CSC570E Machine Learning
Homework 2

Applying k-Nearest Neighbors and Naïve Bayes classifiers to predict crime rate

(due by the end of the day on Wednesday, July 10th)

Using Boston data set, fit k-NN and naïve Bayes classification models in order to predict whether a given suburb has a crime rate above or below the median.

1. Load Boston.cvs data set. It records 14 variables for 506 neighborhoods around Boston:
   crim: per capita crime rate by town.
   zn: proportion of residential land zoned for lots over 25,000 sq.ft.
   indus: proportion of non-retail business acres per town.
   chas: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
   nox: nitrogen oxides concentration (parts per 10 million).
   rm: average number of rooms per dwelling.
   age: proportion of owner-occupied units built prior to 1940.
   dis: weighted mean of distances to five Boston employment centers.
   rad: index of accessibility to radial highways.
   tax: full-value property-tax rate per \$10,000.
   ptratio: pupil-teacher ratio by town.
   black: *1000(Bk - 0.63)^2* where *Bk* is the proportion of blacks by town.
   lstat: lower status of the population (percent).
   medv: median value of owner-occupied homes in \$1000s.
2. Which suburbs have low crime rates?
3. Create a binary variable, crim1, that contains 1 if crim contains a value above its median, and a 0 if crim contains a value below its median. You can compute the median using the median() function.
4. Explore the data in order to investigate the association between crim1 and the other features. Which of the other features seem most likely to be useful in predicting crim1? Decide which attributes you are going to use to predict crim1.
5. Set the seed of the random number generator to a fixed integer so that you can reproduce your work.
6. Normalize the attribute values
7. Randomize the order of the rows in the dataset
8. Split the data into a training set and a test set. Use a test set of 100 rows.
9. Perform kNN on the training data, with several values of *K*, in order to predict crim1. Which value of *K* seems to perform the best on this data set?
10. Run a Naïve Bayes classifier to predict crim1. Remember that the Naïve Bayes does not require that attributes be normalized and the rows be randomized.
11. Compare the kNN classifier with the Naïve Bayes in terms of error rates, false positives and false negatives.

Your submission must consist of two text files:
- a text file, description.txt, no longer than a page: Your answers to the questions above
- a script with history of your session. Save the session into script using: savehistory("script.Rhistory"). You can manually edit the script in order to remove any unnecessary commands, such as trials and errors. The order of the commands must follow the order stated above.