Big Data Laboratory

Week 4

Divya K Raman, EE15B085

The databricks notebook has been exported as a html file and submitted on moodle.

Questions:

1. Regression
   First we typecast all columns to datatype float. Data is randomly split into train and test in the ratio 0.8 : 0.2. The columns are then normalised. All this is put into a pipeline and linear regression is done. RMSE that we get is 18.58 and r2 is 0.002972. We then do cross validation and vary the value of the regularising parameter from 0.001 to 0.7. We obtain a final rmse of 18.62 and r2 of 0.0028.

   The rmse is high and r2 is low. This is because latitude and longitude are not good descriptors of altitude.

2. Clustering
   The dataset is split into 70 percent train and 30 percent test. All features except UNS which is the label are considered. Features are normalised and seed is set to 1L. Cross validation is done over the number of cluster centres which are set to range from 2 to 5. Clustering Evaluator is used to evaluate the performance of our model. We get a squared Euclidean distance of 0.35. Another way to evaluate how well the model performs is to calculate cluster purity using UNS label. Also, this can be viewed as a classification problem and we can apply logistic regression to classify. That might give us better results because supervised methods generally tend to work better than unsupervised methods.