Aashsih Kumar pcr902

January 7, 2020

Contents

1 How Variance influences Concentration

1.1 Empirical Frequency

I perform an experiment of choosing 20 random variables (with 0 for failure and 1 for success) 1,000,000 times. Then, calculate $\frac{1}{20}\sum_{i=1}^{20} X_i \geq \alpha$ for $\alpha \in (0.05, ..., 0.1)$ for each experiment. Then, calculate the empirical frequency for each α. See `Part1.py` file in `code.zip`

1.2 Markov bound

X_i can be 0 or 1 and probabilities of success and failure are 0.05 and 0.95 respectively. Expectation E[X] = 0.05 * 1 + 0.95 * 0 = 0.05

According to the markrovs inequality

$$P(\frac{1}{20}\sum_{i=1}^{20} X_i \geq \alpha) \leq \frac{1}{20} * \frac{\sum_{i=1}^{20} E[X_i]}{\alpha} \tag{1}$$
$$\leq \frac{0.05}{\alpha} \tag{2}$$

Hence, Markrov Bound is $\frac{0.05}{\alpha}$

1.3 Chebychev bound

let $Y = \frac{1}{20}\sum_{i=1}^{20} X_i$ be the random variable

$$E[Y] = E[\frac{1}{20}\sum_{i=1}^{20} X_i] \tag{3}$$
$$= 0.05 \tag{4}$$

here, $E[X^2] = E[X]$ as $X = X^2 \in 0, 1$

$$Var[\frac{1}{20}\sum_{i=1}^{20} X_i] = \frac{1}{20^2}\sum_{i=1}^{20} Var[X_i] \tag{5}$$
$$= \frac{1}{20^2}\sum_{i=1}^{20}(E[X_i^2] - (E[X_i])^2) \tag{6}$$
$$= \frac{1}{20}(0.05 - (0.05)^2) \tag{7}$$
$$= 0.002375 \tag{8}$$

then, according to the chebychev inequality

$$P(|Y - E[Y]| \geq \epsilon) \leq \frac{Var[Y]}{\epsilon^2} \tag{9}$$
$$P(|Y - 0.05| \geq \epsilon) \leq \frac{0.002375}{\epsilon^2} \tag{10}$$
$$P(Y \geq \epsilon + 0.05) \leq \frac{0.002375}{\epsilon^2} \tag{11}$$

By comparing $P(Y \geq \alpha)$ with equation 11,
$\alpha = \epsilon + 0.05$
$\epsilon = \alpha - 0.05$
So, $P\left(\frac{1}{20}\sum_{i=1}^{20} X_i \geq \alpha\right) \leq \frac{0.002375}{(\alpha-0.05)^2}$
Hence, chebychev Bound is $\frac{0.002375}{(\alpha-0.05)^2}$

1.4 Hoeffding Bound

According to Hoeffding's inequality, we have

$$P(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geq \epsilon) \leq e^{-2n\epsilon^2} \tag{12}$$

$$P(\frac{1}{n}\sum_{i=1}^{n} X_i \geq \epsilon + \mu) \leq e^{-2n\epsilon^2} \tag{13}$$

$$P(\frac{1}{n}\sum_{i=1}^{20} X_i \geq \epsilon + 0.05) \leq e^{-2*20*\epsilon^2} \tag{14}$$

By comparing $P\left(\frac{1}{20}\sum_{i=1}^{20} X_i \geq \alpha\right)$ with equation 14,
$\alpha = \epsilon + 0.05$
$\epsilon = \alpha - 0.05$
So,

$$P\left(\frac{1}{20}\sum_{i=1}^{20} X_i \geq \alpha\right) \leq e^{-2*20*(\alpha-0.05)^2} \tag{15}$$

Hence, Hoeffding bound is $e^{-2*20*(\alpha-0.05)^2}$

1.5 Plot for Biased Experiment : Empirical Frequency, Markov bound, chebychev bound and hoeffding bound

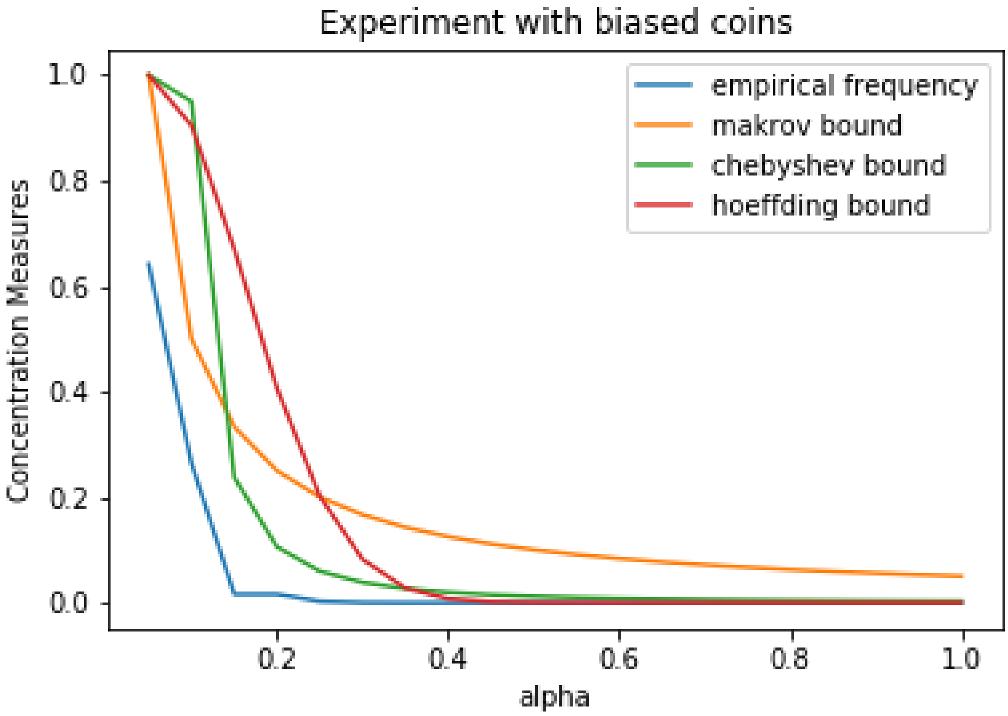

Figure 1: Biased Experiment : Empirical Frequency, Markov bound, chebychev bound and hoeffding bound

1.6 Plot for Unbiased Experiment : Empirical Frequency, Markov bound, chebychev bound and hoeffding bound

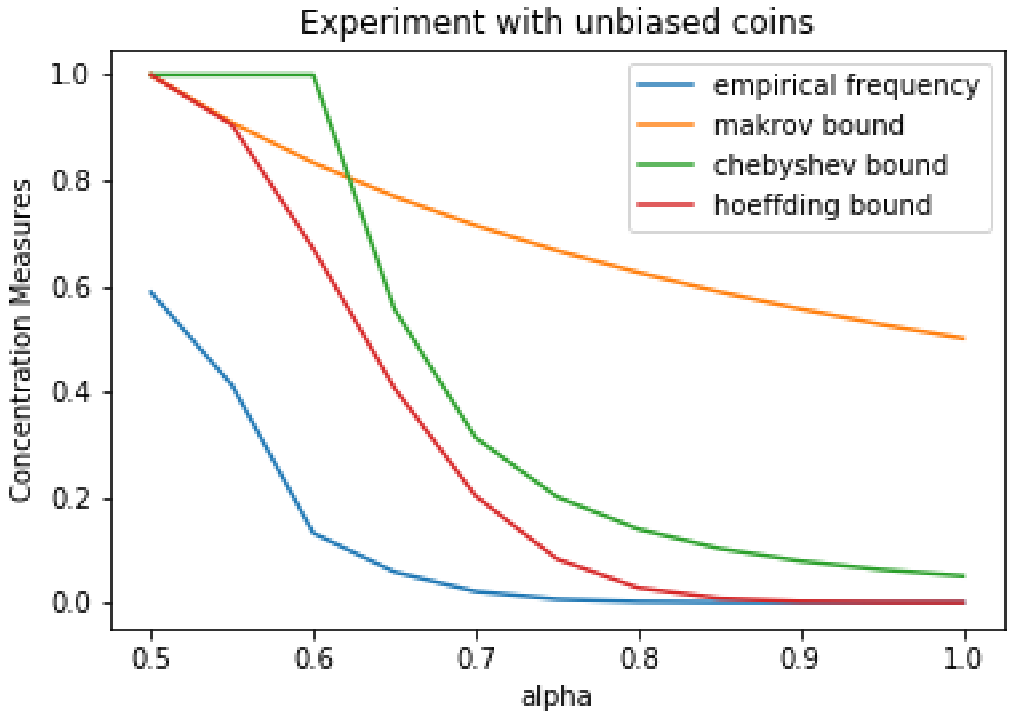

Figure 2: Unbiased Experiment : Empirical Frequency, Markov bound, chebychev bound and hoeffding bound

1.7 Comparison between above plots

In the biased experiment, the markov bound, chebychev bound and hoeffding bound are more tighter compared to the unbiased experiment. for alpha atleast 0.5 the chebychev and hoeffding bound are almost same in the biased experiment. whereas in the unbiased experiment only hoeffding bound coincides with the empirical frequency and that too for alpha greater than 0.8 .

2 A bit more on the VC-Dimensions

2.1 VC-dimension of the hypothesis space $\mathcal{H}_d$ of binary decision trees of depth d

The size of the hypothesis is 2^{2^d}, where d is depth of the tree. We know from previous assignment that

$$\begin{aligned} VC-dimesion &\leq \log_2 M, \quad M \quad size \quad of \quad hypothesis & (16)\\ &\leq \log_2(2^{2^d}) & (17)\\ &\leq 2^d & (18) \end{aligned}$$

And the number of points that can be shattered by the tree with depth is atleast the number of leaves i.e, 2^d. Hence, VC-dimension of the hypothesis space $\mathcal{H}_d$ of binary decision trees of depth d is 2^d.

2.2 VC-dimension of the hypothesis space $\mathcal{H}$ of binary decision trees of unlimited depth

The size of the hypothesis with unlimited depth is infinite. So, for any number of points, the set of tree can shatter them. so, the Vc-dimension of the hypothesis space $\mathcal{H}$ of binary decision trees of unlimited depth would be infinity.

3 Separating Hyperplanes

Using Theorem 3.21 from Lecture Notes

$$\mathcal{P}\left(\exists h \in \mathcal{H}_\gamma : L_{FAT}(h) \geq \hat{L}_{FAT}(h,S) + \sqrt{\frac{8\ln\left(2\left(\left(2n\right)^{d_{FAT}(\mathcal{H}_\gamma)}+1\right)/\delta\right)}{n}}\right) \leq \delta \qquad (19)$$

We can also read the above theorem as : with probability atleast $1-\delta$ for all $h \in \mathcal{H}$, we have

$$L_{FAT}(h) \leq \hat{L}_{FAT}(h,S) + \sqrt{\frac{8\ln\left(2\left(\left(2n\right)^{d_{FAT}(\mathcal{H}_\gamma)}+1\right)/\delta\right)}{n}} \qquad (20)$$

We are given n = 100000, margin,$\gamma = 0.1$, $\hat{L}_{FAT}(h,S) = 0.01$, R = 1 and probability, $1-\delta = 0.99$
then, $\delta = 0.01$
From using Theorem 3.19 from Lecture notes, we can calculate $d_{FAT}(\mathcal{H}_\gamma)$

$$d_{FAT}(\mathcal{H}_\gamma) \leq \frac{R^2}{\gamma^2} + 1 \tag{21}$$
$$\leq \frac{1^2}{(0.1)^2} + 1 \tag{22}$$
$$\leq 101 \tag{23}$$

After substituting the values in equation 20, we get the bound on expected loss

$$L_{FAT}(h) \leq 0.01 + \sqrt{\frac{8 \ln \left(2 \left((2 * 100000)^{101} + 1\right)/0.01\right)}{100000}} \tag{24}$$
$$\leq 0.3247 \tag{25}$$

4 The details of the lower bound

Not Attempted

5 Random Forests

5.1 Normalization

5.1.1 Nearest Neighbour Classification

Yes,Nearest Neighbour Classification is affected by the normalization. It is based on distances. Suppose, we have a dataset with m features where one feature has values in range -1000 to 1000 and all other are from 0-1. In this case, most of the clusters would be based on the values from the feature with range -1000 to 1000 as their values are larger and the difference between 0 to 1 range is very small compared to -1000 to 1000 range. To avoid this we use normalization to scale all the features.

5.1.2 Random Forest Classification

No, Random Forest Classification is not affected by the normalization. It is tree based model and requires partitioning. It is not affected by the monotonic transformations to the individual features. Those transformations would not change anything for Random forest as one feature is never compared to any other feature. Thus, even if normalize the features the output remains same.

5.2 Random Forests in Practice

Accuracy on test data : 0.8762
loss = 0.1238
Following line defines the classifier. Refer `Part5.py` file in `code.zip`

```
classifier = RandomForestClassifier(n_estimators=50)
```