P9120
Topics in Statistical Learning and Data Mining - Final Project

Latent Class Analysis and Clustering on both Continuous and Categorical Cases

Qihang Wu | qw2331

Introduction

- Clustering analysis is commonly used nowadays. As one type of multivariate analyses and unsupervised machine learning algorithms, it aims to group subjects in a way that the objects within the same group will share more similarities than the other groups. The term ”similarity” is defined either by calculating the predefined distances between any two observations or figuring out which observations will come from the same distribution. According to these rationales, there are multiple ways to perform clustering. For instance, K-means and hierarchical clustering are commonly used when a distance (s.t., Euclidean, Manhattan distance) is assumed. An extension of K-means (i.e., fuzzy K-means) is applied if people believe one subject will not just belong to an unique cluster. Other methods like the graph-based clustering is then dividing the vertices in multiple groups so that the number of edges lying between the groups will be minimal.
- However, clustering on categorical data is more difficult than clustering on numeric cases due to some attributes of categorical variables like high dimensionality or the existence of subspace clusters, which will become an issue. More specifically, traditional clustering methods represent each cluster by all dimensions identically. Therefore, it is incapable for higher-dimensional categorical data[1]. Latent Class Analysis (LCA) is a statistical method for identifying unobserved class membership. Different from the traditional methods mentioned above, it uses probabilistic models rather than the predefined distance measures. Meanwhile, it could be applied to categorical and/or continuous variables. Last but not least, it returns probabilities instead of direct class memberships. Thus LCA is widely used in various fields like Psychology and Social Sciences.

Objective

- The goal of this project is to first discuss the traditional methods, i.e., K-means and K-mediods used for the mixed types of data to analyze the customers' behavior and then lay stress on a model-based clustering algorithm, i.e., LCA to generalize the ideas of K-means followed by its application in a real survey data. Methods comparison and evaluation will be at the end of the project.

Data Description

This project contains two data analysis. The first one is the customer personality analysis, which aims to provide a detailed analysis of a company's ideal customers based on their demographic information and behavior. Such analysis helps a business to modify its product according to the

target customers. Then the analysis on personality survey data is on which this project will mainly focus. (Note that all tables and figures mentioned below are in the Appendix in the end.)

- **Customer Personality:** Obtained from Kaggle, the raw data contains 2,240 customers with 29 predictions including some demographics like the year of birth, education, income and also variables related to commercial promotion and buying behaviors. The main data cleaning steps and feature engineering are described below.

  1. Combine several variables or several levels within one variable: (1) 'AcceptedComp' if one customer accept any one of the offer among total 6 campaigns; (2) Combine levels of 'Marital Status' into two levels (i.e., along vs. not alone); (3) Recode the levels of 'Education' for clarification; (4) Sum up number of kids and teens into one variable 'NumChild';
  2. Create some new features: (1) 'Age' from 'Year of birth'; (2) 'Family Size' based on the above two variables 'IsAlone' and 'NumChild';
  3. Impute the only missing data in 'Income' by replacing them with the mean values as the number of missing is minor ($\sim$20) compared to the whole sample size.

  Figure 1 shows the demographic characteristics of total **2,238** customers after cleaning.

- **Big Five Questionnaire:** This data comes from International Personality Item Pool (IPIP) Inventories. The raw data contains over 1M observations with five personality traits: **Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism**[1] (OCEAN). There are 10 questions for each trait with scale 1-5 (s.t., 1=Most Disagree, 3=Neutral, 5=Most Agree). For this analysis, we only use 0.2% observations for computational efficiency and remove observations with any answer to 0[2] and those IPC=1[3]. Final **1,191** observations are used for this project. **Figure 2** shows the geographical distribution of the participants. The top one country with the highest number of participants is the United States.

Statistical Learning

- **Part I:** For the customer data analysis, we use both the K-means with Euclidean distance and K-medoids with Gower distance[4] for the mixed types of data with the optimal number of clusters choosing by Silhouette coefficient (**Figure 3**).

  1. **K-means:** As one type of partitional clustering. The algorithm is simple: randomly choose a number of k centroids, assign all data points to their "nearest" centroids, re-calculate the centroid of each cluster, and iterate the above steps until the clusters do not change. The equation (1) is the cost function used in the iterations[2].

$$\sum_{i=1}^{k} \sum_{\mathbf{x_j} \in S_i} \|\mathbf{x_j} - \mathbf{c_i}\|^2, \tag{1}$$

  where S_i represents the subset of points belonging to the i-th cluster and $\mathbf{c_i}$ is the centroid.

[1] Neuroticism is represented as Emotional Stability in this project.
[2] Score 0 may due to the participants didn't pick or forgot to pick an answer.
[3] IPC: number of records per user's IP address. Higher values indicate shared networks like universities or companies.
[4] Scaled in a numerical range from 0 (identical) to 1 (entirely different).

2. **K-medoids:** In this project, we take the advantage of the PAM (Partitioning Around Medoids). Similar with the K-means, it is also a distance-based clustering. The major differences between these two methods is that (1) in PAM, the medoid of a cluster found is the actual data point, which leads to a better interpretation; and (2) the centers are chosen based on the median of all the attributes (i.e., both continuous and categorical variables). In general, K-medoids is more robust to noise compared with K-means. The dissimilarity matrix is calculated through the Gower distance, which was proposed by Gower in 1971 at first. The similarity measure[3] between any two q-dimentional observations $\mathbf{x_i}$ and $\mathbf{x_j}$ with p categorical variables and $q-p$ continuous variables is defined as

$$D_{\mathbf{x_i},\mathbf{x_j}} = \frac{\sum_{r=1}^{p} \omega_{\mathbf{x_i}\mathbf{x_j}z_r} D_{\mathbf{x_i}\mathbf{x_j}z_r}}{\sum_{r=1}^{p} \omega_{\mathbf{x_i}\mathbf{x_j}z_r}} + \frac{\sum_{r=1}^{q-p} \omega_{\mathbf{x_i}\mathbf{x_j}c_r} D_{\mathbf{x_i}\mathbf{x_j}c_r}}{\sum_{r=1}^{q-p} \omega_{\mathbf{x_i}\mathbf{x_j}c_r}}, \tag{2}$$

where the vector $\mathbf{x}$ can be written as $\mathbf{x} = (z_1, \cdots, z_p, c_1, \cdots, c_{q-p})^T = (\mathbf{z}^T, \mathbf{c}^T)$. $\omega_{\mathbf{x_i}\mathbf{x_j}z_r}$ and $\omega_{\mathbf{x_i}\mathbf{x_j}c_r}$ represent the weights for categorical variable z_r and continuous variable c_r, respectively. Distances are defined either as 0-1 distance for categorical or Manhattan distance for continuous variables.

- **Part II:** For Big Five personality analysis, LCA is implemented with number of classes chosen by the Bayesian information criterion (BIC) (**Figure 4**).

1. **Latent Class Analysis (LCA):** LCA is a way to uncover the hidden patterns of relationships between observations and used for binary data, nominal variables, and ordered categorical variables. The subgroups are called latent classes in the analysis. If there are I indicator variables in the data, s.t., $X_i \in \{1, 2, \cdots, K_i\}$ and $i \in \{1, 2, \cdots, I\}$. ξ is a latent variable represents the class membership, s.t., $\xi \in \{1, 2, \cdots, A\}$. $P(\mathbf{X} = \mathbf{x}|a)$ is then the conditional probability of $\mathbf{X} = \mathbf{x}$ for a given latent variable a. A general form of latent class model will be[4]

$$P(\mathbf{X} = \mathbf{x}|a) = \sum_{a=1}^{A} v_a P(\mathbf{X} = \mathbf{x}|a) = \sum_{a=1}^{A} v_a \prod_{i=1}^{I} P(X_i = x_i|a), \tag{3}$$

where v_a is the probability that an individual in this population is from the latent class a, s.t., $\sum_{a=1}^{A} v_a = 1$ and $\sum_{x_i=1}^{K_i} P(X_i = x_i|a) = 1$ for $i = 1, 2, \cdots, I$. Then if all indicator variables are continuous instead, this type of method is named Latent Profile Analysis. To assess the adequacy of a latent class model, we here use the BIC with the general form like

$$BIC = -2lnL + q[ln(n)], \tag{4}$$

where lnL is the log-likelihood, q is the dimension of data (i.e., the number of parameters in the model), and n is the sample size. Lower BIC score is preferred among all models.

2. **EM Algorithm:** One way to fit a latent class model is through EM algorithm, which maximizes the likelihood function of the complete data including the observed variables and the latent variables. Assume $\mathbf{X}_{obs}$ for the observed data and $\mathbf{Z}$ for the latent class variable, $\mathbf{Y} = (\mathbf{X}_{obs}, \mathbf{Z})$ is the complete data. In the E-step, we calculate the conditional expectation given observed $\mathbf{X}_{obs}$ and $\theta^{(t)}$

$$Q(\theta, \theta^{(t)}) = \mathbf{E}_Z[\ell(\theta, \mathbf{X}_{obs}, \mathbf{Z}|\mathbf{X}_{obs}, \theta^{(t)}], \tag{5}$$

where $\theta^{(t)}$ represents the current understanding of the parameters. In the M-step, we update the parameters $\theta^{(t+1)}$, s.t., $\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}} Q(\theta, \theta^{(t)})$. The goal is to maximize the likelihood of the manifested data $\ell_{obs}(\theta, \mathbf{X}_{obs})$. It guarantees the increase the likelihood at each step, therefore, it will converge to the end.

Results

Customer Data Analysis

- From **Figure 3**, we notice the optimal number with the highest Sihouette width is 3. PAM is performed with a predefined Gower distance matrix. The demographic characteristics and the corresponding purchasing behaviors across different clusters are shown in **Figure 5**. The p values are calculated from one-way ANOVA for continuous variables and Chi-squared test for categorical variables. Results show that there are significant differences exisiting for many variables ($P<0.001$). More specifically, the cluster 1 has highest income compared to the other two but the family size (or the number of children at home) is less than the others. We may also notice that whether the complaint in the past 2 years has no difference across these clusters.

- To further understand their behaviors, we plot the heatmaps (**Figure 6**) with respect to the amount spent on each type of products (s.t., wine, sweet, meat, gold, fruit, and fish)[5] and the number of visits per buying channel (s.t., website or in store), respectively. They show that cluster 2 would buy more wine and meat and cluster 1 prefers to shop directly in stores.

Big Five Personality Analysis

- **Figure 4** indicates that we should choose 5 as the number of classes for the latent class model where the BIC has the lowest value 168501. Alluvial plots (**Figure 7-11**) are shown to describe the performance of each class per trait. The steps for these plots are listed as follows.

  1. Rename the column name for clarification and add the names of the five traits.
  2. Combine the question directions into the dataset and then convert the score correspondingly, e.g., one negative question[6] with answer 2 (i.e., Disagree) will have a score $6-2=4$ in the end while the scores will remain the same for positive questions. In this case, the higher score in one trait means that this subject has larger propensity (e.g., more extraverted or more conscientious).
  3. The thickness of each strip in the plots show the number of participants among all the ones choose the same answer for each question (%). For example, 35% for scoring 1 (i.e., poor) for the question "FeelEmo" in "Agreeableness" in class 5 means that among all participants choosing this answer for this question, 35% of them come from the class 5.

- Based on the above explanations, we may find that the class 5 has the poorest performance (i.e., lowest scores) for nearly every trait while the class 2 almost excel in every aspect, which generally means they are more interested in others, prepared, relaxed, talkative, and creative. Class 3 prefers solitude and cares less about how other people feel but they have relatively stable emotions.

- To further explore the class 2 and class 5, we also make a bar plot for the percentage of scoring 5 for total 50 questions in the dataset. The top 3 questions with "Most Agree" answers in class

[5] Standardize the data with mean zero and unit variance for these amount spent as they are measured in the different ways.

[6] For negative questions like "I don't talk a lot", a higher score like 5 (Most Agree) means this person is less extraverted (code 'direction' as 0) while a high score for a positive question "I am the life of the party" means one is more extraverted (code 'direction' as 1).

2 (**Figure 12**) are "Sympathize with others' feelings", "Interested in people", and "Feel others' emotions", respectively. While the top 3 questions for the class 5 (**Figure 13**) are "I don't like to draw attention to myself", "I'm quite around strangers", and "I'm worry about things".

- The last step is to estimate the Guassian Markov random field for the classes under the assumption that the data is numeric from 1 to 5, we construct the networks based on the graphical lasso (Glasso) with the penalization chosen by the extended BIC (eBIC). The thickness of each edge represents the partial correlation, i.e., thicker edges for stronger correlations. The optimal parameters λ for class 2 and class 5 are 0.21 and 0.26, respectively. Questions within a same trait are labeled in same color (**Figure 14 & Figure 15**). Apparently, the questions within the same trait will have stronger correlations, but we could still observe some interesting patterns across the groups, e.g., in class 5, people who are worried about things also like orders and tend to spend time reflecting on things.

Conclusions & Limitations

- Although there are no standards to evaluate the performance of clustering, the above K-medoids and LCA work well when the categorical variables involved. The small p values for the K-medoids clustering results indicate there are at least one difference between any two clusters. It connects the demographic characteristics and the behavior of customers. Meanwhile, the alluvial plots are intuitive and we could easily compare one latent class with the another one.

- It is known that K-medoids is not sensitive to outliers compared to the K-means, which could cause some problems that put the outliers into one of the clusters and therefore ignore some meaningful clusters to the end. Moreover, similar with the K-means, we will also need to specify the number of value for k in advance before the PAM.

- There are some potential issues in the Big Five Questionnaire data: (1) it is somewhat biased and subjective without scientific evaluations; (2) the lack of information about the participants' demographic, e.g., sex, age, and race since it is an anonymous survey; (3) the question within each trait may be overlapping in some senses so that there might be correlations between variables.

- LCA is believed to render a more convinced and powerful clustering than some traditional methods like K-means. However, this probabilities-based method cannot guarantee correct membership assignments and the name for each latent class is usually hard to identify (i.e., "naming fallacy"). Meanwhile, collapse multiple responses into two or three options may be more appropriate for the final interpretations[5].

- The network estimations may need further modifications: (1) tune the hyperparameter γ for the eBIC; (2) assume the other types of data distributions instead of Guassian; (3) perform community detection to find out which nodes (i.e., questions) will have denser connections than the others based on the estimated networks.

References

[1] Pang, N., Zhang, J., Zhang, C., Qin, X., Cai, J. (2019). PUMA: Parallel subspace clustering of categorical data using multi-attribute weights. Expert Systems with Applications, 126, 233-245. https://doi.org/10.1016/j.eswa.2019.02.030

[2] Fortunato, S. (2010). Community detection in graphs. Physics Reports, 486(3-5), 75-174. https://doi.org/10.1016/j.physrep.2009.11.002

[3] Tuerhong, G., Kim, S. B. (2014). Gower distance-based multivariate control charts for a mixture of continuous and categorical variables. Expert Systems with Applications, 41(4), 1701-1707. https://doi.org/10.1016/j.eswa.2013.08.068

[4] ESHIMA, N. O. B. U. O. K. I. (2022). Introduction to latent class analysis methods and applications. Springer Verlag, Singapor.

[5] Weller, B. E., Bowen, N. K., Faubert, S. J. (2020). Latent Class Analysis: A Guide to Best Practice. Journal of Black Psychology. https://doi.org/10.1177/0095798420930932

Appendix

For codes in details please click **here**.
Figure 1:

Characteristic	N = 2,238[1]
Age	52 (45, 63)
Education	
Associate degree	203 (9.1%)
Bachelor's degree	1,127 (50%)
Below some college	54 (2.4%)
Doctoral degree	484 (22%)
Master's degree	370 (17%)
Is Alone	794 (35%)
Income	51,790 (35,528, 68,307)
Number of Children	
0	638 (29%)
1	1,126 (50%)
2	421 (19%)
3	53 (2.4%)
Family Size	
1	254 (11%)
2	762 (34%)
3	889 (40%)
4	301 (13%)
5	32 (1.4%)
Number of Days Since Last Purchase	49 (24, 74)
Complaint (in last 2 yrs)	21 (0.9%)
Wine[2]	173 (23, 505)
Fruits	8 (1, 33)
Meat Products	67 (16, 232)
Fish Products	12 (3, 50)
Sweet Products	8 (1, 33)
Gold Products	24 (9, 56)
Number of Purchases made with a discount	2 (1, 3)
Customer Accepted the Offer	608 (27%)
Through the Company's Website[3]	4 (2, 6)
Using a Catelogue	2 (0, 4)
Directly in Stores	5 (3, 8)
Website Visits (in last month)	6 (3, 7)

[1] Median (IQR); n (%)
[2] Amount spent on a type of product in last two years
[3] Number of purchases made through a way

Figure 1: Demographics of Customers (n=2238)

Figure 2:

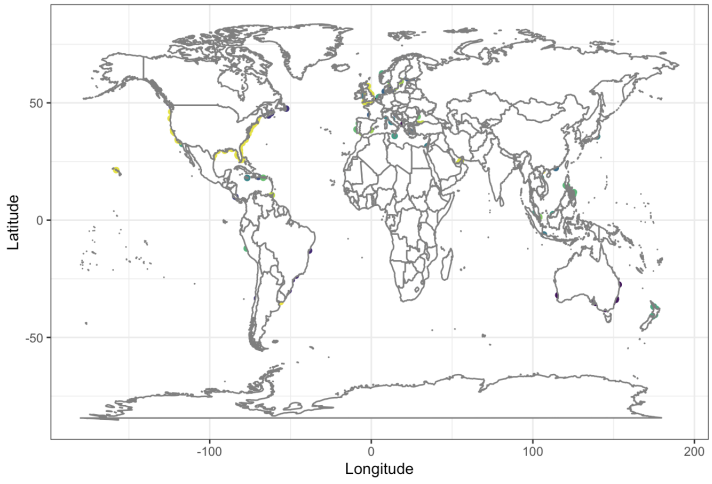

Figure 2: Geographical Distribution of Participants (n=1191)

Figure 3:

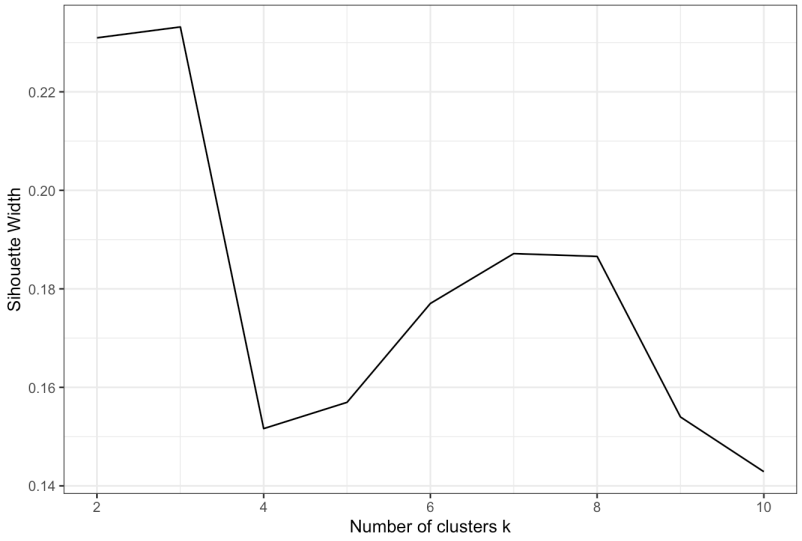

Figure 3: Optimal Number of K-medoids Clusters

Figure 4:

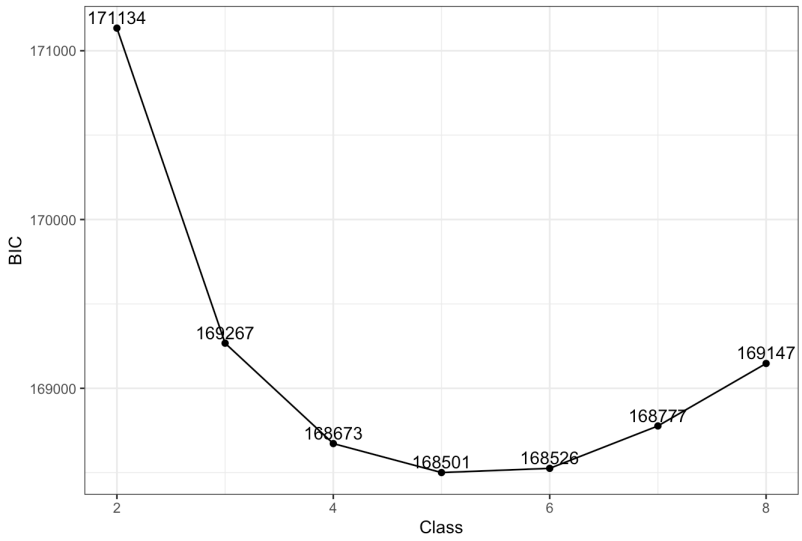

Figure 4: BIC for Latent Class Analysis for Big Five Data

Figure 5:

Characteristic	Overall, N = 2,238[1]	Cluster 1, N = 446[1]	Cluster 2, N = 671[1]	Cluster 3, N = 1,121[1]	p-value[2]
Age	52 (45, 63)	53 (43, 65)	52 (45, 62)	51 (45, 62)	0.2
Education					0.004
Associate degree	203 (9.1%)	38 (8.5%)	53 (7.9%)	112 (10.0%)	
Bachelor's degree	1,127 (50%)	222 (50%)	353 (53%)	552 (49%)	
Below some college	54 (2.4%)	0 (0%)	20 (3.0%)	34 (3.0%)	
Doctoral degree	484 (22%)	115 (26%)	142 (21%)	227 (20%)	
Master's degree	370 (17%)	71 (16%)	103 (15%)	196 (17%)	
Is Alone	794 (35%)	123 (28%)	671 (100%)	0 (0%)	<0.001
Income	51,790 (35,528, 68,307)	77,040 (69,224, 82,783)	46,734 (33,347, 61,825)	44,155 (31,535, 57,937)	<0.001
Number of Children					<0.001
0	638 (29%)	362 (81%)	141 (21%)	135 (12%)	
1	1,126 (50%)	74 (17%)	368 (55%)	684 (61%)	
2	421 (19%)	8 (1.8%)	141 (21%)	272 (24%)	
3	53 (2.4%)	2 (0.4%)	21 (3.1%)	30 (2.7%)	
Family Size					<0.001
1	254 (11%)	113 (25%)	141 (21%)	0 (0%)	
2	762 (34%)	259 (58%)	368 (55%)	135 (12%)	
3	889 (40%)	64 (14%)	141 (21%)	684 (61%)	
4	301 (13%)	8 (1.8%)	21 (3.1%)	272 (24%)	
5	32 (1.4%)	2 (0.4%)	0 (0%)	30 (2.7%)	
Number of Days Since Last Purchase	49 (24, 74)	42 (19, 71)	51 (27, 75)	50 (26, 75)	0.002
Complaint (in last 2 yrs)	21 (0.9%)	1 (0.2%)	7 (1.0%)	13 (1.2%)	0.2
Wine[3]	173 (23, 505)	710 (464, 960)	96 (17, 371)	73 (15, 292)	<0.001
Fruits	8 (1, 33)	48 (23, 93)	6 (1, 21)	4 (1, 15)	<0.001
Meat Products	67 (16, 232)	427 (258, 610)	44 (12, 142)	31 (12, 106)	<0.001
Fish Products	12 (3, 50)	72 (33, 130)	8 (2, 31)	7 (2, 21)	<0.001
Sweet Products	8 (1, 33)	49 (24, 96)	6 (1, 21)	5 (1, 16)	<0.001
Gold Products	24 (9, 56)	57 (32, 114)	22 (8, 50)	16 (6, 40)	<0.001
Number of Purchases made with a discount	2 (1, 3)	1 (1, 1)	2 (1, 3)	2 (1, 3)	<0.001
Customer Accepted the Offer	608 (27%)	342 (77%)	124 (18%)	142 (13%)	<0.001
Through the Company's Website[4]	4 (2, 6)	5 (4, 7)	3 (2, 5)	3 (2, 5)	<0.001
Using a Catelogue	2 (0, 4)	6 (4, 7)	1 (0, 3)	1 (0, 2)	<0.001
Directly in Stores	5 (3, 8)	8 (6, 11)	4 (3, 7)	4 (3, 7)	<0.001
Website Visits (in last month)	6 (3, 7)	3 (2, 5)	6 (4, 7)	6 (5, 7)	<0.001

[1] Median (IQR); n (%)
[2] One-way ANOVA; Pearson's Chi-squared test
[3] Amount spent on a type of product in last two years
[4] Number of purchases made through a way

Figure 5: Demographics of Customers for each K-medoids Cluster (n=2238)

Figure 6:

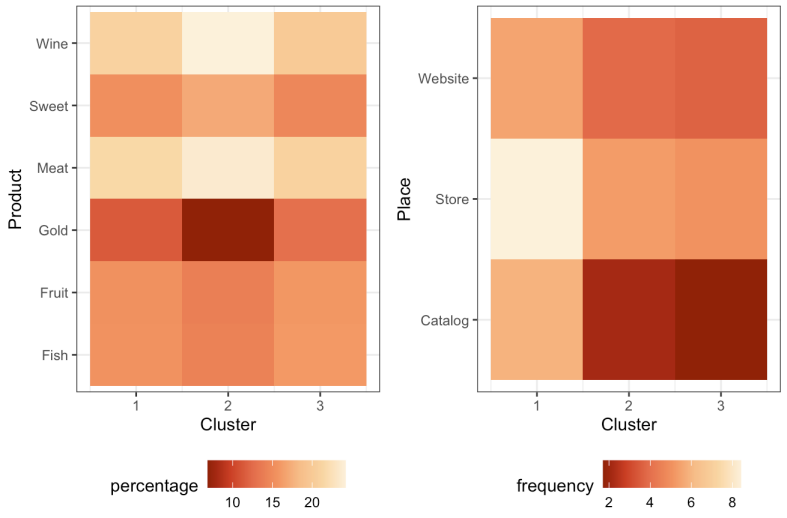

Figure 6: Heatmap of Amount Spent and Number of Visits by K-medoids Clustering

Figure 7:

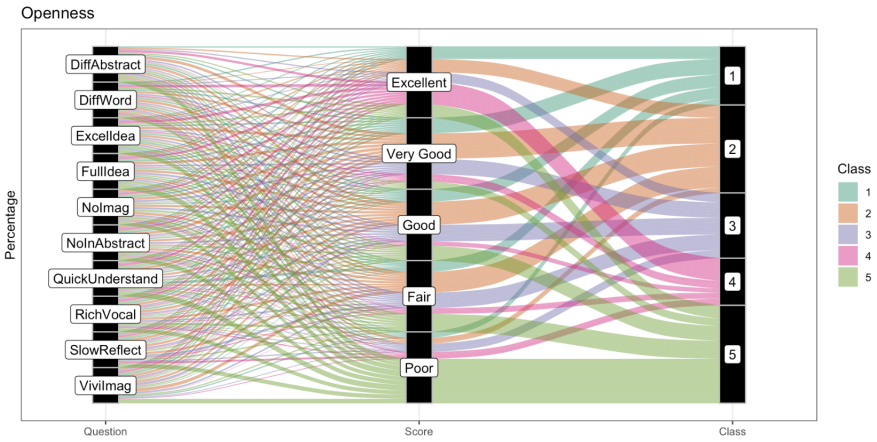

Figure 7: Alluvial Plot for Openness

Figure 8:

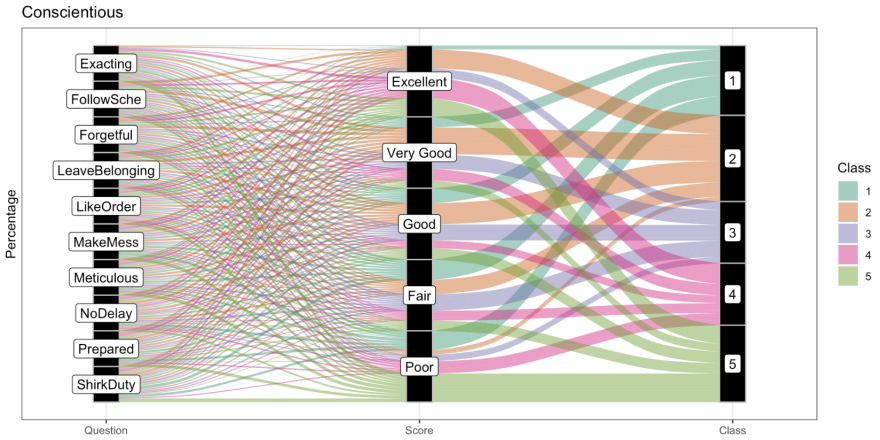

Figure 8: Alluvial Plot for Conscientious

Figure 9:

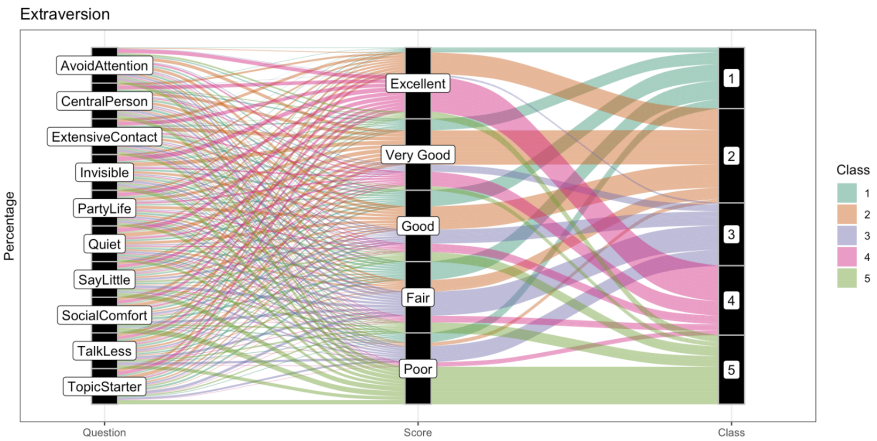

Figure 9: Alluvial Plot for Extraversion

Figure 10:

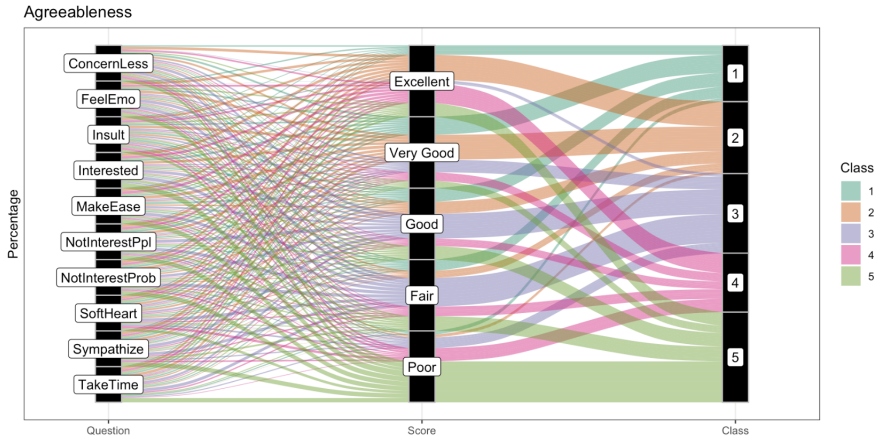

Figure 10: Alluvial Plot for Agreeableness

Figure 11:

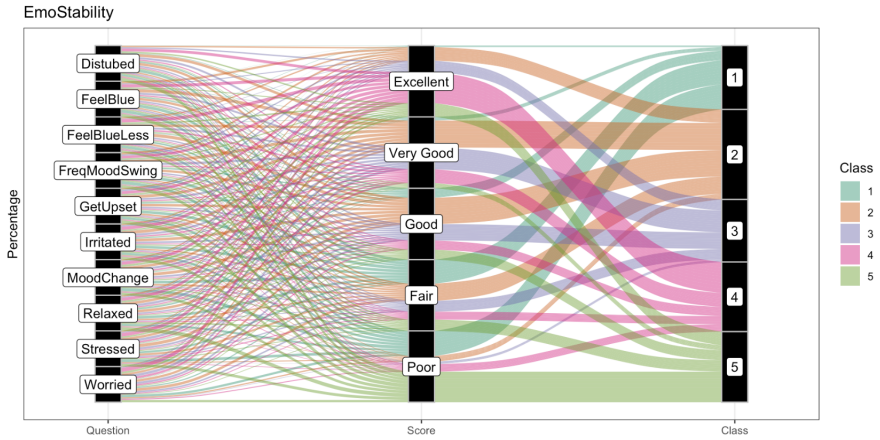

Figure 11: Alluvial Plot for Emotional Stability

Figure 12:

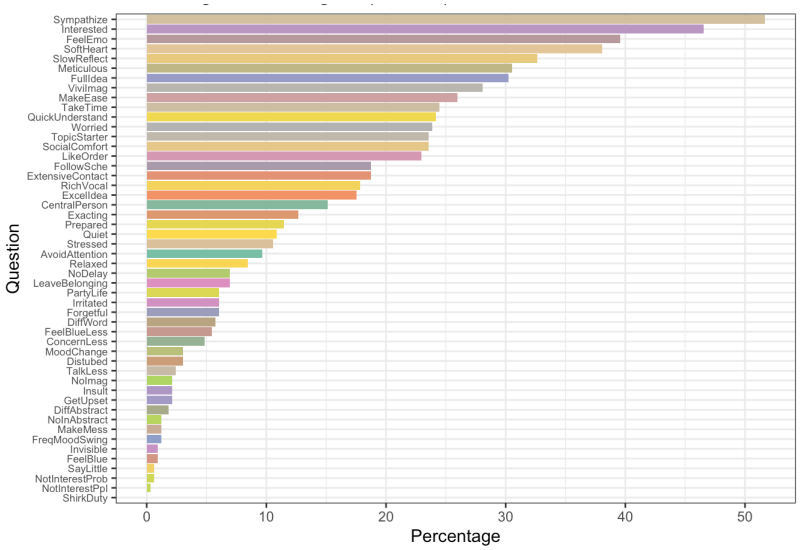

Figure 12: Percentage of Most Agree (score 5) in Class 2

Figure 13:

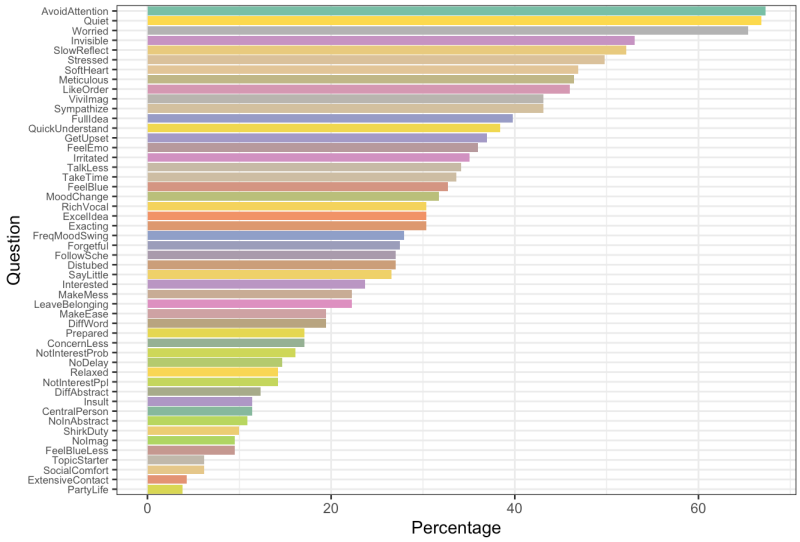

Figure 13: Percentage of Most Agree (score 5) in Class 5

Figure 14:

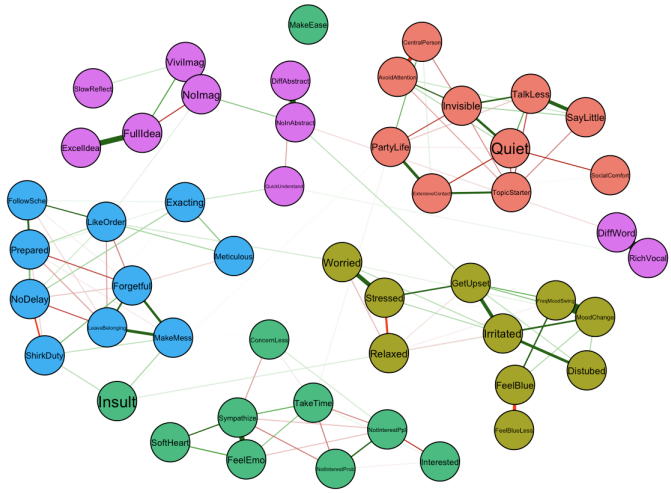

Figure 14: Undirected Graph for Class 2 (lambda=0.21)

Figure 15:

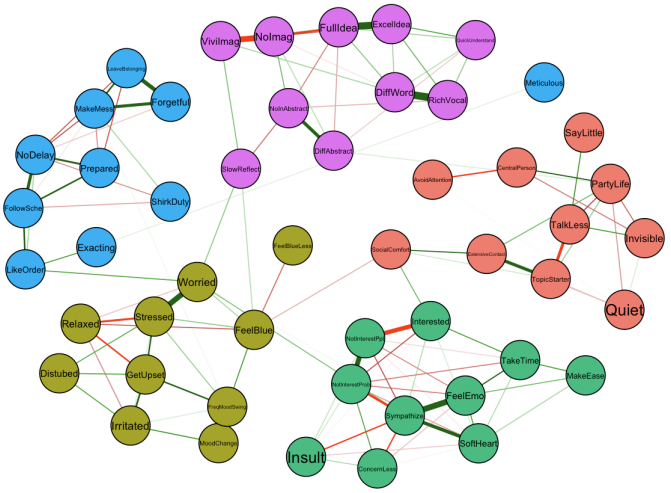

Figure 15: Undirected Graph for Class 5 (lambda=0.26)