CS 5350/6350: Machine Learning Fall 2015

Homework 3

Handed out: Oct 20, 2015
Due date: Nov 3, 2015

1 Warm Up: Feature Expansion

[10 points total] Consider an instance space consisting of points on the two dimensional plane (x_1, x_2). Let $\mathcal{C}$ be a concept class defined on this instance space. Each function $f_r \in \mathcal{C}$ is defined by a radius r as follows:

$$f_r(x_1, x_2) = \begin{cases} +1 & \text{if } x_1^2 + x_2^2 - 2x_1 \leq r^2 \\ -1 & \text{else} \end{cases}$$

This hypothesis class is definitely not separable in $\mathbb{R}^2$. That is, there is no w_1, w_2 and b such that $f_r(x_1, x_2) = sign(w_1 x_1 + w_2 x_2 + b)$ for any r.

1. [4 points] Construct a function $\phi(x_1, x_2)$ that maps examples to a new space, such that the positive and negative examples are linearly separable in that space? That is, after the transformation, there is some weight vector $\mathbf{w}$ and a bias b such that $f_r(x_1, x_2) = sign(\mathbf{w}^T \phi(x_1, x_2) + b)$ for any value of r.

   (Note: This new space need not be a two-dimensional space.)

   Solution: The function $f_r(x_1, x_2)$ is an equation of a circle centered at $(1, 0)$. Hence our radius will be $\sqrt{r^2 + 1}$. The positive examples are within the circle and negative examples outside the circle for any r. This is not linearly separable. If we expand the equation of a circle for any center we can take a new function $\phi(x_1, x_2) = [x_1, x_2, x_1^2, x_2^2]$. In this new space the values are linearly separable for weight vector $\mathbf{w} = [-2, 0, 1]$ and bias $b = -r^2$. $\phi(x_1, x_2) = -sign(w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_2^2 + b)$.

2. [3 points] If we change the above function to:

$$g_r(x_1, x_2) = \begin{cases} +1 & \text{if } x_1^2 - x_2^2 \leq r^2 \\ -1 & \text{else} \end{cases}$$

   Does your $\phi(x_1, x_2)$ make the above linearly separable? If so demonstrate how. If not prove that it does not.

Solution: Yes. My $\phi(x_1, x_2)$ makes the above linearly separable. For $\phi(x_1, x_2) = -sign(w_1x_1 + w_2x_2 + w_3x_1^2 + w_4x_2^2 + b)$ where $w = [0, 0, 1, -1]$ and $b = -r^2$ the above function is linearly separable.

3. [3 points] Does $\phi(x_1, x_2) = [x_1, x_2^2]$ make the function g_r above linearly separable? If so demonstrate how. If not prove that it does not.

Solution: The function $\phi(x_1, x_2) = [x_1, x_2^2]$ does not make the function g_r linearly separable. This is because our function g_r has the equation $x_1^2 - x_2^2 \leq r^2$. We cannot find any weight vector which will give back the original equation on simplification, as we require x_1^2 and x_2^2 but our new feature transformation has only x_1.

2 PAC Learning

1. [15 points] Due to the recent budget cuts the government no longer has any money to pay for humans to monitor the state of nuclear reactors. They have charged you with assessing a Robot's ability to perform this vital task. Every reactor has a different number of binary gauges which indicate whether or not some aspect of the reaction is `normal` or `strange`. The reactor itself can be in one of **five** states – *Normal*, *Meltdown*, *Pre-meltdown*, *Abnormally cool* or *Off*. Each combination of the binary guage settings indicate one of these five reactor states. We want to know if we can train a robot to identify which gauges and gauge combinations are responsible for each reactor state.

   a) [5 points] Suppose that we have N gauges with which to identify reactor states. How large is the hypothesis space for this task? (You may have to make assumptions about the underlying function space. State your assumptions clearly.)

   Solution: We have N gauges for identifying reactor states. A conjunction of these gauges will tell us the reactor states. The hypothesis space for this task will be 3^n as that many conjunctions can be formed using N gauges.

   b) [10 points] The ex-government employee, whose job the robot is taking, trains the robot at a nuclear reactor where there are 20 gauges by showing the robot a set of gauge positions for the five different reactor states. If the robot wants to learn to recognize the reactor's condition with .1 percent error with greater than 99% probability how many examples does the robot need to see?

Solution: Using the formula:

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

$\mid H \mid = 3^{20}, \delta = 99\%, \epsilon = 0.1\%$
number of of examples:

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

$$m > \frac{100}{0.1}\left(\ln 3^{20} + \ln \frac{100}{1}\right)$$

$$m > 1000\,(21.97 + 4.6)$$

$$m > 26570$$

The 26570 examples help the robot train on the five reactor states.

2. [5 points] Is it possible for a learned hypothesis h to achieve 100% accuracy with respect to a training set and still have non-zero true error? If so, provide a description of how this is possible. If not, prove that it is impossible.

Solution: Yes. It is possible for a learned hypothesis h to achieve 100% accuracy with respect to a training set and still have non-zero true error. Lets take an example where my true function is $f(x) = x_2 \wedge x_3 \wedge x_4$. While training on the instance space we eliminate the negative examples and learn a hypothesis $h(x) = x_1 \wedge x_2 \wedge x_3 \wedge x_4$ for all positive examples where all relevant features were also positive. In such a case we see that x_1 is not present in $f(x)$. Thus in general, we will come across a test example which will be positive for $x_1 = 0$. Hence we will have a non-zero true error.

3. [25 points] **Learning decision lists:** In this problem, we are going to learn the class of k-decision lists. A decision list is an ordered sequence of if-then-else statements. The sequence of if-then-else conditions are tested in order, and the answer associated to the first satisfied condition is output. See Figure 1 for an example of a 2-decision list.

   A *k-decision list* over the variables $x_1, \ldots, x_n$ is an ordered sequence $L = (c_1, b_1), \ldots, (c_l, b_l)$ and a bit b, in which each c_i is a conjunction of at most k literals over $x_1, \ldots, x_n$. The bit b is called the *default* value, and b_i is referred to as the bit *associated* with condition c_i. For any input $x \in \{0,1\}^n$, $L(x)$ is defined to be the bit b_j, where j is the smallest index satisfying $c_j(x) = 1$; if no such index exists, then $L(x) = b$.

   We denote by *k-DL* the class of concepts that can be represented by a k-decision list.

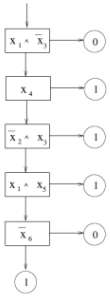


Figure 1: A 2-decision list.

(a) [8 points] Show that if a concept c can be represented as a k-decision list so can its complement, $\neg c$. You can show this by providing a k-decision list that represents $\neg c$, given $c = \{(c_1, b_1), \ldots, (c_l, b_l), b)$.

Solution: A k-decision list is represented by $c = \{(c_1, b_1), \ldots, (c_l, b_l), b)$. The complement of c can be written as $\neg c = \{(c_1, \neg b_1), \ldots, (c_l, \neg b_l), \neg b)$. which is also a k-decision list. The default value of each c_i conjunction is complemented.

(b) [9 points] Use Occam's Razor to show:
For any constant $k \geq 1$, the class of k-decision lists is PAC-learnable.

Solution: To prove that class of k-decision list is PAC-learnable, we have to bound the size of the hypothesis space H. In k-decision list each conjunction has atmost k-literals. No of such conjunctions =

$$|\mathcal{C}| = \sum_{i=1}^{k} \binom{n}{i} .2^i = \mathcal{O}(n^k)$$

now either of the conjunction can have a bit value b as 0 or 1 or it can be absent in the k-decision list. So we have three choices. The conjunctions can also be arranged in any order, which is a permutation. Hence:

$$|\mathcal{H}| = 3^{|\mathcal{C}|}|\mathcal{C}|!$$

$$\log(|\mathcal{H}|) = \log((3^{|\mathcal{C}|}|\mathcal{C}|!)$$

since factorial will be dominant term among all:

$$\log(|\mathcal{H}|) = \mathcal{O}(k.n^k log(n))$$

using Occam's razor: number of training samples

$$M > \frac{1}{\epsilon}(\ln \frac{1}{\delta} + \log(|\mathcal{H}|))$$

using $|\mathcal{H}|$:

$$M > \frac{1}{\epsilon}(\ln \frac{1}{\delta} + \mathcal{O}(k.n^k log(n))$$

which is polynomial in n number of literals. So the k-DL is PAC-learnable.

(c) [8 points] Show that 1-decision lists are a linearly separable functions. (Hint: Find a weight vector that will make the same predictions a given 1-decision list.)

Solution: Let there be l literals that appear in the 1-DL from top to bottom denoted by $x_1, x_2, ..., x_l$. Let $b_1, b_2, ..., b_l$ denotes the polarities associated with them in the 1-DL. we can set $b_i = -1$ if x_i is appears with zero bit or $b_i = 1$ otherwise. Now we can create a linear threshold function:

$$sign(w^T x + \theta) \geq 0$$

that makes the same decisions as the decision list. Let $\mathbf{w} \in \mathcal{R}^l$ be the weight vector.The ith component of the weight vector is $w_i = b_i 2^{l+1-i}$.If the default value of the 1-DL is false then set threshold θ to -1. Otherwise $\theta = 1$.

Note that we built the weight vector based on literals that appear in the decision list. Instead,if we construct the weight vector based on x, we should represent x as a $2n$ dimension vector $[x_1, \neg x_1, x_2, \neg x_2, ..., x_n, \neg x_n]$, and set all weights in $w \in R^{2n}$ corresponding to the literals that do not appear in the 1-DL to zero.

The key point to notice about the design of this linear threshold function is that the weights must decrease geometrically as you go down the decision list so that the current feature's weight dominates all the weights that come after it whenever all the features that came before it were inactive. Note that the threshold has been set so that when all features are inactive, the prediction of the linear threshold function is the same as that of the 1-DL.

4. [20 points, **CS 6350 students only**] Let X be an instance space and let $D_1, D_2, ..., D_m$ be a sequence of distributions over X. Let $\mathcal{H}$ be a finite class of binary classifiers over X and let $f \in \mathcal{H}$.

Suppose we have a sample S of m examples, such that the instances are independent but are not identically distributed. The i^{th} instance is sampled from D_i and then y_i is set to be $f(x_i)$. Let $\bar{D}_m$ denote the average, that is, $\bar{D}_m = \frac{1}{m}\sum_{i=1}^{m} D_i$.

Let $h \in \mathcal{H}$ be a classifier that gets zero error on the training set. That is, for every example $x_i \in X$, we have $h(x_i) = f(x_i)$. Show that, for any accuracy parameter $\epsilon \in (0, 1)$, the probability that the expected error of the learned classifier h is greater than ϵ is no more than $|\mathcal{H}|e^{-\epsilon m}$. That is, show that

$$\mathbb{P}\left[E_{x \sim \bar{D}_m}\left[h(x) \neq f(x)\right] > \epsilon\right] \leq |\mathcal{H}|e^{-\epsilon m}$$

(Hint: You have to use the fact that the arithmetic mean of a set of non-negative numbers greater than or equal to their geometric mean.)

Solution: We know that $\bar{D}_m = \frac{1}{m}\sum_{i=1}^{m} D_i$.
The expectation of a random variable X, where my hypothesis is successful even though it is not equal to the true function, is :

$$\begin{aligned} E_{X\sim\bar{D}_m}[1_{h(x)\neq f(x)}] &= \sum_X \bar{D}_m 1_{h(x)\neq f(x)} \\ &= \sum_X \frac{1}{m}\sum_{i=1}^{m} D_i 1_{h(x)\neq f(x)} \\ &= \frac{1}{m}\sum_{i=1}^{m}\sum_X D_i 1_{h(x)\neq f(x)} \\ &= \frac{1}{m}\sum_{i=1}^{m} E_{X\sim D_i}[1_{h(x)\neq f(x)}] \end{aligned}$$

We have a sample space S such that the instances in the space are independent but not identically distributed. Hence:

$$\begin{aligned} \mathbb{P}[\forall i, h(x_i) \neq f(x_i)] &= \mathbb{P}[h(x_1) \neq f(x_1) \; and \; h(x_2) \neq f(x_2)....h(x_m) \neq f(x_m)] \\ &= \prod_{i=1}^{m} \mathbb{P}[h(x_i) \neq f(x_i)] \end{aligned}$$

We know by the relation of arithmetic and geometric mean:

$$\left(\frac{1}{m}\sum_{i=1}^{m} X_i\right)^m \geq \prod_{i=1}^{m} X_i$$

Using the above relation we get:

$$\prod_{i=1}^{m} \mathbb{P}[h(x_i) \neq f(x_i)] \leq \left(\frac{1}{m}\sum_{i=1}^{m} \mathbb{P}[h(x_i) \neq f(x_i)]\right)^m$$

Using Bernoulli's trial we know that:

$$\mathbb{P}(X) = E[X]$$

Hence:

$$\left(\frac{1}{m}\sum_{i=1}^{m} \mathbb{P}[h(x_i) \neq f(x_i)]\right)^m = \left(\frac{1}{m}\sum_{i=1}^{m} E_{X\sim D_i}[1_{h(x)\neq f(x)}]\right)^m$$

We know that the expectation of a hypothesis to be successful greater than ϵ for a single example, the probability is $1-\epsilon$. Hence for all examples is defined by:

$$\mathbb{P}\left[\left(\frac{1}{m}\sum_{i=1}^{m} E_{X\sim D_i}[1_{h(x)\neq f(x)}]\right)^m > \epsilon\right] = (1-\epsilon)^m$$

As derived in the beginning, we can write:

$$\mathbb{P}\left[E_{x\sim \bar{D}_m}\left[h(x)\neq f(x)\right] > \epsilon\right] = (1-\epsilon)^m$$

We know that $1-x \leq e^{-x}$, therefore:

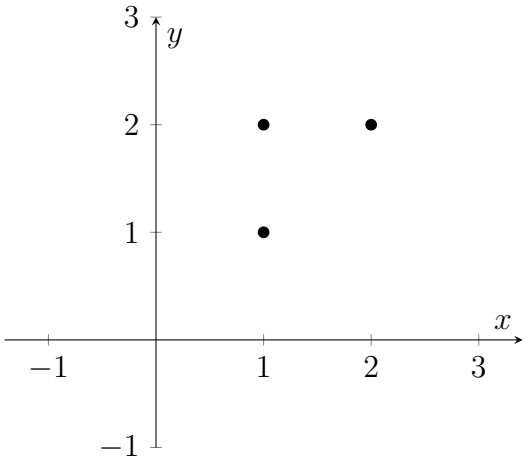
$$\mathbb{P}\left[E_{x\sim \bar{D}_m}\left[h(x)\neq f(x)\right] > \epsilon\right] \leq e^{-\epsilon m}$$

Therefore for all $h \in \mathcal{H}$ we get:

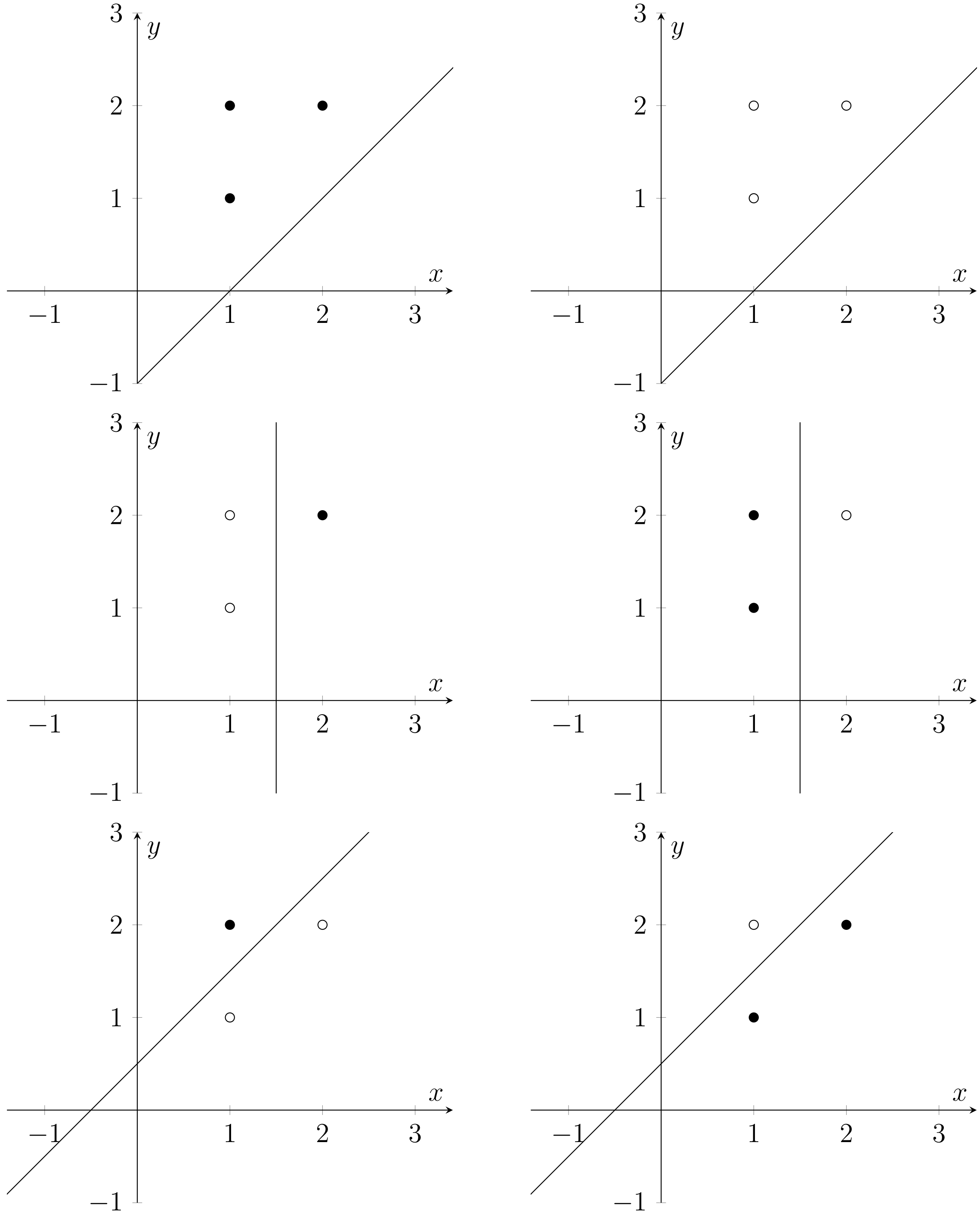
$$\mathbb{P}\left[E_{x\sim \bar{D}_m}\left[h(x)\neq f(x)\right] > \epsilon\right] \leq |\mathcal{H}|e^{-\epsilon m}$$

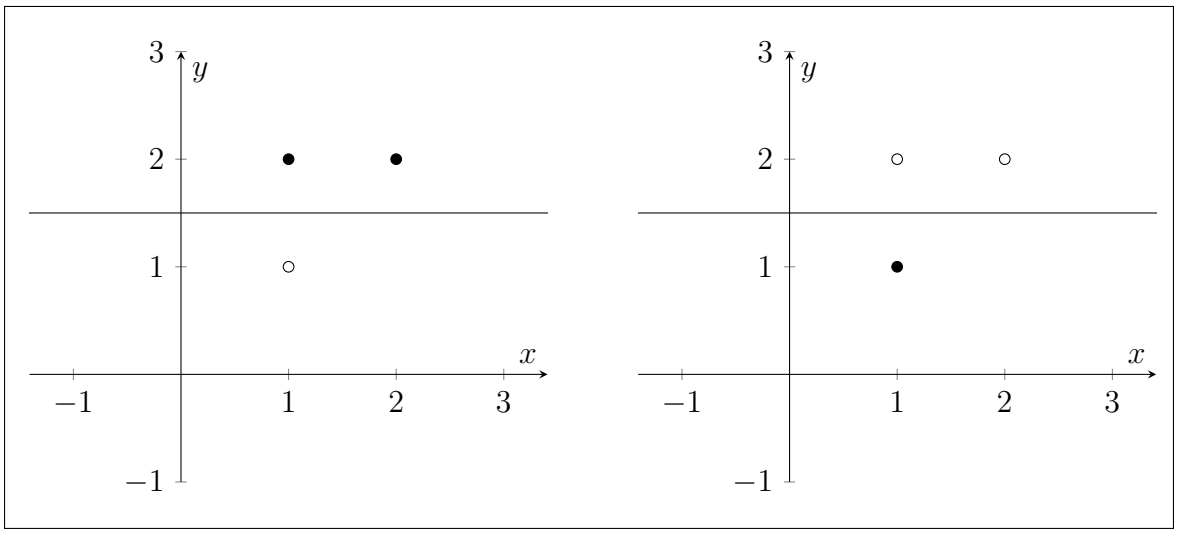
3 VC Dimension

1. [5 points] Assume that the three points below can be labeled in any way. Show with pictures how they can be shattered by a linear classifier. Use filled dots to represent positive classes and unfilled dots to represent negative classes.



Solution:


y
x
3
2
1
−1
−1
1
2
3
y
x
3
2
1
−1
−1
1
2
3
y
x
3
2
1
−1
−1
1
2
3
y
x
3
2
1
−1
−1
1
2
3
y
x
3
2
1
−1
−1
1
2
3
y
x
3
2
1
−1
−1
1
2
3



2. **VC-dimension of axis aligned rectangles in $\mathbb{R}^d$**: Let H^d_{rec} be the class of axis-aligned rectangles in $\mathbb{R}^d$. When $d = 2$, this class simply consists of rectangles on the plane, and labels all points strictly outside the rectangle as negative and all points on or inside the rectangle as positive. In higher dimensions, this generalizes to d-dimensional boxes, with points outside the box labeled negative.

   (a) [10 points] Show that the VC dimension of H^2_{rec} is 4.

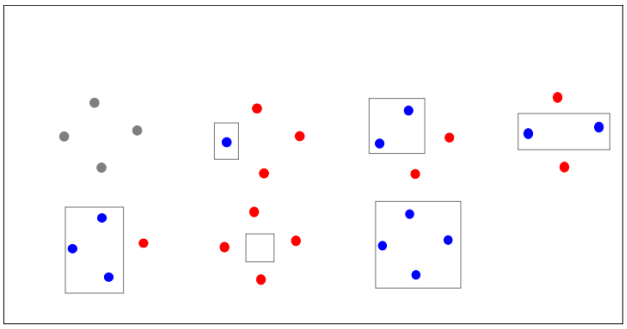


Figure 2: Shattering of 4 points for axis aligned rectangles

Solution: If we take any four points randomly or on the corners of the rectangle, the adversary can label them in such a way that they will not be shatterable. So we will take a points on the edges. As we can see in the figure the points are shatterble. In case of 5 points , our adversary can always label them in such a way that the points will never be shatterable. Therefore the VC dimension of H^2_{rec} is 4.

   (b) [10 poin ts] Generalize your argument from the previous proof to show that for d dimensions, the VC dimension of H^d_{rec} is $2d$.

Solution: Using the above explanation we can see that for a 2 - dimensional of axis aligned rectangles the VC dimension was 4. This is equal to 2 times the dimensional. For an extra point there is always a point which is either inside the rectangle or is marked negative, which cannot be consistent with the axis aligned rectangles. Using the same general understanding for higher dimension we can conclude that for d dimensions, the VC dimension of H_{rec}^{d} is 2d.

3. In the lectures, we considered the VC dimensions of infinite concept classes. However, the same argument can be applied to finite concept classes too. In this question, we will explore this setting.

   (a) [10 points] Show that for a finite hypothesis class $\mathcal{C}$, its VC dimension can be at most $\log_2(|\mathcal{C}|)$. (Hint: You can use contradiction for this proof. But not necessarily!)

   Solution: Considering that for a finite hypothesis class $\mathcal{C}$, its VC($\mathcal{C}$) = d. As there are d points we can label them in 2^d ways.
   The 2^d combinations can be in worst case classified by the size of entire hypothesis class $|\mathcal{C}|$. Therefore $2^d \leq |\mathcal{C}|$. Taking log on both side we get $d \leq \log_2(|\mathcal{C}|)$.

   (b) [5 points] Find an example of a class $\mathcal{C}$ of functions over the real interval $X = [0, 1]$ such that $\mathcal{C}$ is an **infinite** set, while its VC dimension is exactly one.

   Solution: For a class of function over the real interval $X = [0, 1]$ is a left bounded interval $[0, a)$ in it. Here we can see that the function $f(x) = +1 if 0 \leq x < a$ defines the infinite set. We know that for this the VC dimension is 1.

   (c) [5 points] Give an example of a **finite** class $\mathcal{C}$ of functions over the same domain $X = [0, 1]$ whose VC dimension is exactly $\log_2(|\mathcal{C}|)$.

   Solution: We can see that over the same domain $X = [0, 1]$ a finite class $\mathcal{C}$ is a function of Integers. Hence, we can deduce the class as $[0, 1]$ where the function can either be positive for value $x = 1$ or value $x = 0$. There fore there are 2 functions. But only one of them is shatterable, which is VC = $\log_2(2) = 1$. Hence VC dimension for it is exactly $\log_2(|\mathcal{C}|)$.