Anomaly Detection Exercise:

1a. Recall the 68-95-99.7 rule, we only have 0.3% of data points outside of 3 standard deviations above or below the mean. Therefore, we have 0.3% \* 1000000 = 3000 outliers.

1b. No, it does not make sense. The uniform distribution looks like a rectangle, which means it is not true that most of the data points are clustered around the mean. Therefore, we do not have points(outliers) very far from the mean and we have 0% of data points outside of 3 standard deviations above or below the mean. Not a good way to identify outliers. (Similar to question 3 below)

2.

图片包含 图表

描述已自动生成

3.

文本, 信件

描述已自动生成

4. First method: I would compute the points which are greater than 4 standard deviations above or below the mean and consider them as outliers.

Second method: Consider all T points as a new dataset and recompute the points which are greater than 3 standard deviations above or below the mean and consider them as outliers.

Third method: Apply PCA to reduce the dimension of the dataset and then find out the points which are greater than 3 standard deviations above or below the mean and consider them as outliers.

5. Similar to isolation forest, I would draw the dendrogram first and then set a threshold of depth. Any points above that depth threshold are considered anomaly points. Anomalies usually have shorter path lengths.

图示

描述已自动生成

6. How would you define “much farther away?”

I think we can calculate the Euclidean distance of all points from their cluster center and put them in a list. Then we select the top several points (maybe top 1%) with the largest distance as outliers.

For this process, k-means clustering for anomaly detection, how might you decide on the optimal value for k?

We can use the elbow method and find the most appropriate k values corresponding to the changing point.

Final Exercise:

3. The observations that do not get chosen by bootstrap sampling will be considered as out-of-bag samples. And they are used as the test set for calculating CER.

Eg. Sample: 1,2,3,4,5

Bootstrap sample: 1,2,2,3,3

OOB sample: 4,5

We use bootstrap sample to train our model and 4,5 to fit the model and compute CER or accuracy.

5. I design a small game about sampling with replacement and invite some students in my hallway to play it. I first made eight cards with numbers from 1 to 8 representing 8 different observations. I then asked my friends to randomly pick eight cards with replacement and recorded the number they picked.

图片包含 桌子, 电脑, 照片, 不同

描述已自动生成文本, 信件

描述已自动生成

The percentage of original sample show up in our bootstrap sample:

75%, 62.5%, 62.5%, 87.5%, 75%, 37.5%, 50%, 50%

The average is 62.5%, which is very close to the theoretical average 64%!

6. Submitted!

7. I think I would choose the job related to supervised learning and work on improving model prediction for solving real-world problems. Utilizing data power to address real-world challenges is a practical career field I would like to be actively engaged in in the future. I hope I can do a project related to the business market or sales prediction (estimate sales price or stock price prediction), where I can leverage my advanced knowledge in statistical learning and data analytics to help gather, interpret and deliver the insights to the business decision-makers or clients.