CS 5350/6350: Machine Learining Fall 2015

Homework 3
Christopher Mertin

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.)

   - The distribution of the postive cases are in a circle that is centered around $(1, 0)$. This can be mapped into 3-dimensions by using a Gaussian Distribution over the data. An alternative equation to the problem is $(x-1)^2 + y^2 \leq r^2 + 1$ which describes a cirlce of radius $\sqrt{r^2+1}$. Then, a plane at $Z = 0.1$ (or some other small value for Z to bisect the data) can be used to bisect the data. The formula that would be used is

     $$\phi(x_1, x_2) = \exp\left(- \left(\frac{(x_1 - 1)^2}{2\sqrt{r^2+1}} \right) + \left(\frac{x_2^2}{2\sqrt{r^2+1}} \right) \right)$$

     Which will create a Gaussian distribution in the z-dimension for positive examples and wouldn't change anything for the negative examples, which a hyperplane can bisect the two.

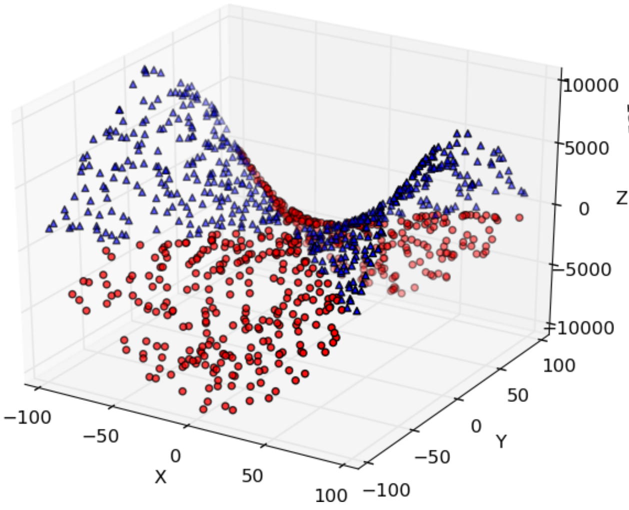


Figure 1: Plot of $x^2 - y^2 \leq 2^2$

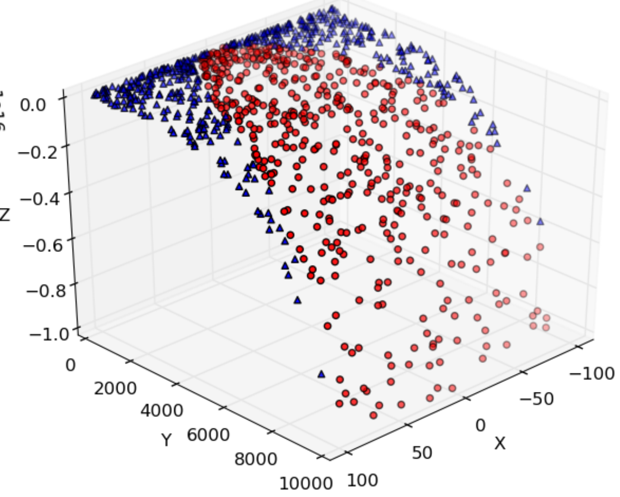


Figure 2: Plot of $\phi(x, y)$

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.

- It is trivially seen that $\phi(x_1, x_2)$ *cannot* make $g_r(x_1, x_2)$ linearlly separable as ϕ was a gaussian distribution in the shape of a cirlce. This allowed to raise the positive points above the axis. g_r is a hyperbola, and therefore cannot be linarly separated by a gaussian distribution.

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.

- This can be shown graphically that this is *not* the case for all r. By taking the plots and plotting them in the 3^{rd} dimension, it can be trivially seen that there is no segment/hyper plane that can be drawn through the data. Figure 1 shows the plot of $x^2 - y^2 \leq 2^2$, while Figure 2 is of taking the data points in Figure 1 and squaring the y value, while keeping the classificataions.

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.)

   - We can *assume* that since this is a nuclear reactor that all of the gauges are important, and thus there are 3^N different conjunctions for determining the state of the nuclear reactor.

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?

   - The following equation represents the relation for PAC learning

     $$\delta \geq He^{-\epsilon R}$$

     Where R is the number of required training examples, ϵ is the error, and δ is the probability. We can solve for R to get the maximum number of training examples.

     $$\frac{1}{\epsilon}\log\left(\frac{\delta}{H}\right) \geq -R$$
     $$\frac{1}{\epsilon}\log\left(\frac{H}{\delta}\right) \geq R$$
     $$\frac{1}{\epsilon}\left[n\log(3) - \log(\delta)\right] \geq R$$

     where δ is defined as $(1 - \text{probability})$ and ϵ is defined as $(1 - \text{accuracy})$. We can plug these in from the stated problem and we get

     $$\frac{1}{0.10}\left(20\log(3) - \log(0.01)\right) \geq R$$
     $$R \leq 266$$

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.

   - Yes, an instance of this would be when the data is overfit to the training data, which would result in 100% accuracy on the training set and a large (non-zero) error on the test set, which would result in 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 3 for an example of a 2-decision list.

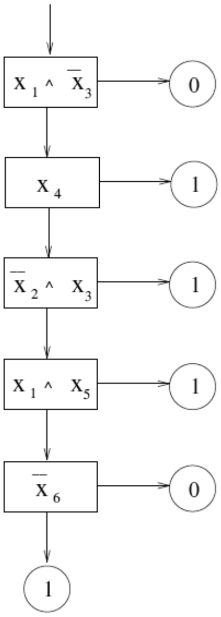


Figure 3: A 2-decision list.

A *k-decision list* over the variables $x_1, \dots, x_n$ is an ordered sequence $L = (c_1, b_1), \dots, (c_\ell, b_\ell)$ and a bit b, in which each c_i is a conjunction of at most k literals over $x_1, \dots, x_n$. 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.

(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 $c = \{(c_1, b_1), \dots, (c_\ell, b_\ell), b\}$.

- To prove that the complement of c is a k-decision list, we can simply take c and negate all of the bits to produce another k-decision list. It would look such as the following: $\neg c = \{(c_1, \neg b_1), \dots, (c_\ell, \neg b_\ell), \neg b\}$, which is a k-decision list.

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

- Occam's Razor states that the least complicated case is most likely true, but in this course we take it as being true. Therefore, we can show this to be PAC-learnable by finding a lower bound on the set such that it is the minimum

k-decision list. This can be done by showing that the dimensions of data needed is *finite.*
It can be trivially seen that the size of the concept class is $|c_k| = \mathcal{O}\left(n^k\right)$. Each of the bits belong in $b_i \in \{0,1\}$, and since there are ℓ of them it results in 2^ℓ different combinations. There are also $\ell!$ orderings of those bits. Finally, the size of the concept class that can describe the data is $2^{|c_k|}$. Therefore, the overall bound is represented by the product of these three which is $\mathcal{O}\left(2^{|c_k|} \cdot \ell! \cdot 2^\ell\right)$. However, the *maximum value* of ℓ is represented by $|c_k|$, so we can simplify our notation as $\mathcal{O}\left(2^{2|c_k|}\,|c_k|!\right)$.
Finally, the size of the number of examples needed in our set is logarithmically proportional to the size of the concept class, so by taking the log we get $\mathcal{O}\left(2n^k \log(2) + kn^k \log(n)\right)$ which is a *finite number* for any $k \geq 1$ and thus 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.)

- To do this, we need to construct our $\mathbf{x}$ vector and then our $\mathbf{w}$ vector where $\mathbf{w}$ is the weight vector defined as $\mathbf{w} = (\theta, w_1, w_2, \ldots, w_\ell)^T$ with θ being the bias. We can define our $\mathbf{x}$ vector as $\mathbf{x} = (1, x_1, x_2, \ldots, x_\ell)^T$ with the leading term being needed to support the bias. x_i can be defined by the value of b_i, where $x_i = -1$ for $b_i = 0$, and $x_i = 1$ for $b_i = 1$. From here, we can create a threshold function to "learn" the k-decision list, which is defined as $\mathbf{w}^T \cdot \mathbf{x} \geq 0$.
  In building $\mathbf{w}$, we can define each element of $w_i = b_i \cdot 2^{\ell+1-i}$ as to separate the individual data points. This would make the first few elements $\mathcal{O}\left(\pm 2^\ell\right)$ and the last elements $\mathcal{O}(\pm 2)$, where the ± comes about due to the individual bit b_i. Following this, we can look at the original 1-DL and if the overall/default bit is false, then we can set $\theta = -1$, else $\theta = 1$. This would overall be equivalent to the 1-DL and would make it liearly separable and learnable.

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.)

- We can say that the expection value of the function is defined as

$$E_{x\sim\bar{D}_m}(\mathbb{I}_{\{h(x)=f(x)\}}) = E_{x\sim\bar{D}_m}(\mathbb{Z}(x))$$

Where the second relation is the same but $\mathbb{Z}(x)$ was introduced to shorten notation. We can say that this relation is equivalent to

$$E_{x\sim\bar{D}_m}(\mathbb{Z}(x)) = \underbrace{\sum_x \bar{D}(x)}_{\frac{1}{m}\sum_{i=1}^m D_i(x)} \ \mathbb{Z}(x) = \frac{1}{m}\sum_x\sum_{i=1}^m D_i(x)\mathbb{Z}(x)$$

From here, we can switch the summations

$$= \frac{1}{m}\sum_{i=1}^m \underbrace{\sum_x D_i(x)\mathbb{Z}(x)}_{E_{x\sim D_i}} = \frac{1}{m}\sum E_{x\sim D_i}\mathbb{Z}(x) \quad (1)$$

From the above problem, we can see that it's an independent distribution of probabilities, though it is not uniform. Therefore, we have the total probability as being

$$P\left(\mathbb{Z}(x_i)\ \forall i\right) = \prod_{i=1}^m P\left(\mathbb{Z}(x_i)\right) \le \left(\frac{1}{m}\sum_{i=1}^m P\left(\mathbb{Z}(x_i)\right)\right)^m \quad (2)$$

Where (2) comes about because $\frac{1}{m}\sum_{i=1}^m x_i \ge (\prod_{i=1}^m x_i)^{\frac{1}{m}}$. By plugging (1) into (2), we get

$$= \left(\frac{1}{m}\sum_{i=1}^m E_{x\sim D_i}\left[\mathbb{Z}(x)\right]\right)^m < (1-\epsilon)^m$$

From here we can say that $(1-x) < e^{-x}$ as $(1-x)$ is a first order taylor series expansion of e^{-x}. Therefore, we reduce it to the following relation

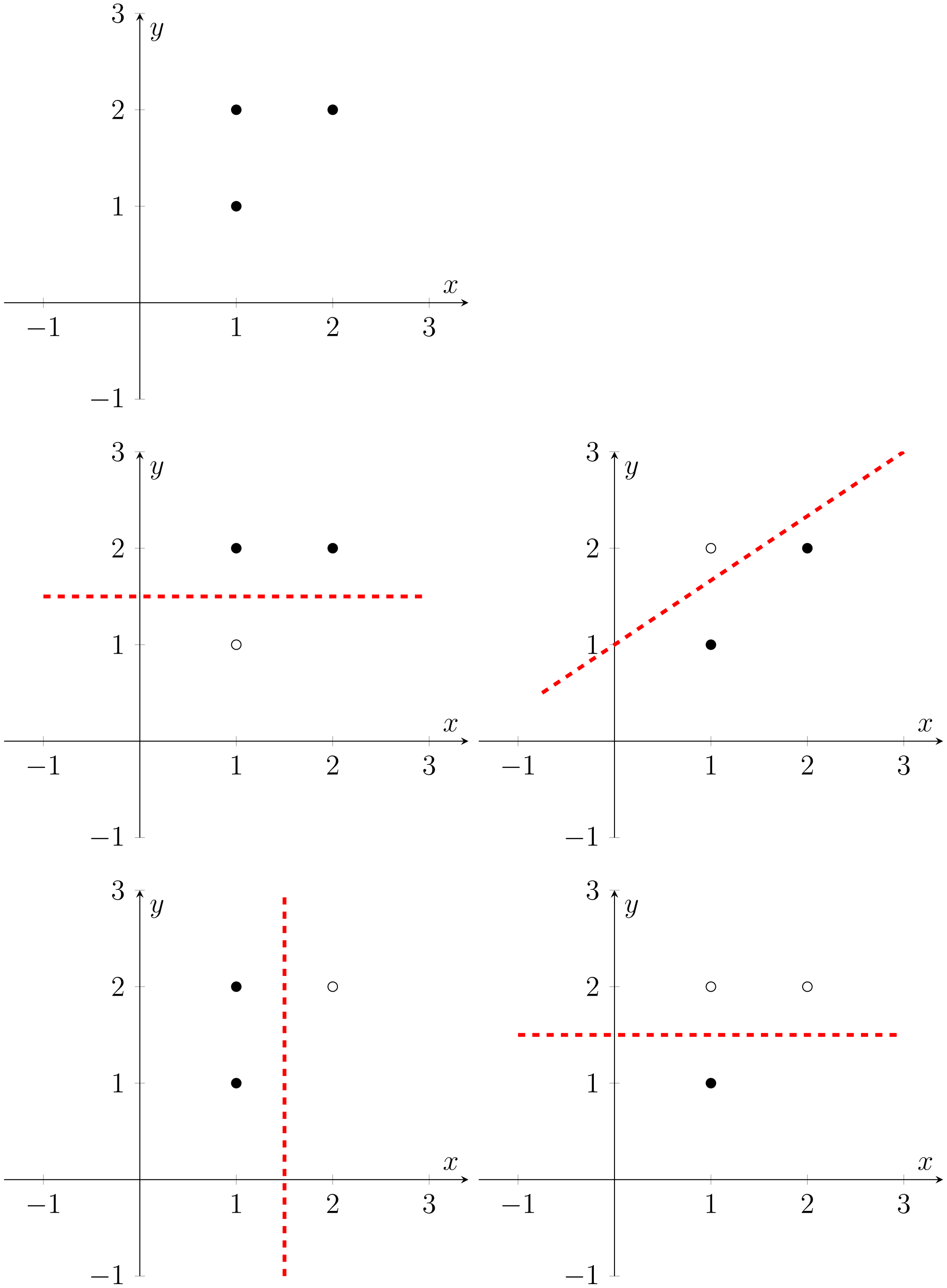
$$< e^{-m\epsilon}$$

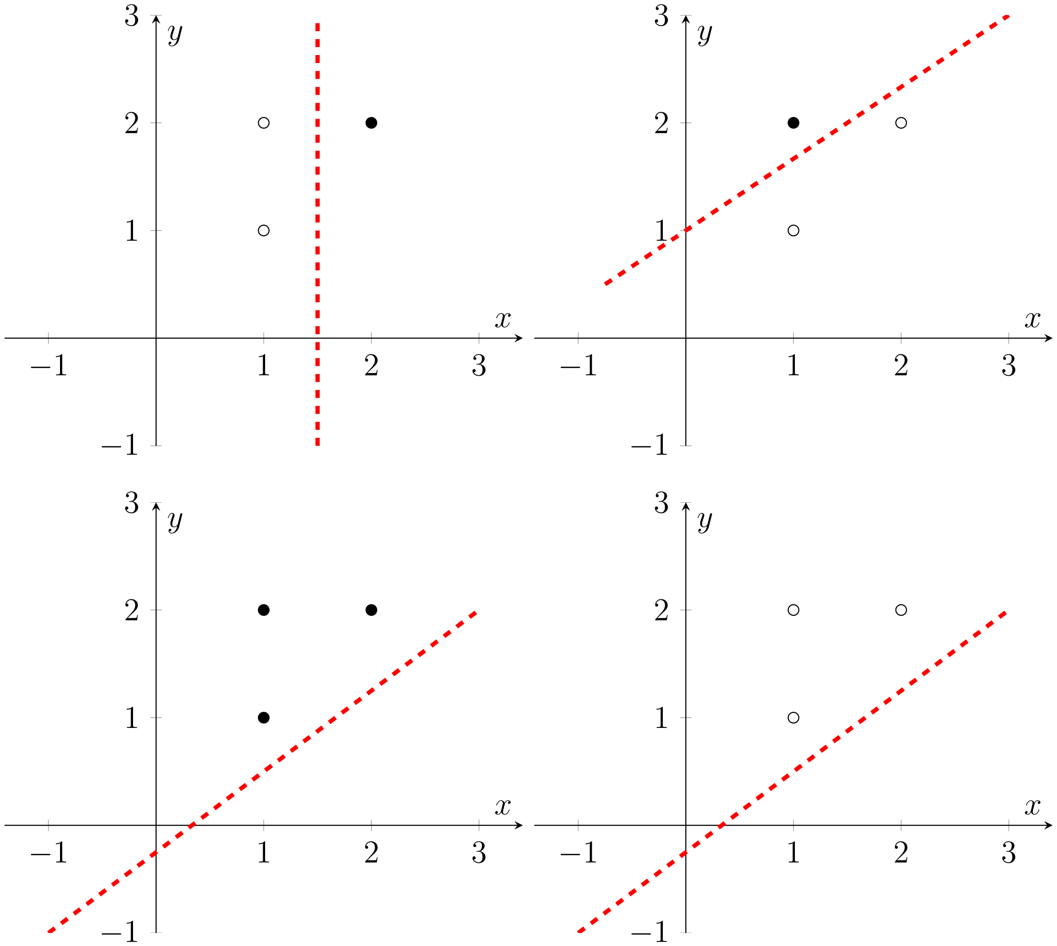
Where we can then apply the *Union Bound*, which states that the probability of at least one event happening is less than the sum of the probabilities of the individual events.

$$< |H|\, e^{-m\epsilon}$$

3 VC Dimension

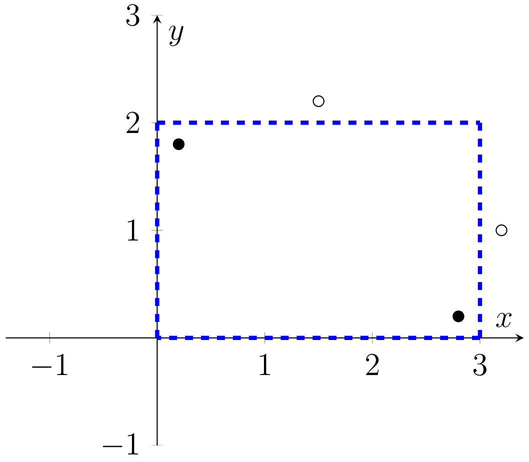
1. [5 points] Assume that the three points below can be labeled in any way. Show with pictures how they can be shattered by a linear classifier. Use filled dots to represent positive classes and unfilled dots to represent negative classes.





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.



- Need 4 because we need to define the outer regions of the box. Since the box is axis-aligned, that means that two of the sides are already defined so we need to restrict them such that the other two are as well. We then need to define the *inside* of the box by having *at least* two points to define the max. 5 points in this instance cannot be shattered.

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

- To generalize the logic in the above step, each dimension needs at least 2 points so that a line can shatter them. In the previous example, the blue line on the right shatters the points $(2.8, 0.2)$ and $(3.2, 1.0)$, as does the line for the other two points. Therefore, for the line to shatter the points, there must be two in each dimension. It is only restrictecd to 2 because the box is axis-aligned meaning that two of the lines run along the axis of the box and are therefore already defined.

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.

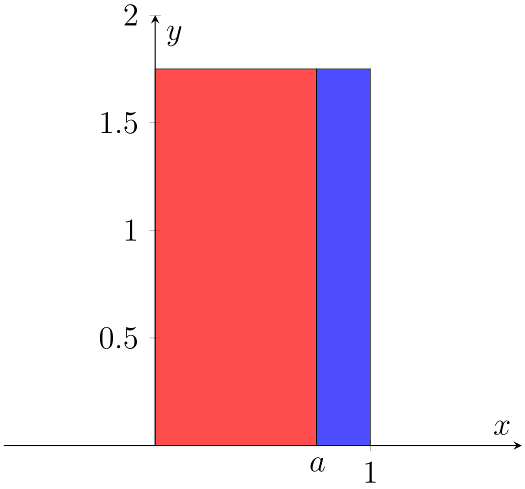
(a) [10 points] Show that for a finite hypothesis class $\mathcal{C}$, its VC dimension can be at most $\log_2(|\mathcal{C}|)$. (Hint: You can use contradiction for this proof. But not necessarily!)

- First, we can assume that $VC(|\mathcal{C}|) = d$, where d is the number of dimensions and $|\mathcal{C}|$ is the size of the hypothesis class. From here, we can come to the conclusion that there are 2^d *unique* hypotheses in the hypothosis space in order to shatter d instances. Therefore, we have the following

$$\begin{aligned} 2^d &\leq |\mathcal{C}| \\ d\log_2(2) &\leq \log_2(|\mathcal{C}|) \\ VC(|\mathcal{C}|) = d &\leq \log_2(|\mathcal{C}|) \end{aligned}$$

  Where the last relation comes about since $VC(|\mathcal{C}|) = d$.

(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.



- Such that the dividing line is at some point $(0, a)$ and $(a, 1)$ where one region is marked as positive and the other as negative. By confining the points to these two regions, there are an infinite number of points in each region but they can be classified by a single line.

(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 $\log_2(|\mathcal{C}|)$.

- We can use the example from question 1 of this section as a case. If you have N literals then you will have a concept class size of 2^N, which would have a VC-dimension of N if you take the size to be $\log_2\left(2^N\right)$.