Assignment 2¶

Due Friday, October 13, 2017

Submission: Put the data and Jupyter notebook files in a folder. Make sure all links to data are relative to the folder so the TAs can run the notebooks.

Use the public dataset that you used for cleaning and EDA in Assignment 1. You *MUST* get approval if you wish to use a *different* dataset.

In this assingment you will classify your data using k-Nearest Neighbors, decision trees, naive Bayes and support vector machines and interpret the results.

Part A - k-Nearest Neighbors (20 points)¶

- Classify your data using k-Nearest Neighbors. Answer the following questions:
  - How well does the kNN classifier perform?
  - Does the k for kNN make a difference? Try for a range of values of k.
  - Does scaling, normalization or leaving the data unscaled make a difference for kNN?

Part B - Decision Trees (20 points)¶

- Generate a Decision Tree with your data. You canuse any method/package you wish. Answer the following questions:
  - How well does the decision tree classifier perform?
  - Does the size of the data set make a difference?
  - Do random forests make a difference?
  - Do the rules make sense? If so why did the algorithm generate good rules? If not, why not?
  - Does scaling, normalization or leaving the data unscaled make a difference?

Part C - Naive Bayes (20 points)¶

- Classify your data using Naive Bayes. Answer the following questions:
  - How well does the naive Bayes classifier perform?
  - Which form of Naive Bayes did you use (Bernoulli, Multinomial or Gaussian)? Why?
  - Does scaling, normalization or leaving the data unscaled make a

difference for Naive Bayes?

Part D - Support Vector Machines (20 points)¶

- Classify your data using Support Vector Machines. Answer the following questions:
  - How well does the SVM classifier perform?
  - Try different kernels. How do they effect its performce?
  - What might improve its performce?
  - Does scaling, normalization or leaving the data unscaled make a difference for SVMs?

Part E - Compare the methods (20 points)¶

Which of the methods (k-Nearest Neighbors, decision trees, naive Bayes and support vector machines) did the best? Why?