CS 5350/6350: Machine Learning Spring 2015

Homework 3

Anirudh Narasimhamurthy(u0941400)

March 8, 2015

1 Warm up: Feature expansion

We are given the set of functions f_r defined by an integer of radius 'r' as follows:

$$f_r(x_1, x_2) = \begin{cases} +1 & x_1^2 + x_2^2 \leq r^2; \\ -1 & \text{otherwise} \end{cases} \tag{1}$$

- This definition implies that all those points that lie within a circle will have a label of '+' and all those points which lie outside the circle will have a label '-' . This is the reason why it is not clearly linearly separable in R^2
- In order to map the examples to a new space such that positive and negative examples are linearly separable, we will map the examples to 3-dimensional space R^3
- As the third dimension x_3 we can consider it as follows:
  $x_3 = \sqrt{x_1^2 + x_2^2}$ This can be used to replace the 'r' term as we have been asked to have ϕ not depend on r. When points are mapped into 3D space we move from the geometry of circle into geometry of cone.
- We also know from our knowledge of LTU, that an example will be classified + or - based on the following expression: $sgn(w^T.x + b)$
- In our case we would initally start with a weight vector w of (0,0,1) with 1 corresponding to x_3 term. Our x which is the feature vector would be (x_1, x_2, x_3) As we perform learning via the the perceptron, we will keep updating the weight vector and the bias term.
- Finally when we have learned, we would have the bias term moving to 'r' and we would have all points with label + lying on the surface of the cone. Thus we would be able to linearly separate the labels in 3D plane.
- We can then write the mapping function as follows: $\phi(x_1, x_2) = 1$ if $x_1^2 + x_2^2 \leq x_3^2$

  And so when a plane cuts this in 3D we have the pluses on one side and the minuses on the other side.

Please note that the above proof is provided considering the fact the initial centers are at 0,0. If we were to assume different centers then our initial equations would be in terms of $(x-a)^2+(y-b)^2 < r^2$ and then we would follow similar approach to perform the linear separation

Another method of proof:

- We could also possibly consider the polar co-ordinates of the circle and proceed with this problem using the parameters θ and R and we could perform the linear separation in two dimensions itself in that case.

2 PAC Learning

Part 1

- **Hypothesis space in case of Naive Robot**

  We are given N available parts and in this implementation we are free to combine any parts. The two possible combinations are that a part can be included or it is not included. Since we have N parts, the size of hypothesis space would be $\underline{2^N}$

- **Size of hypothesis space as per Rule 2**

  In this case there are two possibilities. We can either use the part as full or we can split the product into two parts say p1 and p2. We have already seen the case of using a part in full in the previous part. If we were to split the parts into two pieces, then there are four combinations which we would have.

  - p1 included p2 not included
  - p1 not included p2 included
  - p1 included and p2 included
  - p1 not included and p2 not included

  Thus this would give us 4^N combinations. Thusthe size of hypothesis space would be the combination of both cases. **So the size of hypothesis space would be 6^N.**

  This is under the assumption that p1 and p2 combined together is not equal to part P. If this is the case then I am considering that not including p1 and not including p2 is not same as not including P.

  If that assumption is incorrect, and that not including p1 and not including p2 is same as not including P, since we are effectively not including the part, then the **size of hypothesis space would be** 5^N

- **Number of examples the robot have to see**

  The number of examples the robot would have to see in order to learn any product with 0.01 error and probability 99 % is given by the following expression:

$$m \geq \frac{1}{\epsilon}(\ln(H) + \ln(1/\delta)).$$

  In our case size of the hypothesis class $|H|$ is 2^6 since we are considering the naive robot case. $(1-\delta) = 0.99$ and $\epsilon = 0.01$. Therefore

$$m \geq \frac{1}{0.01}(\ln(2^6) + \ln(1/0.01)).$$

$$m \geq \frac{1}{0.01}(6\ln(2) + \ln(100)).$$

$$m \geq 100(6 * 0.693 + 4.605).$$

$$m \geq 100(4.158 + 4.605).$$

$$m \geq 100(8.763).$$

$$m \geq 876.3.$$

  Thus the number of examples that the robot would have to see in order to learn the product with 0.01 error and 99 % accuracy is atleast 877 examples.

Part 2

In this part we are expected to derive an expression for the number of training examples sufficient to ensure that every hypothesis will have true error no worse than $(1+\gamma)error_D(h)$.

From the Chernoff bounds that have been provided we know that

$Pr[S/m < (1-\gamma)p] \leq e^{-mp\gamma^2/2}$
where S/m is the training error and p is the true error.

$Pr[\frac{S/m}{(1-\gamma)} < p] \leq e^{-mp\gamma^2/2}$

If we apply the Taylor series expansion or multiply and divide by the conjugate of $(1-\gamma)$ then we would have the following:

$Pr[S/m * (1+\gamma) < p] \leq e^{-mp\gamma^2/2}$

The true error lower bounds will now be given by

$Pr[p > S/m * (1+\gamma)] \leq e^{-mp\gamma^2/2}$

$Pr[\text{For all h} \in H, p > S/m * (1+\gamma)] \leq |H| * e^{-mp\gamma^2/2}$

Assume the above probability to be defined by the term δ. Then we will have

$\delta \leq |H| * e^{-mp\gamma^2/2}$

$e^{-mp\gamma^2/2} \geq \frac{|H|}{\delta}$

Taking ln on both sides we would get,

$mp\gamma^2/2 \geq \ln(|H|) + \ln(\frac{1}{\delta})$

Therfore we would have :

$\mathbf{m \geq \frac{1}{p\gamma^2/2} \cdot \ln(|H|) + \ln(\frac{1}{\delta})}$

This is the required expression

3 VC Dimension

1. **Shattering**

   A concept class shatters a set of points if for any labeling of those points there is some function in the class that correctly labels it.

   - In our given problem we are given C to be set of all conjunctions of n Boolean variables
   - We are asked to find the set $S \subseteq \{0,1\}^n$ consisting of exactly n examples that can be shattered by C.
   - Assuming it is conjunctions then we can either have a variable inluded in the conjunction or have it negated.
   - For n=1, we have 2 possible combinations i.e either x1 can be there or $\neg x1$ can be there.
   - For n=2, we have nine different combinations. $x1, x2, \neg x1, \neg x2, x1 \wedge x2, x1 \wedge \neg x2, \neg x1 \wedge x2, \neg x1 \wedge \neg x2, x1 \wedge \neg x1 \wedge x2 \wedge x2$
   - But given our set notation our set S $\subseteq 0, 1^n$ for n=2 would have a combination of 0,1 in the set which would give us the elements 01,10,11,00
   - For n=3, it would be 000,001,010,011,100,101,110,111
   - So if our set set consist of only one of the literals to be 1, then for n=1, we will have 1 example. For n=2 we will have 2 examples and so for n=3 we will have 3 examples and for n=4 we will have 4 examples and so on.

     Thus our set picks exactly n examples out of the set. Hence this could be one of the answers.

2. **Proving VC dimension for a finite concept class C**

   We are given that we have a finite concept class C . We are asked to prove that the VC dimension of this class is at most $\log|C|$

   Proof by contradiction

- Suppose the VC dimension of this class (Lets assume it is given by the variable d) were to be greater than $\log|C|$ then it implies that there must exist a set of d points such that all possible combination of points are correctly labeled '+' and '-' or classified correctly according to their respective labels.
- The number of possible combinations in this case would be 2^d which would be greater than $2^{log|C|}$
- But then for our given concept class C we have only $|C|$ distinct concept classes and hence the d cannot be greater than log $|C|$ and d can be atmost log $|C|$
  Hence the VC dimension of concept class C is atmost log $|C|$

3. **VC dimension of the rectangular region**

   In our given problem an example $x = \{x_1, x_2\}$ in R^2 is labeled as +1 iff $x_1 \geq a$ and $x_2 \leq b$ and is labeled - otherwise.

   And when a=1 and b=4, we see that the rectangle formed is bounded on two sides and is unbounded on the other two side (extending to infinity) owing to the given definition. And all the points inside this rectangle will be labeled as positive. In othr words this can be considered as a half rectangle.

   - Suppose we were to have four different points with all of them forming a diamond shaped and having the same lable then a single rectangle could fir them and make sure that all four examples are labeled positive and all others outside the rectangle is labeled negative.
   - But if we were to consider an example of 3 points lets say(2,1),(3,1) and (4,1) with their labels being + , - and +, then it would be impossible to shatter them with our half rectangles as one of the positive labels would be left out. So clearly we can't shatter more 3 points using axis parallel rectangle.
   - If we consider a set of two points with one point being being labeled -, and another point being labeled + which is slightly above the point with label '-', then again it will be impossible to shatter them with our half rectangle. This is because when we fit our rectangle it would have both the - and + points within it and sicne we can only extend our rectangle in one direction and not draw it as such as a full rectangle, shattering two points is also not possible.
   - In a way our half rectangle is similar to the half interval/half-lines problem. It has a closed interval on one end and open interval on the other. Hence the number of points that can be shattered is atleast and atmost 1. **Therefore VC dimension of this class is exactly 1**

4. **VC Dimension of concept class consisting of union of two intervals**

   We know that for a full interval problem the VC dimension is 2. In our given problem we are asked to determine the VC dimension of union of two intervals on the real line. We are also given the information, that points within either of the intervals will be labeled as positive.

Let us first prove the VC dimension of an interval problem and that result wouild help us in coming up with an answer for the union of two intervals.

- In the interval problem, the hypothesis class is the set of intervals on the real axis [a,b] for some $b > a$. If we had three examples x,y,z where $x < y < z$ and x has a lablel '+', y has a label '-' and z has a label '+', then it would be impossible to shatter them.
- This is because after fitting the interval to make x, +, we would still be able to place y outside the interval and to teh right of x and so it would get the label '-'. But when we have to give the third point z which is to the right of y, a '+' , then we can't do it with a single interval. We would require one more interval.
- Thus VC dimension of a full interval would be 2.
- **Two intervals**
  - When we have union of two intervals, then we would be able to overcome the problem we had in the previous part.
  - With two intervals and 3 examples **(+, -, +)** we could perfectly fit the three points and be able to classify them correctly.
  - When we have four points we would still be able to shatter them with union of two intervals.
  - Even for the worst/adversarial case of four distinct points **(+, -, + , -)** we would be able to fit in the positive examples within the two intervals and the negative examples will lie outside of those intervals. Thus we could shatter atleast 4 points.
  - But if we were to try for 5 points, with the points being v, w, x, y and z with $v < w < x < y < z$ and the labels being +, -, +, -, + respectively, then even with two intervals we would not be able to shatter the points as the 5th point would be outside of our given intervals and would be incorrectly classified as '-'. Hence we can shatter atmost 4 points.
  - So 4 is the exact number of points that we can shatter with union of two intervals.

  Therefore for concept class consisting of union of 2 intervals, the VC dimension is exactly 4

5. **VC dimension of union of k intervals**

- We have seen and proved that VC dimension of the interval problem is 2.
- We have also seen that the VC dimension of union of 2 intervals is 4.
- We have seen that single intervals can label a sequence of two distinct points correctly but cannot label a sequence of 3 distinct points (+, - ,+).
- Therfore the union of k-intervals on the real line would be able to successfully shatter atleast 2* k such distinct points.

- If we were to try to shatter 2k+1 distinct points with k intervals, then we would fail to correctly provide the label for one of the points, as we would be able to fit in k points whose label is '+' within our intervals and the 'k' points whose label is '-' outside each of those different intervals.
- If we were to fit in one more point whose label is '+' and which satisfies our criteria of its value being greater than all the other 2k elements which we had placed in the number line, then we ould require one more interval to correctly classify the point as '+'. Thus we would be able to shatter atmost 2k distinct points.
- However if the point were to have a label '-', then we would be able to shatter the 2k+1 points.
- But since the defintion of VC dimension of the hypothesis space over the instance space X is the size of the largest finite subset of X that is shattered by H. In othe case of union of k-intervals it is 2k. **Hence the VC dimension of the union of k-intervals on the real line is 2k**

6. **Proving $VC(H_1) \leq VC(H_2)$. for the given considerations**

   In our problem we are given two hypothesis classes H_1 and H_2 such that $H_1 \subseteq H_2$.

   Given the following information we can make the following assumption:

   - Assume the VC dimension of H1 to be d1.
   - Then VC dimension of H2 could be d1+ ϵ where ϵ is greater than or equal to 0.
   - This is because set H2 might contain more points and hence it could possibly shatter more points than what could be shattered by H1. Hence its VC dimension would be atleast d1+ ϵ
   - Thus if $|H1|$ is very small when compared to $|H2|$, and the number of points which can be shattered by H2 is morethen the VC dimension of H1 would strictly be lesser than that of H2 i.e $\mathbf{VC(H_1) \leq VC(H_2)}$
   - If $|H1|$ and $|H2|$ are of equal size or if the number of points which both H1 and H2 can shatter are the same, then $\mathbf{VC(H_1) = VC(H_2)}$
   - Combining the previous two equations, we thus get $\mathbf{VC(H_1) \leq VC(H_2)}$

   Hence proved