P9120 Final Project Proposal

Qihang Wu | qw2331

Topics & Objective

My project will mainly focus on one type of multivariate analysis method: Clustering. The idea that underpins this method is to identify the unmeasured (unknown) cluster membership among subjects and therefore the heterogeneity underlying a population. Despite some popular classes of methods, such as K-means and hierarchical clustering, most of them will function well especially when the data is partitioned based on the continuous variables. However, a rigorous method to categorize the subjects with binary or discrete indicators is still underdeveloped and in great demand. For this project, I will briefly discuss the traditional methods first with real data analysis and then lay stress on the model-based clustering (i.e., a Latent Class Analysis, a.k.a., LCA) to generalize the ideas of K-means and its application in the area of psychology, followed by the method comparisons and discussion at the end of the project. A tentative plan with anticipated methods and datasets is described as follows.

Methods plan to investigate

- (Briefly) Partitional clustering
  - Mainly on K-means, modified K-means methods, and K-medoids clustering
  - Data Pre-processing
  - How to choose the number of clusters/criterion: elbow method, silhouette method, and gap statistic will be briefly discussed and compared
  - Partition visualization
- (Briefly) Hierarchical clustering
  - Use different inter-cluster linkages and distance measures
  - (without a predefined number of clusters) Dendrogram visualization
- Principal Component Analysis (PCA) may be used for a clearer visualization and interpretation of the partitional clustering
- (Main) Latent Class Analysis: a statistical method to identify unmeasured (latent) class membership among subjects using either categorical and/or continuous variables (if in this case, named Latent Class Profile)
  - EM algorithm
  - Possibly accompanied with network analysis to a better visualization
  - Applications in the area of psychology

Dataset/simulation plan to use

- Customer Personality Analysis: it helps a business to modify its product based on its target customers
- Big Five Personality Traits: anonymous personality survey data containing around 1,015,341 records and 50 questions (with values from level 1 to 5 for five personalities including extraversion, agreeableness, conscientiousness, emotional stability, and intellect)
- Possibly EM algorithm