Advanced topics in Machine Learning

Assignment 1

Mariuta Nicolae (rqt629) September 13, 2015

# Question 1

# 1.

For this exercise, I will define the samples to have values where 0 means that it is a red ball and 1 if we have a green ball. When extracting one ball from all the 2n from a bin, we have the probabilities:

, and we know that .

Given these, I will have to calculate the probability that if we extract the balls without replacement:

According to Heoffding without replacement, in our case, we have that:

Where:

I am not sure how correct this poof might be, but if I would sample repeatedly with replacement, we will have the probability to get n green balls at n extractions:

So if we make the sampling without replacing, we will have

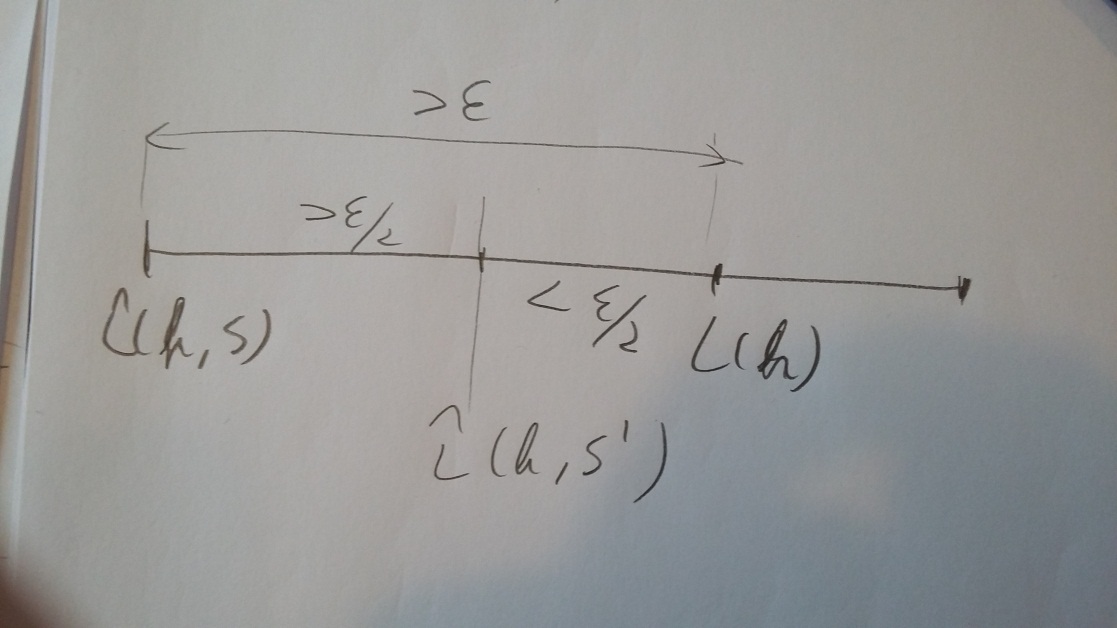
But I think this proof is not complete and correct.

# 2.

For this exercise, I will mostly follow the steps shown in the lecture last week. So we have a sample with

so without empirical loss and we are interested in the expected loss, . And I suppose that we have infinite hypotheses.

**Step1**: we introduce the ghost sample with empirical loss as in the picture:



**Step2**: Now we show that:

Since we have that the equation above becomes:

And I am interested in the event that:

I have to show that the distance is large and that also the distance is large.

Now, I fix for which calculate:

Because and are independent, we have that confirmed so we can get rid of condition

**Step3**: now, I get to the step where I have to use Hoeffding’s inequality:

We have to bound: which is : in our case, because .

I decide to use the method that we sample and we split into: S and S’.

There is a finite number of ways to label 2n points, at most . Let be the maximal number of ways to label 2n points with hypotheses in H.

Let be the number of ways to label and let be the corresponding hypothesis.

So we I can conclude that:

Since we have that as in the previous exercise, we have shown that:

**Now, I put everything together**:

And with probability greater than , for all :

So I have proven that for all that satisfy we have with probability greater than :

and I have also calculated the complete result.

# Question 2

I struggled much to install LIBSVM to make this exercise, but I did not manage to use any functions from the library in Matlab, so I could not make this exercise.

# Question 3

I will start with choosing just 2 more random values for distributions p and q. I will take that and . In this case, we will have that kl-divergence between p and q:

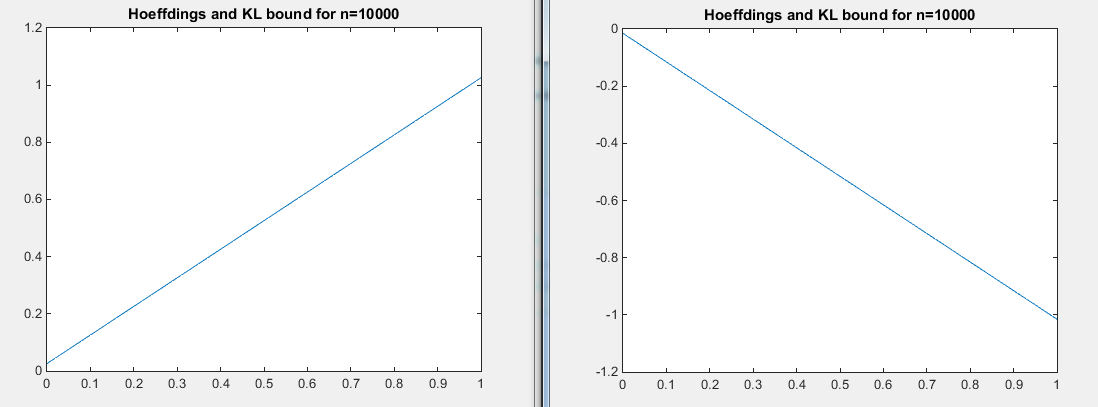
and the divergence between q and p:

Since I provided an example for p and q that have , I proved that kl is asymmetric in its arguments.

# Question 4

I have plotted together the upper and lower bounds for both inequalities, by varying the value of and calculating the values of inequalities for different values of n, by using:

for Hoeffding’s inequality and for KL inequality with values between 0 and 1 for and I only get 2 linear plots for each, which is not really what I was expecting:



I do not understand why there is not any difference between the two lines, but I think I miss something.