Assignment #1

Paulvir Dhanda
CMPT459
Simon Fraser University

June 1, 2020

Question 1

For this question I will talk about an experience as a Data Engineering Co-op. At our company we had live feeds of X-ray data that came into our system and our data pipeline had an ideal pattern of what the readings should look like. At a certain energy level the curve of the ideal reading would be Gaussian. So, when live readings were coming into our system, we needed a way to see if we could trust the sensor data. What we did was taking prior live x-ray readings and did a least squares fitting on all of them. Then we did a Chi-squared goodness of fit test to classify x-rays into two populations: good fits and bad fits, or in other words, x-ray reading we could trust and not trust. This can be seen in Figure 1 where the blue represents good fits and the red represent bad fits.

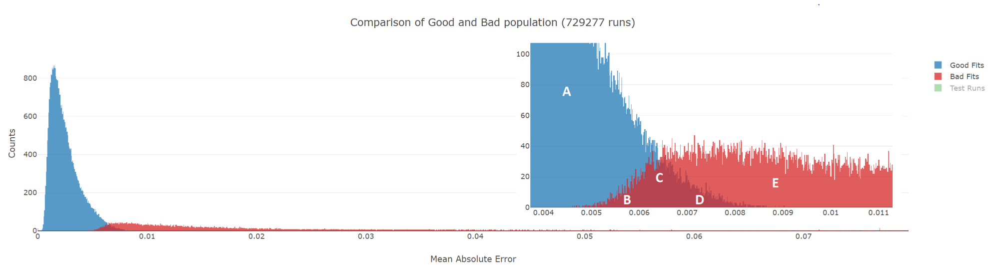

Figure 1: Good Fits vs. Bad Fits with Mean Absolute Error Distribution

Since it would be time consuming to do a chi-squared analysis on every x-ray reading that came into our system, we found a mean absolute error value (some where where blue and red over-lap in Figure 1) that would be a cut off point using the data gathered by the Chi-squared goodness of fit test. We would then use that mean absolute error value as a cut-off point to classify live x-ray data coming into our system.

This is a good example that encompasses pattern mining (knowing our peaks are Gaussian), classification (classifying good fits and bad fits) and other data mining techniques.

Question 2

Question 2.1

For this question the Chi-Square test to see if two samples come from the same distribution come from this link. The formula from this link suggest the formula for the Chi-square test is as follows:

$$X^2 = \sum_{i}^{n} \frac{(S_i - R_i)^2}{S_i + R_i}$$

Where S_i are observation of sample 1 at i and R_i are observations of sample 2 at i. If let Sample 1 be the realizations for P.J. Tucker and Sample 2 be the realizations for Brook Lopez, then:

$$X^2 = 22.5499 \approx 22.55$$

This reference stats that for the Chi-squares test the hypothesis are as follows:

$$H_0 = \text{Assumes that there is no association between the two variables.}$$
$$H_a = \text{Assumes that there is an association between the two variables}$$

And from the same reference, we get our degrees of freedom calculation. If we let $r =$ number of rows and $c =$ number of columns. Then:

$$DF = (r-1)(c-1)$$
$$DF = 4$$

Now, with $X^2 = 22.55$ and a DF = 4 we can do a look up on this chi-square table and see that our test statistic is really large for DF = 4. Hence, we reject H_0 and accept H_a. Concluding that our two samples do observe to have an association.

Question 2.2

In this question, we will calculate the KL-Divergence of PJ Tucker diverging from Brook Lopez. From the lecture slides, the KL-Divergence is given by the following:

$$D_{KL}(p(x)||q(x)) = \text{divergence of q(x) from p(x)}$$
$$D_{KL}(p(x)||q(x)) = \sum_{x \in X} p(x) \ln \frac{p(x)}{q(x)} \text{ , s.t. p,q are two probability distributions}$$

Now for our case, we are measuring the KL-Divergence of PJ Tucker diverging from Brook Lopez. Thus $q(x) =$ distribution of PJ Tucker and $p(x) =$ distribution of Brook Lopez. The following table shows us the calculated values for $p(x)$ and $q(x)$.

```
x           p(x)      q(x)
-----------------------------------
0 - 3       0.166     0.155
3 - 10      0.101     0.093
10 - 16     0.014     0.045
16 - 3pt    0.008     0.020
3pt         0.711     0.687
```

Plug in these values into our equation we get:

$$D_{KL}(p(x)||q(x)) = 0.024414416$$

Question 2.3

For this question, we must answer how Chi-Square and KL-divergence are related to each-other. Using this question as an example, we can see the they are inversely related. That is, when X^2 is large then $D_{KL}(p(x)||q(x))$ is small. This is because a larger X^2 value means that the two distributions are more similar. Same is true for a smaller $D_{KL}(p(x)||q(x))$ value.

Question 3

Question 3.1

Number of Unique Tokens:

$$D1 : 4492$$
$$D2 : 3756$$

Question 3.2

The frequency of the top 100 tokens from each data set can be found in the **-top100.csv* files for their respective data set.

Question 3.3

The word cloud can be found in **.png* files. They were generated using python (for creating the .csv file) and https://www.wordclouds.com/. The size represents the frequency of the word in the data set and the color doesn't represent anything.

Question 3.4

I would propose that the two word-clouds be put on the same image and probably be separated by color. Also, any word that appears in both data sets would be a gradient based on the occurrence in the data sets.