CS146: Machine Learning

Homework 1

Chaojie Feng

Oct.18, 2018

# Problem 1

Please check CCLE for answers

# Problem 2: Entropy and Information

## Problem 2(a)

Since is constant for all k,

Where:

As a result, Gain = H[s] – H[s] = 0

# K-Nearest Neighbor and Cross-Validation

## Problem 3(a)

To minimize training set error, K is selected to be 5. The resulting training error is 4

## Problem 3(b)

If K is too large, test data will absorb not only all error data near to itself, but also correctly labeled data near to it. In voting process, the test data will not be voted correctly for its label.

## Problem 3(c)

# Programming Exercise

## Problem 4(a)

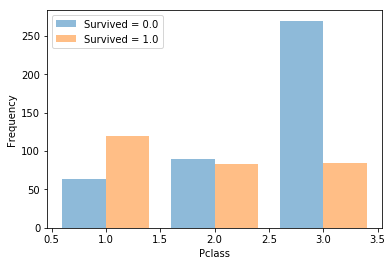


Figure 1. Ticket Class

The first feature is ticket class. 1 is for 1st class, 2 is for 2nd class and 3 is for 3rd class. It can be concluded that survival rate is decreasing as ticket class downgrades, and survival rate for 3rd class is significantly lower than 1st and 2nd class. Also, among all the passengers not survived, 3rd class passengers contributed most. However, survived passengers are distributed almost evenly among classes.

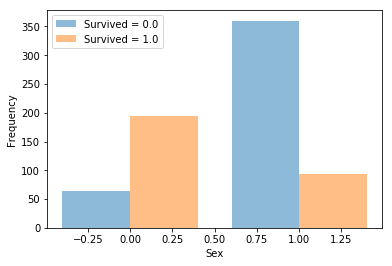


Figure 2. Sex

The second feature is sex. Female is much more likely to survive as male. Also, we can tell that there are more male passengers boarding than female passengers.

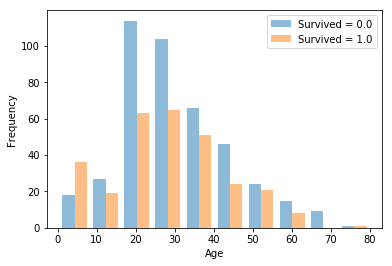


Figure 3. Age

The third feature is age. It is not surprising that young children are more likely to survive because they will have priority. For all other ages, passengers are less likely to survive.

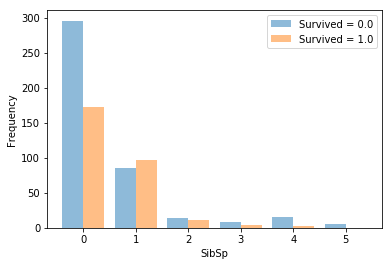


Figure 4. # of siblings and spouses

The fourth feature is # of siblings and spouses aboard. Obviously, as more siblings and spouses come with you, you might get help from them, which makes you more likely to survive.

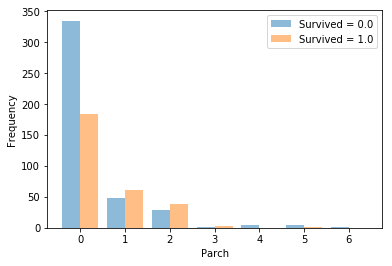


Figure 5. # of parents and children

The fifth feature is # of parents and children passenger brings aboard. Trend is pretty similar as previous case, where a single passenger is less likely to survive.

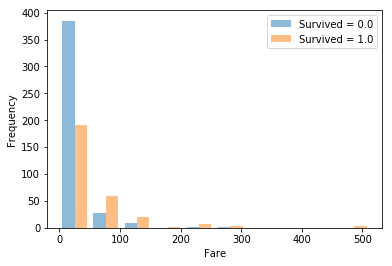


Figure 6. Fare

The sixth feature is fare. It makes sense that as passenger fare increases, the chance of surviving gets greater.

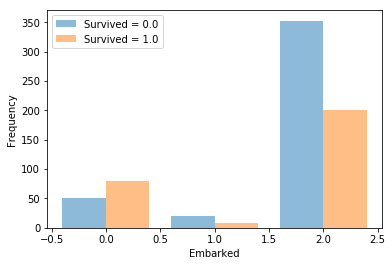
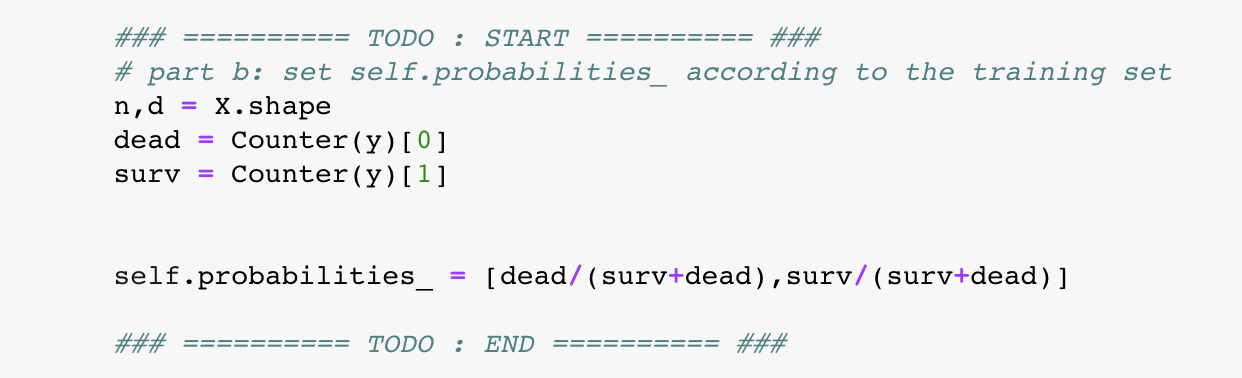


Figure 7. Embarked

The seventh feature is port of embarkation. For those who embark at Cherbourg, they are more likely to survive than those who embarked at other two location.

## The output of error is 0.485 as suggested in problem description



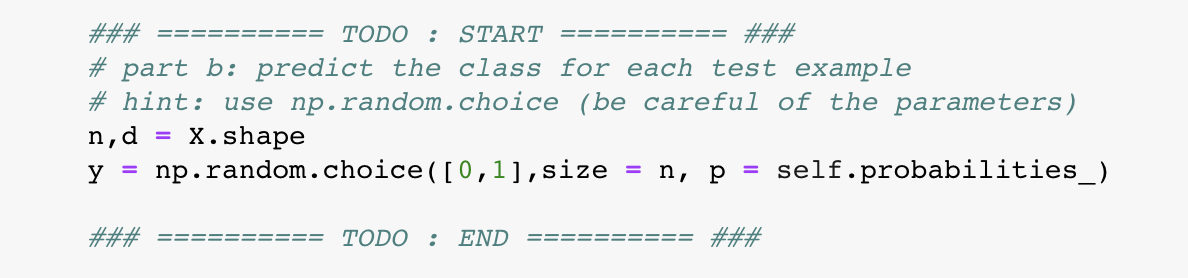


Figure 8. Random Classifier

## The training error is 0.014 for decision tree classifier using entropy criterion. Please see attachment “ dtree.pdf ” for more figures and more detail.

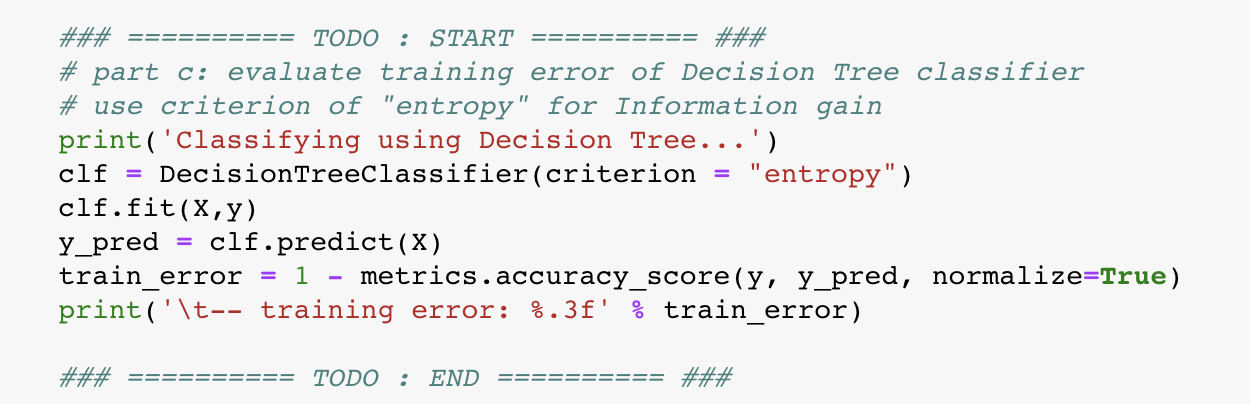


Figure 9. Decision Tree Classifier

1. When k = 3, the training error is 0.167. When k = 5, the training error is 0.201. When k = 7, the training error is 0.240. The section of code is provided below.



Figure 10. KNN Classifier

## Average training and test error of each classifier

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Majority Vote | Random | Decision Tree | KNN |
| Train error | 0.397 | 0.517 | 0.012 | 0.232 |
| Test error | 0.434 | 0.510 | 0.000 | 0.245 |

## Average error is 0.303. It can be seen that the error is very high when K is 1, but after that, K is becoming larger again.

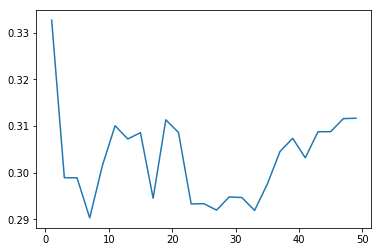


Figure 11. Average error vs. K