CM-146 Problem Set 1

Reinaldo Daniswara
604840665

1. Splitting Heuristic for Decision Trees

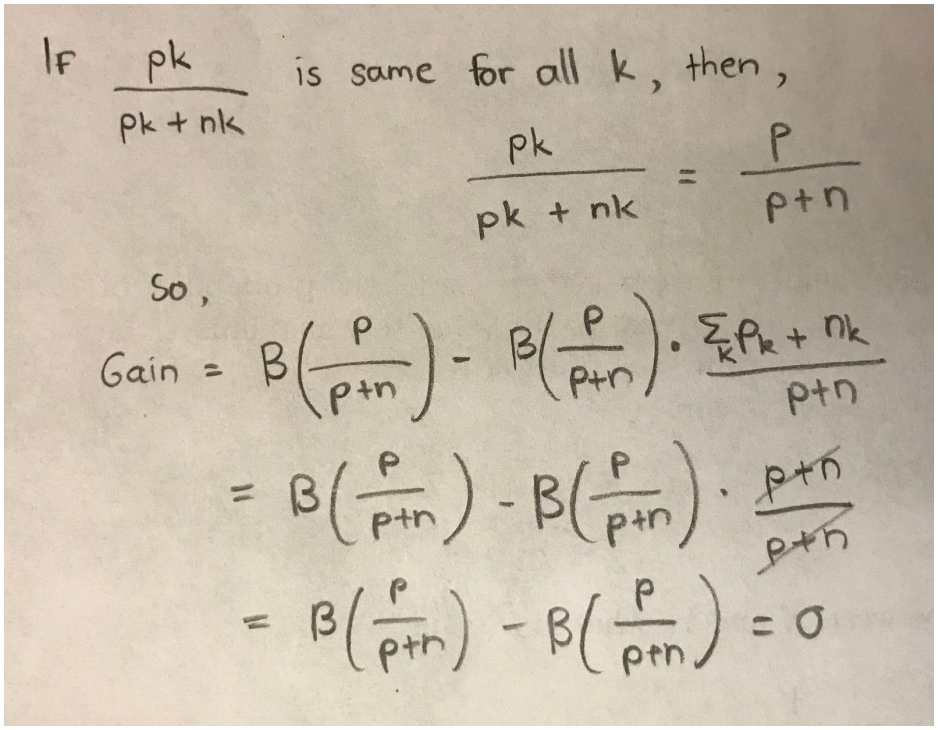
a. The best 1-leaf decision tree make over the 2^n training examples is 2^{n-3} mistakes. With the error rate of $2^{n-3}/2^n = 1/8^{th}$

b. There is no split that reduces the number of mistakes by at least 1. If we split X_i when $i \geq 4$, the proportion will still be the same (⅞), and both trees will still predict 1. Even when we try to split tree into X_1, X_2, X_3 one of the leaf will still have one leave that only contains 1. Thus, will still predict 1 in both leaves.

c. We can solve this problem by Bernoulli approach.
$Y \sim \text{bernoulli}\ (⅞) = (⅛) \log 8 + (⅞) \log (8/7) = 0.543$

d. Yes, there is a split that reduces the entropy of the output Y by a non-zero amount
$(1/8) \log(4) + (3/8) \log(4/3) = 0.406$

2. Entropy and Information



3. k-Nearest Neighbors and Cross-validation

a. Value of k to minimize the training set error of this dataset it k=0. And the error is 0

b. Have too large k value could lead to over classified the class of data, while too small value of k will lead to overfitting.

c. k=5 or k=7, and the resulting error is 4/14.

4. Applying Decision Trees and K-nearest neighbors

4.1. Visualization

- **Pclass**: Class number 1 (Upper Class) has more chance to survive, compare to the class 2 or class 3, while people in class 3 has the lowest chance to survive.
- **Sex**: Female has a higher chance to survive compare to male. 0 more frequent than 1.
- **Age**: If we just look on how many people that survive, people age 20 has the highest number of people survive. However, if we want to look on the frequency of survival, child under 10 years old roughly has a chance of 0.66, while people age 20 years old only 0.3 as the number of 20 years old passenger who is not survive also high.
- **Sibsp**: Passengers with 1 siblings more likely to survive compare to the other passenger that traveling alone if we look at the ratio. If we just examine the number, passenger traveling alone has higher number of survive, but also higher people that not survive.
- **Parch**: Passengers with parents or children also more likely to survive compare to passenger who travel alone.
- **Fare**: Passengers who pay more have a higher chance to survive. This statements relevant with the Pclass feature, as upper class people has more chance to survive.
- **Embarked:** Surprisingly, passengers who embark from 0 have higher survival chance. When I look on the explanation on kaggle.com, it seems 0 is Cherbourg.

4.2. Evaluation

B. Random classifier : 0.485

C. Training Error for decision tree classifier : 0.014

D. Training error using k-nearest neighbor with k=3 is 0.167

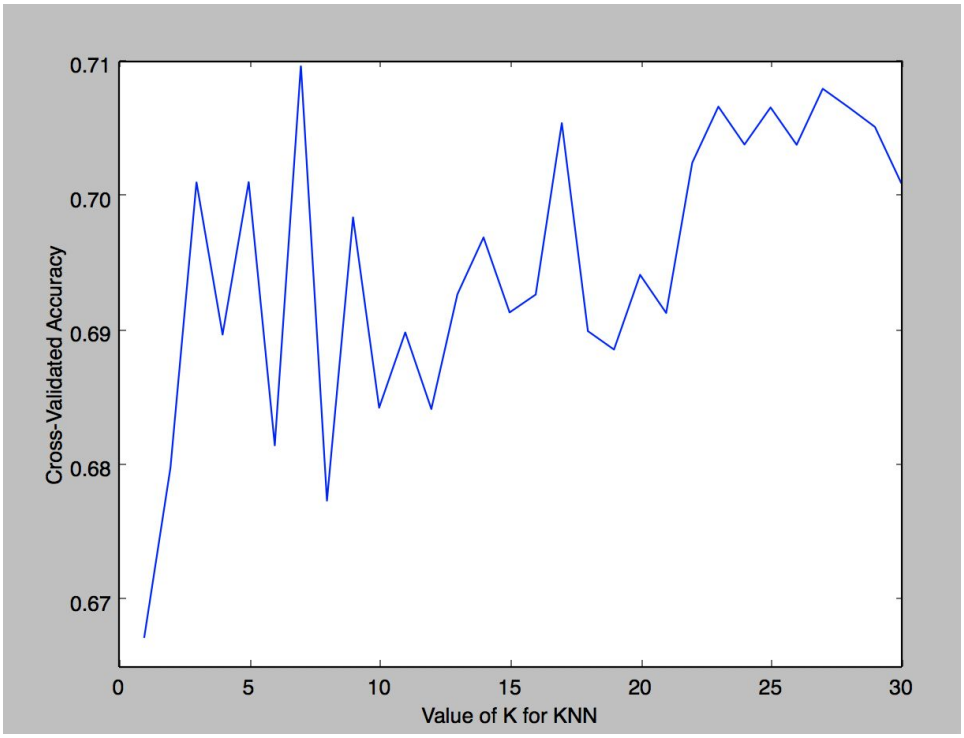
k=5 is 0.201

k=7 is 0.240

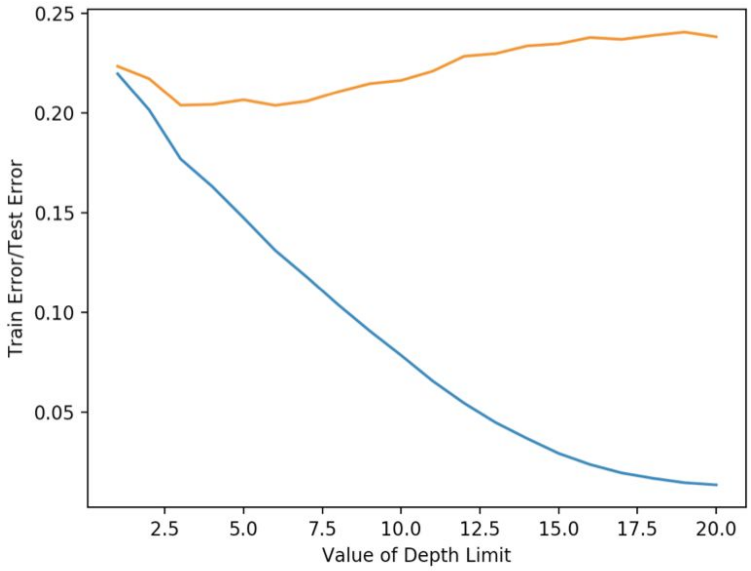
E.

Classifier	Avg Training Error	Avg Test Error
Majority	0.3972	0.4336
Random	0.5167	0.4825
Decision tree	0.0123	0.2413
Knn (k=5)	0.2320	0.3357

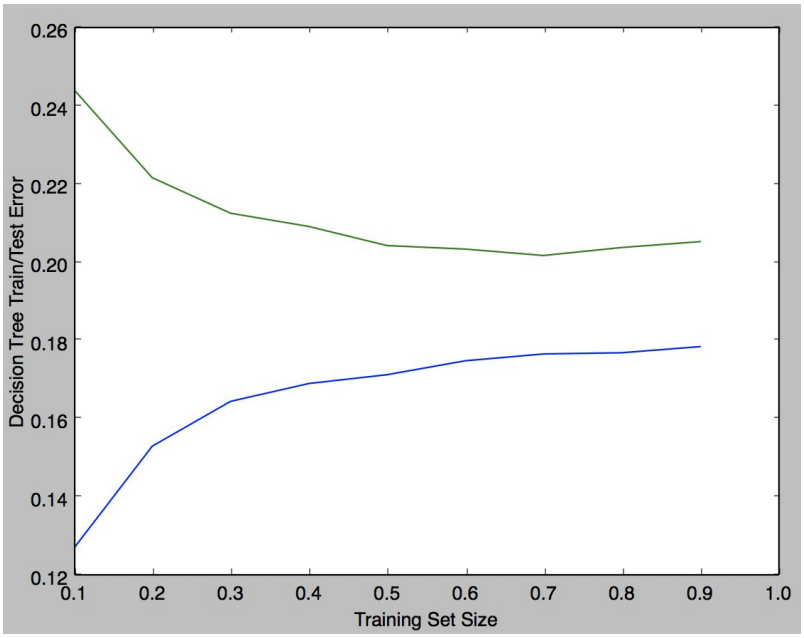
F. Value of K is around 7, with the accuracy of 0.7, then it has about 30% cross validation error



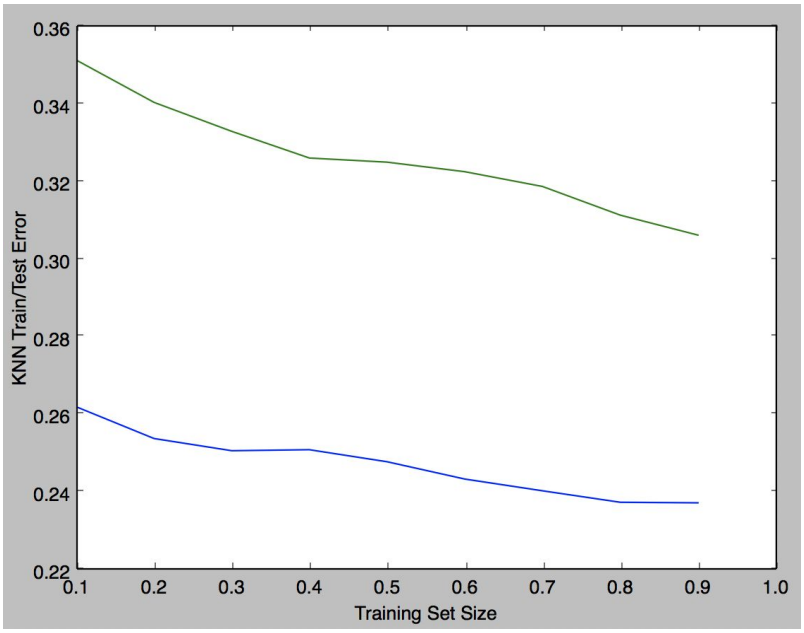
G. From the picture below, yellow line is testing error, and blue line is training error. From the graph that was shown, we can see that the best depth is 3, and has around 0.2043 error or 20%ish error. From the graph also I observer an overfitting at depth 6, as error percentage is increasing after the depth 6.



H.



For the decision tree, it indicates that it has a high variance estimator. Training Error (blue) increases as the increasing of number training data. On the other hand, test error (green) decreases as the increasing of number training data.



On the other hand, for KNN, it shows a decreasing for both training error and testing error.