# Week 5 Assignment: Predictive Modeling Using Clustering

Jonathan Ibifubara Pollyn

College of Science, Engineering and Technology, Grand Canyon University

DSC-530: Predictive Modeling

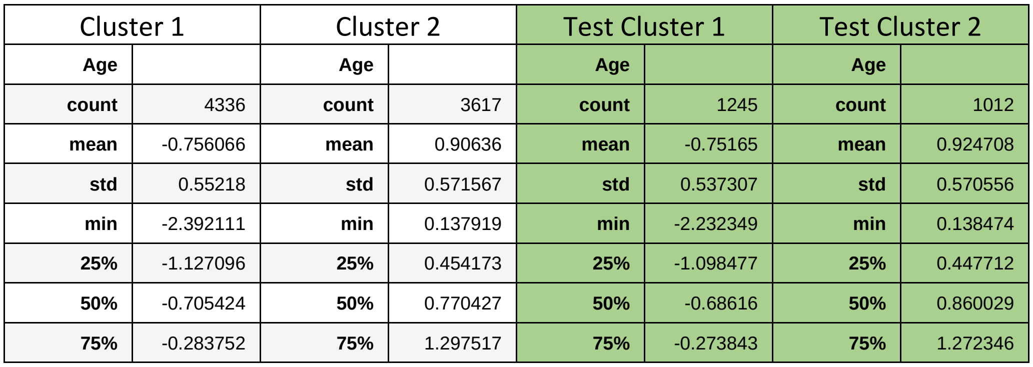
Filippo Posta

October 06, 2021

## Part 1: Operational Tasks

**Construct a table of statistics summarizing your clusters. Describe what these two clusters consist of and test validation.**

The statistic table shows that cluster one has 4336 records out of the 7953 records, while the remaining 3617 records go to cluster two. Cluster one has an age mean below the sex age with a mean of approximately -0.756 compared to cluster two, which is above with a value of about 0.906. with cluster one having an age standard deviation of 0.552 of the reported sex age and cluster two having 0.571, there is not much difference. The statistic table also shows a significant difference in the maximum and minimum age for the two clusters. Cluster one has a minimum sex age below the minimum age -2.392 and a maximum above with a value of 0.032. However, cluster two has maximum and minimum age above with a value of 2.773 and 0.137, respectively. We should validate the training model by running the same process using test data whenever we perform clustering. After running the model using the test data, the statistic table below shows that the result is validated with the test data showing cluster one has a mean below the age with a value of 0.751 and standard deviation and maximum of 0.53 and 0.034, respectively. Cluster two also shows the same result as the training model with a mean above the age with the value of 0.924 with standard deviation and maximum age of 0.924 and 2.715, respectively. The below table shows data collected for both the training and test model.



| Cluster 1 | | Cluster 2 | | Test Cluster 1 | | Test Cluster 2 | |
|---|---|---|---|---|---|---|---|
| **Age** | | **Age** | | **Age** | | **Age** | |
| **count** | 4336 | **count** | 3617 | **count** | 1245 | **count** | 1012 |
| **mean** | -0.756066 | **mean** | 0.90636 | **mean** | -0.75165 | **mean** | 0.924708 |
| **std** | 0.55218 | **std** | 0.571567 | **std** | 0.537307 | **std** | 0.570556 |
| **min** | -2.392111 | **min** | 0.137919 | **min** | -2.232349 | **min** | 0.138474 |
| **25%** | -1.127096 | **25%** | 0.454173 | **25%** | -1.098477 | **25%** | 0.447712 |
| **50%** | -0.705424 | **50%** | 0.770427 | **50%** | -0.68616 | **50%** | 0.860029 |
| **75%** | -0.283752 | **75%** | 1.297517 | **75%** | -0.273843 | **75%** | 1.272346 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| max | 0.032502 | max | 2.773368 | max | 0.035395 | max | 2.715456 |

## Construct a table of statistics summarizing your clusters of K=3. Describe which records belong to each cluster and validation.

When the k-mean is three, the cluster behaved differently, with clusters one and three having a mean above the accepted sex age of 1.240 and 0.016 respectively and their maximum sex ages of 2.773 and 0.559, respectively. Cluster two have both mean and maximum age value below the expected value of -1.134 and -0.600 respectively. Validating the table below shows that the models have been validated, and both models behaved similarly.

| Cluster 1 | | Cluster 2 | | Cluster 3 | | Test Cluster 1 | | Test Cluster 1 | | Test Cluster 1 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Age | | Age | | Age | | Age | | Age | | Age | |
| count | 2262 | count | 2521 | count | 3170 | count | 942 | count | 660 | count | 660 |
| mean | 1.24052 | mean | -1.133912 | mean | 0.016572 | mean | -0.049848 | mean | 1.237049 | mean | 1.237049 |
| std | 0.459489 | std | 0.401279 | std | 0.330724 | std | 0.350032 | std | 0.455188 | std | 0.455188 |
| min | 0.665009 | min | -2.392111 | min | -0.494588 | min | -0.583081 | min | 0.65387 | min | 0.65387 |
| 25% | 0.875845 | 25% | -1.44335 | 25% | -0.283752 | 25% | -0.376922 | 25% | 0.860029 | 25% | 0.860029 |
| 50% | 1.086681 | 50% | -1.021678 | 50% | 0.032502 | 50% | -0.067684 | 50% | 1.169267 | 50% | 1.169267 |
| 75% | 1.508353 | 75% | -0.810842 | 75% | 0.348755 | 75% | 0.241553 | 75% | 1.478505 | 75% | 1.478505 |
| max | 2.773368 | max | -0.600006 | max | 0.559591 | max | 0.550791 | max | 2.715456 | max | 2.715456 |

## Construct a table of statistics summarizing your clusters of K=4. Describe which records belong to each cluster.

When the k-mean is four, the cluster shows that clusters two and three have a mean and maximum sex age above the accepted sex age with the value of 1.533 and 0.540, respectively. Cluster one and four have a mean value below the accepted sex age with a value of -1.330 and -0375, respectively. After running the validation (using test data), it shows that all clusters have been validated as the statistic table shows that the values are in agreement between models.

After performing the model with various numbers of K, the k=3 is better because they obtained a more considerable optimal value of k, as shown from the statistical table.

## PART 2: Mathematical and Statistical Basis

**Discuss the clustering issues described in Section 2, including variable versus data clustering, hierarchical clustering, and oblique principal component clustering.**

Liu, G., & Yang, H. (2017). Self-organizing network for variable clustering. *Annals of Operations Research, 263*(1-2), 119–140. https://doi.org/10.1007/s10479-017-2442-2

The author proves that clustering based on variables versus data clustering are based on data organized into rows of samples and variables in a typical table format. The author stipulated that "Clustering" is frequently referred to as "data clustering." As a result, data clustering aims to find patterns or features shared by samples in the same row. The author emphasized that data clustering organizes data samples into identical subsets, where samples are within a cluster are more closely separated than samples inside other clusters. Also, the author stated that clustering variables is more concerned with the variables in the rows. In the case of huge data, there are typically more variables than samples. Detecting subsets of homogenous variables and then grouping them into the same groups, where variables have greater collaboration than those in other groups, is necessary for complex relationship patterns among variables.

The author stipulated that prior research has combined hierarchical clustering with Pearson's correlation and mutual information for variable clustering. It evaluates the linear dependency between two variables by comparing Pearson's correlation coefficients. Variable correlations are characterized and quantified using mutual information, which means that each variable in the initial stage of an algorithm uses just one singleton cluster. After that, the two clusters that are the closest to one another are combined into a single cluster. The recursive merging moves up the hierarchy until the stopping requirement is met, including the maximum clusters or the maximum group-average difference.

The author stated that oblique principal component clustering is a common technique for variable clustering that uses latent-variable techniques. Saying that suppose Xnp is equal to one of the following values: An xinT data matrix represents the results of an experiment, where each row represents one sample, and each column represents one variable. It's possible to get zero mean and unit standard deviation for the variables in the data matrix X is without sacrificing generality. The author went on to assert that data is transformed into orthogonal space using principal component analysis (PCA), which uses a sparse collection of principal components (PCs) to maintain most information in the raw data. Latent variables are linear projections of the original variables, which are what principal components are.

**Evaluate the self-organizing network discussed in Section 3.2, and in particular the force equations, for its applicability to the clustering issue discussed in the paper**

Liu, G., & Yang, H. (2017). Self-organizing network for variable clustering. *Annals of Operations Research*, *263*(1-2), 119–140. https://doi.org/10.1007/s10479-017-2442-2

The author referred to data in Figure 3b which tries explains asymmetric nonlinear interdependence. The author stated that algorithms that use similarity to cluster data are inapplicable here the referred data. That nonlinear interdependences among variables are not entirely considered when employing latent-variable approaches such as oblique PCA or component analysis. The author stated that organization network method can be created to deal with these issues to group variables with nonlinear and asymmetrical interrelationships. The author's current research extends to earlier work on organization recurrence network architecture to self-organizing clustering of extremely redundant variables. Author went on to say that there has not been much research done on clustering variables with nonlinear and asymmetrical interdependences in literature. The author proposed variables as nodes in the network and nonlinear interdependencies between variables as the weights of edges, ranging from 0 to 1. To simplify, consider G as the weighted directed network with nodes and edges whose weights equal 1. The author stated that there are two types of forces in a spring-electrical system: attracting and repulsive. Repulsive forces occur between two nodes, whereas attractive forces exist between nodes with a nonlinear dependency relationship. As the nonlinear dependency between two nodes (variables) decreases, the repulsive force increases. This is because a stronger repulsive force will be required to separate two nodes.

**How do these model parameters affect the model-driven predictive model of space-time vectorcardiogram (VCG) signals?**

Liu, G., & Yang, H. (2017). Self-organizing network for variable clustering. *Annals of Operations Research*, *263*(1-2), 119–140. https://doi.org/10.1007/s10479-017-2442-2

The author test and verified the proposed technique with a real-world case study that uses the extracted parameters from vectorcardiogram (VCG) signal representation models for myocardial infarction prediction modeling. For 3-lead VCG signals, the author established a sparse basis function model that reduces the number of basic functions used while still providing adequate explanatory power. Since huge quantities of data are reduced to a small number of model parameters (such as weights, shifts, and scaling factors in basis functions), while still retaining the information, this technique is known as data reduction. The author stated that to create a lasso-penalized logistic regression model for predicting cardiac diseases, they employed model parameters and their derivatives as predictors. However, the author stated that results from experiments reveal that parametric characteristics are highly correlated, leading to sensitive prediction models (i.e., increased variances of estimation). It is the first-time topological network methods have been developed for dealing with redundancy and relevance in variables while also enhancing predictive modeling performance.  The author then stated that the modeling of vectorcardiogram (VCG) signals using multiscale is a sparse basis function model for 3-dimensional VCG signals. With the multiscale modeling the author stated, enormous quantities of data may be condensed down to a small number of model parameters while retaining all relevant information. Predictive models for myocardial infarctions will be further developed utilizing a low-dimensional collection of model parameters rather than the original data.

# References List

Larose, C. D., Larose, D. T., & Larose, Chantal D., Author. (2019). *Data science using python and r*. John Wiley & Sons,inc,.

Liu, G., & Yang, H. (2017). Self-organizing network for variable clustering. *Annals of Operations Research*, *263*(1-2), 119–140. https://doi.org/10.1007/s10479-017-2442-2