrow R_S$ 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 by first sorting the points based on x value then iterating through the sorted list to identify intervals:

$curRegion.label = 0$, $curRegion.min = nil$, $curRegion.max = nil$, $prevPt = nil$

$Z' = sort(\{X, Y\})$ by $x_i$ #Now $Z'$ is sorted by $x_i \in Z'$

foreach $x_i \in Z'$:

  If $curRegion = 0 \& y_i = 1$ then:

    $curRegion.label = 1$, $curRegion.max = x_i$

  if $curRegion = 1 \& y_i = 0$ then:

    $curRegion.max = prevPt$, emit $curRegion$

    $curRegion.label = 0$, $curRegion.min = nil$, $curRegion.max = nil$

  $prevPt = x_i$

if $curRegion.label = 1$, return $curRegion$

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 $(b, c)$

Let $R_1 = [a, b]$, $R_2 = [c, d]$ and $R_3 = (b, c)$ and it is clear that $R_i \cap R_j = \emptyset$, where $i \neq j$

Notice that if $Pr[R_3] = \epsilon'$, then for an i.i.d sampling of m points, the probability of having no points fall into that region is $(1 - \epsilon)^m \leq \exp(-m\epsilon)$

Based on construction, $R_S \subseteq R' = (R_1 \cup R_2 \cup R_3)$. Furthermore, errors are also subset of $R'$

Therefore, if $Pr[R_1 \cup R_2 \cup R_3] = Pr[R_1] + Pr[R_2] + Pr[R_3] \leq \epsilon \Rightarrow \Re[R_S] \leq \epsilon$ (A)

Next, consider if $Pr[R_i] \geq \epsilon/2$ for all of the regions.

Consider subintervals where $Pr[r_k] = \epsilon/4$ bordering either side of and within $R_1$ and $R_2$, there are 4 such regions, and for $\Re[R_S] > \epsilon$, $R_S$ must at least miss one of them or there must be no points sampled in $(b, c)$

$$
\begin{aligned}
Pr_{S\sim D^m}[\Re[R_S] > \epsilon] &\leq Pr_{S\sim D^m}[\cup_{k=1}^4 \{R_S \cap r_k = \emptyset\} \bigwedge (S \cap R_3 = \emptyset)] \\
&\leq \textstyle\sum_{k=1}^4 Pr_{S\sim D^m}[\{R_S \cap r_k = \emptyset\}] + \exp(-m\epsilon/2) && \text{(by union bound)} \\
&\leq 4(1 - \epsilon/4)^m + \exp(-m\epsilon/2) \\
&\leq 4\exp(-m\epsilon/4) + \exp(-m\epsilon/2) \\
&\leq 4\exp(-m\epsilon/4) + \exp(-m\epsilon/4) \\
&= 5\exp(-m\epsilon/4)
\end{aligned}
$$

Setting $r.h.s. \leq \delta \Rightarrow m \geq (4/\epsilon)\log(5/\delta)$

Now consider if one of $Pr[R_i] < \epsilon$, let us call this interval $k$. By (A), $\sum Pr[R_i] > \epsilon$ or

$Pr[\Re[R_S] > \epsilon] = 0$, therefore the probabilistic mass of that interval must be distributed to the other two intervals for the problem to be nontrivial. Now consider the same construction of subintervals on the ends of each target interval, except for the smaller interval $Pr[r] = Pr[R_k]/2$ (this construction is valid for the false positive interval between the two target intervals as we can alter construct of the proof to the scenario where sampled points miss both subintervals), while we distribute the rest of the probabilistic mass $(\epsilon/2 - Pr[R_k])$ to the other subintervals as well (B). Same assumption of having to miss one of the subintervals or not sampling an points in the false positive region applies. Therefore, if the smaller region is $R_1$ or $R_2$

$$\begin{aligned}
&Pr_{S\sim D^m}[\Re[R_S] > \epsilon] \leq Pr_{S\sim D^m}[\cup_{k=1}^{4}\{R_S \cap r_k = \emptyset\} \bigwedge (S \cap R_3 = \emptyset)] \\
&\quad \leq 2(1 - Pr[R_k)/2)^m + 2(1 - [\epsilon/4 + (\epsilon/2 - Pr[R_k])/2]) + \exp(-m\epsilon/2) \\
&\quad \leq 2\exp(-m\ Pr[R_k]/2) + 2\exp(-m[\epsilon/4 + (\epsilon/2 - Pr[R_k])/2]) + \exp(-m\epsilon/2)
\end{aligned}$$

The same algorithm will generalize to the union of $p$ intervals. Using the same construction as before, there will be $2p$ subintervals at either end of the target regions, each with $Pr[r_l] = \epsilon/2p$ and $p-1$ regions between each with $Pr[R'_n] = \epsilon/p$ (the case where the regions and subintervals are smaller can be treated similarly as the 2 interval case).

Therefore

$$\begin{aligned}
&Pr_{S\sim D^m}[\Re[R_S] > \epsilon] \leq Pr_{S\sim D^m}[\cup_{l=1}^{2p}\{R_S \cap r_l = \emptyset\} \bigwedge (\exists n, S \cap R'_n = \emptyset)] \\
&\quad \leq \textstyle\sum_{l=1}^{2p} Pr_{S\sim D^m}[\{R_S \cap r_l = \emptyset\}] + (p-1)\exp(-m\epsilon/p) \\
&\quad \leq 2p(1 - \epsilon/2p)^m + (p-1)\exp(-m\epsilon/p) \\
&\quad \leq 2p\exp(-m\epsilon/2p) + (p-1)\exp(-m\epsilon/p) \\
&\quad \leq (3p-1)\exp(-m\epsilon/2p)
\end{aligned}$$

Setting $r.h.s. \leq \delta \Rightarrow m \geq (2p/\epsilon)\log[(3p-1)/\delta]$

## 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 $x_i \neq x_j$, $i \neq j$ 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 $\theta$ 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 $\theta$, $g$ would classify the points in the opposite way as $f$. Note, however, that the classification for $\theta$ 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