CS 5350/6350: Machine Learining Fall 2015

Homework 3

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

General Instructions

- You are welcome to talk to other members of the class about the homework. I am more concerned that you understand the underlying concepts. However, you should write down your own solution. Please keep the class collaboration policy in mind.
- Feel free ask the instructor or the TAs questions about the homework.
- Your written solutions should be brief and clear. You need to show your work, not just the final answer, but you do *not* need to write it in gory detail. Your assignment should be **no more than 10 pages**. Every extra page will cost a point.
- Handwritten solutions will not be accepted.
- The homework is due by midnight of the due date. Please submit the homework on Canvas.
- Some questions are marked **For 6350 students**. Students who are registered for CS 6350 should do these questions. Of course, if you are registered for CS 5350, you are welcome to do the question too, but you will not get any credit for it.

Note

Do not just put down an answer. We want an explanation. No points will be given for just a statement of the results of a proof. You will be graded on your reasoning, not just on your final result.

Please follow good proof technique; what this means is if you make assumptions, state them. If what you do between one step and the next is not trivial or obvious, then state how and why you are doing what you are doing. A good rule of thumb is if you have to ask yourself whether what you're doing is obvious, then it's probably not obvious. Try to make the proof clean and easy to follow.

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: There are many options for functions here I will present one. Take $\phi(x_1, x_2) = [x_1^2, x_2^2, x_1]$. To show that this function is an appropriate transformation, we need to produce a w and a b such that the original function is equivalent to $sign(w^T \phi(c_1, x_2) - b)$. One such w is the vector $[-1, -1, 2]$ and $b = -r^2$. This makes

   $$\begin{aligned} sign(w^T \phi - r^2) &= sign(w_1 x_1^2 + w_2 x_2^2 + w_3 x_1 - b) \\ &= sign(-x_1^2 - x_2^2 + x_1 + r^2) \end{aligned}$$

   To see this in another way think about operations on the original function. Let's just try and manipulate the inequality into the $sign()$ function:

   $$\begin{aligned} x_1^2 + x_2^2 - 2x_1 &\leq r^2 \\ x_1^2 + x_2^2 - 2x_1 - r^2 &\leq 0 \end{aligned}$$

   but we want that $sign(w^T \phi) \geq 0$ not $sign(w^T \phi) \leq 0$ for positive labels. So we multiply by a negative

   $$-x_1^2 - x_2^2 + 2x_1 + r^2 \geq 0$$

   from this we get the above transformation and weight vector

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: My example does make it linearly separable. Take $w = [-1, 1, 0]$ and $b = -r^2$ and we are done.

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: No it does not. This function is equivalent to $sign(w_1x_1 + w_2x_2^2 - b) \geq 0$ assume it makes the function linearly separable then there exists some w and b for which this holds and the original inequality is also true for all choices of r and x_1, x_2. Let's take an example. Without loss of generality set $r = 2$. Now we have a hyperbola in $\mathbb{R}^2$ defined by $x_1^2 - x_2^2$. Let's take two points ripe for counter example and assume that there is a w which make it so. If we find a contradiction then we will have proved that this is impossible. I choose to use the two points $[-2, 0]$ and $[2, 0]$. Notice that both of these points are on the boundary (good counterexample place :). So take the definition of linearly seperable and let's see what that says about these two points:

$$
\begin{aligned}
w_1x_1 + w_2x_2^2 - b &\geq 0 \\
w_1x_1 &\geq b - w_2x_2^2 \\
-2w_1 &\geq b && \text{[-2,0]} \\
2w_1 &\geq b && \text{[2,0]}
\end{aligned}
$$

The only way this could work is if $w_1 = 0$ but that can't be because then we can't describe the original equation and we have our contradiction.

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: The Hypothesis space depends on the assumptions underlying what function of the gauges to take. If we assume that all gauges are relevant and needed then it is 5^{2^N} because there are 5 reactor states is which are a function of some combination of the gauges which have binary states. If we assume that they are not all relevant then it is 5^{3^N} because each gauge can be 1,0 or non-existent.

   b) [10 points] The ex-government employee whose job the robot is taking trains the robot at a nuclear reactor where there are 20 knobs by showing the robot a set

of knob positions for 5 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: The number of examples needed to attain this accuracy can be found by using:

$$\begin{aligned} m &\geq \frac{1}{\epsilon}\left(\ln|H| + \ln\left(\frac{1}{\delta}\right)\right) \\ &= \frac{1}{.001}\left(\ln\left(5^{2^{20}}\right) + \ln\left(\frac{1}{.01}\right)\right) \\ &= 2434724717.88 \end{aligned}$$

This means that we need 2,434,724,718 examples or more.

In the case where we assume that all gauges are not relevant the answer is greater than $99,951,397,526$

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, There are many examples of this.

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.

   (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: $\neg c = \{(c_1, \neg b_1), (c_2, \neg b_2), ..., (c_l, \neg b_l), \neg b\}$

   (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 show this we will show $\log(|H|)$ is bounded by a polynomial. The number of conjunctions with at most k terms using n variables is $\sum_{i=1}^{k} \binom{n}{i}$. This has an upper bound of n^k. Each decision has two possible leaves and can be ordered any way. The number of k-desicion lists is then $\mathcal{O}(n^k!)$. Since $\log n!$ has an upper bound of $n \log n$, $\log(|H|)$ is $\mathcal{O}(n^k \log n^k)$, thus k-decision lists are PAC learnable.

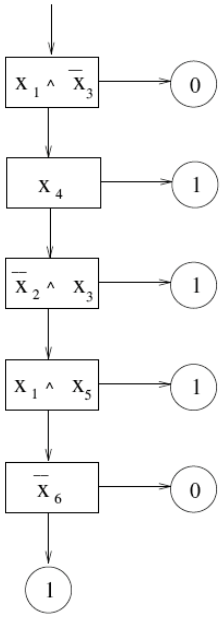


Figure 1: A 2-decision list.

(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: Because it is a 1-decision list each decision will have only one variable. Let n be the decision index with the first decision having $n = 0$. Assume the leaf nodes are $b_n \in \{1, -1\}$ and there are N total decisions. If a variable x_i appears in the list then $w_i = b_n 2^{-n}$ otherwise $w_i = 0$. The bias term is $b_N 2^{-N}$

4. [20 points, **CS 6350 students only**] Let X be a domain 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 are getting 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, $\bar{D}_m = \frac{1}{m}\sum_{i=1}^{m} D_i$. Fix an accuracy parameter $\epsilon \in (0, 1)$ and show that:

$$\mathbb{P}\left[\exists h \in \mathcal{H} \text{ s.t. } L_{\bar{D}_m, f}(h) > \epsilon \text{ and } L_{S,f}(h) = 0\right] \leq |\mathcal{H}| e^{-\epsilon m}$$

(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:

Proof. We begin by assessing a single hypothesis. We can define $L_{Di}(h) = \epsilon_i$ which says that the loss of our desired hypothesis on Dataset D_i is epsilon. With this we can look at the probability that our hypothesis on a particular example matches that

of $f(x_i)$. $\mathbb{P}[h(x_i) = f(x_i)] = 1 - \epsilon_i$. Because each of these distributions is drawn independently we know that $\mathbb{P}[A \cap B] = \mathbb{P}(A)\mathbb{P}(B)$ and:

$$\mathbb{P}[h(x_i) = f(x_i)] \forall x_i = \prod_i (1 - \epsilon_i)$$

Now apply the arithmetic-geometric mean inequality:

$$\begin{aligned}
\sqrt[m]{\prod_i (1 - \epsilon_i)} &\leq \frac{1}{m} \sum_i (1 - \epsilon_i) \\
\prod_i (1 - \epsilon_i) &\leq \left(\frac{1}{m} \sum_i (1 - \epsilon_i) \right)^m \\
&\leq \left(1 - \frac{1}{m} \sum_i (\epsilon_i) \right)^m \qquad (1)
\end{aligned}$$

Remember the definition of $e^{-x} = \lim_{m \to \infty} (1 - \frac{x}{m})^m$. With this let's define $\epsilon = \sum_i^m \epsilon_i$ now:

$$\sqrt[m]{\prod_i (1 - \epsilon_i)} \leq e^{-m\epsilon}$$

Now remember this is for a single hypothesis. To account for the entire hypothesis class use the Union bound:

$$\begin{aligned}
\mathbb{P}\left[\bigcup_{h_j \in \mathcal{H}} h_j(x) = f(x) \right] &\leq \sum_{j=1}^{|\mathcal{H}|} e^{-m\epsilon} \\
&\leq |\mathcal{H}| e^{-\epsilon m}
\end{aligned}$$

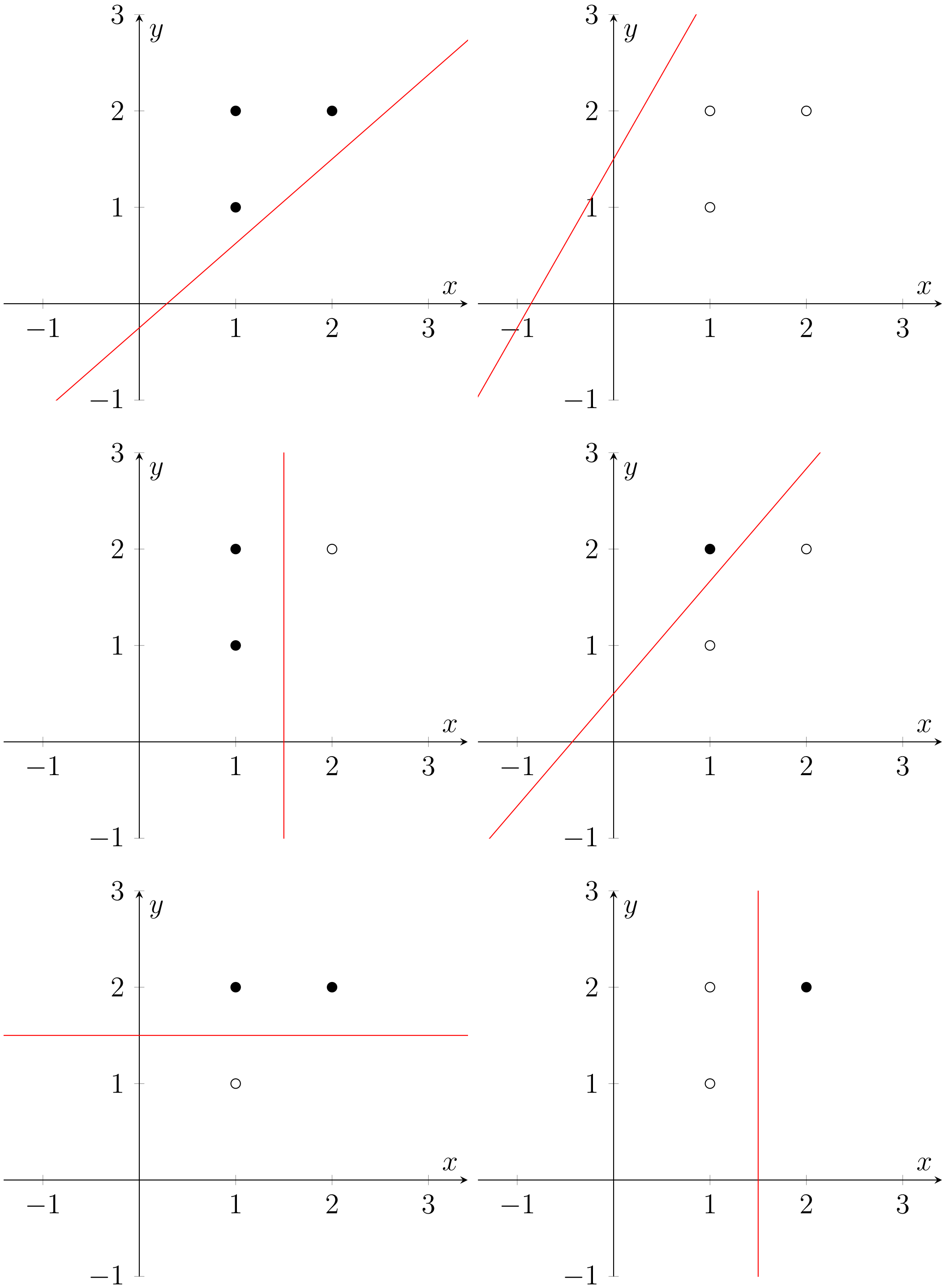
And we are done!

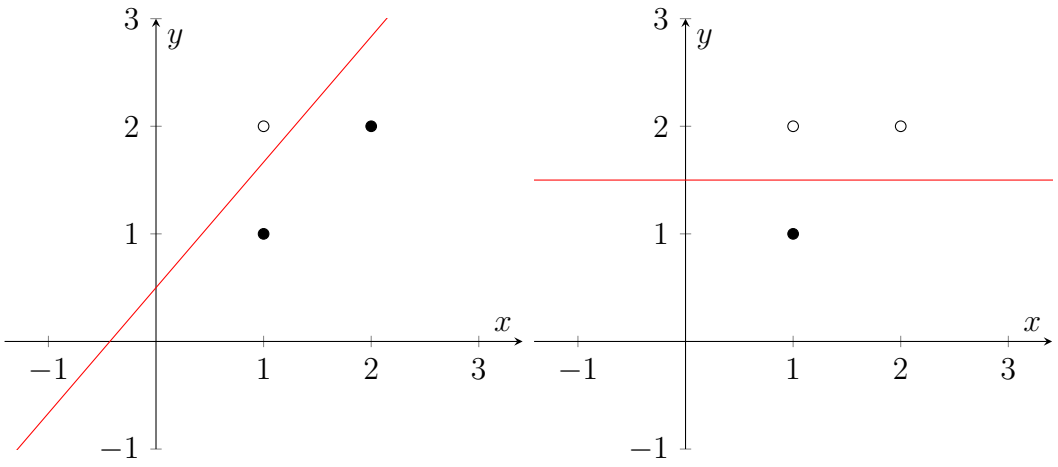
Note: (*) Remember the definition of $e^{-x} = \lim_{m \to \infty} (1 - \frac{x}{m})^m$ to get this inequality step.

□

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.


y
x
−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.

   Solution: To prove this requires two steps. The first is to demonstrate that there is a set of four points which can be shattered by axis aligned rectangles. Then we need to show that no set of five points can be shattered by these rectangles.

   For the first step, you could draw out an example data set that can be shattered. However, we will construct one with words. Draw a rectangle and take four points, where each point lies on a different edge of the rectangle. Then this data set is shatterable because each point lies on a different boundary of the rectangle so I can draw a rectangle the touches any set of these points and excludes an arbitrary number of the other ones. Thus making it shatterable. This gives us that $VC(H^2_{rec}) \geq 4$

   Now we need to show how NO set of five points can be shattered by this classifier. Take any set of points and find the max and min along each axis. Use these to define a box. Then because we have taken all of our extreme points we know that the fifth point must lie inside of our box. Make this fifth point negative. Therefore, no five points can be shattered by four d rectangles

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

   Solution: Draw a d-rectangle and take $2d$ points, where each point lies on a different d-rectangle of the rectangle (Note: We can define a higher dimensional rectangle by its d-rectangles and there are $2d$ of these for a d dimensional rectangle. It is analogous to a face in 3D and an edge in 2D). Then this data set is shatterable because each point lies on a different boundary of the d-rectangle so I can draw a d-rectangle the touches any set of these points and excludes an arbitrary number of the other ones. Thus making it shatterable.

Now to show that we cannot shatter $2d+1$ points take again a d-dimensional box set the limits of the d-dimensions as the max and min along each axis. I can do this for $2d$ points. Then the last point must lie inside my box by definition and we can make it negative showing that it can't be shattered.

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: (Direct proof) If we take some set of possible inputs $S \in \mathcal{X}$ then $|C_S| \leq |C|$ because we have restricted the domain of our functions. Now we know that $|C| \leq 2^{|C|}$ by the definition of shattering. Therefore $|C_S| \leq |C| \leq 2^{|C|}$. Now because C_s has fewer possible hypotheses than C we know that $VC(C_s) \leq VC(C)$. This tells us that examining the VC dimension will preserve this inequality. Now simply notice that the VC dimension is the size of the set of points that can be shattered. To extract point size from the hypothesis space we use the log so: $VC(C_S) \leq VC(C) \leq \log_2(|C|)$

   (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: The class of right facing intervals is infinite because I can create a different classifier starting at any possible point on [0,1] and [0,1] is isomorphic to $\mathbb{R}$. To see this take $\tan(x)$ $x \in (0, 1)$. It has VC dimension 1 because I can classify any point by choosing my interval to the right or left of the point. But I cannot classify two points because we can always choose the left most point to be positive and the right most to be negative.

   (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: Take the set of functions $x >= .5$ and $x <= .5$. There are two functions here. The elements grater than .5 can be labeled positive or negative therefore $|\mathcal{C}| = 2$. This class has VC dimension 1 because any point can be placed to the left or right of .5 and for any labeling I can choose what value my function outputs. The VC dimension cannot be 2 because we can always choose the left or right most point to be positive and the right or left most to be negative.