Advanced topics in Machine Learning

Assignment 3

Mariuta Nicolae (rqt629) September 27, 2015

# Question 1

## 1.

For this exercise, I think I should split the equation into proving:

and

and by proofing these 2 equations, then I will also have that .

So first of all: I would say it is quite obvious, because we cannot label n objects in larger number of ways than the size of hypotheses.

Secondly, the fact that was actually shown in the lecture class because there is a maximum of of hyper planes that can split n points.

As a conclusion, I have shown it is true that. I do not know if the assignment requires more complex proof, but I cannot think of other way to show this.

The VC dimension of H is :

## 2.

For proving that I will just use the fact presented during the lecture that:

which means that:

and

By using these in the equation that has to be proved, I will get:

By dividing the equation above by we will have:

Which is true for any , so I have proved that .

## 3.

I will start with basic steps of the induction and I calculate the result for :

Now for :

Now that the equation is true for and then I will suppose that it is true for d = k :

Now, I calculate for :

Since it is true that

Then it also happens that :

So by induction it can be proved that:

## 4.

I will use the intuition shown from the lecture that:

And in order to proof this, I will consider n+1 objects and if we take the first object, it can be either *in* and then there are ways to select the remaining sets, or if it is *out* then there are possibilities to select the remaining set.

From this intuition, we have that:

and at the previous exercise I have shown that

so I can bound:

## 5.

So the VC generalization bound is:

By using the result obtained above, It results that:

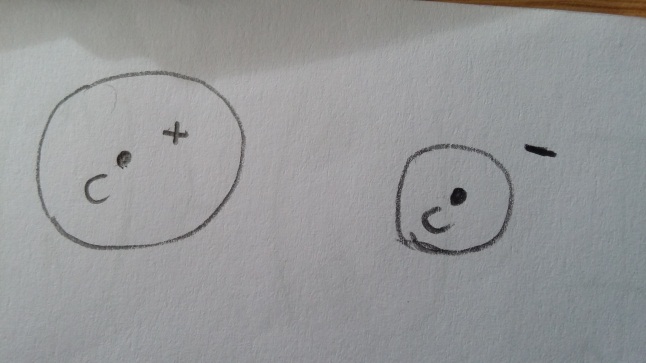
I think that the bound is more meaningful (more tight) if d is not very large in comparison to N.

# Question 2

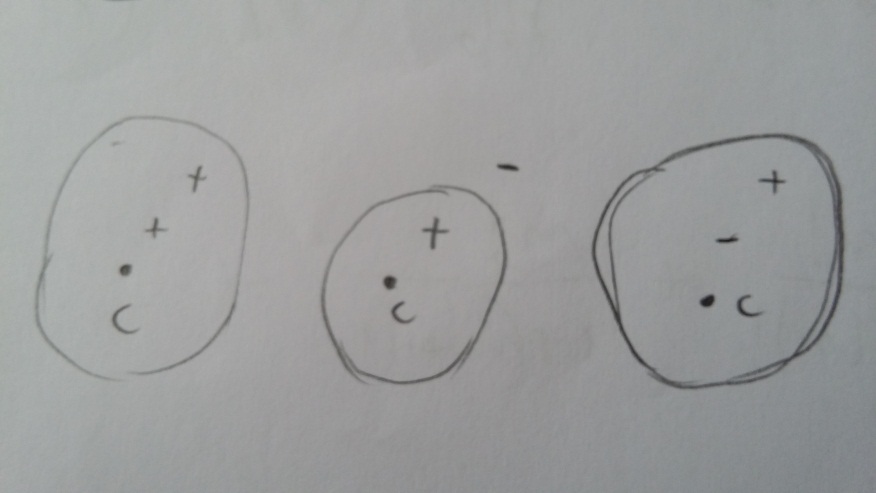
## 1.

In this case, a function to label a point p calculates the distance between the center of a circle c and the point p, then compares to radius r. From what I understood, the VC dimension is the maximum number of point p that can be arranged so the function can shatter them.

For dimension it appears we can classify correctly no matter what the sign is:



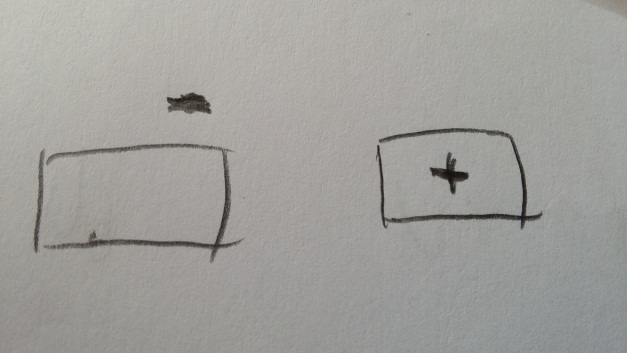
But for it appears we have a case when we cannot draw a circle to make the classification correctly as it can be seen in the second case shown in my drawing:



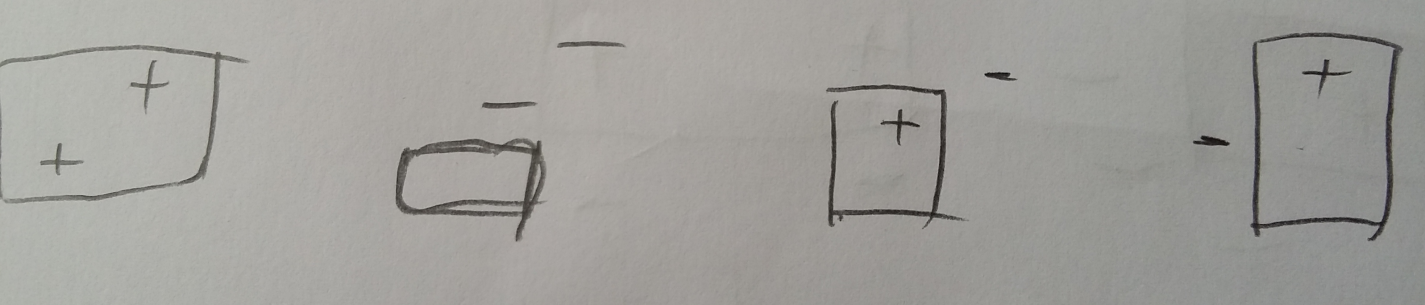
So, I conclude that since only for one point we can draw a circle that can classify it correctly, no matter what is the sign of the point.

## 2.

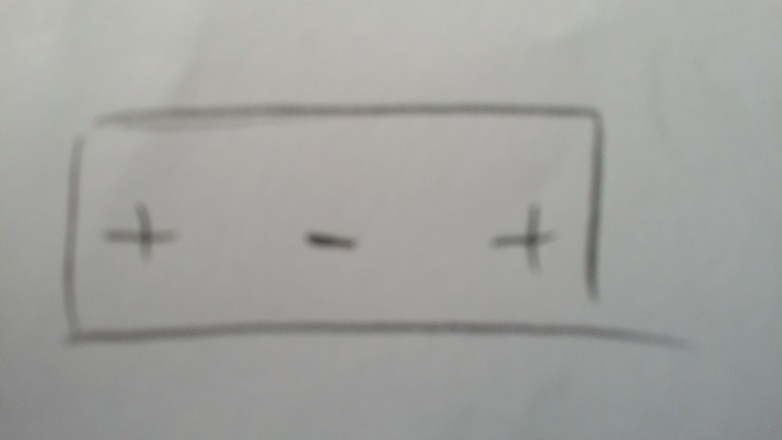
In this case, we obviously can make a convex set to classify correctly one point by drawing one convex set:



In case we have 2 points, we can classify correctly no matter how the points are displayed:



And in case of 3 points, I found a case, when the points are collinear, there is a negative one at the middle and the other 2 points are positive and there is no way to draw a convex set without including the negative point:



So I can conclude that the VC dimension in this case is

## 3.

According to the result obtained at Question 1.5., we have that:

In case of learning with convex sets where , we will have:

And in case of circle set, we will have:

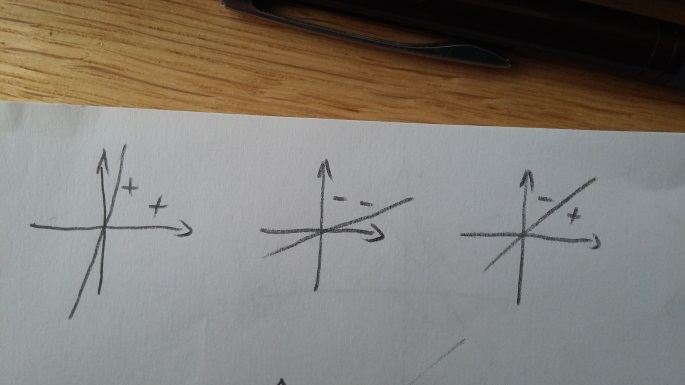
With these results, I would say that for the distribution we will have worse overfitting since the distance will be larger. But I think that for any case where we will have that the convex set produces less overfitting that the circles.

## 4.

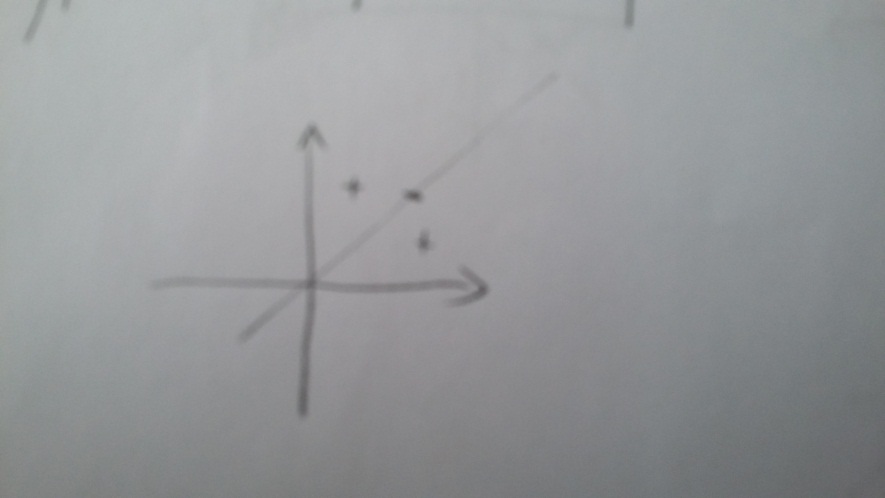
I would expect that

## 5.

For homogeneous separating hyperplanes, we will be able to shatter 2 points:



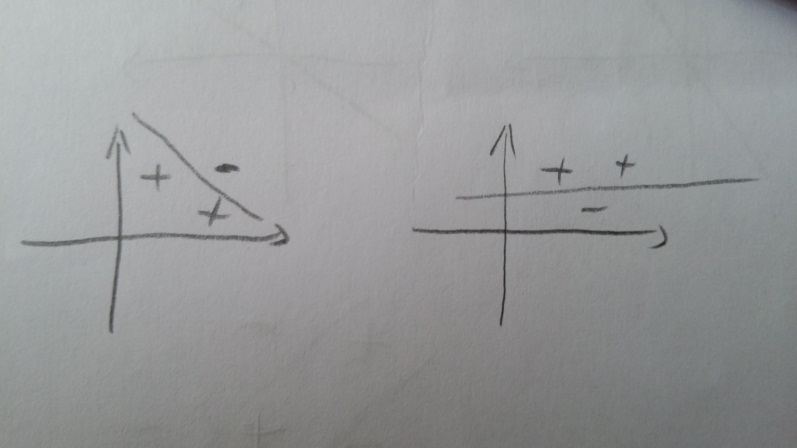
But for 3 points, there will be a case when we cannot classify using hyperplanes:



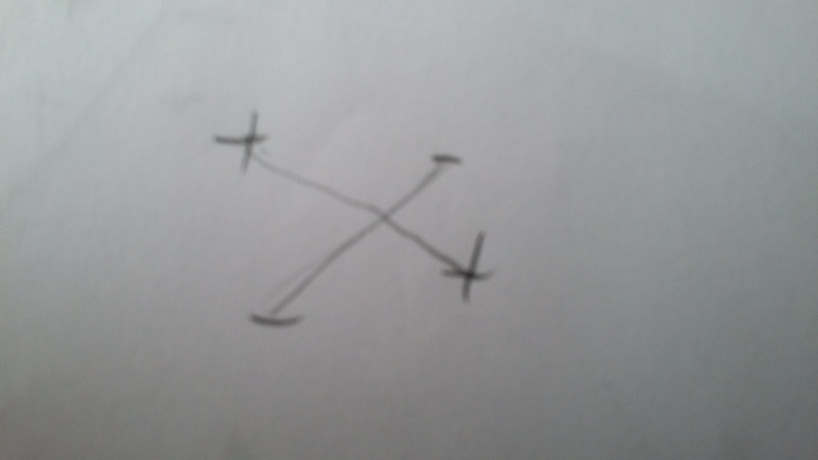
So the VC dimension in this case is

## 6.

For using general separating planes, we will be able to shatter 3 points:



But in the case of having 4 points, there is a case where we cannot separate them using hyperplane:



So the VC dimension in this case is

## 7.

I will use again the bound from Question 1.5:

In this case, with probability we will have:

I do not know how to further solve this equation, but by calculating N, we will obtain the number of samples needed in this case.

# Question 3

By looking at the derivation shown in the class, for a general radius R we will have:

In this case, we slice the hypotheses space H into a nested sequence of subspaces where for all i<d we define. By Theorem 24 from lecture notes, we have and then we have by Theorem 23:

We take and we will have that

I am not really sure how to continue this, but I think I should use the Occam’s Razor and obtain:

# Question 4

## 1.

At assignment 1, we have the space of strings of length d and it was shown that:

and there was also the space of functions that map strings from to values {0,1}. Since the functions can split all words into category 0 or 1, I can say that the VC dimension of is equal with the size of hypotheses.

For d = 0, we have

For d = 1, we have

So I can say that

Then the VC dimension of is

## 2.

Now we have the infinite set:

So the VC dimension in this case is:

## 3.

In the Assignment 1, by using Occam’s Razor for infinite sets and obtained:

And with high probability bound we have:

Using the VC lower bound, we will have that:

It appears we have no contradiction between these 2 bounds and the result from exercise 2.

# Question 5

For the KL inequality, we have that:

And for the Occam’s razor bound we have:

By denoting the right hand side of kl inequality by , we obtain with probability greater than :

And we also have the Pinsker’s inequality from the lecture notes (1.11):

If we apply 1.11 into the equation from above, we will have that:

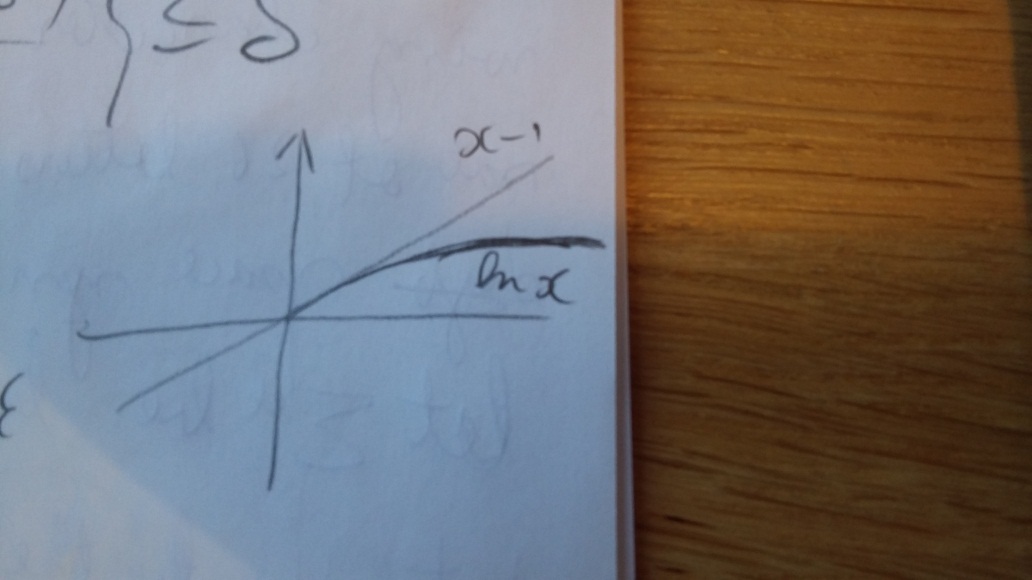
But we have from the Occam’s razor that:

I do not know how to continue this, but my guess is that I have to combine the last 2 equations somehow and obtain a general result.

# Question 6

## 1.

Graphically, lnx and x-1 look like this:



The logarithm grows a lot slower than the linear function x-1 so this will always be true, but I will try to also prove this mathematically.

In order to prove:

I will move all to the left of the inequality:

By calculating the derivative of the function , I obtain:

The derivative is always negative for so the function is decreasing.

For x = 0. The value of function which has the limit so the function is always negative, so the inequality:

is true for any .

## 2.

In order to prove that I will use the fact that for discrete case we have:

I will follow the suggestion in the assignment and I will try to prove that:

Now, I am getting something like in the previous exercise, so I think I can write that:

Now I do not know how to continue this, but I think it should be a trick that will lead to the result that:

and then will have proved that .