Kenneth Le

UID:804953883

Problem Set 1

CS M146

Problem Set 1 Problem 4 Write Up

(a.) For the 14 features here are the general observations:

Workclass: Most people are self-employed in this census.Private workers seem to have the highest ratio of people making >50k while Government Workers have the inverse relationship making <50k.

Education: Higher levels of education result in higher ratio of people making over 50k

Marital\_status:Divorcees have the highest ratio of people making over 50k

Occupation: Exec-managerial occupations have the highest ratio of people making over 50k. Most people work in “Other-service”

Relationship:Husbands have the highest ratio of people making over 50k

Race: Most people in this census are white. Asian-pac-islander have the highest ratio of people making over 50k

Native\_country: Most people are from the US in this census and have the highest ratio of people making over 50k

Age:Younger people of age~20 have the lowest ratio of people making over 50k while age~40-50 have the highest.

Fnlwgt:Fraction of people represented is relatively consistent across all values in regards to people making more/less than 50k

Education\_nums:The general trend is that people who have been educated longer have a higher probability of making >50k. Most people have 10 years of education.

Capital\_gain:There is a strong relationship between people who make more capital gains will make >50k. Most people do not have any capital gains.

Capital\_Loss: Most people do not have capital-loss but shows that it is relatively balanced with increase of capital-loss.

Hours Per Week:Most people work ~40 hours a week.People who work longer hours tend to make more than >50k

Sex: Most of the people in this census are women with the highest ratio of people earning >50k.

(c.) When implementing the DecisionTreeClassifier method with the ‘entropy’ criterion we find that the training error is 0.00.

(d.)Training errors for the k nearest neighbors classifier where k = 3,5,7 are:

-- k = 3 training error: 0.153

-- k = 5 training error: 0.195

-- k = 7 training error: 0.213

(e.)The subsequent average training error,test error, and f1 score for majority,random,decision tree, and k nearest neighbors classifiers was:

-- Majority Training Error = 0.240

-- Majority Test Error = 0.240

-- Majority F1 Score = 0.760

-- Random Training Error = 0.375

-- Random Test Error = 0.382

-- Random F1 Score = 0.618

-- Decision Tree Training Error = 0.000

-- Decision Tree Test Error = 0.205

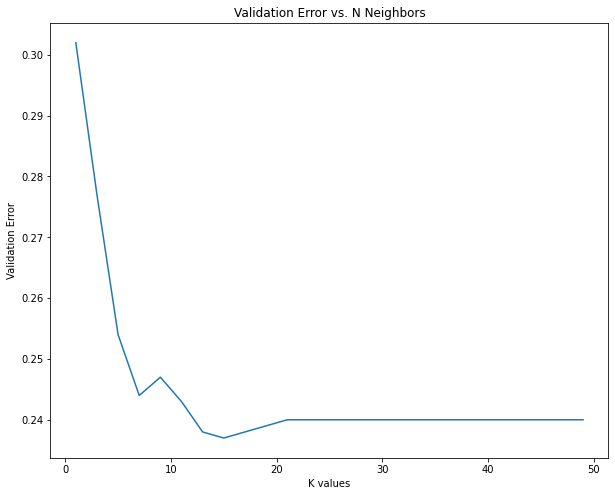
-- Decision Tree F1 Score = 0.795

-- KNeighbors Training Error = 0.202

-- KNeighbors Test Error = 0.259

-- KNeighbors F1 Score = 0.741

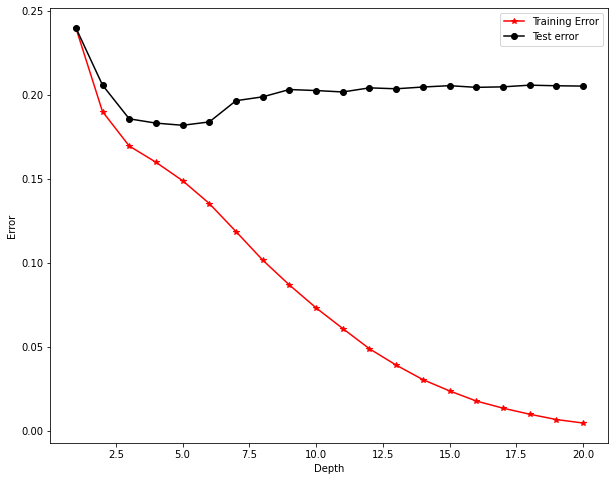
(f.) The plot for the k nearest neighbors classifier is shown below where the number of neighbors k is along the x axis and the y axis is the validation error.



The minimum validation error of 0.237 is when k = 15. As such , the best value for k is 15 as this is the right balance in between overfitting and underfitting.

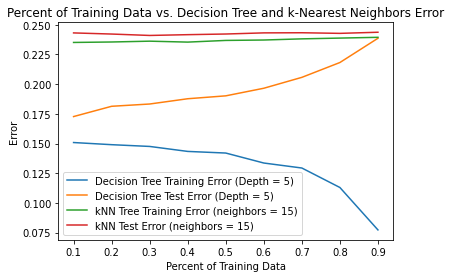
(g.) Down below we have the two curves representing the training error labeled in red with an ‘\*’

And the test error labeled in black with an ‘o’.



From this we can see that using cross validation the best depth is when it is equal to 5 as this is where the test data is the least at a value of 0.182. In the case where depth = 5 we can see that the difference between the training and test data is not so large indicating that it is not overfitting where if we increase the depth from this point we can see that it begins to overfit at more drastically as the difference between the test and training data becomes larger while the test error gets worse and then doesn’t improve.

(h.) From this graph we can see the training and test error of the classifiers. With the decision tree we see that as the percent of training data that we use increases along the x axis, the training error decreases while the test error increases. In the case of k-Nearest Neighbors, the training and test error are not affected by the change in the amount of training data.



i.)After standardizing the data I found that all the graphs with regards to the KNN classifier has changed where in the graph from part f. I found that the new best value of k is 27 with error of 0.173 and that the error for the last graph above has reduced as well. This is due to the fact that the different features are now standardized so that each of them are brought down to the same scale reducing the effect of outliers and distances over different scales from before.