CS6350 - Homework/Assignment-4

Arnab Das(u1014840)

October 30, 2016

1: PAC-Learning

(1.a) Rule-1 states that you are free to combine any of the parts as they are. Given, there are N available parts, then for every part there are two options - either the part is chosen or it is discarded. It is analogous to a N-bit vector, where each part corresponds to a bit in the vector. The decision of choosing the part can be thought of as setting 1 to that bit position while the choice of discarding a part is equivalent to seeting 0 to that bit position. For such a set-up the size of the hypotheses space is equal to the possible numbers that the n-bit vector can accomodate. Hence,

Size of hypotheses space by Rule-1 = 2^N (Answer).

(1.b) Here both Rule-1 and Rule-2 are followed. Rule-2 suggests that the original parts are allowed to be broken into two distinct pieces before using. Thus, in combination of Rule-1 and Rule-2, there will be now four choices for every part, that is , the part is not chosen, or the part is chosen as original without cut, or the part is cut and the first piece is chosen, or the part is cut and the second piece is chosen. For the analogy of the n-bit vector, each bit can now take 4 values instead of just two. Hence,

Size of hypotheses space by Rule-2 = 4^N (Answer).

(1.c) Number of available parts = 6. Hence, N = 6.

Size of hypotheses space = $|H| = 4^6$

The allowed error, $\epsilon = 0.01$

Given, the probability with which the robot wants to predict correctly within ϵ is 99%. Then $(1 - \delta) = 0.99$.

Then, δ = 1 - 0.99 = 0.01

Note that, in this case the learning is not agnostic, since the true concept is contained in the hypotheses space. Hence, we need to use the non-agnostic learning formula for finding the lower bound on the number of examples:

$$m > \frac{1}{\epsilon}\left(ln(|H|) + ln(\frac{1}{\delta})\right) \tag{1}$$

Plugging in the values for $|H|, \epsilon, \delta$ we get the value of m as:

$$m > 1292.293$$

Rounding off to the ceil:

$$m > 1293$$

(Answer)

(2) Suppose H denotes the hypotheses space. Consider a hypothesis h $\in$ H. Let us define a random Variable 'X' , such that:

$$X = 0; h(x) = c(x)$$
$$X = 1; h(x) \neq c(x)$$

where x belongs to the distribution D over which the instance space is defined, and c the target concept.
If we select 'm' random independent examples from the distribution D and the mark the outcomes as $X_1, X_2, \ldots, X_m$, then $\frac{\sum X_i}{m}$ denotes the error fraction over the sample space. The **training error**, $error_s(h)$, is defines as the fraction of the training examples misclassified by the hypothesis h. Hence,

$$error_s(h) = \frac{\sum X_i}{m} \tag{2}$$

The true error, or generalization error, $error_D$ over the entire distribution D from which the examples are randomly drawn is defined as

$$error_D(h) \equiv \underset{x \in D}{P}[c(x) \neq h(x)]$$

The expected value of a discrete random variable, is the probability weighted average of all possible values, that is over the entire distribution. Hence, the from the above definition of $error_D$, we can say it is the expected value of $error_s$, the training error.

For PAC learnability, we would like that the true error is not worse than ϵ plus the training error. Then we can characterize the event $(error_D - error_s > \epsilon)$ as a bad event and would like the probability of such an event be upperbounded by a small probability δ. We can write the probability of the above event as following:

$$P[error_D > error_s + \epsilon]$$

In this problem, we are given that the error term is a multiplicative terms relative to the training error, that is true error is no worse than $(1+\epsilon)error_s$. Then we can write the above as:

$$P[error_D - (1+\epsilon)error_s > 0]$$

Rearranging the terms:

$$P[error_D > (1+\epsilon)error_s]$$

$$P[\frac{1}{1+\epsilon}error_D > error_s]$$

$$P\left[error_D\left(1 - \frac{\epsilon}{1+\epsilon}\right) > error_s\right] \tag{3}$$

Recalling the Chernoff bound, where for $X_1, X_2, \ldots, X_m$ being the outcomes of m independent trials, and the probability of a $P[X_i = 1] = p$ and $P[X_i = 0] = (1-p)$, then the expectation $E\left[\frac{\sum X_i}{m}\right]$=p, and the chernoff bound governs the probability that $\frac{\sum X_i}{m}$ will differ from p by some factor $0 < \gamma < 1$, as:

$$P\left[\frac{\sum X_i}{m} > (1+\gamma)p\right] \leq \exp(\frac{-mp\gamma^2}{3})$$

$$P\left[\frac{\sum X_i}{m} < (1-\gamma)p\right] \leq \exp(\frac{-mp\gamma^2}{2})$$

Recalling equation(2) and the definition of true error, we see that equation(3) is in exact formation of the second chernoff bound described above such that
$p = error_D$ and $\gamma = \frac{\epsilon}{1+\epsilon}$. Since $0 < \epsilon < 1$, then $0 < \gamma < 1$, which satisfies the condition on the form factor of the Chernoff bound. Thus, using the second chernoff bound in our problem, we get:

$$P\left[error_D\left(1 - \frac{\epsilon}{1+\epsilon}\right) > error_s\right] \leq \exp(\frac{-m.error_D.\left(\frac{\epsilon}{1+\epsilon}\right)^2}{2}) \tag{4}$$

The above was determined for a single hypethesis $h \in H$. For the entire hypotheses space the above equation has to be adjusted to :

$$P\left[error_D\left(1 - \frac{\epsilon}{1+\epsilon}\right) > error_s\right] \leq |H| \exp\left(\frac{-m.error_D.\left(\frac{\epsilon}{1+\epsilon}\right)^2}{2}\right) \tag{5}$$

We use the result of equation(5) to determine the number of training examples required to reduce this probability of failure below a desired level of δ. Hence:

$$|H| \exp\left(\frac{-m.error_D.\left(\frac{\epsilon}{1+\epsilon}\right)^2}{2}\right) \leq \delta$$

Taking natural logarithms on both sides and reaarnging the terms to single out m(number of examples for training), we get:

$$m \geq \frac{2(1+\epsilon)^2}{\epsilon^2.error_D}\left(ln|H| + ln|\frac{1}{\delta}|\right) \tag{6}$$

(Answer).

2: VC Dimensions

---

(1)

(2.a) $H^X_{=k} = h \in 0, 1^X : |x : h(x) = 1| = k$ that is, the set of all functions that assign the value 1 to exactly k elements. Consider 1 sample point. The adversary can mark this point as a '1' label or a '0'. The class defintiion says that for a given fixed 'k', the hypotheses contained in this space are the ones that mark exactly k points as 1. Suppose k=1. The possible dichotomies of a single point is two, that is either the point is labeled as 1 or labeled as 0. Then, for k=1, it works for the dichotomy where the label is 1, but for label 0, it does not have a hypotheses in this class. The class for k=0, has a hypotheses for label=0, but it doesn't havs a hypotheses for the label being 1. Thus, for a specific fixed k, it cannot shatter a single point. Hence, **VC dimension of this class is 0**. (answer)

(2.b) $H^X_{\leq k} = h \in 0, 1^X : |x : h(x) = 1| \leq k or |x : h(x) = 0| \leq k$
That is a class of functions for a given fixed k, there exists hypothesis that can label atmost k elements as 1 or a 0. Let us consider a few rudimentary cases to identify the formal pattern:
case-1: single point - If this point is labeled 0, then there exists a $h \in H_1$ that can label a single point as 0, due to the fact $|x : h(x) = 0| \leq 1$, and holds true for all $k \geq 1$.
If this point is labeled 1, then also there exists a $h \in H_1$ that can label a single point as 1 due to the fact $|x : h(x) = 1| \leq 1$, and holds true for all $k \geq 1$.
case-2: 2 points - Consider 2 points. 2 points can have the following 4 labels: (0,0), (0,1), (1,0) and (1,1). In all 4 cases the $|x = 1|$ is atmost 2, and hence satisfied by $h \in H_2 : |x : h(x) = 1| \leq 2$ and also $|x = 0|$ is atmost 2 and hence satisfied by $h \in H_2 : |x : h(x) = 0| \leq 2$.
As the pattern is very evident that if the number of points, n, is less than or equal to k, then $H^X_{\leq k}$ can shatter the points.
Now let us consider, that the number of points n is greater than k, say k+1, for a given fixed k. There exists a labelling where all the points are labelled as 1, such that $|x = 1| =$ k+1. But the hypotheses class we have has hypothesis that can label **atmost** k points as 1. Hence, in this case we have a breakpoint and so the (k+1) points cannot be shattered.
Hence, **VC dimension is k**. (answer)

(3) Instance space consisting of real numbers and a hypothesis space H consisting of two disjoint intervals, defined by $[a, b]$ and $[c, d]$.
In figure-1, we describe the possible shatterings of set of points on the real number line. For 5 points we can show a labelling, where two disjoint intervals cannot shatter the set of points.Hence, **VC dimension is 4** . (Answer).

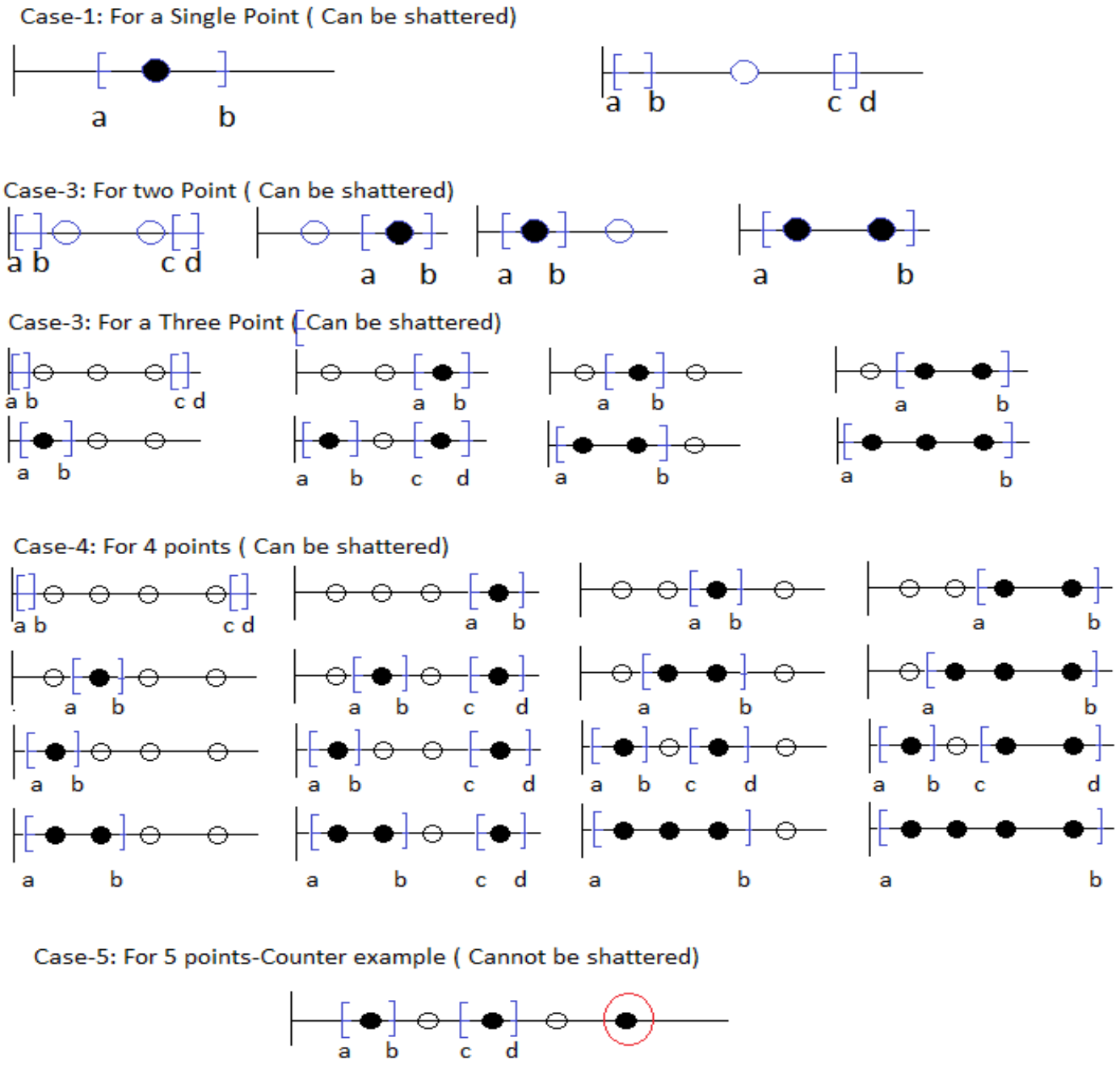

Figure 1: Shattering of real number instance space by two disjoint intervals