CMSC 828G Homework 1

Data Mining,CMSC 828G, Spring 2002

Due Date: Thursday February 21, at the start of class

Problem 1

An *outlier* is a data point that is sometimes defined as "an extreme observation that does not seem consistent with the rest of the data." Outliers can arise because of data entry errors or may be a real measurement for a particularly unusual object or person. The interpretation of outliers can be subjective and context-dependent. One person's outlier may be another person's Nobel prize!

In data mining it is always a useful exercise to check for outliers before embarking on any analysis of the data, since the presence of outliers can significantly skew the results and interpretation of any analysis.

Part A

Imagine you have been given a large medical data set for a data mining project. Discuss briefly how you might check for outliers (i.e., potentially multiple outliers) if the data consists of:

- a single real-valued variable,
- a set of 20 real-valued variables

What if some (or all) of the variables were categorical? How might this affect the problem of finding outliers?

Part B

Imagine we look at a single variable in the medical data set mentioned above, and we estimate the mean and the median, where the mean is the arithmetic average of the values and the median is defined as the value for which 50% of the values are higher and 50% are lower.

How sensitive are each of these methods to outliers? Feel free to use a sketch/plot and/or some simple equations to provide an example to justify your answer.

Part C

Another related idea is known as "robust statistics." For example, one type of robust estimate of the mean for a single variable is the "trimmed α mean" which is defined as the mean of the observations that like between the α and $100 - \alpha$ percentiles of the data, i.e., the $\alpha = 5\%$ trimmed mean is the arithmetic average of the 90% of values that lie above the lowest 5% of the values and below the top 95% of the values. Clearly this will help give an estimate of the mean that is less sensitive to outliers.

Propose how this idea might be extended to multiple variables (multiple dimensions). How would your proposed method handle the data shown in Figure 2.6 in the text??

Part D

Given your answers to the questions above, do you think that it is reasonable to expect that the problem of outlier detection can be completely automated? Is it possible to detect outliers without having a prior model or set of expectations for the data? Discuss your answer.

Problem 2

Part A

In class we discussed fitting a model of the form

$$Y = aX + b$$

to a data set D consisting of pairs of X and Y values, $D = \{(x(1), y(1)), \ldots, (x(n), y(n))\}$. Assume that our goal is to find the parameters a and b that minimize a score function

$$S(a, b) = \sum_{i=1}^{n} \Big(y(i) - (ax(i) + b) \Big)^2$$

the sum of squared errors between the observed "target" $y(i)$ and the predicted value $x(i)$. Use basic calculus to derive closed-form solutions for the values of a and b that minimize $S(a, b)$. Clearly show all steps in your derivation.

Note that a and b when viewed as unknowns are called *parameters* of the model, and the values of a and b that we calculate from the data D are known as *parameter estimates*.

Part B

Let us say that now we suspect that D might contain some outlier measurements, randomly distributed among a few of the y and x values. Suggest (briefly) two different ways we might address this problem (by modifying in some manner the parameter estimation methodology from Part A). What might happen if we ignored outliers? Use a sketch of a graph to explain your answer (hint: consider Figure 2.6 in the text).

MATLAB Exercises

To do this part of the assignment you first need to make sure you can get MATLAB up and running. Please consult the class Web page for details.

If you are on one of the unix machine where you can access the junkfood cluster, add the pathname `/fs/www/class/spring2002/cmsc828g/hws/hw1` to your path in MATLAB (on a junkfood machine). You can do this from the command line, or by choosing File and next Set Path. If you are using a PC or MAC, you will need to copy the files from the above directory to your machine. You can access them from the class web page, `www.cs.umd.edu/class/spring2002/cmsc828g/hws/hw1`

A. Demo Scripts

Type `pimademo` at the command line in MATLAB and a script will be executed to illustrate some basic data analysis and visualization capabilities in MATLAB, using the Pima Indians data set. Please go through this carefully and make sure you understand what is being plotted at each step.

Now type `censusdemo` at the command line in MATLAB and go through the demo for the Census data.

Feel free to explore other aspects of the data beyond what is shown in the demo, e.g., call the functions in the script with other variables, use other visualization functions in MATLAB to look at the data (e.g., bar charts, pie-plots, etc.).

B. Data Exploration

For each of the datasets "census" and "pima" do the following:

- Load the each data set by typing `load pima` or `load census`.
- Find 3 interesting histograms (for three variables in each data set), plot them, and briefly discuss why they are interesting. To do this you will have to figure out how to call the dmhist.m and dmhist2.m functions (that are automatically included in your path from above).
- Find 2 interesting scatter plots (for two sets of two variables in each data set), plot them, and briefly discuss why they are interesting. Again you will need to use dmscatter.m and/or dmscatter2.m for this purpose (or write your own simple functions if you wish).

Reading and Discussion

Write about one page that compares and contrasts the two assigned papers by Friedman and by Shneiderman. Discuss at least two points, conclusions or recommendations on which the authors might agree and at least two points on which the authors might disagree.