CS 5350/6350: Machine Learning Fall 2018

Homework 4

Abhinav Kumar (u1209853)

Handed out: October 26, 2018
Due date: November 13, 2018

1 PAC Learning

1. A factory assembles a product that consist of different parts. Suppose a robot was invented to recognize whether a product contains all the right parts. The rules for making products are very simple: 1) you are free to combine any of the parts as they are 2) you may also cut any of the parts into two distinct pieces before using them.

   You wonder how much effort a robot would need to figure out the what parts are used in the product.

   - [5 points] Suppose that a naive robot has to recognize products made using only rule 1. Given N available parts and each product made out of these constitutes a distinct hypothesis. How large would the hypothesis space be? Brief explain your answer.

     For each part because of rule 1, we have only 2 options - either the part is present or absent. Hence, size of hypothesis space = total number of combinations possible $= 2^N$

   - [5 points] Suppose that an experienced worker follows both rules when making a product. How large is the hypothesis space now? Explain.

     For each part because of both the rules, we have only 3 options - either the part is absent, present without divided, present with divided. An important point to note here is that the parts are obtained in exactly the same way irrespective of the cut (as discussed on instructure). Hence, size of hypothesis space = total number of combinations possible $= 3^N$

   - [10 points] An experienced worker decides to train the naive robot to discern the makeup of a product by showing it the product samples he has assembled. There are 6 available parts. If the robot has to learn any product at 0.01 error with probability 99%, how many examples would the robot have to see?

     Let H be any hypothesis space. With probability $1 - \delta$, a hypothesis $h \in H$ that is consistent with a training set of size m will have an error $< \epsilon$ on future

examples if $m > \frac{1}{\epsilon}\left(ln|H| + ln\frac{1}{\delta}\right)$
Assuming both the rules are followed, $|H| = 3^6$ and $\epsilon = \delta = 0.01$
Substituting the values, $m > 1119.68$ and therefore $m = 1120$ examples.

2. [20 points] Consider the class C of concepts of the form $(a \leq x \leq b) \wedge (c \leq y \leq d)$ where a, b, c, d are integers in the interval $(0, 20)$. Each concept in this class consists of a rectangle with integer valued boundaries and labels all points inside the rectangle as positive and everything outside as negative.

   Give an upper bound on the number of randomly drawn examples needed to ensure that for any function c in the set C, a consistent learner that uses $H = C$ will, with probability 99% produce a classifier whose error is no more than 0.01.

   Hint: To answer this question, you could use the fact that in a plane bounded by the points $(0, 0)$ and (n, n), the number of distinct rectangles with integer-valued boundaries in the region is $\left(\frac{n(n+1)}{2}\right)^2$.

   Let H be any hypothesis space. With probability $1 - \delta$, a hypothesis $h \in H$ that is consistent with a training set of size m will have an error $< \epsilon$ on future examples if $m > \frac{1}{\epsilon}\left[ln|H| + ln\frac{1}{\delta}\right]$

   Here,ize of the concept class $|H| = \left(\frac{n(n+1)}{2}\right)^2$

   Hence, $m > \frac{1}{\epsilon}\left[ln\left(\frac{n(n+1)}{2}\right)^2 + ln\frac{1}{\delta}\right]$
   We have, $n = 20$ and $\epsilon = \delta = 0.01$
   Substituting the values, $m > 1529.93$ and therefore $m = 1530$ examples.

2 Shattering

[15 points] Suppose we have a set X_n consists of all binary sequences of a length n. For example, if $n = 3$, the set would consist of the eight elements {`000`, `001`, `010`, `011`, `100`, `101`, `110`, `111`}.

Consider a set of functions H_n that we will call the set of *templates*. Each template is a sequence of length n that is constructed using `0`, `1` or `-` and returns $+1$ for input binary sequences that match it and -1 otherwise. While checking whether a template matches an input, a `-` can match both a `0` and a `1`.

For example, the template `-10` matches the binary strings `010` and `110`, while `-1-` matches all strings that have a `1` in the middle position, namely `010`, `011`, `110` and `111`.

Does the set of templates H_n shatter the set X_n? Prove your answer.

It is assumed that the set of templates H_n can also consist of n bits and can't exceed n bits.

Lets check with $n = 1$. Then, $X_1 = \{0, 1\}$. There are four possibilities of labelling -

Label of 0	Label of 1	Matching function
+1	+1	-
-1	+1	1
+1	-1	0
-1	-1	Nothing possible

Clearly set of templates H_1 cannot shatter X_1 itself. So, clearly in general H_n cannot shatter X_n.

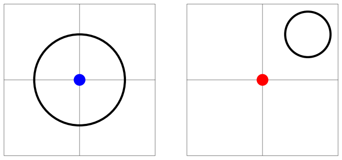
3 VC Dimensions

1. Consider learning problems where examples are points in the two dimensional plane. What is the VC dimension of the following concept classes? (In each case you need to prove your answer by showing both the upper and lower bounds.)

   (a) [15 points] The concept class H_c consisting of circles, with points strictly outside being negative.

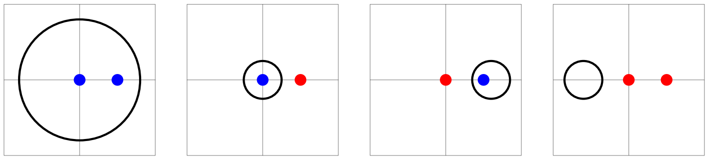
   It is assumed that there is only a single circle available at each point. Not multiple circles. Note that the blue point corresponds to figure with label +1 while red corresponds to points with label -1.

   We start with a single point in a plane. There can be only two cases - either the point is labelled +1 or -1



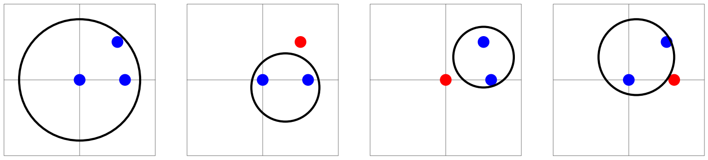
   Clearly, these can be correctly classified. Hence, VC dimension is atleast 1.

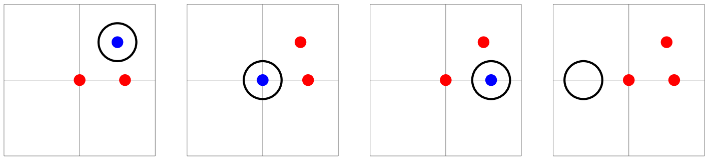
   Lets go to two points case. Each point can be labelled as +1 or -1. So, there are 4 such cases.



   Clearly, these can be correctly classified. Hence, VC dimension is atleast 2.

   Lets go to three points case. Each point can be labelled as +1 or -1. So, there are 8 such cases.

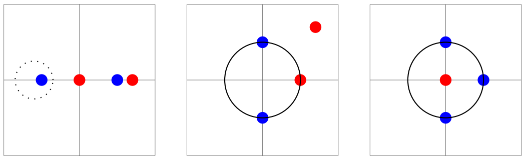




Clearly, these can be correctly classified. Hence, VC dimension is atleast 3.

Lets go to four points case. Each point can be labelled as +1 or -1. There are two arrangement of points possible -

- Atleast 3 points are collinear - Label points which are collinear as +1 and -1 in an alternate fashion.
- 3 points are non-collinear and 4th point lies outside the circle formed by those 3 points - Assign one of the points in the circle and point outside as -1 and remaining two points as +1.
- 3 points are non-collinear and 4th point lies inside the circle formed by those 3 points - Assign all points on the circle as +1 and point inside as -1.



Hence, any arrangement of points in dimension 4 can not be shatterred by this class. Hence, final VC dimension is 3

(b) [15 points] The concept class H is defined as follows: A function $h \in H$ is specified by two parameters a and b. An example $\mathbf{x} = \{x_1, x_2\}$ in $\Re^2$ is labeled as + if and only if $x_1 \geq a$ and $x_2 \leq b$ and is labeled $-$ otherwise.

For example, if we set $a = 1, b = 4$, the grey region in figure 1 is the region of $\mathbf{x} = \{x_1, x_2\}$ that has label +1.

What is the VC dimension of this class?

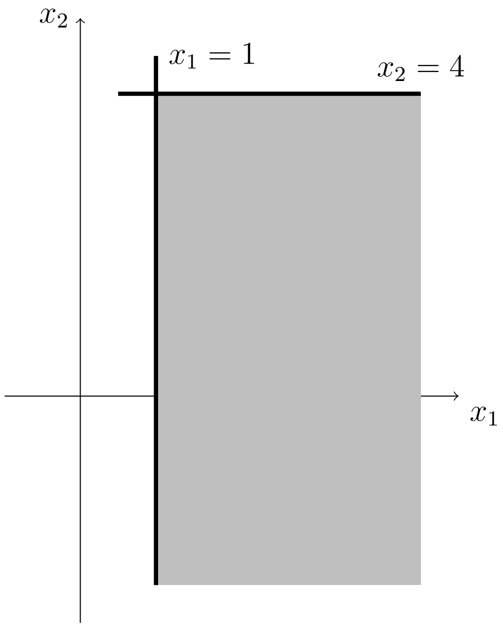
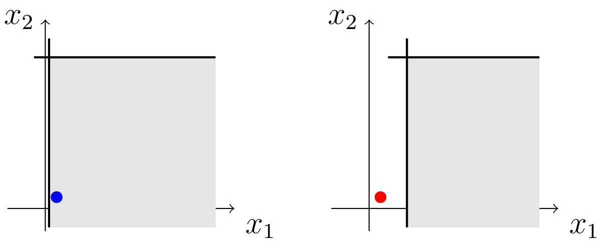


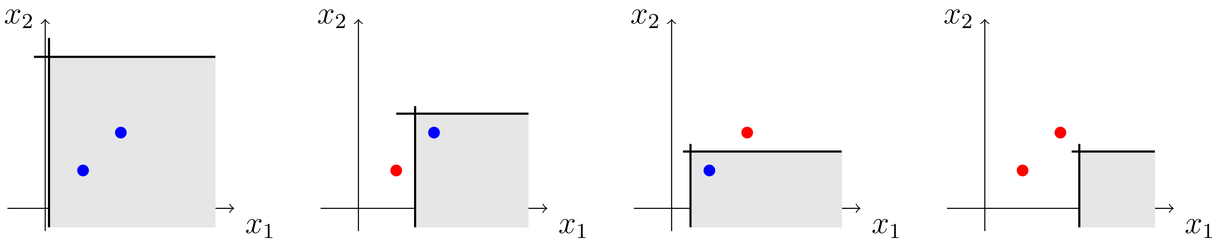
Figure 1: An example with $a = 1, b = 4$. All points in the gray region (extending infinitely) shows the region that will be labeled as positive.

The blue point corresponds to figure with label +1 while red corresponds to points with label -1. Now, start with a single point.



Clearly, these can be correctly classified. Hence, VC dimension is atleast 1.

Lets go to two points case. Each point can be labelled as +1 or -1. So, there are 4 such cases.

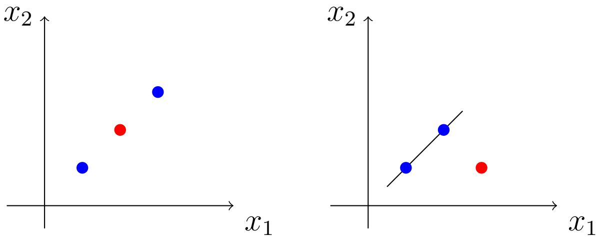


Clearly, these can be correctly classified. Hence, VC dimension is atleast 2.

Lets go to three points case. We have two arrangements of points possible -

- Three points are collinear- Label point in the middle as -1 and others two as +1.

- Three points form a triangle - Label point which lies towards the rightmost of the line joining the other two points as -1 and other as +1



Hence, points in dimension 3 can not be shatterred by this class of functions. Hence, final VC dimension is 2

2. [**For 6350 Students,** 15 points] Let two hypothesis classes H_1 and H_2 satisfy $H_1 \subseteq H_2$. Prove: $VC(H_1) \leq VC(H_2)$.
   Let $VC(H_1) = d$. Then, $\exists$ an arrangement of d points which can be shatterred by H_1.

   Since, $H_1 \subseteq H_2$ implies that every function in H_1 is also present in H_2 and so the arrangement of d points which can be shatterred by H_1 is also shatterred by H_2.

   Hence, $VC(H_2)$ is atleast d or, $d \leq VC(H_2)$.
   But, $VC(H_1) = d$.
   Hence, $VC(H_1) \leq VC(H_2)$

4 AdaBoost

[15 points] You are given the following examples in the table below

$\mathbf{x} = [x_1, x_2]$	y
$[1, 1]$	-1
$[1, -1]$	1
$[-1, -1]$	-1
$[-1, 1]$	-1

Suppose you are also given the following 4 weak classifiers (i.e. rules of thumb).

$$\begin{aligned} f_a(\mathbf{x}) &= \operatorname{sgn}(x_1) \\ f_b(\mathbf{x}) &= \operatorname{sgn}(x_1 - 2) \\ f_c(\mathbf{x}) &= -\operatorname{sgn}(x_1) \\ f_d(\mathbf{x}) &= -\operatorname{sgn}(x_2) \end{aligned}$$

1. Step through the full AdaBoost algorithm by choosing f_t from the above weak classifiers in each round. Remember that you need to choose a hypothesis from f_a, f_b, f_c, f_d whose weighted classification error is **less than half**.

2. At the end, state the final hypothesis $H_{final}(x)$.

To get you started, we have chosen f_a as the first hypothesis and show the values of ϵ_1, α_1, Z_1, D_1 in the table below. Fill in the values of D_2. For ease of grading, please follow this template below: Report the hypothesis you choose, its predictions for all the examples, and the values of ϵ_t, α_t, Z_t, D_t, and D_{t+1} for *three subsequent rounds*

Round 1: Choose $h_1(\mathbf{x}) = f_a(\mathbf{x}) = \text{sgn}(x_1)$.

$\mathbf{x} = [x_1, x_2]$	y_i	$f_a(x)$	D_1	$D_1(i)y_ih_t(\mathbf{x_i})$	D_2
$[1, 1]$	-1	1	1/4	-1/4	
$[1, -1]$	1	1	1/4	1/4	
$[-1, -1]$	-1	-1	1/4	1/4	
$[-1, 1]$	-1	-1	1/4	1/4	

$\epsilon_1 = 1/4, \alpha_1 = \frac{\ln 3}{2}, Z_1 = \frac{\sqrt{3}}{2}$

In each round, we have to choose one hypothesis without using any arithmetic operation. $f_a(\mathbf{x})$, $f_b(\mathbf{x})$ and $f_d(\mathbf{x})$ are hypothesis whose error is less than half.

Let $f_1 = f_a(\mathbf{x}), f_2 = f_b(\mathbf{x}), f_3 = f_d(\mathbf{x})$ and $f_4 = f_a(\mathbf{x})$

$\mathbf{x}$	y_i	$f_1(\mathbf{x})$	D_1	$f_2(\mathbf{x})$	D_2	$f_3(\mathbf{x})$	D_3	$f_4(\mathbf{x})$	D_4
$[1, 1]$	-1	1	0.25	-1	1/2	-1	0.3	1	0.16
$[1, -1]$	1	1	0.25	-1	1/6	1	0.5	1	0.27
$[-1, -1]$	-1	-1	0.25	-1	1/6	1	0.1	-1	0.50
$[-1, 1]$	-1	-1	0.25	-1	1/6	-1	0.1	-1	0.05

$\epsilon_1 = 0.25, \alpha_1 = 0.549, Z_1 = 0.866$
$\epsilon_2 = 0.16, \alpha_2 = 0.804, Z_2 = 0.745$
$\epsilon_3 = 0.10, \alpha_3 = 1.098, Z_3 = 0.6$
$\epsilon_4 = 0.16, \alpha_4 = 0.804, Z_4 = 0.745$

Final Prediction $= \text{sgn}\left[0.549\,\text{sgn}(x_1) + 0.804\,\text{sgn}(x_1 - 2) - 1.098\,\text{sgn}(x_2)\right]$
A code has been written which does these calculations.