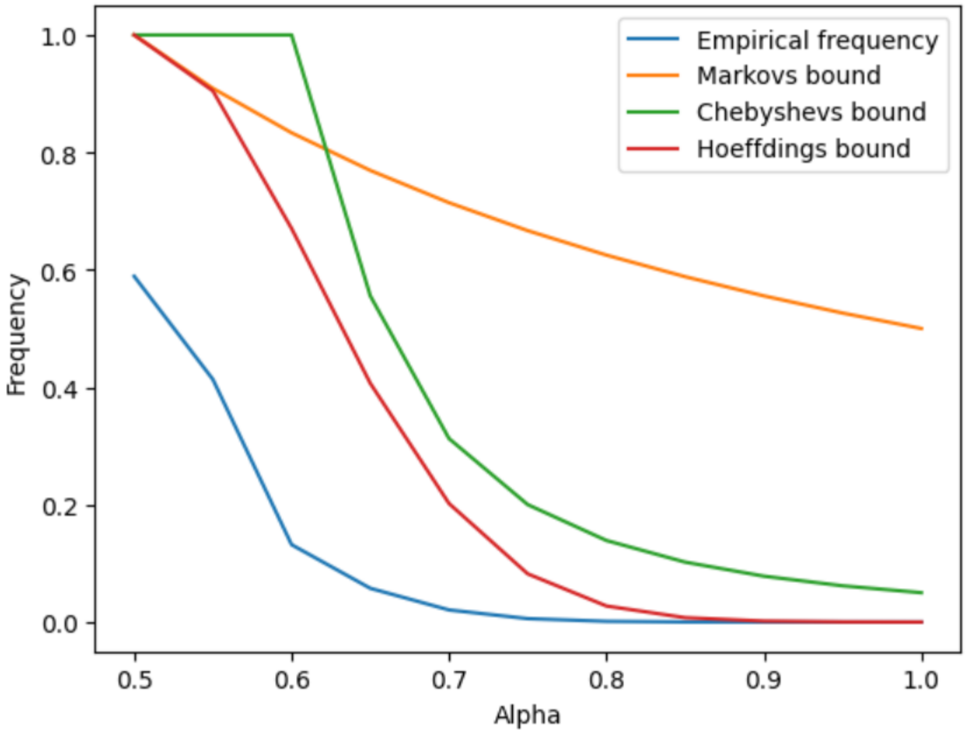
Machine Learning A (2023)
Home Assignment 2

Alexander Husted | wqg382

Contents

1 Markov's, Chebyshev's, and Hoeffding's Inequalities

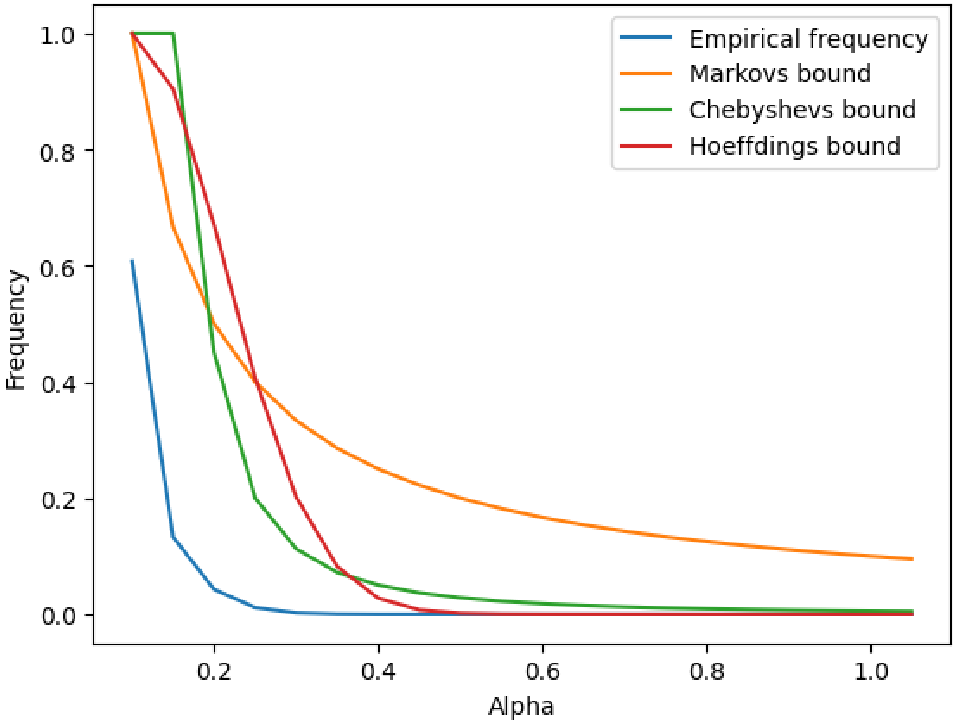


Question 2
If we incremented alpha by 0.01 instead of by 0.05, we would still get the same curve, but smoother. A smoother curve would not help us understand the empirical frequency as alpha approches one, but only increasse the runtime.

Question 6
We see that the tightest bound is Hoeffdings bound.

Question 7
For $\alpha = 1$ we get an empirical frequency of 0.
For $\alpha = 0.95$ we get an empirical frequency of 0.000002.

Question 1.b



Question 1.c

We see that the Emipirical frequency and the bounds decresses much faster as alpha goes from bias to 1, when bias is set to 0.1. In both plot the Empirical frequenct and hoeffdings bound converges to 0.

2 The Role of Independence

Imagine that we have a set of n led-lamps, the proberbility p of the first lamp turning green is $\frac{1}{2}$ otherwise it turns red. The lamps are dependt of each other st. if the first lamp turns red the remaining $n-1$ lamps also turns red.

Let $red = 0$ and $green = 1$ such that $X_i \in \{0,1\}$
Then the true expected value is $E(X_i) = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{1}{2}$
But if we try to run the expiriment we get that

$$\hat{\mu}_n = \begin{cases} \frac{1}{n}\sum_{i=1}^{n} X_i = 1 & if\ X_1 = 1 \\ \frac{1}{n}\sum_{i=1}^{n} X_i = 0 & Otherwise \end{cases}$$

We put the numbers into the formular:

$$P(|\mu - \hat{\mu}_n| \geq \frac{1}{2}) = 1 \Rightarrow$$

$$1 = \begin{cases} P(|\frac{1}{2} - 1| \geq \frac{1}{2}) = P(| - \frac{1}{2}| \geq \frac{1}{2}) & if\ X_1 = 1 \\ P(|\frac{1}{2} - 0| \geq \frac{1}{2}) = P(|\frac{1}{2}| \geq \frac{1}{2}) & Otherwise \end{cases}$$

Which is True, since the posibility of $\frac{1}{2}$ being greater or equal to $\frac{1}{2}$ is 1.

3 The effect of scale (range) and normalization of random variables in Hoeffding's inequality

We have to show that Corollary 2.5

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

Follows Theorem 2.3

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

We see in the lecture notes that $\hat{\mu}_n$ converges to μ at the rate of n^{-1}, therefor lets choose $\epsilon = \epsilon^* n^{-1} = \frac{\epsilon^*}{n}$

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

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

We know that $E(X_i) = \mu$

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

Thus we have that collary 2.5 follows Theorem 2.3.

4 Linear Regression

4.1 Question 1: Show your implementation

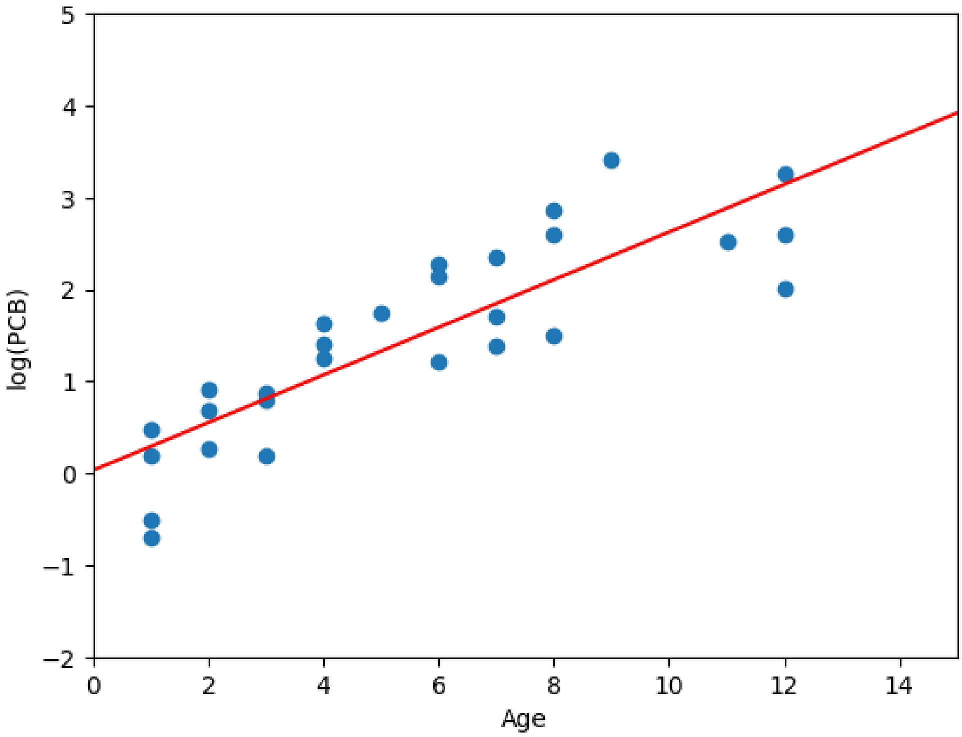
```
def linreg(X, y):
    w = np.dot(X.T, X)
    w = np.linalg.inv(w)
    w = np.dot(w, X.T)
    w = np.dot(w, y)
    return w
```

4.2 Question 2: Non-linear model

I got he following parameters:
a = 0.25912824
b = 0.03147247

The mean squared error is 34.83556116722035

4.3 Question 4 : Plot the data and the model output



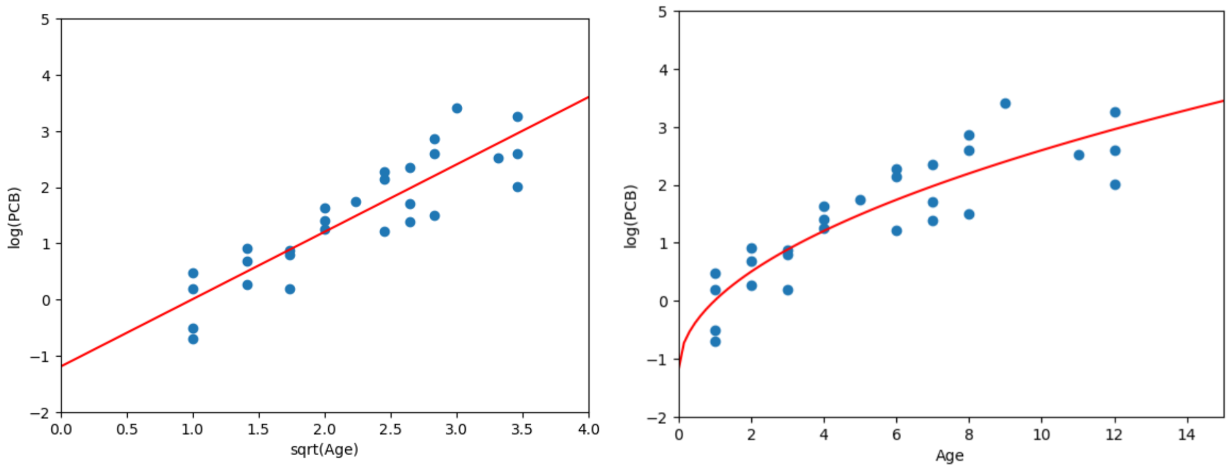
4.4 Question 5 : Compute the coefficient of determination

$$R^2 = 0.3570135731609865$$

R^2 tells us how much of variability can be explained by the model. Meaning that if 100% of the variability can be explained by the model ($R^2 = 1$), we would have a perfect fit and our line would go through all points in the plot. On the other hand $R^2 = 0$ means that there is no linear relationship between the points.

R^2 cannot be equal to zero. R is the correlation between the x and y, if you square this you get R^2. R can be a negative number, but when squared it becomes positive. This can also be seen in the formular where both the numerator and the denominator is a sum of squared (and therefore positive) numbers.

4.5 Question 6 : Build a non-linear model



(a) With transformed x-axis (b) Without transformed x-axis

$$MSE = 28.084390174944378$$

$$R^2 = 0.4816250669292409$$

We see that after we transformed x, our model explains more of the variability in the data. Meaning that the models fits better to the data points, which we know since R^2 is larger. We also have a smaller deviation from the mean, because the MSE is lower.