**MATH 3210**

**Assignment #5 - Experimental Design Exercise**

**Scenario:** This exercise explores the research – problem solving process. The assignment has two stages. In the first stage you’ll experiment with the algorithm, familiarizing yourself with its behavior. Then, you’ll use it to cluster an unknown dataset in the second stage.

**Stage 1:** You must design your own experiment to answer the questions for this stage. You will use the K-Means Clustering algorithm and the Iris dataset for this stage. Your experiment should attempt to answer the following questions:

* How effective is the K-Means Clustering algorithm in clustering data?
* Is this effectiveness sensitive to the choice of the value K?
* Is this effectiveness sensitive to the starting values for the cluster means?
* How tolerant is the K-Means Clustering algorithm to noise?

You must design and create the datasets you need to complete your experiment for this stage from the original Iris dataset. You may need several datasets to answer the assignment questions. You may look at my Iris datasets for the K-Means Clustering algorithm to understand the dataset format required by the program, but you must make your own datasets for this assignment. Consult the header of k\_means\_multi.cpp for the file structure of the control and input datasets.

What you need for this exercise:

* The k\_means\_multi.exe executable
* Your designed datasets [using iris.xls as a starting set]
* Your control.dat control file

**Stage 2:** Design and conduct your own experiment to cluster the data in Seeds.xls spreadsheet using the K-Means mulit-attribute Clustering algorithm. Your results from stage 1 will help you design your experiment for this stage and address the analysis technique in your analysis report.

**Write a COMPLETE Analysis Report describing your analysis and recommendations for your work in Stage 2 only.**

**Disclaimer:** There is no guarantee that the Seeds data will form interesting clusters [or any clusters], but that’s what data mining is all about.