M146 HW1

Zheyi Wang 705147852

October 28 2018

1 Splitting Heuristic for Decision Trees

check CCLE

2 Entropy and Information

(a) $H(S_i) = B(\frac{P_i}{P_i+n_i}) = H(S)$ if $i \leq k$.
Before splitting by X_j , $H(S) = B(\frac{P}{P+n})$.
After splitting by X_j ,
$H(S') = H(S_1)P(S_1) + H(S_2)P(S_2) + ... + H(S_k)P(S_k)$
$= H(S)P(S_1) + H(S)P(S_2) + ... + H(S)P(S_k) = H(S)(P(S_1) + P(S_2) + ... + P(S_k))$
$= H(S)$
Therefore, the information gain of this attribute is 0.

3 k-Nearest Neighbors and Cross-validation

(a) Since, all training data points are the closest neighbor of itself, when $k = 1$, all the training data points has the same class value as its closest neighbor (itself), which has the minimum training error zero.

(b) For example, if we let k equal to the total number of training data points, we will always get same prediction, which is equal to the majority value of the training dataset. Obviously, this way of predicting result is unreliable. Therefore, it is not good to use a too large values k. Too small value of k will overfit the training data, which is also not good in many situations.

(c) When $k = 1$, error rate= $\frac{10}{14}$ When $k = 3$, error rate= $\frac{6}{14}$ When $k = 5$, error rate= $\frac{4}{14}$ When $k = 7$, error rate= $\frac{4}{14}$ When $k = 9$, error rate= $\frac{14}{14}$ When $k = 11$, error rate= $\frac{14}{14}$ Therefore, $k = 5$ or $k = 7$ has the minimizes leave-one-out cross-validation error= $\frac{2}{7}$ for this dataset.

4 Programming exercise

(a) plot 1: People from upper class have higher survival rate than lower class. plot 2: Women have higher survival rate than men. plot 3: Children, teenagers and seniors have higher survival rate than others. plot 4: People with 1 or 2 with sibiling or spouse have higher survival rate than others. plot 5: People with 1 or 2 parent or child have higher survival rate than others. plot 6: People who paid higher fare have higher survival rate. plot 7: People embarked in C have higher survival rate.

(b) output of execution:

```
Classifying using Majority Vote...
        -- training error: 0.404
Classifying using Random...
        -- training error: 0.485
```

(c) output of execution:

```
Classifying using Decision Tree...
        -- training error: 0.014
```

(d) output of execution:

```
Classifying using k-Nearest Neighbors...
        -- k = 3 training error: 0.167
        -- k = 5 training error: 0.201
        -- k = 7 training error: 0.240
```

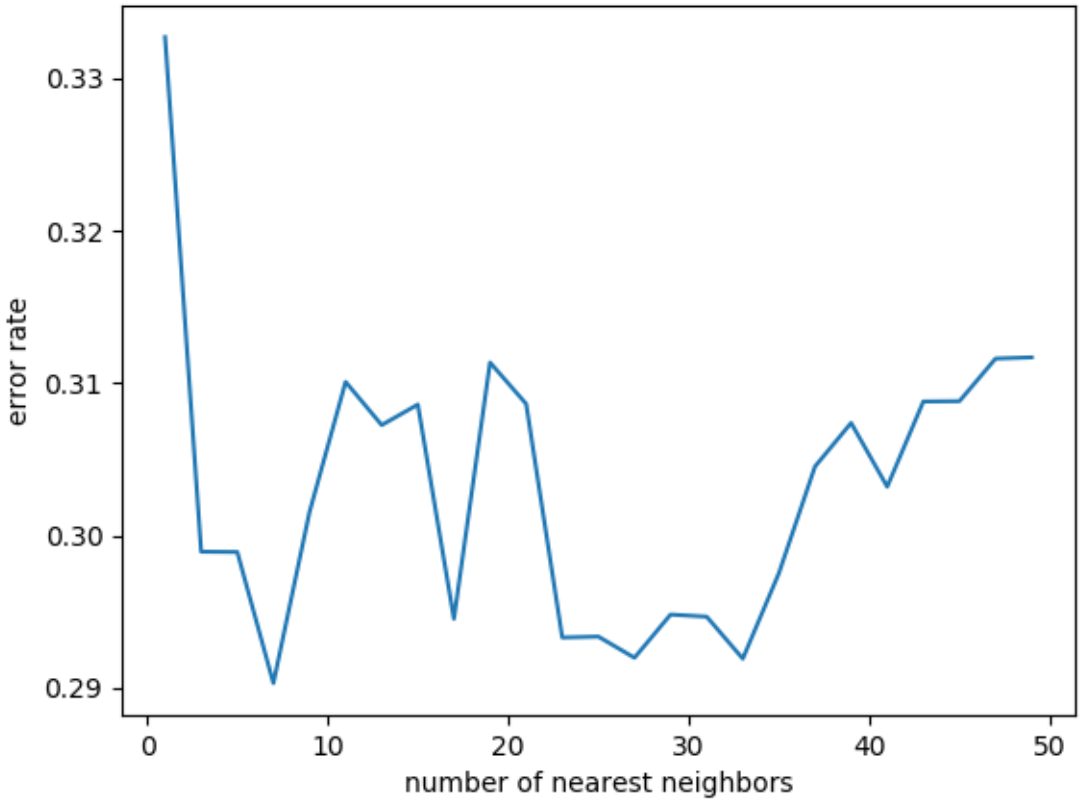
(e) output of execution:

```
Investigating various classifiers...
        -- Average result of major classifier training error = 0.404, testing error = 0.407
        -- Average result of random classifier training error = 0.489, testing error = 0.487
        -- Average result of decision dree classifier training error = 0.012, testing error = 0.241
        -- Average result of K=5 nearest neightbors training error = 0.212, testing error = 0.315
```

(f) output of execution:

```
Finding the best k for KNeighbors classifier...
        -- best number of neighbors: 7.000
```

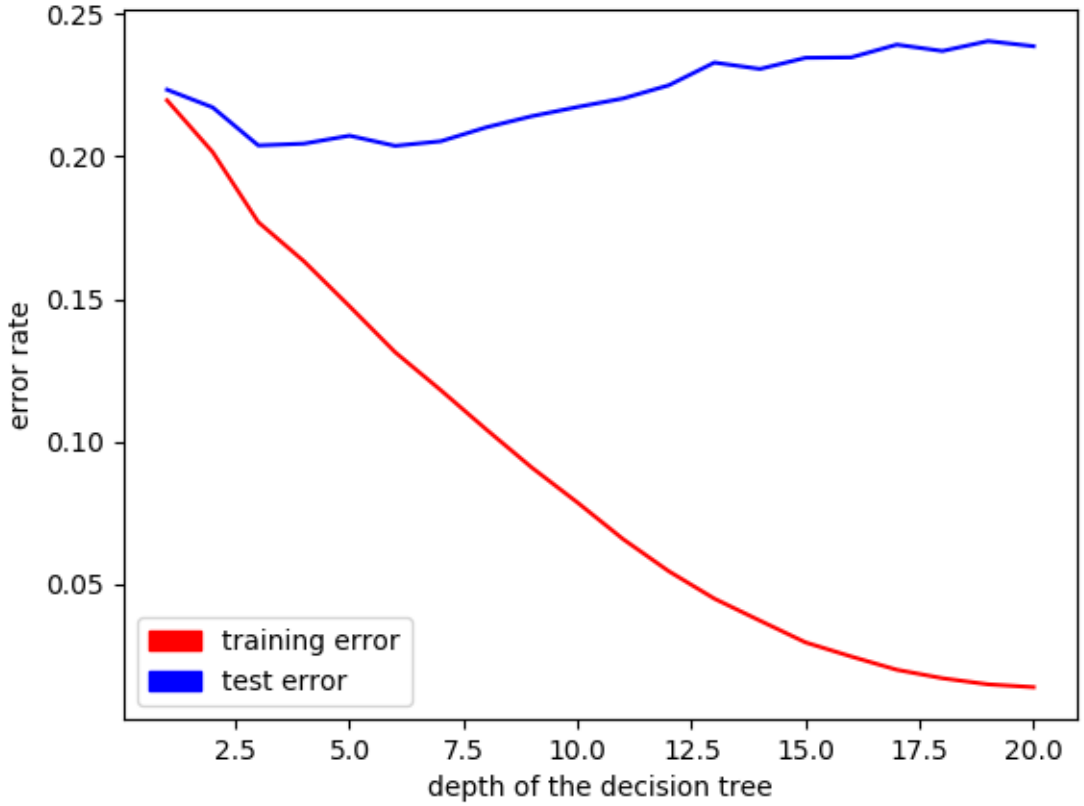
error rate vs k graph



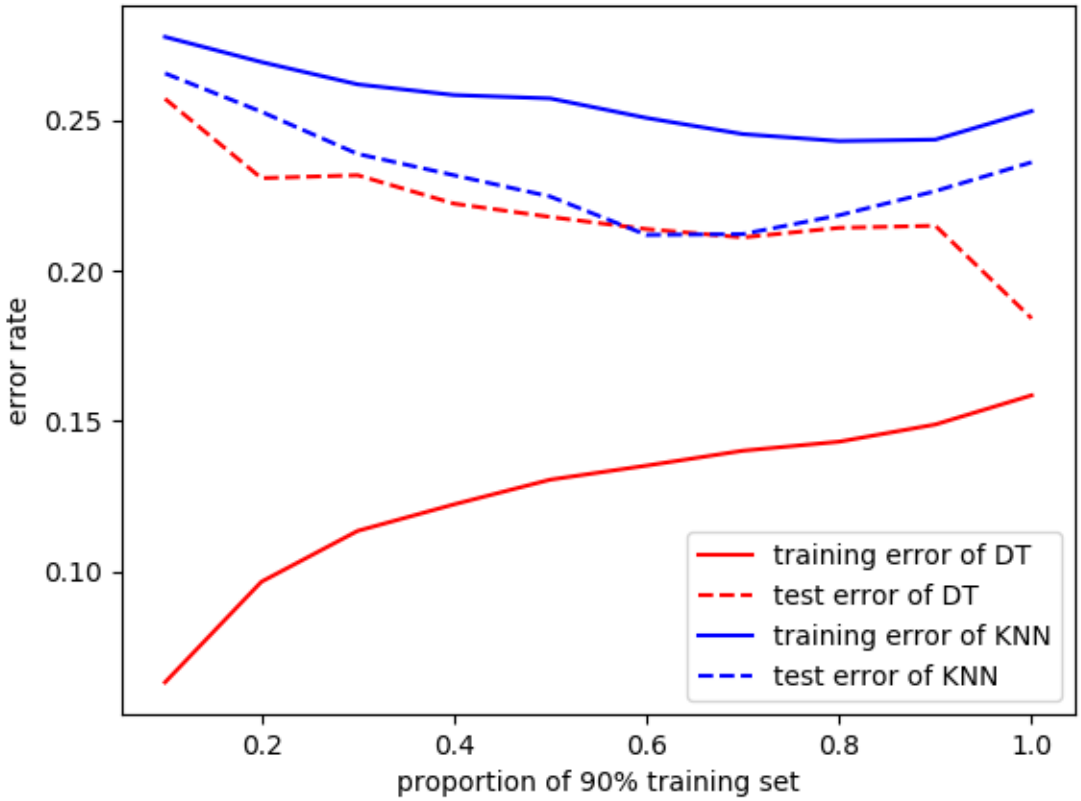
(g) output of execution:

```
Investigating depths...
        -- best depth limit: 5.000
```

error rate vs depth graph



(h) error rate vs % of 90% training set



In terms of the given dataset, decision tree performs better than KNN algorithm.