IS 621: Business Analytics and Data Mining

Assignment 3: Learning versus Design

Instructions:

- Soft deadline for this assignment (recommended due date): **Monday, March 2, 2015 at 11:59 p.m. EST**.
- Hard deadline for this assignment (penalties apply if late): **Monday, March 9, 2015 at 11:59 p.m. EST**.
- Late assignments will receive 50% of credit earned and can be submitted until 4:00 p.m. Thursday, March 12, 2015. Full sample solutions are posted in the course at that time, so assignments submitted after 4:00 p.m. will receive no credit.
- Solutions should be typed. Your submission should be electronic and through Blackboard. Multiple files, clearly named, are acceptable.
- Assignments that either cannot be opened correctly or are illegible will receive no credit. If you have any concerns about this, please ask me before the deadline.
- The quality of your solution presentation is as important as the correct answer. For full credit you must show your steps and give clear, thorough answers. In code, this means good commenting.

Background

In this assignment we will focus on two of the most basic classification techniques: Naïve Bayes and Nearest Neighbor Classification. You will first implement basic versions of each technique by hand. Then you will have the chance to deploy solutions using R and Azure Machine Learning. Finally, you will apply these techniques to two use cases.

Data

Data for the assignment are attached in Blackboard as CSV files. These include…

- A sample training data set and testing data set for use with the Naïve Bayes problem (1,3,4)
- A sample training data set and testing data set for use with the Nearest Neighbor problem (2,3,4)
- Training, public testing, and private testing data sets for the Juror Problem (5)
- Training, public testing, and private testing data sets for the Diabetes Problem (6)

Restrictions

No outside packages should be necessary for problems 1-3. For problems 4-6 you have instructions on which packages to use. If you know others, you can use them as well.

Tasks to Complete:

1. (Programming) Implement a Naïve Bayes algorithm. You can assume all variables are categorical for this exercise. (This is not a requirement for Naïve Bayes, but we don't want to get bogged down by density estimations and the like.)

   Your code should accept as input two data frames. The first should be the training data set, with the classification column specified in a manner of your choosing. The second data frame is the learning data that should have predictions assigned to each observation. You may make the additional assumption that the columns of the two data frames are identical, optionally including the correct answer for the learning data.

   The output should be a single data frame that contains the scored learning data (the original data frame plus the assigned classification values).

   For this task, you should submit

   - Your R code that implements the Naïve Bayes algorithm
   - Your comments on how well the code performs on the sample data set provided

2. (Programming) Implement a nearest-neighbor algorithm. You can assume all variables are continuous and you will use the standard Euclidean metric. (Neither of these are strictly necessary, but we'll keep it simple.)

   Your code should accept as input two data frames, same as in problem 1, as well as a constant that gives the number of neighbors to be used. The first data frame is the training data set, with the classification column specified in a manner of your choosing. The second data frame is the learning data that should have predictions assigned to each observation. You may make the additional assumption that the columns of the two data frames are identical, optionally including the correct answer for the learning data. You should try several different values of k.

   For this task, you should submit

   - Your R code that implements the Naïve Bayes algorithm
   - Your comments on how well the code performs on the sample data set provided

3. (Deployment) Your code for problems 1 and 2 can be used in a cloud-based machine learning environment that supports R. For this course, we are using Microsoft Azure Machine Learning (AzureML). Code a simple experiment in AzureML for each of the two problems. The basic steps are as follows:

   a. Upload the training and testing data sets to AzureML from your local machine
   b. Drag the two data sets onto your canvas
   c. Drag an Execute R Script module onto your canvas
   d. Put your code into the Execute R Script module and modify it appropriately to accept the datasets as inputs and output the scored learning data
   e. Drag a Convert to CSV module onto the canvas and connect the output data set to it

   This concludes the experiment! You should make sure you have invited me to your workspace (please use mfschulte@live.com as my e-mail address for the invitation). I will grade your experiment directly in AzureML.

4. (R Package Exploration) This problem lets you explore existing implementations of Naïve Bayes and Nearest Neighbor algorithms in R.

   First, explore the implementation of Naïve Bayes included in the package **e1071**. Use it to classify the sample learning data used in problem 1.

   Next, explore the package **class** and its implementation of the Nearest Neighbor algorithm.

   For this task, you should submit:

   - Your R code implementing the Naïve Bayes algorithm in **e1071**
   - Your comments on the effectiveness of the **e1071** approach and whether the results are identical to your own hand implementation results in problem 1
   - Your R code implementing the Nearest Neighbor algorithm in **class**
   - Your comments on the effectiveness of the **class** approach and whether the results are identical to your own hand implementation results in problem 2

5. (Application) Apply the Naïve Bayes approach to classify the observations in the data set on potential jurors (available in jury-learning-data-public.csv and jury-learning-data-private.csv) using the survey training data (available in jury-training-data.csv). Try both your own implementation of Naïve Bayes and the package implementation examined in problem 4. Comment on how well your analysis works by examining the confusion matrix.

   For this task, you should submit:

   - Your scored private dataset
   - Your analysis of your results

6. (Application) Apply the Nearest Neighbor approach to classify the observations in the data set on diabetes in Pima Indians (in pima-learning-data-public.csv and pima-learning-data-private.csv) using the training data (in pima-training-data.csv). Try both your own implementation of Nearest Neighbor and the package implementation examined in problem 4. Comment on how well your analysis works by examining the confusion matrix. Be sure to try several values of k.

   For this task, you should submit:

   - Your scored private dataset
   - Your analysis of your results