CM146, Homework#1

Zhaoxing Deng, 005024802

January 30, 2018

1 Splitting Heuristic for Decision Trees

- (a) X is labeled 0 iff $X_1, X_2, X_3$are all 0. |Y=0| = 2^{n-3},|Y=1| = $2^n - 2^{n-3}$. Since |Y=1|>|Y=0|, all vectors will be labeled by 1.This 1-leaf decision will make 2^{n-3}mistakes.
- (b) No. The error rate will remain $2^{n-3}/2^n = \frac{1}{8}$
- (c) $E[Y] = (1/8)log(8) + (1/7)log(8/7) = 0.543$
- (d) Yes. Spliting with any X_1, X_2, X_3, $E[Y] = (1/2)*0+1/2[(1/4)log(4)+ 3/4log(4/3)] = 0.406$

2 Entropy and Information

- (a) Since $p = \sum p_k$,$n = \sum n_k$, $\frac{p_k}{p_k+n_k}$is the same for all k, we can find out that $\frac{p_k}{p_k+n_k} = \frac{p}{n} = c$. The weighted average of $S_k$is $\sum \frac{p_k+n_k}{p+n} B(\frac{p_k}{p_k+n_k}) = \frac{p+n}{p+n} B(\frac{p}{p+n}) = B(\frac{p}{p+n})$. $Gain = B(\frac{p}{p+n}) - B(\frac{p}{p+n}) = 0$

3 k-Nearest Neighbors and Cross-validation

- (a) Since a point can be its own neighbor, when k=1,the training error is zero. Thus, k=1 minimizes the trainning error.
- (b) Using too large k will make some data useless for the prediction. Using too small k will probably cause overfitting, when the predition will be affected by the noise.

4 Programming exercise : Applying decision trees and k-nearest

4.1 Visualization

- (a)

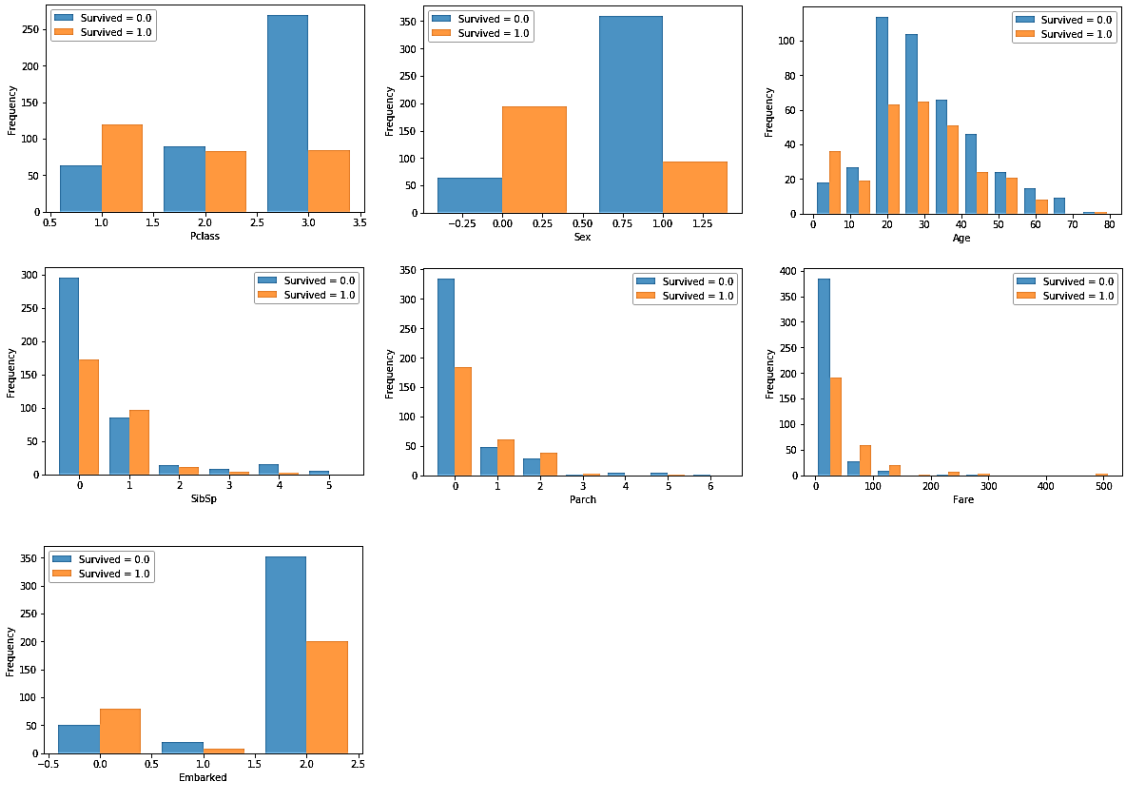

Figure 1: Histograms for each feature

  - Pclass: The first class had the highest survival rate. The third class had the lowest survival rate.
  - Sex: Females had higher survival rate than males.
  - Age: Children (under 10) had the highest survival rate, People between 20 and 30 had the lowest survival rate.
  - SibSp: Passengers with 1 sibling had the highest survival rate.
  - Parch: Passengers with parents and children had much higher survival rate than others.
  - Fare: Passengers who paid more had higher survival rate.
  - Embarked: Passngers who embarked had higher survival rate.
  -

- (b) Classifying using Random...
  - -- training error: 0.485
- (c) Classifying using Decision Tree...

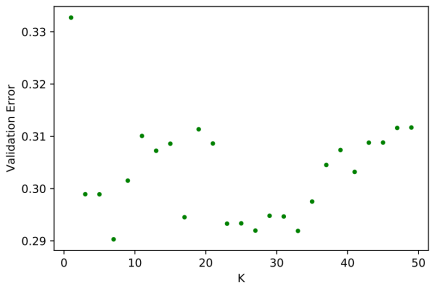

Figure 2: The validation error against the number of neighbors, k

  - -- training error: 0.014

- (d) Classifying using k-Nearest Neighbors...
  - -- 3-NN training error: 0.167
  - -- 5-NN training error: 0.201
  - -- 7-NN training error: 0.240
- (e) Investigating various classifiers...
  - -- MajorityVote training error: 0.404 testing error: 0.407
  - -- Random training error: 0.489 testing error: 0.487
  - -- DecisionTree training error: 0.012 testing error: 0.241
  - -- 5-NN training error: 0.212 testing error: 0.315
- (f) When k is very small, overfitting happens. Then as k goes larger, the error decreases. When K = 7, the error is the smallest. Then the error gets larger. The best value of K is 7.
-
- (g) The best depth limit is 6, since it has the lowest testing error: 0.204. There is overfitting after depth limit 6, because the testing error goes up while trainning error goes down.
-
- (h) As the trainning percentage goes up, both K-NN trainning error and test error goes down, while the decision tree trainning error goes up and test error goes down. There are gaps between Trainning errors and Test errors, so more data will help.
-

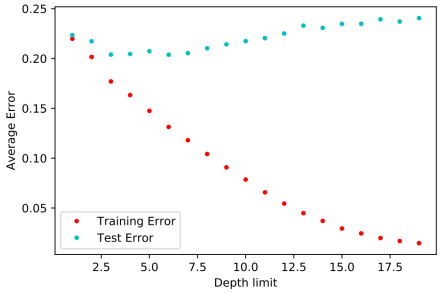

Figure 3: The average training error and test error against the depth limit.