CS M146 Fall 2018 Homework 1

Jun Kai Ong

October 28, 2018

2 Entropy and Information

$$Let\ H_B(S) = entropy\ before\ the\ split$$

$$= B\left(\frac{p}{p+n}\right)$$

$$= B(S)$$

$$= -S\log_2 S - (1-S)\log_2(1-S)$$

$$Let\ H_B(S_i) = entropy\ of\ a\ single\ subset\ after\ the\ split$$

$$= B\left(\frac{p_k}{p_k+n_k}\right)$$

$$\sum_{i=1}^{k} H_B(S_i) = \sum_{i=1}^{k} \frac{p_k+n_k}{p+n} B\left(\frac{p_k}{p_k+n_k}\right)$$

$$= \frac{1}{p+n}(\sum_{i=1}^{k} -p_k \log_2 \frac{p_k}{p_k+n_k} - \sum_{i=1}^{k} n_k \log_2(1-\frac{p_k}{p_k+n_k}))$$

$$= -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2(1-\frac{p}{p+n})$$

$$= -\frac{p}{p+n}\log_2\frac{p}{p+n} - (1-\frac{p}{p+n})\log_2(1-\frac{p}{p+n})$$

$$= H_B(S)$$

$$Information\ Gain, G = \sum_{i=1}^{k} H_B(S_i) - H_B(S)$$

$$Since \sum_{i=1}^{k} H_B(S_i) = H_B(S)$$

$$\therefore G = 0$$

3 k-Nearest Neighbors and Cross-validation

(a) If k = 1 and assuming every point can be its own neighbor, the resulting training error will be 0.

(b) If large values of k are used, the distinction between the classes will be unclear as the labels of the points will now be determined through global majorities instead of local majorities. If small values of k are used, our predicted labels might be easily affected by noise and outliers. Thereby, causing our model to be inaccurate and overfitted.

(c) k = 5 or k = 7 would minimize the cross validation error. The resulting error for both values of k would be 4/14.

4 Applying decision trees and k-nearest neighbors

(a) **Ticket Class**: Upper class passengers have higher chances of survival compared to lower class passengers

Gender: Women have a higher rate of survival compared to men

Age: Younger and older passengers are more likely to survive

# of Sibling/Spouse Passengers who have one or two siblings or spouses have a higher chance of survival

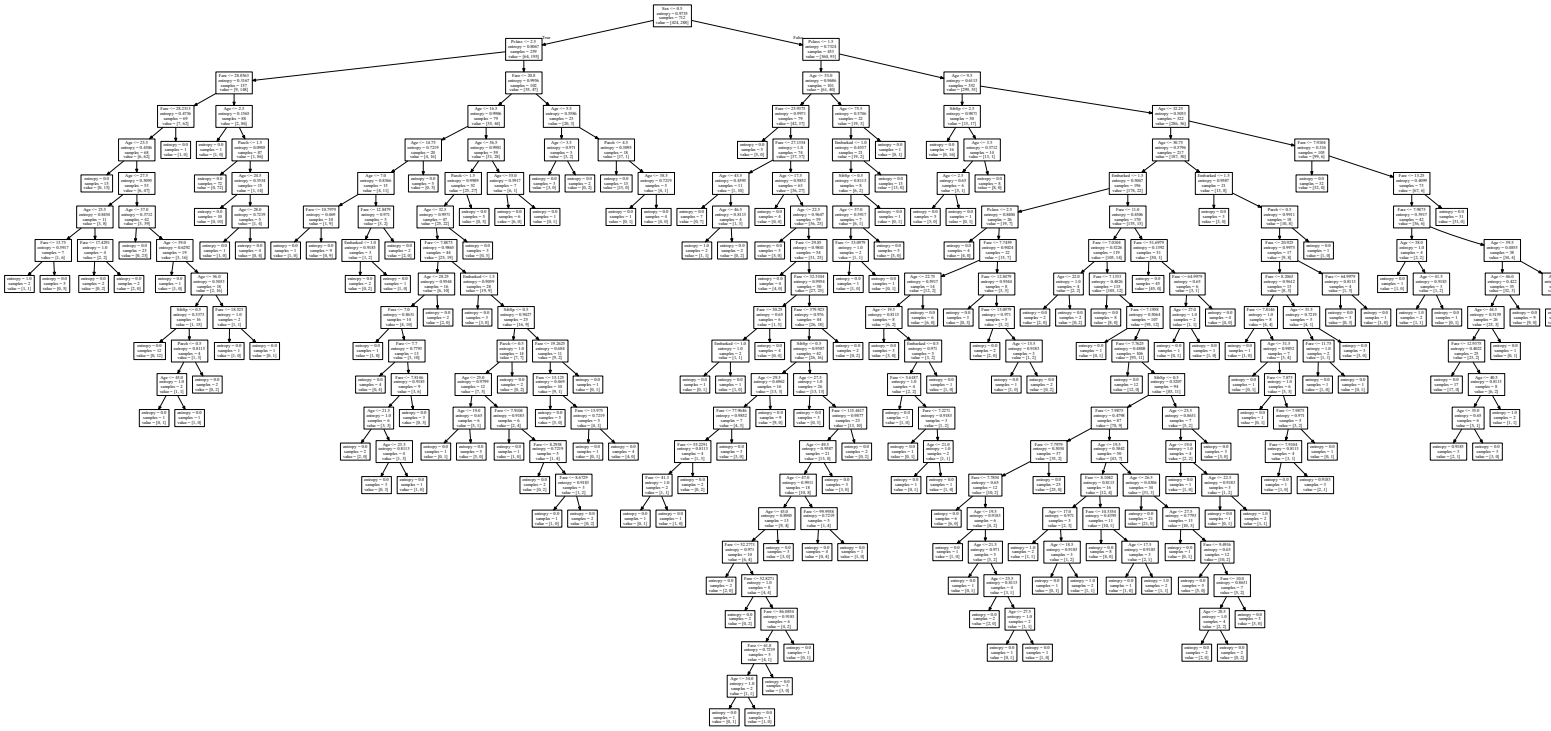
# of parents/children aboard the Titanic:Passengers with one or two parents or children abroad the titanic have a higher chance of survival

Fare: Passengers who pay higher fares have higher chances of survival. The survival chances of those who bought the cheapest fare is almost half of all of the other fares

Port of Embarkation: (0 = Cherbourg, 1 = Queenstown, 2 = Southampton) Passengers who embarked from Cherbourg are two times more likely to survivce compared to passengers who embraked from Queenstown and Southampton.

(b)
```
-- probability for 0 is 0.596
-- probability for 1 is 0.404
-- training error: 0.485
```

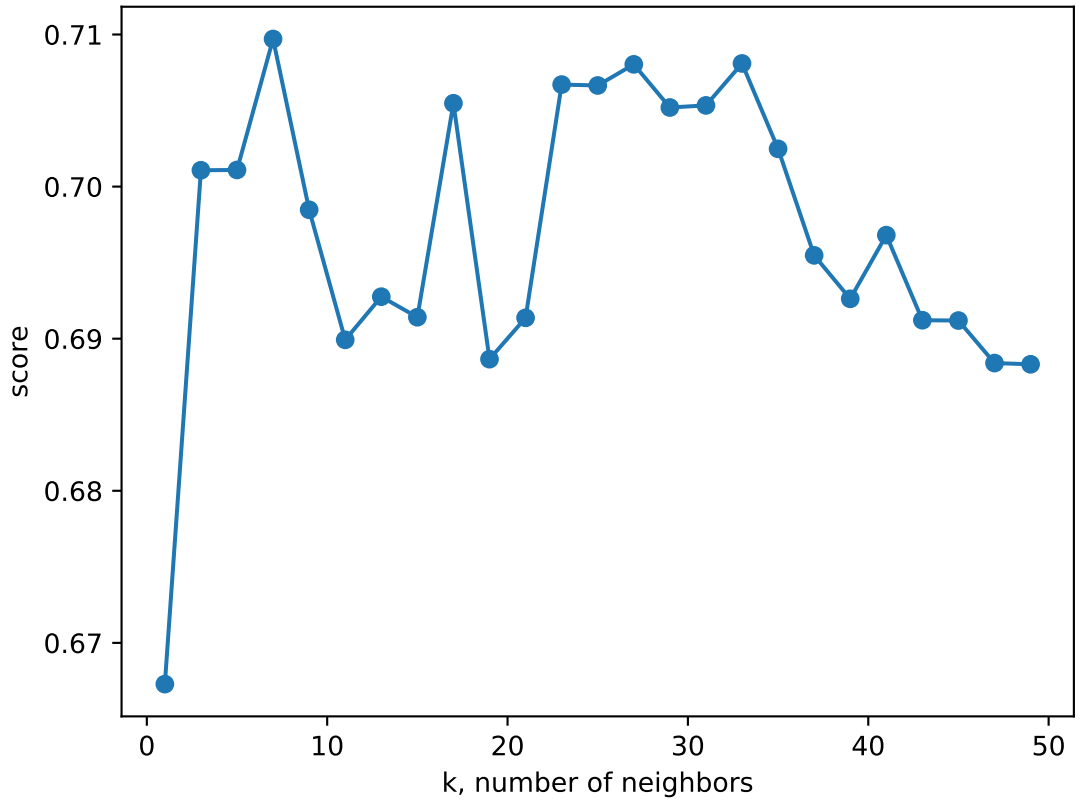
(c) `-- Training error for Decision Tree: 0.014`



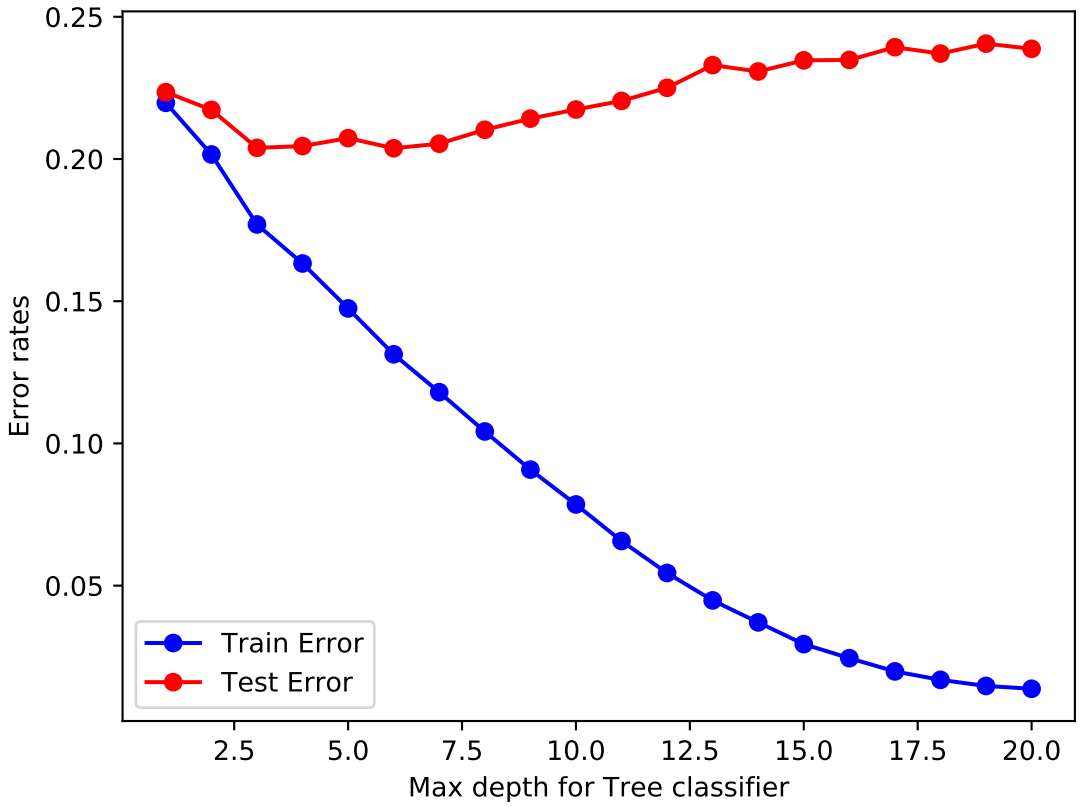
(d)
```
Classifying using k-Nearest Neighbors...
-- Training error for 3-neighbors: 0.167
-- Training error for 5-neighbors: 0.201
-- Training error for 7-neighbors: 0.240
```

(e)
```
-- MajorityVoteClassifier has training error:  0.404 and test error:  0.407
-- RandomClassifer has training error:  0.489 and test error:  0.487
-- Decision Tree has training error:  0.0115 and test error:  0.241
-- KNN-5 has training error:  0.212 and test error:  0.315
```

(f) From the graph below, we conclude that the value of k that gives us the highest score is 7.



(g) From the graph, the best maximum depth is 3 and potentially 6. This is because the test error at 3 is the lowest. Overfitting happens for depth limits higher than 3. This can be identified from the increasing trend of the test errors and the decreasing trend of the training error.



(h) From the graph below, it appears that in general more data leads to better trained models as shown by the general decreasing trend of all four plots. However, there appears to be fluctuations and variance to the data points potentially due to inherent biases in the training data.