BINF 8500

Assignment 2 – Kmeans

Ruijie Xu

The results of the kmeans wen exactly as I expected. As the number of clusters k increases, the WCSS as decreases. However, the rate of decreasing was not constant. The rate was first decreased rapidly, then to the point that the decrease is negligible. Increase in number of clusters is to increase the fitting of the clustering model. Since Kmeans measures the similarities among datapoints in a cluster, the small clusters indicate higher similarities. However, very small clusters (in other word, more number of clusters) can lead to overfitting. As we can imagine, with the number of cluster equals to the number of data point would result in a 0 WCSS, because there are no differences produced with only one point per cluster. This approach, therefore, lost the purpose of the clustering algorithm. We want to locate number of clusters (k) that provides enough similar in between clusters, but also avoid the danger of overfitting, this would come to one of the commonly used techniques, which is to identify number of clusters from the elbow of the WCSS curve. From the results of my kmeans implementation (shown below), k with value 3-5 for the bacteria data points, for example, seems like the most optimal number of clusters to choose. The number of k’s differs for different dataset, for example, for the dataset Bacteria+Archaea, a k above 4 would seem like a bad choice.

AIC and BIC are also a good elevation scale for the best number of clusters to choose. Both methods use a penalty system, so the standard is more sensitive to the increase number of k’s. AIC straightly added penalty with respect to the scale of dimensions and number of k’s. Thus, from the plots below, we observe an increase in AIC as number of k’s increases. However, because all three testing data sets have points with same number of dimensions, the penalty does not add as much influence to the increase as BIC, which is a more sensitive penalization system. BIC has not only taken dimensions into account, but also number of datapoints with a log scale. Thus, we can observe the increases in Bacteria and Archaea datasets are significantly higher in BIC compare to that from Bacteria+Archaea dataset. This difference is due to the smaller sample sizes of these two datasets. Smaller dataset has more danger of overfitting; thus, BIC can serve as a better evaluation compare to AIC under this situation.

In general, the changing pattern of these three evaluation systems fit with my expectation. The differences between WCSS and AIC/BIC (penalty added), differences between AIC/BIC across datasets (BIC increases faster in smaller datasets) had all explained in the plots below.

