Advanced topics in Machine Learning
Assignment 1

Mariuta Nicolae (rqt629) September 13, 2015

Question 1

1.

For this exercise, I will define the samples Z_i to have values $\{0,1\}$ 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:

$P\{Z_i = 1\} = 1 - \varepsilon$, $P\{Z_i = 0\} = \varepsilon$ and we know that $0 < \epsilon \leq \frac{1}{2}$.

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

$$P\{\frac{1}{n}\sum_{i=1}^{n} Z_i = 1\}$$

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

$$P\left\{\frac{1}{n}\sum_{i=1}^{n} Z_i - \mu > \epsilon\right\} \leq e^{-2n\epsilon^2}$$

Where:

$$\mu = \frac{1}{2n}\sum_{i=1}^{2n} Z_i = \frac{1}{2n}2n(1-\varepsilon) = 1 - \varepsilon$$

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:

$$\hat{\mu} = (1 - \varepsilon)^n$$

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

$$\hat{\mu} \leq (1 - \varepsilon)^n = e^{-n\varepsilon}$$

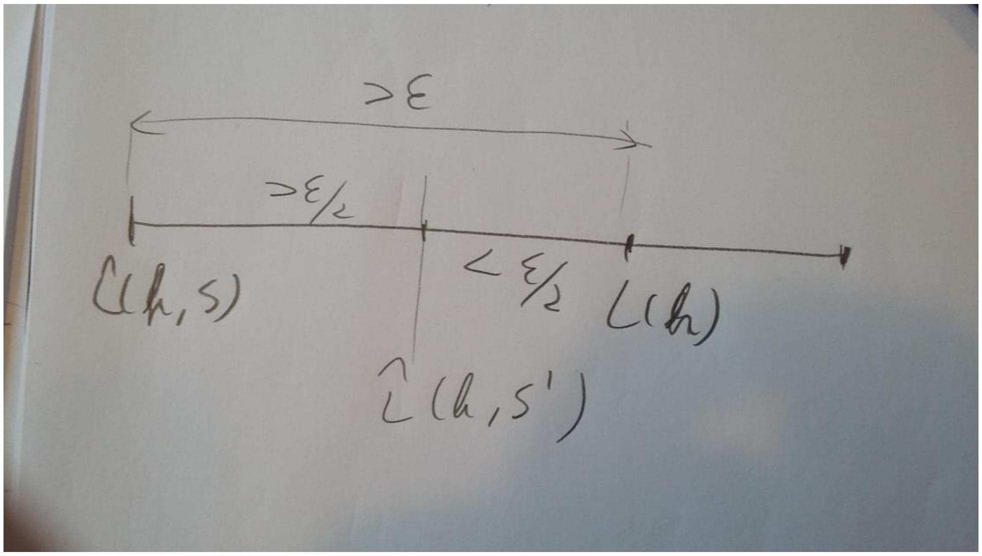
But I think this proof is not complete and correct.

2.

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

$\hat{L}(h, S) = 0$ so without empirical loss and we are interested in the expected loss, $L(h)$. And I suppose that we have infinite hypotheses.

Step1: we introduce the ghost sample with empirical loss $\hat{L}(h, S')$ as in the picture:



Step2: Now we show that:

$$P\left\{sup_{h\in H}\left(L(h)-\hat{L}(h,S)\right)>\varepsilon\right\}\leq 2P\{sup_{h\in H}\left(\hat{L}(h,S')-\hat{L}(h,S)\right)>\frac{\varepsilon}{2}\}$$

Since we have that $\hat{L}(h,S)=0$ the equation above becomes:

$$P\{sup_{h\in H}L(h)>\varepsilon\}\leq 2P\{sup_{h\in H}\hat{L}(h,S')>\frac{\varepsilon}{2}\}$$

$$P\left\{sup_{h\in H}\hat{L}(h,S')>\frac{\varepsilon}{2}\right\}\geq P\{sup_{h\in H}\hat{L}(h,S')>\frac{\varepsilon}{2}\ \ AND\ \ sup_{h\in H}L(h)>\varepsilon\}$$

And I am interested in the event that: $L(h)>\varepsilon$

I have to show that the distance $L(h)-0$ is large and that also the distance $\hat{L}(h,S')-0$ is large.

$P\left\{sup_{h\in H}\hat{L}(h,S')>\frac{\varepsilon}{2}\ \ AND\ \ sup_{h\in H}L(h)>\varepsilon\right\}=\ P\{sup_{h\in H}L(h)>\varepsilon\}*P\left\{sup_{h\in H}\hat{L}(h,S')>\frac{\varepsilon}{2}\middle|\ sup_{h\in H}L(h)>\varepsilon\right\}$

Now, I fix h^* for which $L(h^*)-\hat{L}(h^*,S)>\ \varepsilon$ calculate:

$$P\left\{sup_{h\in H}\hat{L}(h,S')>\ \frac{\varepsilon}{2}\middle|\ sup_{h\in H}L(h)>\varepsilon\right\}\geq P\{\hat{L}(h^*,S')-\hat{L}(h^*,S)>\ \frac{\varepsilon}{2}\ |\ L(h^*)-\hat{L}(h^*,S)>\ \varepsilon\}\geq$$

$$P\left\{L(h^*)-\hat{L}(h^*,S')\leq\ \frac{\varepsilon}{2}\ \ |\ L(h^*)-\hat{L}(h^*,S)>\ \varepsilon\right\}$$

Because S' and S are independent, we have that $L(h^*)-\hat{L}(h^*,S')\leq\frac{\varepsilon}{2}$ confirmed so we can get rid of condition $L(h^*)-\hat{L}(h^*,S)>\ \varepsilon$

$$P\left\{L(h^*) - \hat{L}(h^*, S') \leq \frac{\varepsilon}{2} \mid L(h^*) - \hat{L}(h^*, S) > \varepsilon\right\} = P\left\{L(h^*) - \hat{L}(h^*, S') \leq \frac{\varepsilon}{2}\right\} =$$

$$= 1 - P\left\{L(h^*) - \hat{L}(h^*, S') > \frac{\varepsilon}{2}\right\} \geq 1 - e^{-2n\left(\frac{\varepsilon}{2}\right)^2} \geq \frac{1}{2}$$

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

We have to bound: $P\left\{sup_{h\in H}\left(\hat{L}(h, S') - \hat{L}(h, S)\right) > \frac{\varepsilon}{2}\right\}$ which is : $P\left\{sup_{h\in H}\left(\hat{L}(h, S')\right) > \frac{\varepsilon}{2}\right\}$ in our case, because $\hat{L}(h, S) = 0$.

I decide to use the method that we sample S^{2n} and we split into: S and S'.

$$P\left\{sup_{h\in H}\left(\hat{L}(h, S')\right) > \frac{\varepsilon}{2}\right\} = \sum_{S^{2n}} P\{S^{2n}\} * P\left\{sup_{h\in H}\left(\hat{L}(h, S')\right) > \frac{\varepsilon}{2}\middle| S^{2n}\right\} \leq$$

$$\leq sup_{S^{2n}} P_{split}\left\{sup_{h\in H}\left(\hat{L}(h, S')\right) > \frac{\varepsilon}{2}\middle| S^{2n}\right\}$$

There is a finite number of ways to label 2n points, at most 2^{2^n}. Let $m_H(2n)$ be the maximal number of ways to label 2n points with hypotheses in H.

Let $M(S^{2n})$ be the number of ways to label S^{2n} and let $h_1, h_2, \ldots, h_{M(S^{2n})}$ be the corresponding hypothesis.

$$m_H(2n) = max_S \;\; M(S^{2n})$$

$$P_{split}\left\{sup_{h\in H}\left(\hat{L}(h, S')\right) > \frac{\varepsilon}{2}\middle| S^{2n}\right\} = P_{split}\left\{max_{h\in\left\{h_1, h_2, \ldots, h_{M(S^{2n})}\right\}}\left(\hat{L}(h, S')\right) > \frac{\varepsilon}{2}\middle| S^{2n}\right\}$$

$$\leq \sum_{i=1}^{M(S^{2n})} P\left\{\left(\hat{L}(h, S')\right) > \frac{\varepsilon}{2}\middle| S^{2n}\right\} \leq M(S^{2n}) \; max_{h_i} P\left\{\left(\hat{L}(h, S')\right) > \frac{\varepsilon}{2}\middle| S^{2n}\right\}$$

So we I can conclude that:

$$P\left\{sup_{h\in H}\left(\hat{L}(h, S')\right) > \frac{\varepsilon}{2}\right\} \leq M(S^{2n}) sup_{2n} sup_{h\in H} P_{split}\left\{\left(\hat{L}(h, S')\right) > \frac{\varepsilon}{2}\middle| S^{2n}\right\}$$

Since we have that $\hat{L}(h, S) = 0$ as in the previous exercise, we have shown that:

$$P\left\{\left(\hat{L}(h, S')\right) > \frac{\varepsilon}{2}\middle| S^{2n}\right\} \leq e^{-n\varepsilon}$$

Now, I put everything together:

$$P\left\{sup_{h\in H}\left(\hat{L}(h,S')\right) > \varepsilon\right\} \leq 2P\left\{sup_{h\in H}\left(\hat{L}(h,S')\right) > \frac{\varepsilon}{2}\right\} \leq 2m_H(2n) \leq e^{-n\varepsilon}$$

And with probability greater than $1-\ \delta$, for all $h\epsilon H$:

$$L(h) \leq \frac{\sqrt{\ln\left(\frac{2m_H(2n)}{\delta}\right)}}{n}$$

So I have proven that for all $h\epsilon H$ that satisfy $\hat{L}(h,S)=0$ we have with probability greater than $1-\ \delta$:

$L(h) \leq O\left(\frac{\ln\left(\frac{2m_H(2n)}{\delta}\right)}{n}\right)$ 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 $p=0.2$ and $q=0.4$. In this case, we will have that kl-divergence between p and q:

$$kl(p||q) = 0.2\ln\left(\frac{0.2}{0.4}\right) + (1-0.2)\ln\left(\frac{1-0.2}{1-0.4}\right) = 0.2ln0.5 + 0.8\ln\left(\frac{0.8}{0.6}\right) = -0.138 + 0.23 = 0.092$$

and the divergence between q and p:

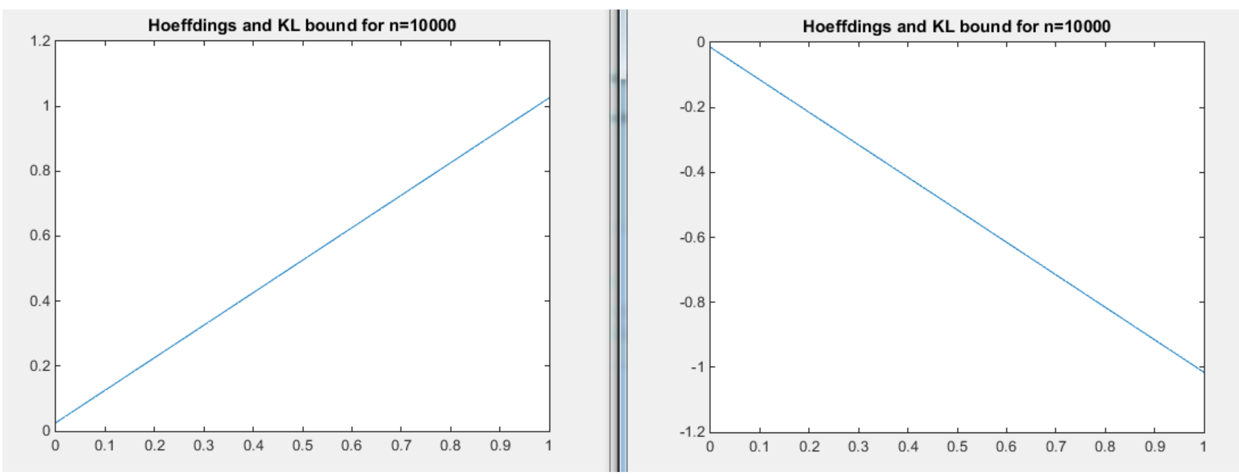
$$kl(q||p) = 0.4\ln\left(\frac{0.4}{0.2}\right) + (1-0.4)\ln\left(\frac{1-0.4}{1-0.2}\right) = 0.4ln2 + 0.6\ln 0.25 = 0.277 - 0.172 = 0.105$$

Since I provided an example for p and q that have $kl(p||q)\ \neq kl(q||p)$, 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 $\hat{p}$ and calculating the values of inequalities for different values of n, by using:

$p \leq \hat{p} + \frac{\sqrt{\ln\left(\frac{1}{\delta}\right)}}{2n}$ for Hoeffding's inequality and $|p - \widehat{p}| \leq \frac{\sqrt{\ln\left(\frac{n+1}{\delta}\right)}}{2n}$ for KL inequality with values between 0 and 1 for $\hat{p}$ 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.