Assignment 5: Learning Theory Solutions

10-601: Machine Learning (Spring 2012)

TA: Robert Fisher

Out: Wed, Feb 29

Due at the beginning of class on Wed, March 7

- We prefer typed solutions, but if you cannot do that, please make sure your handwriting is neat and legible. We cannot give credit to any part of the solution that is not legible.
- Answer the questions in the order they are stated.
- Policy on collaboration: Please refer to course website: page 'Policies'.
- Policy on late homework: Please refer to course website: page 'Policies'.
- For questions and clarifications, contact Robert (rwfisher@cs.cmu.edu).
- Be sure to write your Andrew ID and name on the top of every page.
- Robert will be holding extra office hours on Monday, March 5 from 3:30-5:00 and Tuesday, March 6 from 1:00-3:00 for anyone needing help with this homework. He also has regular office hours on Thursday, March 1 from 1:30-3:00.

Warning: You may find some of the problems similar to the ones in the textbook. However, these problems are likely to be different! Make sure you read what's written on this assignment.

## Q1. (30 points)

Consider a learning problem in which X = $\Re$ is the set of real numbers, and C = H is the set of intervals over the reals, H={(a≤x≤b)}| a, b}.

(a) Give a lower bound on the probability that a hypothesis consistent with m examples of this target concept will have an error at least ε. Solve this using the VC dimension.

For a line interval, VC(H) =2, because any three points in a row, if they have a +;-;+ labeling, cannot be classified with a single hypothesis.
We know that if number of samples is above 1/ε (4 log(2/δ)+8VC(H)log(13/ε)), then with probability 1-δ we can say our error is below ε.

$m>1/\varepsilon\ (4\log(2/\delta)+8VC(H)\log(13/\varepsilon))$

$\leftrightarrow (m^*\varepsilon)- 8VC(H)\log(13/\varepsilon)> 4\log(2/\delta)$

$\leftrightarrow 2^{(1/4)((m^*\varepsilon)-8VC(H)\log(13/\varepsilon))} > 2/\delta$

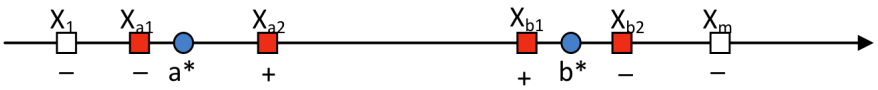
$\leftrightarrow 2^{-(1/4)((m^*\varepsilon)-8VC(H)\log(13/\varepsilon))} < \delta/2$

$\leftrightarrow 2^{1-(1/4)((m^*\varepsilon)-8VC(H)\log(13/\varepsilon))} < \delta$

So in this example, the probability is at least $2^{1-(1/4)((m^*\varepsilon)-8VC(H)\log(13/\varepsilon))}$

(b) Find a simpler way to solve this, based on the geometric properties of the problem and ignoring the VC dimension. For this problem, consider only the error of the most specific consistent hypothesis. Also, assume that the data only comes from an interval of length L, such that the decision boundary of the true concept is fully contained in this interval. The data is uniformly distributed over this interval. Your bound needn't be perfectly tight, but say so in your analysis if you think the bound is loose.

Imagine the m points (X1,..Xm) that are consistent with the hypothesis [a*,b*].




Consider 4 important points:
Xa1 the largest sample which is smaller than a*.
Xa2 the smallest sample which is larger than a*.
Xb1 the largest sample which is smaller than b*.
Xb2 the smallest sample which is larger than b*.

If we are considering the most specific hypothesis, our error will be the probability that a data point falls in the space between a* and Xa2, and b* Xb1. Denote c = |a*-Xa2| and d = |b*-Xb1|. Under our distribution, the probability that some point falls in the error interval on the left is precisely c/L. Likewise it is d/L for the right.

The probability that the total error is at least ε can be lower bounded by the probability that one of these two intervals has at least ε probability mass in it. Likewise it can be upper bounded by the probability that both error intervals contain epsilon mass. The probability that one interval represents an error rate of at least ε is the probability that $\varepsilon \leq c/L$ or $\varepsilon L \leq c$. This is exactly the probability that none of our m training points fell into this interval of length c. The probability that one point does not fall in this interval is (1-c/L) = (1 – εL/L) = (1- ε). Likewise the probability that none of the m points fall into this interval is $(1 - \varepsilon)^m$. This gives us the following bounds on the probability that (ε < error):

$$\mathbf{(1 - \varepsilon)^m \leq \delta \leq 2(1 - \varepsilon)^m}$$

For example, when given a training set of size 50, the probability that the error rate of the final classifier is greater than 5% is somewhere in the range [7.7%, 15.4%].

## Q2. (30 points)

Consider the space of instances corresponding to all points in the x, y plane. Derive the VC dimension of the following hypothesis spaces:

(a) $H_R$ = the set of all rectangles parallel to the axes, in the x, y plane. That is, $H_R$ = {((a<x<b) (c<y<d)) | a, b, c, d}.

VC(rectangles) = 4.

Consider the points (-1, 0), (1, 0), (0, 1), (0,-1). By taking the $2^4$ rectangles with –a, b, -c and d in {0.5, 1.5}, we can shatter these points.

However, we cannot shatter 5 points for the following reason. Consider 5 generic points

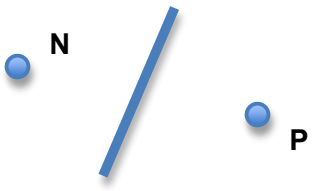
p1 = (x1; y1); p2 =(x2; y2); p3 = (x3; y3); p4 = (x4; y4); p5 = (x5; y5). Let lx = argmin(xi), ux = argmax(xi), ly = argmin(yi), uy = argmax(yi), and let pinside = (xinside,yinside) denote a point pi  {p1, p2, p3, p4, p5} \  {plx; pux;

ply ; puy}. Then it is impossible to realize a classification where $p_{lx}$, $p_{ux}$, $p_{ly}$, and $p_{uy}$ are positive but $p_{inside}$ is negative.

(b) $H_N$: This hypothesis class operates in r-dimensional real space and is parameterized by two points in that space, P = ($P_1$, $P_2$… $P_r$,) and N = ($N_1$, $N_2$… $N_r$,). All data-points that are closed to P than to N (in terms of Euclidean distance) are labeled 1, and all points closer to N are labeled 0.

Hint: Think about this problem in 1 or 2 dimensions, and then try to generalize using a result from class or from the textbook.

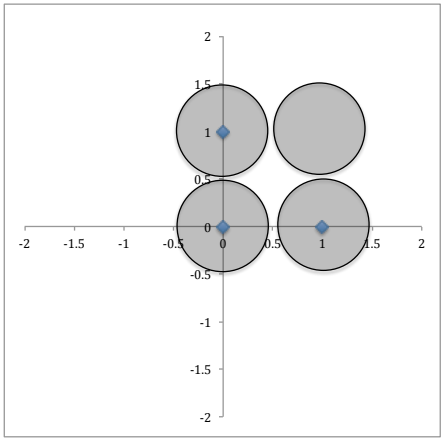
We can consider plotting the Voronoi diagram of P and N in two dimensions. This would look like this:

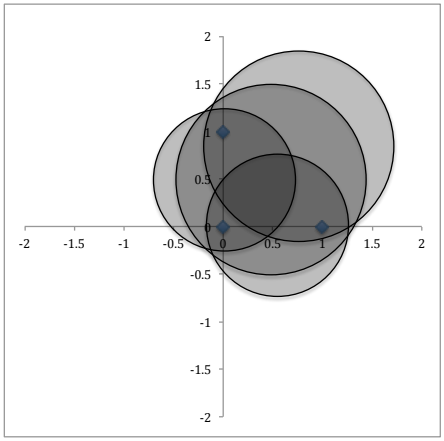



We see that we are dealing with a linear decision boundary. As discussed in class and in the textbook, the VC-dimension of this classifier in r-dimensional space is r + 1.

(c) $H_C$ = Circles in the x, y plane. Points inside the circle are classified as positive examples.

VC(circles) = 3.
By letting S3 = {(0,0),(1,0),(0,1)}, we can easily realize all the dichotomies of S using hypotheses from $H_C$, as shown in the following figure. This means VC ($H_C$) ≥ 3.

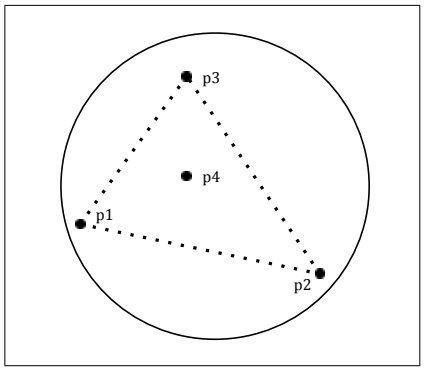


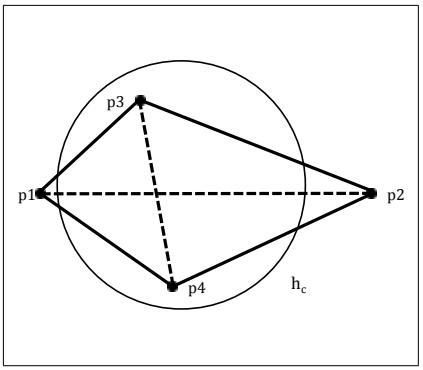

Now we show that V C(HC) < 4.
For any set of 4 points S4 = {pi = (xi,yi) | i = 1,2,3,4}, if any 3 of the points are co-linear, it is easy to see that it is impossible for any circle to classify

the middle point as negative while classifying the other two as positive. If no three of them are co-linear, we could form a triangle using p1, p2, and p3 as vertices. Since no three vertices are co-linear, p4 is either inside or outside of the triangle. We discuss both cases in turn.

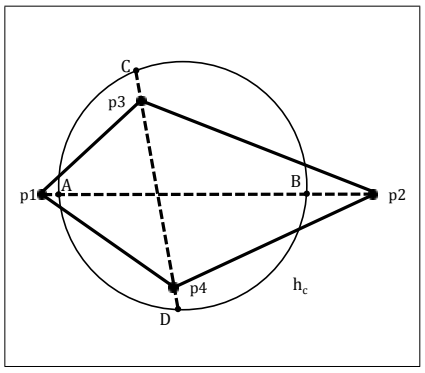
Case 1: If p4 lies inside the triangle, it is clearly impossible for any circle to classify p4 as negative and all the rest as positive. This is because if all the 3 points lie in a circle, so do their convex combinations. See Figure (a) for illustration.



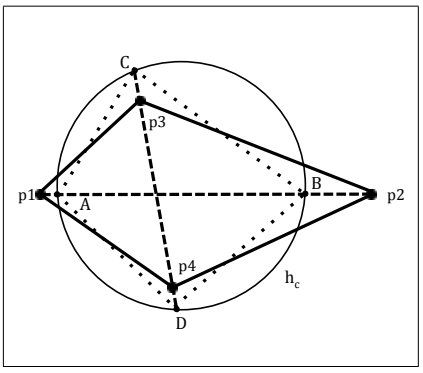

(a) $p_4$ lies inside the triangle $\triangle p_1 p_2 p_3$




(b) A circle $h_c$ that classifies $p_1, p_2$ as negative and $p_3, p_4$ as positive




(c) Extend $\overline{p_3 p_4}$ so that it intersects with $h_c$




(d) Connect $A, D, B, C$ into a cyclic quadrilateral

Case 2: If p4 lies outside the triangle, the points can form a quadrilateral as shown in Figure 3b. Since the degree of all the 4 angles in a quadrilateral sum up to 360◦, there must be a pair of diagonal vertices whose angles sum up to be greater than or equal to 180◦. Without loss of generality, let's assume that p1 and p2 are such a pair of diagonal vertices. Now we are going to show that it is impossible to find a circle that classifies both p1 and p2 as negative while classifying the other two as positive.

We prove by contradiction. Assuming that there is a circle $h_c$ that classifies p3,p4 as positive and p1,p2 as negative (see Figure b), we will show that ∠p1 + ∠p2 < 180◦. Since p3 and p4 are both in the circle, the line p3-p4 should also lie entirely in the circle. We extend line p3-p4 so that it intersects with $h_c$. On the other hand, since p1 and p2 are not in the circle

but there is some point p on the line p1p2 (i.e. the intersection point of p1p2 and p3p4) is in the circle, so p1p2 must intersect the circle. Label the intersections of $h_c$ and p1p2 as A and B, and label those of hc and p3p4 as C and D (see Figure c). Connect A, D, B, and C to form a quadrilateral. Notice that ADBC is a cyclic quadrilateral, and we know that the opposite angles of a cyclic quadrilateral always sum up to 180◦. That is to say, ∠A + ∠B = 180◦. However, since CD > p3p4 while the distance between A and CD is less than that between p1 and p3p4, it is easy to see that ∠p1 < ∠A. Similarly we have ∠p2 < ∠B. Therefore, we have ∠p1 + ∠p2 < ∠A + ∠B = 180◦. This contradicts the assumption that ∠p1 + ∠p2 ≥ 180◦.
**Thanks to Xiao Xinpan for solution**

You don't have to give a formal proof. An informal explanation and/or demonstration is enough. In each case, if you cannot find the exact VC dimension, give the largest set you can find that is shattered.

## Q3. (20 points)

Consider a class C of concepts of the form (a ≤ x ≤ b)  (c ≤ y ≤ d), where a, b, c and d are integers in the interval [0, 99]. Note each concept in this class corresponds to an axis-parallel rectangle with integer-valued boundaries on a portion of the x, y plane.

Hint: Given a region in the plane bounded by the points (0, 0) and (n-1, n-1) the number of distinct rectangles with integer-valued boundaries within this region is:

$$\left(\frac{n(n+1)}{2}\right)^2$$

(a) Give an upper bound on the number of randomly drawn training examples, sufficient to assure that for any target concept c in C, any consistent learner using H=C will with probability 90%, output a hypothesis with error at most 10%.

In this example, the |H| is (100*101/2)^2 = 25502500. Since we want 90% confidence, δ=0.1 and since our error should be below 10%, we have ε=0.1. Plugging in the values into equation 7.2 we will get
M ≥1/ε(ln |H|+ ln(1/δ))=10*(ln(25502500)+ln(10)) = 193.56.
So we need at least 194 examples for learning with the desired level of accuracy and confidence.

(b) Now suppose the rectangles boundaries a, b, c and d take on real values instead of integer values. Update your answer to the first part of this question

Now our hypothesis space size is infinite, so we need to calculate the VC dimension. We claim that VC dimension of this hypothesis space is 4 (See question 2.a for proof). We still have ε=0.1 and δ=0.1. We use equation 7.8 and plug in the values:
$M \geq 1/\varepsilon\ (4\log(2/\delta) + 8VC(H)\log(13/\varepsilon)) = 10*(4\log(20)+8*4*\log(130)) = 2420.034$.
You can see that now we will need at least 2421 samples to get the desired level of accuracy and confidence.

## Q4. (20 points)

(a) Let us consider two arbitrary hypothesis spaces for concept learning, $H_1$ and $H_2$, which operate over the same instance space X. The VC-dimensions of these spaces are $VC(H_1) = a$ and $VC(H_2) = b$. Now consider a third hypothesis space $H_3 = H_1 \cup H_2$. Namely, this is the set of all hypotheses that belong to $H_1$ *or to* $H_2$.

a. Provide a lower-bound $VC(H_3)$ in terms of $a$ and $b$. This lower bound should hold for any choice of hypothesis spaces, $H_1$ and $H_2$.

The bound we will use is max(a,b). It is clear that this is a lower bound for any such hypothesis, $H_3$. Denote c = max(a,b). This means that there is a set of size c, which is shattered by either $H_1$ or $H_2$, using up to $2^c$ hypotheses from that class. Those hypotheses are also elements of $H_3$, so $H_3$ can also shatter this set of size c. This means that $c \leq VC(H_3)$.

b. Come up with two specific examples of $H_1$ and $H_2$ that realize this lower bound. Giving the specific examples will be easier if you consider simple hypothesis spaces, e.g. those that operate over a one-dimensional Instance space X.

Now consider two hypothesis classes in 1-dimension. The first hypothesis class, $H_1$, is the interval classifier, will label all points in the interval [i, j] as 1 when using parameters i and j. The second hypothesis class $H_2$, is the linear classifier in 1-dimension, meaning points will receive a label of 1 if $0 \leq w*x+z$. Both hypothesis classes can shatter any set of two unique points in 1-dimension. However, if we have a set of size 3, call these points x < y < z, then neither hypothesis class can realize the labeling x = 1, y = 0, z = 1. If two of

these three points are the same, then neither classifier can label one of the two identical points as 1 and the other as 0. Therefore, there is no set of size 3 that can be shattered by this hypothesis class.

Therefore, the VC-dimension of $H_3 = H_1 \cup H_2$ is 2, which is max(a,b).

(b) Let H be a hypothesis class for concept learning, such that |H| = c for a given constant, c. Give an upper bound on VC(H) in terms of c. Provide your reasoning.

If the hypothesis class only contains c elements, then given m points, it can label these points in at most c ways. In order to shatter a set of size m, we must be able to label it in $2^m$ different ways, in order to capture every possible labeling. Therefore, it would be impossible for H to shatter a set of size m if $2^m > c$. The VC-dimension of H must be at most floor[$\log_2(c)$]. Note that for some constant hypothesis classes, this upper bound is not tight.