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.)

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.

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.

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.)

   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?

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.

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.

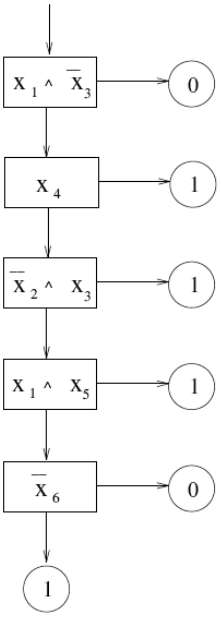


Figure 1: A 2-decision list.

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 $c = \{(c_1, b_1), \ldots, (c_l, b_l), b)$.

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

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, $\bar{D}_m = \frac{1}{m}\sum_{i=1}^{m} D_i$.

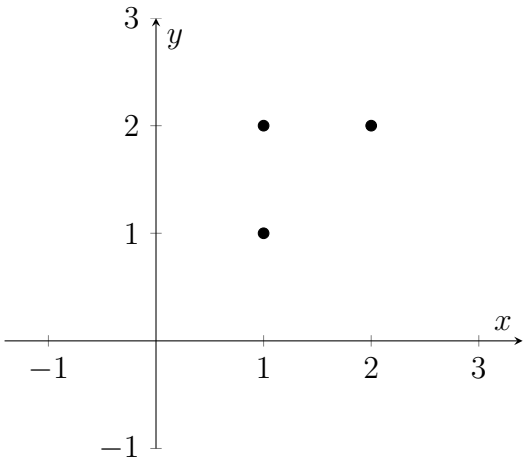
   Let $h \in \mathcal{H}$ be a classifier that gets zero error on the training set. That is, for every example $x_i \in X$, we have $h(x_i) = f(x_i)$. Show that, for any accuracy parameter $\epsilon \in (0, 1)$, the probability that the expected error of the learned classifier h is greater than ϵ is no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

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.

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

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!)

   (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.

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