IS 467

Assignment 1

Submission Instructions

1. Save your solutions with **clearly marked questions and their numbers**, clear and succinct writing, all software **output** into a single PDF or Word file.
2. Submit your file online at the course website at **http://d2l.depaul.edu** and double-check it.
3. Keep a copy of all your submissions!
4. If you have questions about the homework, email me BEFORE the deadline.

Problem 1 (25% points):

Answer each of the following questions with a few sentences:

a. What is the difference between *classification* and *clustering* (5%)?
b. In a data table, what do the *columns* represent and what do the *rows* represent (5%)?
c. After loading new data into SPSS, describe two tasks you might do to clean your data (5%).
d. Explain which type of data mining algorithm (also called data mining functionality) would you use to answer each of these questions and why?
    i. What are five groups of customers who buy similar things (5%)?
    ii. I sell milk – can I predict if a user will buy that based on the other things they bought (5%)?

Problem 2 (25% points):

Explain in few words whether or not each of the following activities is a data mining task and why.

a. Dividing the customers of a company according to their gender (5%).
b. Computing the total sales of a company (5%).
c. Sorting a student database based on student identification numbers (5%).
d. Estimating the probability of the outcomes of tossing a (fair) pair of dice (5%).
e. Predicting the future stock price of a company using historical records (5%).

Problem 3 (50% points):

Fisher's iris data consists of measurements of iris flower plants, specifically the sepal length, sepal width, petal length, and petal width of 150 specimens. There are 50 specimens from each of three species: Setosa, Versicolour and Virginica.

First download the iris dataset from the UCI Machine Learning Repository:

http://archive.ics.uci.edu/ml/datasets/Iris. Data gathering is a process, which is why you're going to download it from this popular data site rather than have me give you a clean copy. In this case, that means clicking the link for the 'data folder' and downloading *iris.data* (use the right-click menu and you'll see 'Save As' or 'Save Target' or 'Save Link' or something similar depending on your machine).

The data file has raw CSV data (each row of data is a line of text, column values are separated by a comma). When you use SPSS to open a data file, by default it looks for its own format (.sav files). To find the file you downloaded, click the *Files of Type* dropdown and choose *All Files* from the bottom of the list. SPSS will guide you through importing the data with a series of windows. The defaults should work for this data but make note of the options. In particular, make sure to check the box indicating that the first line of the file does not include column labels because the data start on the first line of the file.

We want labels for the variables to help us understand what we're doing in the analysis, though, so we'll add them. The correct variable names are listed in the *iris.names* file from the same place you downloaded the data. In that file, look under "Attribute Information" (attribute is yet another name for column/feature/variable in this case). You can manually add these names to the data file by opening it in a text editor and adding a line with the variables at the beginning, but that won't work when the data is too large and we might as well do it in SPSS. The simplest way to do this is to use the *Variables* tab in the data window. You can just type the names in the boxes. Note that you can't use spaces (use underscore, "_", instead). Check the data types of the variables (e.g. nominal or scale values, etc.). In this case the automatic import has done fine, but it's important to check and make sure. You can save the fixed-up data as a new SPSS file (.sav) if you want to be able to work with it later without redoing the renaming.

NOW, use SPSS to answer the following questions (see the first lab for help):

a. Visualize the relationship between the two sepal variables, sepal length and sepal width, using a scatter plot. Use different colors or symbols per class so you can see how the classes are related to this pair of variables. We talked in class about how classifiers work, broadly-speaking. Do you think that a classification algorithm using these two variables will be successful in classifying data with respect to the class labels we have? Explain why or why not and include the plot image with your answer (10%).
b. Repeat part (a) for the petal variables (10%).
c. Create a histogram for each of the four variables. Histograms in SPSS are just a different graph type from scatterplots. Describe what you can tell about the distribution of each variable (10%).
d. Determine if there are any outliers in the data with respect to the sepal length (10%).
e. Repeat d. for the petal length (10%).