Stat 5353 - Fall 2017
Final Report

Dalton Cole
Adam Harter
Samuel Richter

December 3, 2017

1 Introduction

In this experiment, k-means clustering was applied to the UCI heart disease data set[1], with the factors being the distance formula used and the number of clusters formed. The three levels for the distance formula used were Euclidean distance, cosine distance, and Jaccard distance. The three levels for the number of clusters were 2, 3, and 5. The response variable of interest is the resulting sum squared error of the clusters.

2 Design

The design of the experiment was a completely randomized design. To achieve this, every possible combination of the factors was run in random order, with each combination being run twice. The initial starting points for each cluster was also initialized randomly. "Environmental Error" is introduced by randomly choosing data points to train on and data points to test against.

3 Procedure and Data Collection

To collect data, k-means clustering was performed used the Python library nltk[2]. The data for the k-means clustering itself was taken from the UCI heart disease data set and each point contained 14 used attributes, which are as follows:

1. Age
2. Sex
3. Chest Pain Type
4. Resting Blood Pressure
5. Cholesterol Level
6. Fasting Blood Sugar > 120 mg/dl
7. Resting ECG Results (normal, ST-T wave abnormality, left ventricular hypertrophy)
8. Maximum heart rate achieved
9. Exercise induced Angina
10. ST Depression Induced by Exercise Relative to Rest
11. Slope of Peak Exercise ST Segment (up-sloping, flat, down-sloping)

[1] http://archive.ics.uci.edu/ml/datasets/Heart+Disease
[2] http://www.nltk.org/_modules/nltk/cluster/kmeans.html

12. Number of Major Vessels Colored by Flourosopy
13. Thalassemia
14. Diagnosis of Heart Disease

4 Analysis of Results

Table 1: ANOVA Table

Source	d.f.	SS	MS	F-ratio
Model	8	16247.444	2030.93	44.4189
Error	9	411.500	45.72	Prob > F
Combination Total	17	16658.944		<.0001

Table 2: Effect Tests

Source	Nparm	DF	SS	F Ratio	Prob > F
Number of Clusters	2	2	15786.111	172.6306	<.0001
Distance Metric	2	2	100.000	1.0936	0.3757
Distance Metric * Number of Clusters	4	4	374.222	2.0462	0.1711

Table 3: Experimental Data

Distance Metric	Number of Clusters	Number Correct
Euclidean	2	180
Euclidean	3	139
Euclidean	5	105
Cosine	2	184
Cosine	3	149
Cosine	5	104
Jaccard	2	176
Jaccard	3	151
Jaccard	5	108
Euclidean	2	180
Euclidean	3	138
Euclidean	5	99
Cosine	2	186
Cosine	3	137
Cosine	5	101
Jaccard	2	174
Jaccard	3	136
Jaccard	5	128

Table 4: Tukey's test ($\alpha = 0.05$)

LSMean[i] \ LSMean[j]

Mean[i] - Mean[j] Std Err Dif Lower CL Dif Upper CL Dif	2	3	5
2	0 0 0 0	38.3333 3.903 94 27.4335 49.2332	72.5 3.903 94 61.6002 83.3998
3	−38.333 3.903 94 −49.244 −27.433	0 0 0 0	34.1667 3.903 94 23.2668 45.0665
5	−72.5 3.903 94 −83.4 −61.6	−34.167 3.903 94 −45.067 −23.267	0 0 0 0

Level				Least Sq Mean
2	A			180.00000
3		B		141.66667
5			C	107.50000

5 Conclusion

The ANOVA table can be found in Table 1. Using $\alpha = 0.05$, the different combinations of factors was significant, as Prob > F < .0001, which is less than 0.05. Using the effect tests table, found in Table 2, several things can be concluded. There is no interaction between the Distance Metric and the Number of Clusters, as Prob > F = 0.1711, which is greater than 0.05. The Distance Metric was also not significant as Prob > F = .3757, which is greater than 0.05. The number of clusters, however, was significant, as Prob > F < .0001, which is less than 0.05. Each number of clusters was statistically distinct and having two clusters produced the most accurate predictor, as shown in Table 5 and Table 4. Tukey's test was performed on the number of clusters instead of a regression as the number of clusters is discrete, and not continuous.

Table 5: Number Correct by Cluster Size

Number of Clusters	Number Correct
2	1080
3	849
5	645

A List of Data

TODO: Put data here