CSCC11 Discussion Questions A1

By: Yulun Wu and Abu Bakar Al-Hilal

Part 1, 1.1.5:

In our opinion we would say in most cases MAE is the better of the two metrics. Since MSE squares the errors it places a larger weight on them, not only that but this causes outlines to have extremely high weight placed on them. Thus MSE is highly susceptible to noise therefore it is a less accurate representation of the model as compared to MAE.

Part 1, 1.2.4 (b):

Linear regression:
MSE1 := test error (MSE):: 0.004576720874727214
MAE1 := test error (MAE): 0.05028702371274829

K-means regression:
MSE2 := test error (MSE):: 0.0046098700716455
MAE2 := test error (MAE): 0.049999556105782675

From above we can see that MAE1 > MAE2 which would indicate that cluster based linear regression with an optimal k is more accurate. Note the optimal k was 2 which means regular linear regression was already a pretty good model. However we also see that MSE2 > MSE1 suggesting that splitting the data into clusters caused more data points to be viewed as outliers on their respective cluster regression lines as opposed to the singular regression line. This likely happened at the border between the two kmeans where a bunch of data points were only slightly closer to one mean than the other which resulted in those points being treated like outliers in their cluster regression whereas in regular regression they were treated as regular points. Upon running the program a few more times we found that MSE1 can be greater than MSE2 so it seems it just depends on the random state of the training and testing data, also based on the randomness of K-means++. The difference in MAE on each run is so marginal it’s negligible so in conclusion we cannot really say that wether K-means or simple linear regression is better.