CSM146 HW1

Terry Ye

October 26, 2018

1 Question 1

Question answered in CCLE

2 Question 2

If the ratio is $\frac{p_k}{p_k+n_k}$ for all k subsets, then the ratio of every subset is the same as ratio of total example set S which is $\frac{p}{p+n}$. The entropy for the total sample is $H(S) = B(\frac{p}{p+n})$ And the entropy of every sample set is $H(S_k) = B(\frac{p}{p+n})$. The information gain is $Gain(S) = H(S) - \sum \frac{S_k}{S} * H(S_k)$. $H(S_k)$ is the same as H(S) and $\sum \frac{S_k}{S} = 1$ because the total number of examples are divided into subsets and the total number is the same. So $Gain(S) = H(S) - H(S) = 0$.

3 Question 3

(a) When k = 1, the nearest point in the training sample is itself. So the training set error is always 0 because we only calculate difference between the point and itself.

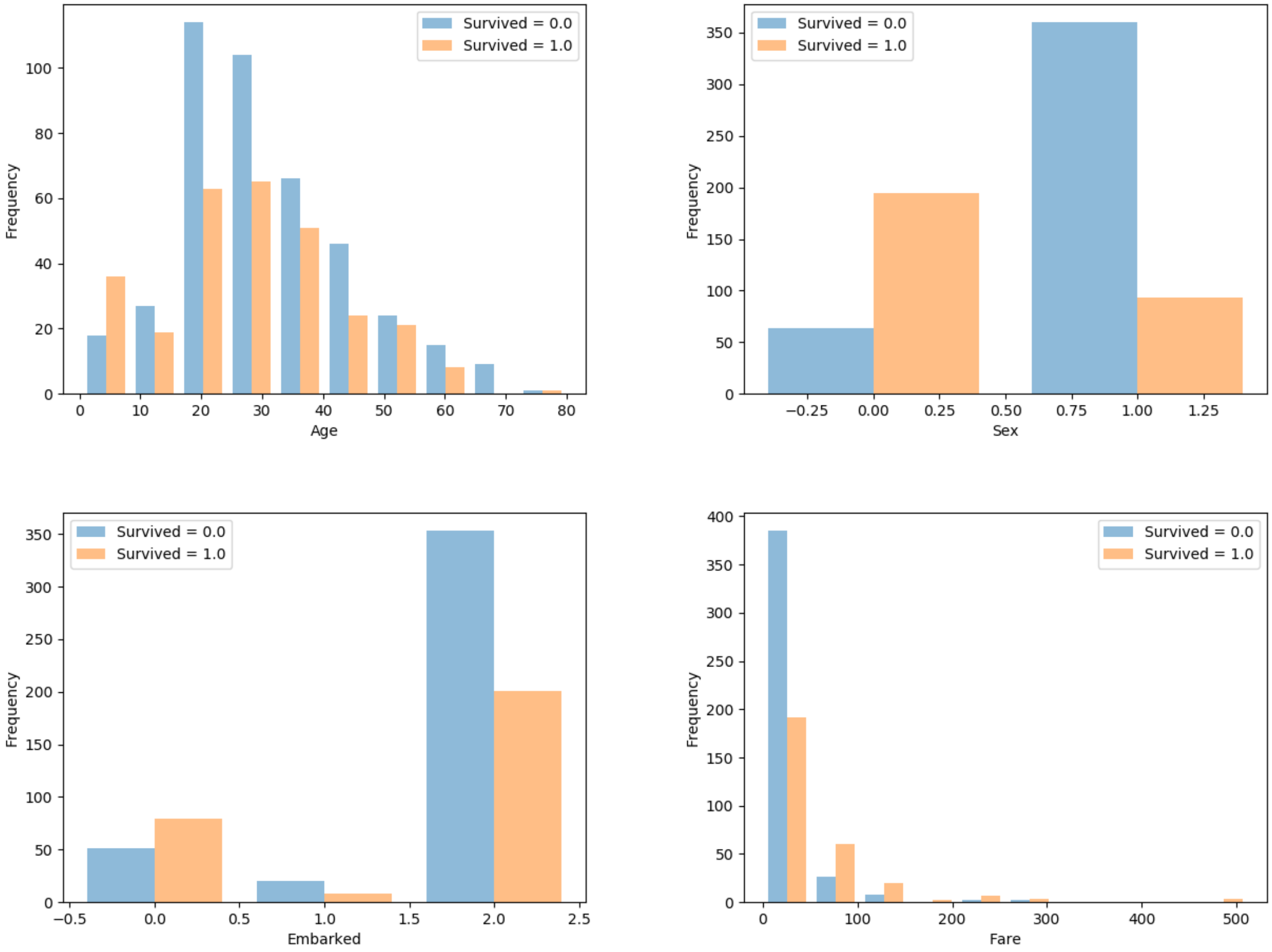
(b) When K is too big in the data set, the model will always just tend to predict the majority labels in the data set and lose the purpose of training . If K is too small, the result is prone to noise points or outliers that may outvote the correct data points in some region.

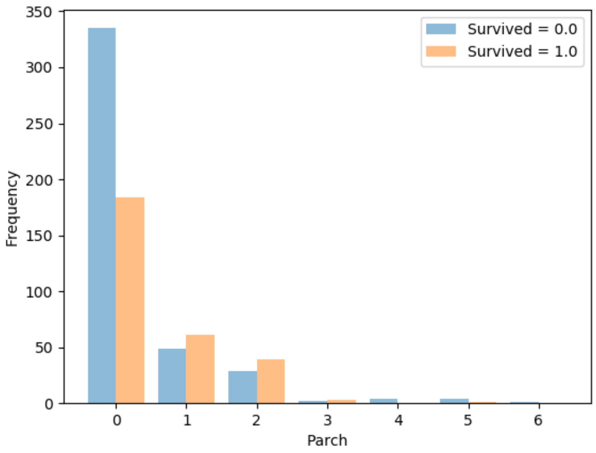
(c) When K = 5 or K = 7, the classifier gives wrong results for 2+ and

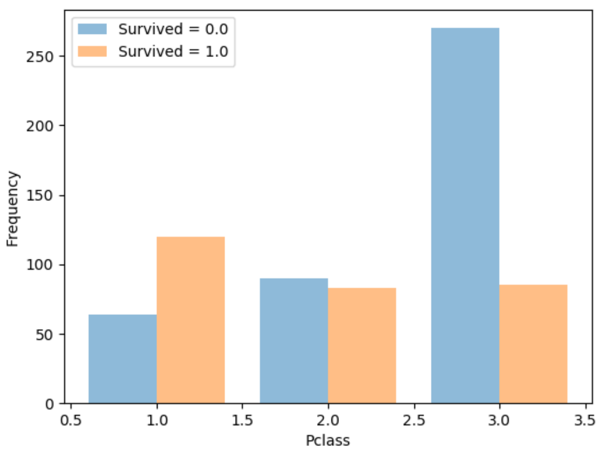
2- on the outside while gives correct results for the rest. The error rate is $4/14 \approx 0.2857$ which is the lowest among all k values.

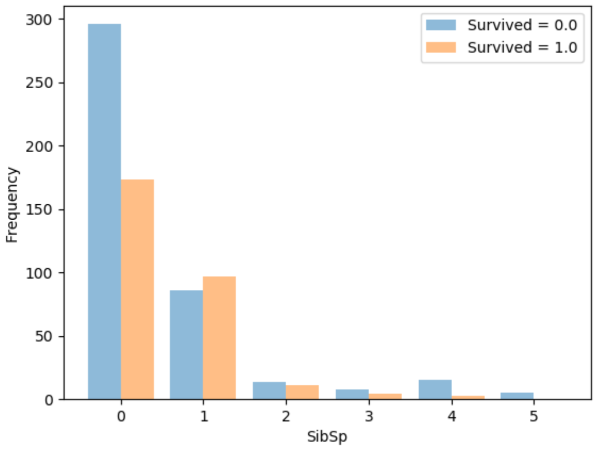
4 Question 4

4.1 Visualization









Age: kids below 10 years old have a higher survival rate than the others.

Sex: women have a higher survival rate than men.

Embark: people embarked at port C have a higher survival rate than the other two.

Fare: people paying less than 50 has lowest survival rate while the other groups have similar.

Parch: people with 0 parent/children have a lower survival rate than the others.

Pclass: people in the upper class have a higher survival rate than people in the other two classes.

SibSp: people with 1 sibling have a higher survival rate than the others.

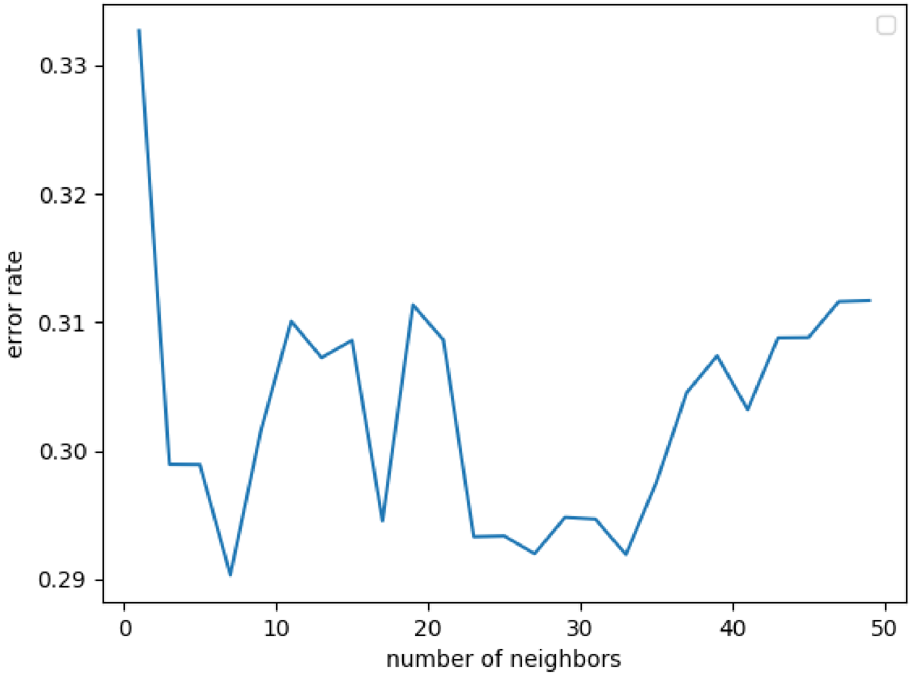
4.2 Evaluation

(b) I did it.

(c) The training error using decision tree is 0.014.

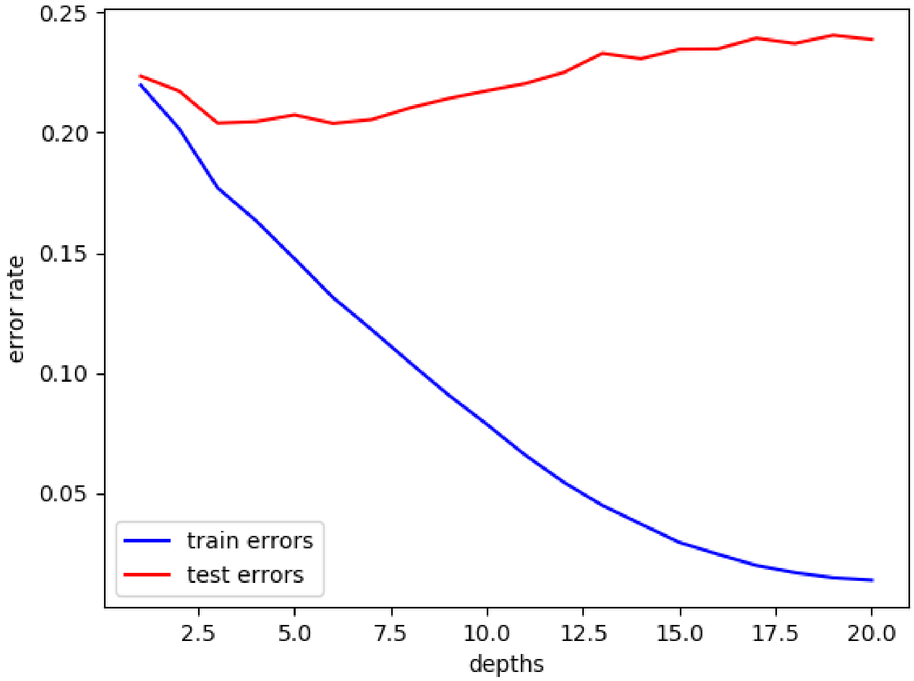
(d) The training error using KNN is
0.167 when k = 3
0.201 when k = 5
0.240 when k = 7

(e) Majority Classifier: training error = 0.404 test error = 0.407
Random Classifier: training error = 0.489 test error = 0.487
Decision tree classifier: training error = 0.012 test error = 0.241
KNN classifier: training error = 0.212 test error = 0.315



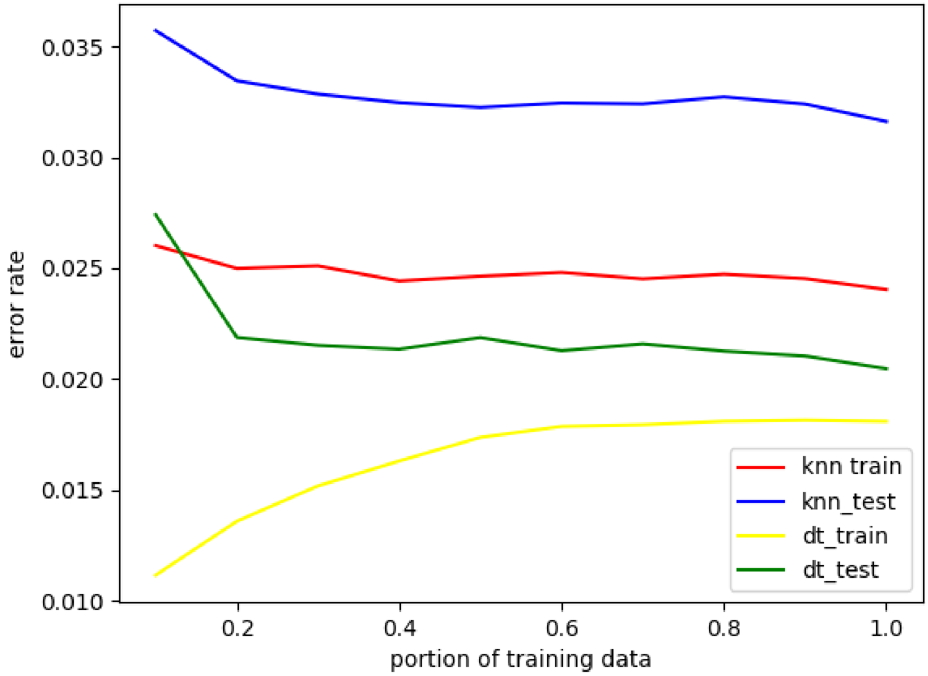
(f)

There is no specific feature of the graph, but we can see that really small and big k tends to have a higher error rates. The best error rate is achieved when K=7 and that is 0.290



(g)

The best depth limit is 3 for this data set. Yes there is overfitting in the pattern. As the depth limits grows, the error rate for training data set gets lower and lower to close to 0, but the error rate for test data set gets higher on the other hand.



(h)

This is the graph that run 10 errors for each split ratio to mitigate the bias of just 1 single run for each split ratio. Generally, the error rate is decreasing if the portion of training data is increasing because the more data it learns, the better its model is to predict the label. DecisionTree train is the only one that has an increasing error rate when training data size increases.