Stat 5353 - Fall 2017
Project Proposal

Dalton Cole
Adam Harter
Samuel Richter

October 13, 2017

For this project, we propose applying k-means clustering to the UCI heart disease data set [1]. Our factors will be: distance formula used and number of clusters formed. The three levels of distance formulas will be: euclidean distance, manhattan distance, and chebychev distance. The three levels of the number of clusers will be 2, 3, and 5. The purpose of the number of clusers is that the data set has 5 different levels of heart disease. Zero represents no heart disease, and 1-5 represents different levels of heart disease. Since this is a completely randomized design, we will have to replicate the experiment twice. The initial starting points for each cluster will be determined randomly, thus introducing "environmental" error. The response variable of this experiment is the resulting sum squared error of the clusters.

We will be using the nltk libary in Python to perform k-means [2]. After creating the model, we will test the model against a set aside test set. The sum of squared error or the actual value and the predicted value will be the final result. Table 1 shows the corresponding ANOVA table.

Table 1: ANOVA Table

Source	d.f.	SS	MS	F-ratio
Treatment Combinations	8	$SS_{TreatComb}$	$MS_{TreatComb}$	F_{Trt}
Number of Clusters (NC)	2	SS_{NC}	MS_{NC}	F_{NC}
Distance Formula (DF)	2	SS_{DF}	MS_{DF}	F_{DF}
NC * DF	4	SS_{NCDF}	MS_{NCDF}	F_{NCDF}
Error	9	SSE	MSE	
Total	17	SSTotal		

References

[1] `http://archive.ics.uci.edu/ml/datasets/HeartDisease`.

[2] `http://www.nltk.org/_modules/nltk/cluster/kmeans.html`.