EECE 540 **Spring 2014**

Project 3: Statistics

Due: April 16, 2014 50 points

In this project we will be using the header files named “Matrix.cpp”, "RandomNumbers.cpp" and "Stats_Support.cpp", along with their associated headers. This code is what has been used in the homework, and in the examples in class. Also a binary matrix file, called “StatsData.mtx” is on the website. You will need to download it.

This project will involve writing programs that perform statistical analysis of data, establishing which sets are related and which are not. Also we will be doing an experiment that will demonstrate the concept of a Confidence Interval.

0. Coding Basics :

a) Commenting (1)
b) Formatting and Indention (1)
c) Function and Variable Naming (1)
d) Proper Layout, Loop's, If's and such. (2)

1. Regression and Correlation:

a) Using the matrix read function from Project 2 and the functions in the “Stats_Support.h” file, compute the Linear Regression (LR) parameters and the correlation coefficient (CC) between the each of the first four rows and the fifth row from "StatsData.mtx". In other words, compute the LR parameters and CC between row1, and row 5, then between row2 and row 5, row 3 and row 5 and then finally between row 4 and row 5. (5)

b) Select the two rows with the higher CC from the independent rows (row 1, row 2, row 3 and row 4). Then perform a multivariate regression for the data in the file "StatsData.mtx". The fifth row should be the dependent variable and the two selected rows are to be treated as the independent variables. Be sure to compute the Coefficient of Determination (CD) and then compute its square root, which is similar to the CC.

The selection of the two rows, does not need to be done in software, but rather can be "hard coded" into the program. (10)

c) -Based on the weights, and CC computed from the 5 cases of regression analysis performed, what can be said about the data?
-What data and terms appear to be related to the dependent variable and which is not related?
-Are the weights truly helpful in determining which variable is more important to the dependent variable? (5)

2. Histograms, PDF's and Confidence Intervals:

a) Assuming that the mean and variance of the entire row is basically the same as the parameters of the hidden process, compute the 90% Confidence Interval (CI) for each row, in “StatsData.mtx” assuming a sample size of 81, (Note sqrt(81) = 9). (4).

b) Then compute the mean for each 81 point subsection of each row. There will be 2000 of these 81 point subsections in each row. These will be referred to here after as the Short Interval Means (SIM's). Compute the mean and variance of the 2000 SIM's and compare this to the predicted mean and variance. The mean of the SIM's should be the same as the mean for the row, while the variance should be the variance of the row divided by length (81 points) of the short intervals. (5)

c) Count the number of times the SIM's fall inside the 90% CI bound for the each row. Convert this to an estimate of probability and compare it to 90%. How well do they match, noting that some of the distributions are not Gaussian?. (2)

d) Produce a histogram of the data in each row of the matrix. Setting the number of bins to the square root of the number of samples, and using the range of the data in each row create the histogram for each. Use the histograms to plot an estimate of the probability density of each row. Also plot the matching Gaussian distribution for the row, based on the mean and variance of the entire row. (5)

e) Finally compute the Absolute-Sum-Difference (ASD) between the histogram estimate and the Gaussian PDF. The formula for which is given here.

$$ASD = \sum_{n=1}^{N} \left| HE_n - PDF_n \right|$$

where HE_n is the Histogram Estimate at bin n, and PDF_n is the PDF at bin n.

Based on the histogram plots, and the ASD, what type of distribution is each row, and can the ASD be used as a measure of how Gaussian a set of data is? (9)