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:

$$m_H(n) \leq \mathrm{M} \quad \text{and} \quad m_H(n) \leq 2^n$$

and by proofing these 2 equations, then I will also have that $m_H(n) \leq \min\{\mathrm{M}, 2^{\mathrm{n}}\}$.

So first of all: $m_H(n) \leq \mathrm{M}$ 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 $m_H(n) \leq 2^n$ was actually shown in the lecture class because there is a maximum of 2^n of hyper planes that can split n points.

As a conclusion, I have shown it is true that$m_H(n) \leq \min\{\mathrm{M}, 2^{\mathrm{n}}\}$. 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 :

$$\mathrm{d}_{VC}(H) = \max\{n : m_H(n) = 2^n\}$$

2.

For proving that $m_H(2n) \leq m_H(n)^2$ I will just use the fact presented during the lecture that:

$$m_H(n) = \sum_{i=0}^{\mathrm{d}} \binom{n}{i}$$

which means that:

$$m_H(2n) = 2\sum_{i=0}^{\mathrm{d}} \binom{n}{i}$$

and

$$m_H(n)^2 = \sum_{i=0}^{\mathrm{d}} \binom{n}{i} * \sum_{i=0}^{\mathrm{d}} \binom{n}{i}$$

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

$$2\sum_{i=0}^{\mathrm{d}}\binom{n}{i} \leq \sum_{i=0}^{\mathrm{d}}\binom{n}{i} * \sum_{i=0}^{\mathrm{d}}\binom{n}{i}$$

By dividing the equation above by $\sum_{i=0}^{\mathrm{d}}\binom{n}{i}$ we will have:

$$2 \leq \sum_{i=0}^{\mathrm{d}}\binom{n}{i}$$

Which is true for any ≥ 2 , so I have proved that $m_H(2n) \leq\ m_H(n)^2$.

3.

I will start with basic steps of the induction and I calculate the result for $\mathrm{d} = 0$:

$$\sum_{i=0}^{0}\binom{n}{i} = \binom{n}{0} = \frac{n!}{0!\,(n-0)!} = 1 \leq\ n^0 + 1$$

Now for $\mathrm{d} = 1$:

$$\sum_{i=0}^{1}\binom{n}{i} = \binom{n}{0} + \binom{n}{1} = 1 + \frac{n!}{1!\,(n-1)!} = 1 + n \leq\ n^1 + 1$$

Now that the equation is true for $\mathrm{d} = 0$ and $\mathrm{d} = 1$ then I will suppose that it is true for d = k :

$$\sum_{i=0}^{\mathrm{k}}\binom{n}{i} \leq n^{\mathrm{d}} + 1$$

Now, I calculate for $\mathrm{d} = \mathrm{k} + 1$:

$$\sum_{i=0}^{\mathrm{k+1}}\binom{n}{i} = \sum_{i=0}^{k}\binom{n}{i} \ + \binom{n}{i+1} = \sum_{i=0}^{k}\binom{n}{i} + \frac{n!}{(i+1)!\,(n-i-1)!}$$

Since it is true that

$$\sum_{i=0}^{\mathrm{k}}\binom{n}{i} \leq n^{\mathrm{d}} + 1$$

Then it also happens that :

$$\sum_{i=0}^{k}\binom{n}{i} + \frac{n!}{(i+1)!\,(n-i-1)!} \leq n^{\mathrm{d+1}} + 1$$

So by induction it can be proved that:

$$\sum_{i=0}^{\mathrm{k}} \binom{n}{i} \leq n^{\mathrm{d}} + 1$$

4.

I will use the intuition shown from the lecture that:

$$\binom{n+1}{k} = \binom{n}{k} + \binom{n}{k-1}$$

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 $\binom{n}{k-1}$ ways to select the remaining sets, or if it is *out* then there are $\binom{n}{k}$ possibilities to select the remaining set.

From this intuition, we have that:

$$m_H(n) = \sum_{i=0}^{\mathrm{d}} \binom{n}{i}$$

and at the previous exercise I have shown that

$$\sum_{i=0}^{\mathrm{k}} \binom{n}{i} \leq n^{\mathrm{d}} + 1$$

so I can bound:

$$m_H(n) \leq n^{\mathrm{d}} + 1$$

5.

So the VC generalization bound is:

$$L(h) \leq \hat{L}(h,S) + \sqrt{\frac{8}{N}\ln\left(\frac{4m_H(2N)}{\delta}\right)}$$

By using the result obtained above, It results that:

$$L(h) \leq \hat{L}(h,S) + \sqrt{\frac{8}{N}\ln\left(\frac{4(N^{\mathrm{d}}+1)}{\delta}\right)}$$

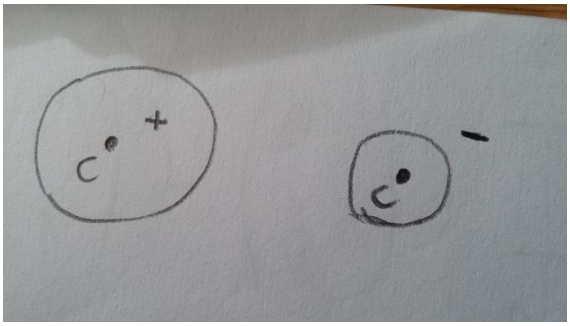
I think that the bound is more meaningful (more tight) if d is not very large in comparison to N.

Question 2

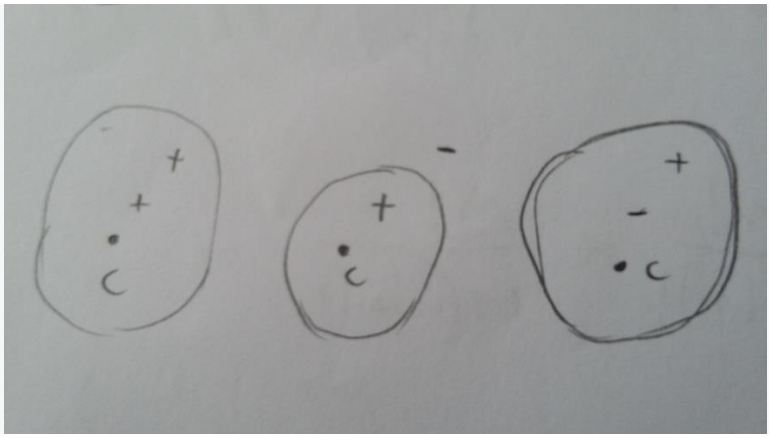
1.

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 $\mathrm{d}_{VC}(H) = 1$ it appears we can classify correctly no matter what the sign is:



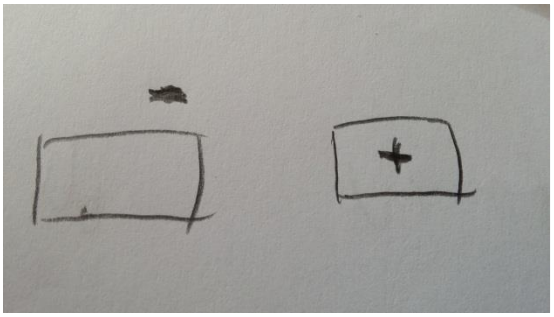
But for $\mathrm{d}_{VC}(H) = 2$ it appears we have a case when we cannot draw a circle to make the classification correctly as it can be seen in the second case shown in my drawing:



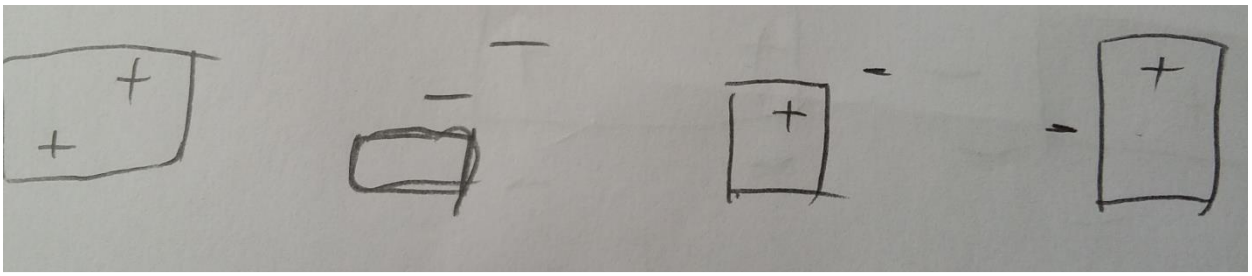
So, I conclude that $\mathrm{d}_{VC}(H) = 1$ since only for one point we can draw a circle that can classify it correctly, no matter what is the sign of the point.

2.

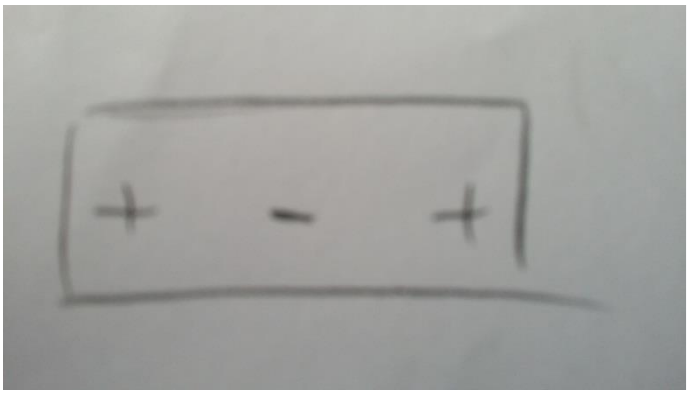
In this case, we obviously can make a convex set to classify correctly one point by drawing one convex set:



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 $\mathrm{d}_{VC}(H) = 2$

3.

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

$$L(h) - \hat{L}(h,S) \leq \sqrt{\frac{8}{N}\ln\left(\frac{4(N^{\mathrm{d}_{vc}}+1)}{\delta}\right)}$$

In case of learning with convex sets where $\mathrm{d}_{vc} = 2$, we will have:

$$L(h) - \hat{L}(h,S) \leq \sqrt{\frac{8}{N}\ln\left(\frac{4(N^2+1)}{\delta}\right)}$$

And in case of circle set, we will have:

$$L(h) - \hat{L}(h,S) \leq \sqrt{\frac{8}{N}\ln\left(\frac{4(N+1)}{\delta}\right)}$$

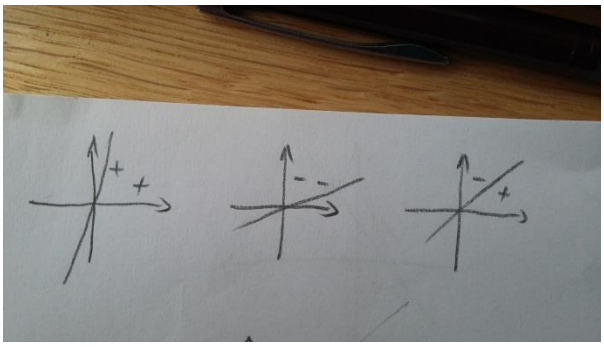
With these results, I would say that for the distribution $\delta = 1$ we will have worse overfitting since the distance $L(h) - \hat{L}(h,S)$ will be larger. But I think that for any case where $\delta < 1$ we will have that the convex set produces less overfitting that the circles.

4.

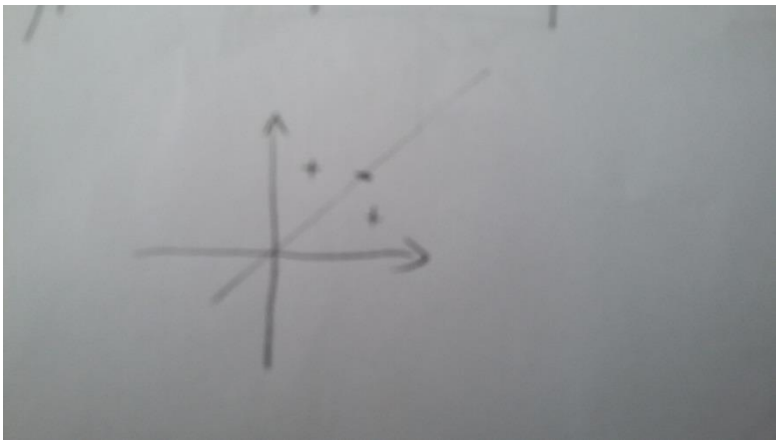
I would expect that

5.

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



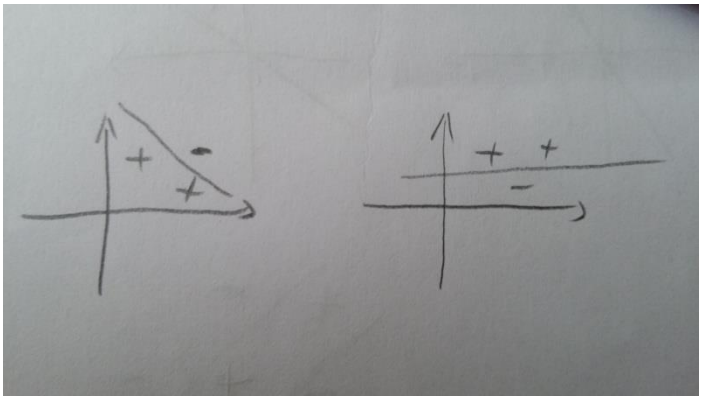
But for 3 points, there will be a case when we cannot classify using hyperplanes:



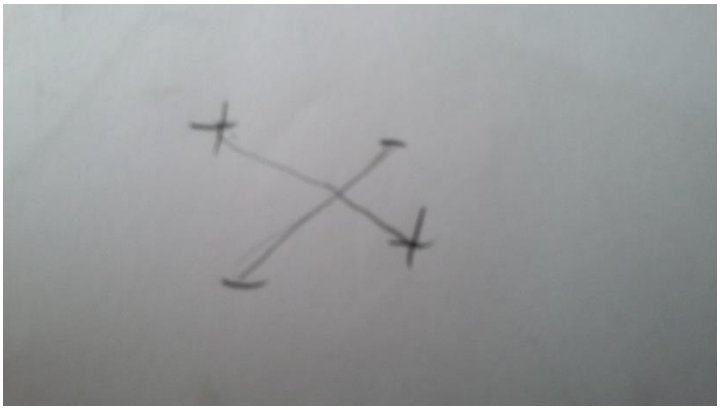
So the VC dimension in this case is $\mathrm{d}_{VC}(H) = 2$

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 $\mathrm{d}_{VC}(H) = 3$

7.

I will use again the bound from Question 1.5:

$$L(h) - \hat{L}(h,S) \leq \sqrt{\frac{8}{N}\ln\left(\frac{4(N^{\mathrm{d}_{vc}}+1)}{\delta}\right)}$$

In this case, with probability $1-\delta = 0.95 => \delta = 0.05$ we will have:

$$0.05 \leq \sqrt{\frac{8}{N}\ln\left(\frac{4(N^{10}+1)}{0.05}\right)} => 0.05^2 \leq \frac{8}{N}\ln\left(\frac{4(N^{10}+1)}{0.05}\right)$$

$$\frac{N}{8}0.05^2 \leq \ln\left(\frac{4(N^{10}+1)}{0.05}\right) => \frac{N}{8}0.05^2 \leq \ln(4) + \ln(N^{10}+1) - \ln(0.05)$$

$$\frac{N}{8}0.05^2 \leq 1.38 - 2.99 + \ln(N^{10}+1) => \frac{N}{8}0.05^2 + 1.61 \leq \ln(N^{10}+1)$$

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:

$$\mathrm{d}_{VC}(H) < R^2\mathrm{d} + 1$$

In this case, we slice the hypotheses space H into a nested sequence of subspaces $H_1 \subset H_2 \subset \cdots H_{\mathrm{d}-1} \subset H_{\mathrm{d}} = H$ where for all i<d we define$H_i = H_{\rho=i}$. By Theorem 24 from lecture notes, we have $\mathrm{d}_{VC}(H) < R^2\mathrm{i} + 1$ and then we have by Theorem 23:

$$P\left\{\exists h \in H_i : L(h) > \hat{L}(h,S) + \sqrt{\frac{8}{n}\ln\left(\frac{2\left((2n)^{1+R^2\mathrm{i}}+1\right)}{\delta_i}\right)}\right\} \leq \delta_i$$

We take $\delta_i = \frac{R^2}{i(i+1)}\delta$ and we will have that $\sum_{i=1}^{\infty}\frac{R^2}{i(i+1)} = R^2$

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

$$P\left\{\exists(w,b)\in H_i:\ L(w,b)>\ \hat{L}(w,b,S)+\sqrt{\frac{8}{n}\ln\left(\frac{2\left((2n)^{1+||R^2w^2||}+1\right)}{\delta_i}\right)}\right\}\leq\delta_i$$

Question 4

1.

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

$$|\Sigma^{\mathrm{d}}|=27^{\mathrm{d}}$$

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

For d = 0, we have $|H_0|=2$

For d = 1, we have $|H_0|=2^{27}$

So I can say that $|H_{\mathrm{d}}|=2^{27^{\mathrm{d}}}$

Then the VC dimension of H_{d} is $\mathrm{d}_{VC}(H_{\mathrm{d}})=2^{27^{\mathrm{d}}}$

2.

Now we have the infinite set:

$$H=\bigcup_{\mathrm{d}=0}^{\infty}H_{\mathrm{d}}$$

So the VC dimension in this case is:

$$\mathrm{d}_{VC}(H_{\mathrm{d}})=\sum_{\mathrm{d}=0}^{\infty}H_{\mathrm{d}}=\sum_{\mathrm{d}=0}^{\infty}2^{27^{\mathrm{d}}}$$

3.

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

$$P\left\{\exists h \in H:\ L(h) > L'(h,S) + \frac{\sqrt{\ln\left(\frac{1}{p(h)\delta}\right)}}{2n}\right\} \leq \delta$$

And with high probability bound $1-\delta$ we have:

$$L(h) \leq L'(h,S) + \frac{\sqrt{\ln\left(\frac{1}{p(h)\delta}\right)}}{2n}$$

Using the VC lower bound, we will have that:

$$E\left[sup_h L(h) - \hat{L}(h,S)\right] \geq 0.25$$

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:

$$P\{kl(\hat{p}||p) \geq \varepsilon\} \leq (n+1)e^{-n\varepsilon}$$

And for the Occam's razor bound we have:

$$P\left\{\exists h \in H : L(h) > L'(h,S) + \frac{\sqrt{\ln\left(\frac{1}{p(h)\delta}\right)}}{2n}\right\} \leq \delta$$

By denoting the right hand side of kl inequality by δ, we obtain with probability greater than $1-\delta$:

$$kl(\hat{p}||p) \leq \frac{\frac{\ln(n+1)}{\delta}}{n}$$

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

$$kl(p||q) \geq \frac{1}{2}(|p-q| + |(1-p)-(1-q)|)^2 = 2(p-q)^2$$

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

$$\widehat{|p-p|} \leq \sqrt{\frac{kl(\hat{p}||p)}{2}} \leq \sqrt{\frac{\frac{\ln(n+1)}{\delta}}{2n}}$$

But we have from the Occam's razor that:

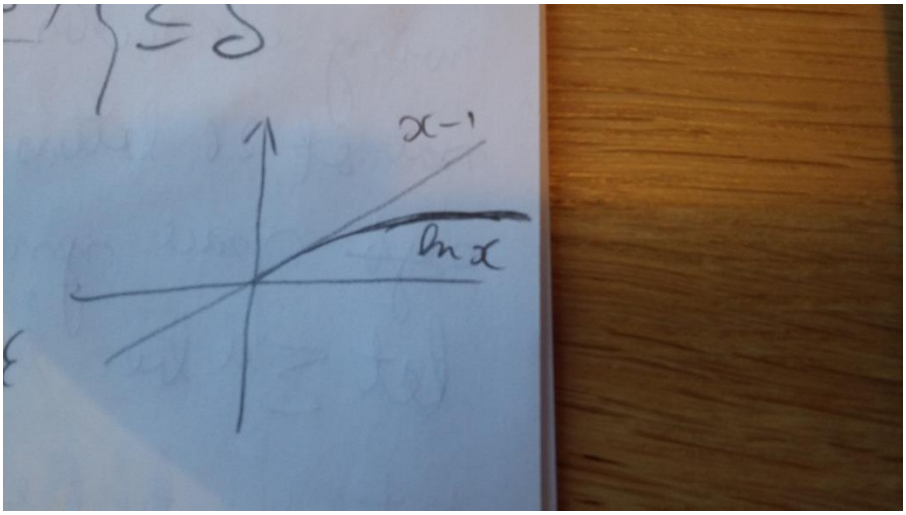
$$\hat{p} - p \leq \frac{\sqrt{\ln\left(\frac{1}{p(h)\delta}\right)}}{2n}$$

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:

$$lnx \leq x - 1$$

I will move all to the left of the inequality:

$$lnx - x + 1 \leq 0$$

By calculating the derivative of the function $f(x) = lnx - x + 1$, I obtain:

$$f(x)' = \frac{1}{x} - 1 = \frac{1 - x}{x}$$

The derivative is always negative for $0 < x < \infty$ so the function is decreasing.

For x = 0. The value of function $f(0) = ln0 + 1$ which has the limit $-\infty$ so the function $f(x) = lnx - x + 1$ is always negative, so the inequality:

$$lnx \leq x - 1$$

is true for any $0 < x < \infty$.

2.

In order to prove that $kl(p||q) \geq 0$ I will use the fact that for discrete case we have:

$$kl(p||q) = \sum_x p(x) \ln\left(\frac{p(x)}{q(x)}\right) = \sum_x p(x) lnp(x) - \sum_x p(x) lnq(x)$$

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

$$-\sum_x p(x) \ln\left(\frac{p(x)}{q(x)}\right) \leq 0$$

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

$$-\sum_x p(x) \ln\left(\frac{p(x)}{q(x)}\right) \leq -\sum_x p(x)\left(\frac{p(x)}{q(x)} - 1\right)$$

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

$$\sum_x p(x) lnp(x) \geq \sum_x p(x) lnq(x)$$

and then will have proved that $kl(p||q) \geq 0$.