Home Assignment 2
Machine Learning A

FEDERICO FIORIO

September 20, 2022

1 Illustration of Markov's, Chebyshev's, and Hoeffding's Inequalities (23 points)

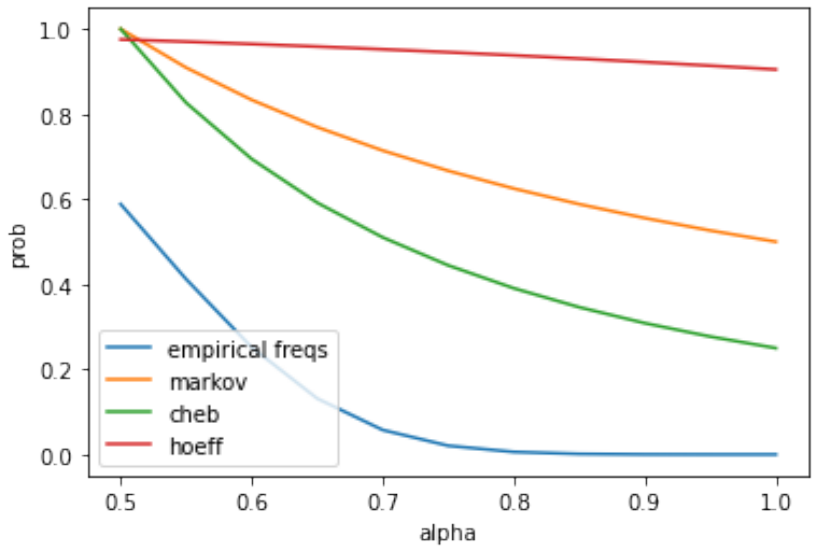

Figure 1: the empirical frequency of observing vs alpha

2. Explain why the above granularity of α is sufficient.
It is sufficient because the values for the empirical frequencies only increase by 0.05, because have 20 random variables, so when we take this value: $\frac{1}{20}\sum_{i=1}^{20} X_i$ it will only takes values up to 1 with an increase of 0.05, we cannot have values like 0.512 for the empirical frequency, so an α of value 0.51

would be useless

6. Compare the four plots

In blue we can see the empirical frequencies and with the other colors we can see the concentration measure for these frequencies.
In other words I computed different bounds, and while Markov's (yellow) tells us the bound for the probability of a random variable being higher than a specified constant alpha, Chebychev's and Hoeffding's are related to the deviation between empirical error and expected error, and in particular, Chebychev's says that no more than a certain fraction of values can be more than a certain distance from the mean, while Hoeffding's says that as N increases, it becomes exponentially unlikely that the training error deviates from the expected error with ϵ tolerance.
So different curves for different meaning, but they are all related to alpha and a probability, and they all can be plotted together.

7. For $\alpha = 1$ and $\alpha = 0.95$ calculate the exact probability P

$\alpha = 1$, I obtain prob = 1e-06
$\alpha = 0.95$, I obtain prob = 1.6e-05

1.1 2.b

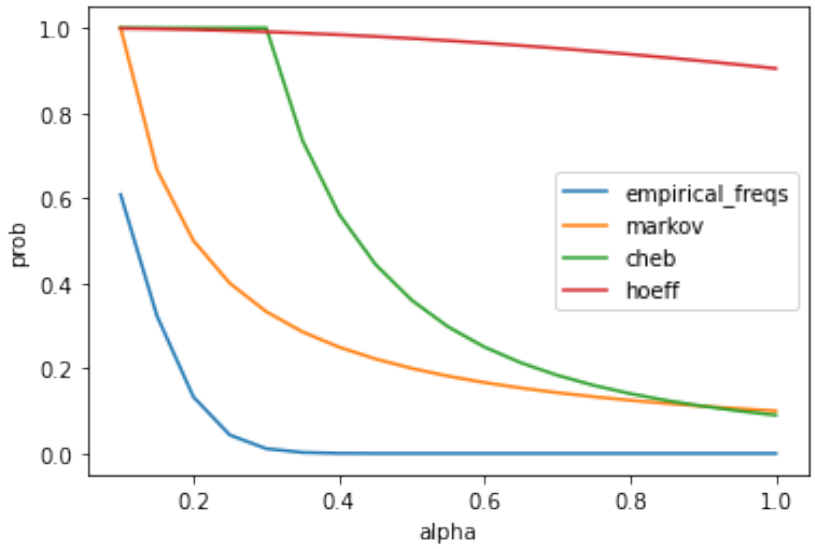

Figure 2: the empirical frequency of observing vs alpha

6. Compare the four plots

The same said before it's valid also here, as we can see, here we have a lower expectation values for the random variables, so Markov's bound became tighter, because it's less likely that we will have a value higher than a specific constant.
For Chebychev we also have a strong decrease, because of the lower variance compared to the previous example.(where the line is straight it's because the values where higher than 1 and then were rounded to 1)
Hoeffding's relates to the number of N so it's stays the same since the random variables were still 20 to be considered.

7. For $\alpha = 1$ and $\alpha = 0.95$ calculate the exact probability P

$\alpha = 1$, I obtain prob = 0.00
$\alpha = 0.95$, I obtain prob = 0.00

2 The Role of Independence (14 points)

let's say that I have $X_1, X_2...X_3$ identically distributed but dependent r.v.
the distribution is a Bernoulli distribution which have p = 0.4, so value 1 with prob 0.4 and value 0 with prob. 0.6.
the E$[X_1] = 0.4$ and that is the same for all the other variables because they follow the same distribution.
Now let's say that $P(X_2|X_1) = 1, P(X_3|X_2) = 1...P(X_n|X_{n-1}) = 1$ this means that for n that goes to infinity the mean will not converge to the expected value, instead the mean will tend to 1.
the final result $P(|\mu - mean| \geq 1/2) = 1$ for n that goes to infinity will be $P(|0.4 - 1| \geq 1/2) = 1$ thus the example will be proven, independence is crucial for convergence of empirical means to the expected values.

3 Tightness of Markov's Inequality (14 points)

I can define a r.v. X which has value 1 with probability 1/k and value 0 with probability 1 - 1/k.
This means that the r.v. X has $E[X] = \frac{1}{k}$
$\mathbb{P}(X \geq \epsilon^*) = \frac{E[X]}{\epsilon^*}$
$\mathbb{P}(X \geq \epsilon^* * E[X]) = \frac{1}{\epsilon^*}$
$\mathbb{P}(X \geq \epsilon^* * \frac{1}{k}) = \frac{1}{\epsilon^*}$
taking $\epsilon = k$
$\mathbb{P}(X \geq k * \frac{1}{k}) = \frac{1}{\epsilon^*}$

$\mathbb{P}(X \geq 1) = \frac{1}{k}$
which is exactly the value 1 with prob. 1/k of my r.v.
So for $\epsilon = k$, with the defined random variable X, I get a tight Markov's bound.

4 The Effect of Scale (Range) And Normalization of Random Variables in Hoeffding's Inequality (14 points)

$\mathbb{P}(|\sum_{i=1}^{n} X_i - E[\sum_{i=1}^{n} X_i]| \geq \epsilon) \leq 2e^{-2\epsilon^2 / \sum_{i=1}^{n}(b_i - a_i)^2}$

$\mathbb{P}(|\sum_{i=1}^{n} X_i - E[\sum_{i=1}^{n} X_i]| \geq \epsilon) \leq 2e^{-2\epsilon^2/n}$

replace ϵ with $n\epsilon$

$\mathbb{P}(|\sum_{i=1}^{n} X_i - E[\sum_{i=1}^{n} X_i]| \geq n\epsilon) \leq 2e^{-2n\epsilon^2}$

removing the abs. value positive case:

$\mathbb{P}(\sum_{i=1}^{n} X_i - E[\sum_{i=1}^{n} X_i] \geq n\epsilon) \leq 2e^{-2n\epsilon^2}$

I obtain the same bound by diving what's inside the probability by 1/n, because 1/n is positive and I multiply it in both terms of the inequality.
I was able to substitute the μ because we have in the corollary 2.5 of the book the assumption that the variables are identically distributed.

$\mathbb{P}(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geq \epsilon) \leq 2e^{-2n\epsilon^2}$

5 Logistic Regression

5.1 Cross-entropy error measure (11 points)

With w and x as vectors:

a)

$\prod_{n=1}^{N} \mathbb{P}(y_n|x_n)$

that is the equivalent of minimizing a more convenient quantity:

$-ln(\prod_{n=1}^{N} \mathbb{P}(y_n|x_n))$

$\sum_{n=1}^{N} ln(\frac{1}{\mathbb{P}(y_n|x_n)})$

then using the error measure of page 91, we obtain 3.7 at page 90 of Logistic regression chapter from Abu-Mostafa's book:

$\sum_{n=1}^{N} [\![y_n = +1]\!] ln(\frac{1}{h(x_n)}) + [\![y_n = -1]\!] ln(\frac{1}{1-h(x_n)})$

b)

If I merge the two cases, I will obtain that: $h(x) = \theta(w^T x)$ and $1 - h(x) = 1 - \theta(w^T x) = \theta(-w^T x)$ and so we get : $\mathbb{P}(y|x_n) = \theta(y w^T x)$

$\sum_{n=1}^{N} [\![y_n = +1]\!] ln(\frac{1}{h(x_n)}) + [\![y_n = -1]\!] ln(\frac{1}{1-h(x_n)})$

$\sum_{n=1}^{N} ln(\frac{1}{\mathbb{P}(y_n|x_n)})$

$\sum_{n=1}^{N} ln(\frac{1}{\theta(y_n w^T x_n)})$

substituting the functional form $\theta(y_n w^T x_n)$ produces the training error measure for logistic regression:

$\sum_{n=1}^{N} ln(1 + e^{-y_n w^T x_n})$

And since we are minimizing it should be the same to add the constant 1/N:

$\frac{1}{N} \sum_{n=1}^{N} ln(1 + e^{-y_n w^T x_n})$

5.2 Logistic regression loss gradient (13 points)

$E_{in} \nabla(w) \frac{1}{N} \sum_{n=1}^{N} ln(1 + e^{-y_n w^T x_n})$

$\frac{1}{N} \sum_{n=1}^{N} \frac{1}{1+e^{-y_n w^T x_n}} e^{-y_n w^T x_n} * (-y_n x_n)$

$\frac{1}{N} \sum_{n=1}^{N} \frac{e^{-y_n w^T x_n}}{1+e^{-y_n w^T x_n}} * (-y_n x_n)$

substitute the logistic function

$\frac{1}{N}\sum_{n=1}^{N}\theta(-y_n w^T x_n) * (-y_n x_n)$

Argue that a 'misclassified' example contributes more to the gradient than a correctly classified one:

$\nabla E_{in}(w) = -\frac{1}{N}\sum_{n=1}^{N}\frac{y_n x_n}{1+e^{y_n w^T x_n}}$

In the case of 'misclassified' example, the y and w will have different signs, so the resulting sign from their multiplication is negative.
This means that the $e^{y_n w^T x_n}$ will become part of the numerator and the contribution for the gradient it's more than in the case of correctly classified sample, because in this last case the denominator will decrease the overall value of the gradient (in fact it won't become part of the numerator).

point 2 of the exercise, repeat it but with [0,1] as errors:

we know that given $\mathbb{P}(y|x) = h(x)$ if $y = 1$ else $1 - h(x)$ if $y = 0$

if we substitute h(x) by its value $\theta(w^T x)$ we can use the fact that $1 - \theta(s) = \theta(-s)$.

So if y = 1 we get $\theta(w^T x)$ else if y = 0 we get $1 - \theta(w^T x)$ which is $\theta(-w^T x)$

If we rely on Max likelihood: $\prod_{n=1}^{N}\mathbb{P}(y_n|x_n)$ and we minimize a more convinient quantity: $\frac{1}{N}\sum_{n=1}^{N} ln(\frac{1}{\mathbb{P}(y_n|x_n)})$

now substituting with what we have defined:
$\frac{1}{N}\sum_{n=1}^{N}[\![y = 1]\!]\ \ln(\frac{1}{\theta(w^T x_n)}) + [\![y = 0]\!]\ln(\frac{1}{\theta(-w^T x_n)})$

Let's compute the gradient for both the 2 cases and let's compare it to what we have in the slides:

$\nabla(w)\ln(\frac{1}{\theta(w^T x_n)}) = \nabla(w)\ln(\frac{1+e^{w^T x}}{e^{w^T x}})$
without making all steps, the gradient is:
$-\frac{x}{e^{w^T x_n}+1}$

the same for: $\nabla(w)\ln(\frac{1}{\theta(-w^T x_n)}) = \nabla(w)\ln(\frac{1+e^{-w^T x}}{e^{-w^T x}})$
without making all steps, the gradient is:

$-\frac{xe^{w^T x_n}}{e^{w^T x_n}+1}$

Now, given what we have in the slides: $\nabla(w) = -\frac{1}{N}\sum_{n=1}^{N}[y_n - \theta(w^T x)]x_n$
if we take the case in which $y_n = 1$ we get $(1 - \theta(w^T x))x_n = \theta(-w^T x)x_n$

$\theta(-w^T x)x_n = \frac{e^{-w^T x}}{1+e^{-w^T x}}x_n = \frac{x_n}{1+e^{w^T x_n}}$
This last quantity is what we have obtained before with the gradient of the positive case with y = 1 except for a -, which we can take from the formula in the slides, so with y = 1 the two elements correspond.

Given what we have in the slides: $\nabla(w) = -\frac{1}{N}\sum_{n=1}^{N}[y_n - \theta(w^T x)]x_n$
if we take the case in which $y_n = 0$ we get $(0 - \theta(w^T x))x_n = -\theta(w^T x)x_n$

$-\theta(w^T x)x_n = -\frac{e^{w^T x}}{1+e^{w^T x}}x_n = -\frac{x_n e^{w^T x}}{1+e^{w^T x_n}}$
Which is the same quantity obtained in the gradient in the negative case, when y = 0.

so by doing a recap:

$$\nabla(w) = -\frac{1}{N}\sum_{n=1}^{N}[\![y = 1]\!]\frac{x_n}{1 + e^{w^T x_n}} + [\![y = 0]\!] - \frac{x_n e^{w^T x}}{1 + e^{w^T x_n}}$$

And they correspond to what we have obtained from the gradient of the starting quantity.

Argue that a 'misclassified' example contributes more to the gradient than a correctly classified one:

in the case of 'missclassification' in the case we have a 1 and we predict a low prob for the positive class, logistic function goes to something similar or near 0, so we have a bigger error w.r.t. the case in which we have 1 and we predict a prob. close to 1, analogously for the case in which we have 0.

5.3 Log-odds (11 points)

with w and x as vectors
$w^T x + b = ln(\frac{\mathbb{P}(Y=1|X=x))}{\mathbb{P}(Y=0|X=x))}$

now consider that:
$\mathbb{P}(Y|x) = h(x)$ for $y = +1$
$\mathbb{P}(Y|x) = 1 - h(x)$ for $y = -1$
let this h(x) be the logistic function, we obtain:

$\mathbb{P}(Y = 1) = \frac{e^{w^T+b}}{1+e^{w^T+b}}$
$\mathbb{P}(Y = 0) = 1 - \frac{e^{w^T+b}}{1+e^{w^T+b}} = \frac{1+e^{w^T+b}-e^{w^T+b}}{1+e^{w^T+b}} = \frac{1}{e^{w^T+b}}$

$ln(\frac{\mathbb{P}(Y=1|X=x))}{\mathbb{P}(Y=0|X=x))} = ln(\frac{\frac{e^{w^T+b}}{1+e^{w^T+b}}}{\frac{1}{e^{w^T+b}}}) = ln(e^{w^T+b}) = w^T + b$

And now it's proven that using the logistic function, $w^T + b$ are encoded with the log-odds, if we would have used another function this final result wouldn't have been possible.