Written Homework # 2 for CSCI3320

Prof. John C.S. Lui
CSE Department
CUHK

Deadline: Feb 21, 2020. 5:00 pm.
Objectives: Understand VC dimension and PAC.

1. (**VC Dimension**) Let's consider the space of instances $\mathcal{X}$ corresponding to all points in x, y plane. Find the VC dimensions for the following hypothesis spaces:

   (a) H_c = circles in the (x, y) plane. Points inside the circle are classified as positive examples.

   (b) H_c = circles in the (x, y) plane. Points inside the circle are classified as negative examples.

   Solution: The above two questions are in fact the same. The VC dimension is 3. The reason is simple. Because your hypothesis can control both the *center* and *radius* of a circle. Consider one point and you can definitely differentiate a positive point from a negative point. Now you can consider two points. In this case, you have four configurations and as long as these two points are in general position, you can separate positive from the negative points. This also hold when you are given three points, in which case you have eight configurations. You can show that you can shatter all these eight points.

2. (**VC Dimension**) What is the VC dimension of intervals in R (or real numbers) in which the target function is specified by an interval (or two points in the $R-$ space), such that points within the interval are labelled as positive while points outside the interval are labelled as negative?

   Solution: VC dimension in this case will be 2 since there is only a single interval of positive examples that could be within that interval. One can also construct an example that if we have a set of 3 points within an interval such that they are labeled as $+$, $-$, and $+$. In this case, the target function *cannot* shattered these 3 points. This goes to show that the VC dimension is 2.

3. (**PAC**) Find the minimum number of sample points we need for so that the hypothesis is PAC-learnable.

   Our domain $\mathcal{X}$ is the real line. Our concept class $\mathcal{C}$, is the set of all positive half-lines, which are lines that separate the real line into two halves: all the points to the left of the line are negatively labelled and all the points to the right of the half-line are positively labelled. This is illustrated by the following figure.

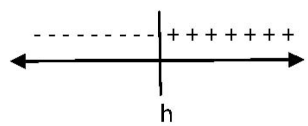

Figure 1: Simple learning hypothesis h

   Given a set of labelled points, a valid hypothesis h could be to find the place where the examples transition from negative labels to positive labels, and place the half-line anywhere in the transition region (please refer to Figure 1), i.e., we can place the half-line midway between the right-most negative and left-most positive examples. So, how many samples do we need to ensure that h will be $\epsilon-$ good with high probability $1-\delta$?

   Solution: As seen in Figure 2, the region of error is the region between c and h: If we get any points there, h will say its negative while c would say its positive. We can also have c to the right of h, where a similar argument can be made due to symmetry.

   Now, we want the error region to be at most ϵ. More precisely, we want the probability of a point falling into this region to be less than ϵ. In

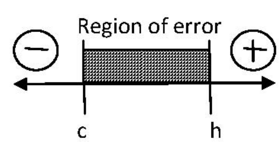

Figure 2: Region of error

   general, we can consider all error cases via the following figure:

   Errors can be divided into two cases (or events), which are:

   - $b+$: Hypothesis h is more than ϵ to the right of the target concept c.
   - $b-$: Hypothesis h is more than ϵ to the left of the target concept c.

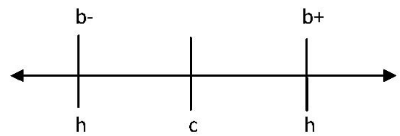

Figure 3: Two Bad Events

As long as those two events do not happen with high probability, then we can say that h is $\epsilon-$good. Since these two events are symmetric, we only need to look at one of them at a time. This is illustrated via the following figure: Figure 4 shows the $b+$ event and a region of probability ϵ to the

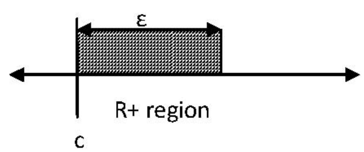

Figure 4: Region with probability ϵ

right of c, which is called $R+$. In this region, the probability mass between c and the edge of the region is ϵ. We want to bound the probability of h ending up outside the region $R+$. If theres even a single point that falls in $R+$, which is a positive example, since it falls to the right of c, then h will fall to the left of that point and also be inside the region $R+$. Therefore, the only way for the bad event to happen is if no points fall inside the R+ region. That is:

$$\begin{aligned} Pr[b+] &\leq Pr[x_1 \notin R^+ \wedge \cdots \wedge \notin R^+] \\ &= Pr[x_1 \notin R^+] \cdots Pr[x_m \notin R^+] = (1-\epsilon)^m \ \leq \ e^{-\epsilon m}. \end{aligned}$$

For the above derivation, we use the assumption that all x_i are independent and identically distributed (IID), where the probability of x_i falling in the region $R+$ is ϵ. We use the inequality $1 + x \leq e^x$ to obtain the result. Then:

$$\begin{aligned} Pr[h \text{ is } \epsilon- \text{ bad}] &\leq Pr[b + \vee b-] && (1) \\ &\leq Pr[b+] + Pr[b-] && (2) \\ &\leq 2e^{-\epsilon m} && (3) \end{aligned}$$

Solving m, we have:

$$m \geq \frac{1}{\epsilon} \ln \frac{2}{\delta}.$$

Therefore, if we have at least $m \geq \frac{1}{\epsilon} \ln \frac{2}{\delta}$ samples, our algorithm is PAC-learnable. Solving for ϵ, we can say that with at least $(1 - \delta)$ probability, we have $err(h) \leq \frac{1}{m} \ln \frac{2}{\delta}$.