

## **ANL305**

# **Association and Clustering**

---

### **Group-Based Assignment**

#### **July 2019 Presentation**

<b>Name of Group Leader</b>	<b>PI. Number</b>
Ho Zhong Ta Benjamin	W1711125
<b>Names of Group Members</b>	<b>PI. Numbers</b>
Lee Si Qi	K1780967
Goi Jin Yi	Z1780890

Date: 14 October 2019

---

## Question 1a

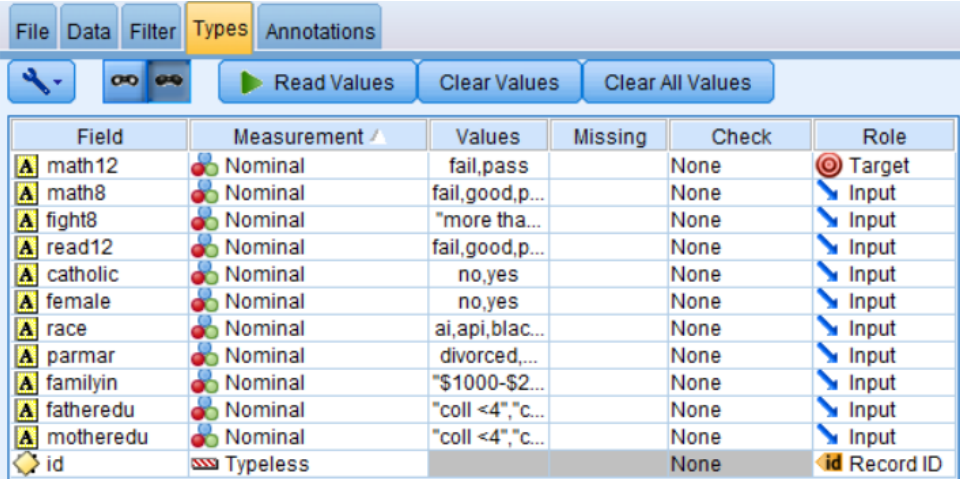
### Suitability of Association Rule Mining (ARM)

ARM seeks to discover itemsets with frequent co-occurrence within a dataset. Since the data mining objective of High School ABC is to distinguish the factors affecting students' math scores in the final exam of Grade 12, ARM can be applied to effectively identify association rules between Grade 12 math scores and variables such as parents' highest education level, marital status, family income, race, gender, performance in reading test and involvement in a fight.

Considering Apriori algorithm's inability to handle numeric inputs, it will be ideal to apply ARM on this dataset since it consists purely nominal data, where data transformation via data binning or discretization is not required.

### Question 1b i)

#### Data Preparation



The screenshot shows a software interface with tabs: File, Data, Filter, Types (selected), and Annotations. Below the tabs are buttons: Read Values, Clear Values, and Clear All Values. A table lists attributes with their measurement types, values, missing data, check status, and roles.

Field	Measurement	Values	Missing	Check	Role
math12	Nominal	fail,pass		None	Target
math8	Nominal	fail,good,p...		None	Input
fight8	Nominal	"more tha...		None	Input
read12	Nominal	fail,good,p...		None	Input
catholic	Nominal	no,yes		None	Input
female	Nominal	no,yes		None	Input
race	Nominal	ai,api,blac...		None	Input
parmar	Nominal	divorced,...		None	Input
familyin	Nominal	"\$1000-\$2...		None	Input
fatheredu	Nominal	"coll <4","c...		None	Input
motheredu	Nominal	"coll <4","c...		None	Input
id	Typeless			None	Record ID

Figure 1. Roles of Attributes Applied in Types Tab for the Dataset

As the main purpose of this study is to determine the factors affecting the Grade 12 math scores, the results of *math12* will be set as the only consequent. Thus, the role of *math12* will be set as 'Target'. The roles of attributes influencing students' math scores from *math8* to *motheredu* are set as 'Input' to ensure that the 9 variables will only be considered as antecedent. Role of *Id* field will be changed to 'RecordID'.

## Pre-modelling

Fields Model Expert Annotations

Model name: ☐ Auto ☐ Custom

☒ Use partitioned data

Minimum antecedent support (%): 25.0

Minimum rule confidence (%): 85.0

Maximum number of antecedents: 5

☒ Only true values for flags

Optimize: ☒ Speed ☐ Memory

Figure 2. Model Parameters in Apriori Node

As the default parameter setting (10% minimum antecedent support, 80% minimum rule confidence) will generate a large number of rules (i.e. 303), we propose to increase minimum antecedent support to 25% and minimum rule confidence to 85% to ensure only meaningful rules are generated. In this way, only rules with stronger co-occurring relationships will be displayed.

## Results

Model Settings Summary Annotations

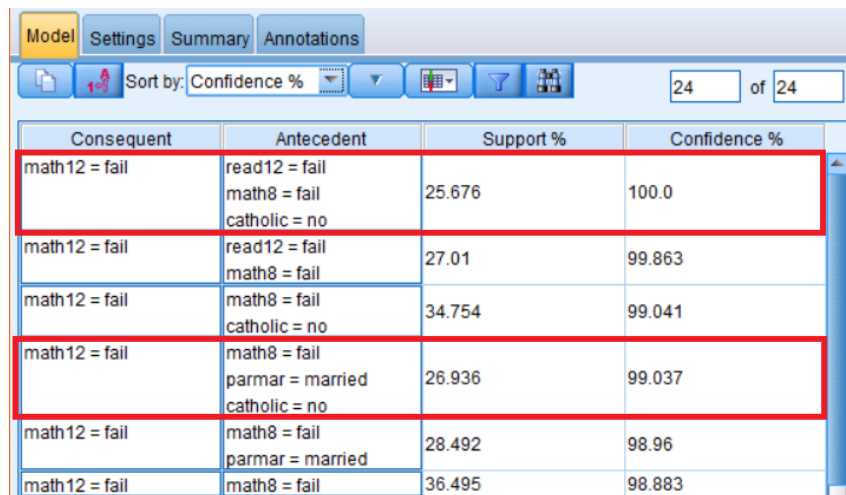
Sort by: Confidence % 24 of 24

Consequent	Antecedent	Support %	Confidence %
math12 = fail	read12 = fail math8 = fail catholic = no	25.676	100.0
math12 = fail	read12 = fail math8 = fail	27.01	99.863
math12 = fail	math8 = fail catholic = no	34.754	99.041
math12 = fail	math8 = fail parmar = married catholic = no	26.936	99.037
math12 = fail	math8 = fail parmar = married	28.492	98.96
math12 = fail	math8 = fail	36.495	98.883
math12 = fail	math8 = fail fight8 = never	25.491	98.547
math12 = fail	read12 = fail catholic = no	31.79	95.804
math12 = fail	read12 = fail	33.346	95.444
math12 = fail	read12 = fail parmar = married	25.861	94.986
math12 = pass	math8 = pass read12 = pass fight8 = never parmar = married	25.565	93.768
math12 = pass	math8 = pass read12 = pass fight8 = never	29.604	92.991
math12 = pass	math8 = pass		

Figure 3. Modelling Results using Parameter Setting Seen in Fig. 2

There are **24** rules generated.

### Question 1b ii)



The screenshot shows a software interface with tabs for 'Model', 'Settings', 'Summary', and 'Annotations'. Below the tabs is a toolbar with icons for file operations and a 'Sort by: Confidence %' dropdown. A table displays the results, with columns for 'Consequent', 'Antecedent', 'Support %', and 'Confidence %'. The table contains seven rows of data. The first, third, and fourth rows are highlighted with red borders. The first row shows 'math12 = fail' as the consequent and 'read12 = fail', 'math8 = fail', and 'catholic = no' as antecedents, with a support of 25.676 and confidence of 100.0. The third row shows 'math12 = fail' as the consequent and 'math8 = fail' and 'catholic = no' as antecedents, with a support of 34.754 and confidence of 99.041. The fourth row shows 'math12 = fail' as the consequent and 'math8 = fail', 'parmar = married', and 'catholic = no' as antecedents, with a support of 26.936 and confidence of 99.037. The second, fifth, sixth, and seventh rows are not highlighted.

Consequent	Antecedent	Support %	Confidence %
math12 = fail	read12 = fail math8 = fail catholic = no	25.676	100.0
math12 = fail	read12 = fail math8 = fail	27.01	99.863
math12 = fail	math8 = fail catholic = no	34.754	99.041
math12 = fail	math8 = fail parmar = married catholic = no	26.936	99.037
math12 = fail	math8 = fail parmar = married	28.492	98.96
math12 = fail	math8 = fail	36.495	98.883

Figure 4. Extract of Modelling Results, Highlighted Fields which includes 'catholic'

#### Rule 1

As seen in Fig 1d, *read12=fail*, *math8=fail* and *catholic=no* co-occurs with *math12=fail*. Out of all 2,699 Grade-12 students, 25.676% of them have failed their Grade-8 math test and Grade-12 reading test, and are not from a catholic high school. The Apriori modelling results shows that students with these 3 attributes will definitely fail their Grade-12 math test.

Rule 1 implies that students who possessed the aforementioned attributes have a higher tendency to fail their Grade-12 math test. This provides valuable insights for the teachers in High School ABC, enabling them to allocate their distribution of resources effectively and prioritise on students who are predicted to perform weaker in their math test. As such, this would eventually improve the school's overall math scores.

#### Rule 2

Similarly, Fig 1d shows that *math8=fail*, *parmar=married* and *catholic=no* co-occurs with *math12=fail*. As indicated in the dataset, 26.936% of the student population does not come from a catholic high school, failed their Grade-8 math test and their parents marital status is married. Rule 2 implies that students with this profile are 99.037% likely to fail their Grade-12 math test.

With this profile identified, High School ABC could formulate a teaching strategy to strengthen the mathematics foundation for students who fall in this category and anticipate their learning needs to improve on their math performance.

### **Question 1c)**

No, it is inconclusive that *catholic=no* caused a student to fail *math12*. The association rule generated merely suggests that there is a relation between *catholic=no* and *math12* but it does not suggest that students who do not come from a catholic high school caused them to fail math at Grade 12.

### **Difference between Association and Causation**

Association analysis uses a common technique called Apriori algorithm to discover the co-occurrence of attributes within a large dataset (Tan, Chan, Lee, & Li, 2019). The dependencies between the data are known as association rules, where the strength of association is determined by support and confidence values. In other words, Apriori algorithm generates association rules that satisfies the user-specified minimum support and minimum confidence constraints.

Causation relationship, on the other hand, can be learnt through controlled experiments which are often expensive and impossible to conduct (Li, et al., 2013). Causal relationship discovery adopts Causal Bayesian Network (CBN) theory and studies how the variation of one attribute cause changes to other attributes, providing useful prediction and reasoning for decision-making (Li, et al., 2013). While causal relationships imply associations, the reverse is not necessarily true (Li, et al., 2013).

To sum up, association analysis indicates statistical relationship and strength of association between variables A and B while causation identifies a cause and effect relationship where the occurrence of the first causes the other. As associations do not always imply causality, it is inconclusive to say *catholic=no* inherently caused a student to fail *math12*.

### **Question 1d)**

#### **Difference between CARMA and Apriori**

One of the distinct differences between CARMA and Apriori is that the former allows rules with multiple consequents and the latter only allows one consequent (Tan, Chan, Lee, & Li, 2019). As this problem seeks to understand the potential factors that affects students' math grades, *math12* is the only attribute we are interested as consequent. Hence, it does not make any difference on whether CARMA or Apriori is applied to solve this problem as both algorithms would result in one consequent.

Generally, both algorithms are equally suitable to be applied in this issue as the dataset consists of purely nominal data. However, as CARMA can only handle binary inputs, data preparation is required to transform the data into ‘flag’ format — data with two distinct values before it can generate insightful rules. There are 8 out of 11 fields (i.e. *math8*, *fight8*) consist of data with more than two possible values, making it an inappropriate data format for CARMA analysis.

In contrast, the nature of the dataset makes Apriori a better approach to solve the problem as it reduces the time and effort needed to transform the data.

### Data Preparation for ‘fight8’ Field

The possible values for *fight8* includes *never*, *once or twice*, and *more than twice*. This field can be transformed into two distinct value — yes or no to determine if students were involved in fights in Grade-8 of their studies.

Before	After
<i>never</i>	no
<i>once or twice</i>	yes
<i>more than twice</i>	

Figure 5: Possible Values for ‘Fight8’ Field before and after Data Preparation

*Fight8* field with records indicating ‘never’ can be replaced with ‘no’ while records with ‘once or twice’ and ‘more than twice’ can be replaced with ‘yes’. In this way, data format for *fight8* field can be specified as flag instead of nominal, which is the ideal format required by CARMA analysis.

## Question 2a

### Suitability of Applying Clustering

Clustering analysis is the task to group records displaying a high degree of similarity into the same cluster. By using clustering, it enables decision makers to explore the variables for data analysis for effective decisions. This includes identifying extreme values and outliers, in the attempt to identify the most representative cluster (Tan et al., 2019). As clustering is useful to identify attributes that share common characteristics, it is a suitable approach to determine groupings of pulsars (Tan et al., 2019).

However, as clustering does not validate the identified groupings and requires close examination of cluster profiles, subject matter expert is required to justify their interpretation theoretically or empirically (Tan et al., 2019).

## Question 2b i)

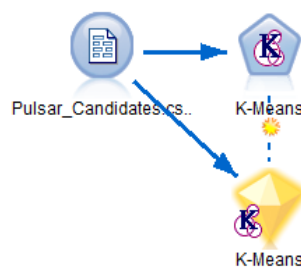


Figure 6. Stream of Constructed K-Means Model

Field	Measurement	Values	Missing	Check	Role
# Mean	Continuous	[5.8125,19...		None	Input
# SD	Continuous	[24.77204...		None	Input
# Kurtosis	Continuous	[-1.876011...		None	Input
# Skewness	Continuous	[-1.791885...		None	Input
# Mean_DM-S...	Continuous	[0.213210...		None	Input
# SD_DM-SNR	Continuous	[7.370432...		None	Input
# Kurtosis_DM...	Continuous	[-3.139269...		None	Input
# Skewness_...	Continuous	[-1.976975...		None	Input
Class	Flag	1/0		None	Target

Figure 7. Roles of Fields to be used for Modelling

## Question 2b ii)

Mislabelling Rates are defined as instances where a fake pulsar had been mislabelled as a "class=1" or real pulsar mislabelled as a "class=0".

$$\text{Mislabelling Rate} = \frac{\text{No. of Wrongly Labelled Data Points}}{\text{No. of Real Value}} \times 100\%$$

When the Number of Clusters is 2 (K=2),

\$KM-K-Means			
Class	cluster-1	cluster-2	
0	15166	1093	
1	365	1274	

Figure 8. Matrix Node when K = 2

The real value of Class 0 is **16,259** and Class 1 is **1,639**.

As majority for cluster 1 is **Class 0**,

$$\text{Mislabelling Rate for Class 1} = \frac{365}{1,639} \times 100\% = 22.27\%$$

As majority for Cluster 2 is **Class 1**,

$$\text{Mislabelling Rate for Class 0} = \frac{1,093}{16,259} \times 100\% = 6.72\%$$

When the Number of Clusters is 3 (K=3),

\$KM-K-Means			
Class	cluster-1	cluster-2	cluster-3
0	9463	979	5817
1	337	1236	66

Figure 9. Matrix Node when K = 3

Since the real value of Class 0 is **16,259** and Class 1 is **1,639**,

As the majority for Cluster 1 and 3 is **Class 0**,

$$\text{Mislabelling Rate for Class 1} = \frac{337 + 66}{1,639} \times 100\% = 24.59\%$$

As the majority for Cluster 2 is **Class 1**,

$$\text{Mislabelling Rate for Class 0} = \frac{979}{16,259} \times 100\% = 6.02\%$$

When the Number of Clusters is 4 (K=4),

\$KM-K-Means				
Class	cluster-1	cluster-2	cluster-3	cluster-4
0	10220	18	4887	1134
1	375	1008	48	208

Figure 10. Matrix Node when K = 4

Since the real value of Class 0 is **16,259** and Class 1 is **1,639**,

As the majority for Cluster 1, 3 and 4 is **Class 0**,



$$\text{Mislabelling Rate for Class 1} = \frac{375 + 48 + 208}{1,639} \times 100\% = \mathbf{38.5\%}$$

As the majority for Cluster 2 is **Class 1**,

$$\text{Mislabelling Rate for Class 0} = \frac{18}{16,259} \times 100\% = 0.11\%$$

### **Question 2b iii)**

In my opinion, the K-value of 2 is the optimal K-value to be used to study this problem. K-value of 2 is the optimal value to be used as it has the majority cluster representing Class 1 with the lowest mislabelling rate and the highest Average Silhouette Measure

For this problem, clustering using K-means is used to determine if the pulsars are real or fake by using the candidate's moving features. In order to determine the optimal K-value to study this problem, we should study the clusters' Silhouette measure of cohesion and separation, and their mislabelling rates.

The Silhouette measure of cohesion and separation is a method of evaluating cluster quality by fusing the two concepts of cluster cohesion and cluster separation (Rousseeuw, 1987). The resulting number, known as the Average Silhouette measure, is given in a range of -1 to 1, with -1 being the lowest clustering quality to 1 being highest. The three figures below, Fig 11, Fig 12 and Fig 13, illustrates the individual Average Silhouette Measure for the K-values of 2, 3 and 4 respectively.

When the number of clusters is 2 (K=2),

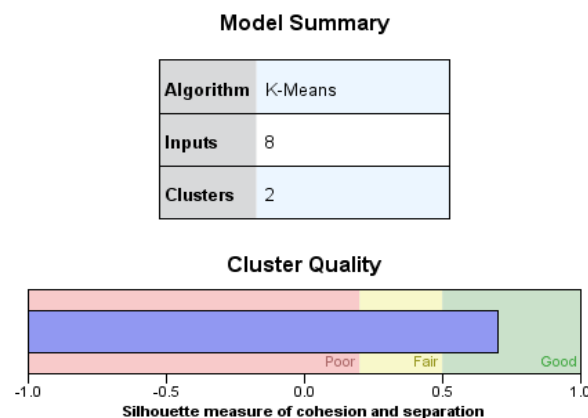


Figure 11. Cluster Quality when K = 2

Average Silhouette Measure = **0.7**

When the number of clusters is 3 ( $K=3$ ),

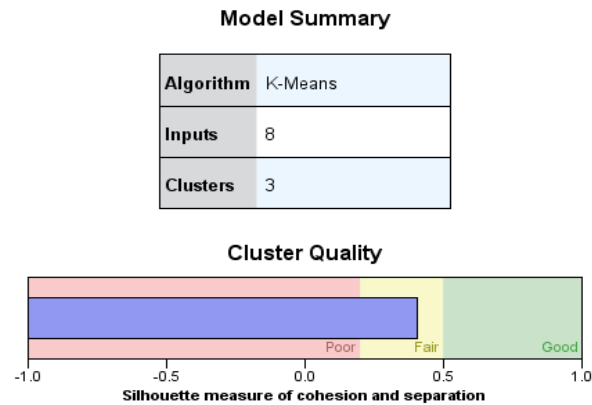


Figure 12. Cluster Quality when  $K = 3$

Average Silhouette Measure = **0.4**

When the number of clusters is 4 ( $K=4$ ),

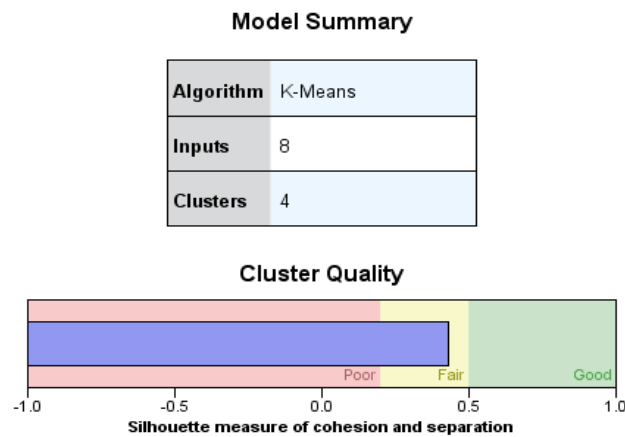


Figure 13. Cluster Quality when  $K = 4$

Average Silhouette Measure = **0.4**

Based on their Average Silhouette Measure, the K-value of 2 has the highest Average Silhouette Measure of 0.7, falling in the “good” region, while 3 and 4 has similar Average Silhouette Measure of 0.4, falling in the “fair” region. If we are solely determining the optimal K-value on Silhouette Measure, the K-value of 2 will be optimal.

After studying the mislabelling rates of each cluster for K-value 2, 3 and 4, K-value of 2 is to be able to determine a real pulsar with significantly low percentage error. The mislabelling rates are determined by comparing the minority attribute (Class 0 or Class 1) of a cluster, with

the real value given. These values are considered as mislabelled as the majority attribute of cluster determines the main attribute of that cluster.

After studying the results of the K-means cluster in (ii), it can be concluded that Cluster 2 of all three K-values depicts the cluster which represents Class 1 pulsars, which are real pulsars as it is the only Cluster with more Class 1 than Class 0.

	<b>K-value = 2</b>	<b>K-value = 3</b>	<b>K-value = 4</b>
<b>Mislabelling Rate of Class = 1</b>	22.27%	24.59%	38.5%

*Table 1. Mislabelling Rates of Class = 1 for Different K-Values*

Table 1 shows the mislabelling rates of “Class =1” for all three K-values. From the mislabelling rates, it is evident that when K-value is 2, Cluster 2 is able to determine Class 1 pulsars with the lowest mislabelling rate as compared to other two K-values. This means that when  $K = 2$ , the moving features of Cluster 2 will strongly suggest that the pulsar is a Class 1 pulsar. Thus, K-value of 2 is the optimal K-value to be used for the study.

With the highest Average Silhouette coefficient and the lowest mislabelling rate, the K-value of 2 is the most optimal value to be used for the analysis. In order to further confirm the optimal K-value to be used, more testing methods such as the Elbow Method and the Gap Statistics should be used.

### **Question 2c**

K-means is a technique that aims to explore and identify the most representative record of clusters (Tan et al., 2019). Even though clusters may contain different densities, K-means assumes that all clusters are divided into equal volumes. Moreover, it addresses overlapping clusters by dividing data space into equal-volume regions. This results in the partition of large clusters and the merging of small clusters. Hence, K-means is unable to detect small clusters as data points are based on Euclidean closeness to cluster centroid, where densities are not taken into consideration.

Therefore, one possible solution proposed by Yordan et. al to resolve K-means inability to detect small clusters will be to use the Maximum A-posteriori Dirichlet Process (MAP-DP)

(Raykov, Boukouvalas, Baig, & Little, 2016). While K-means uses a geometric framework governed by the Euclidean Distance, the MAP-DP uses a probabilistic framework known as the *Dirichlet process mixture models*, which assumes the number of clusters (K) in relation to varying data size (Raykov et al, 2016). The steps for MAP-DP are as such:

**Step 1: Calculate the probability of all cluster assignment variables**

$$p(z_1, \dots, z_N) = \frac{N_0^K}{N_0^{(N)}} \prod_{k=1}^K (N_k - 1)! , \text{ where } N_0^{(N)} = N_0(N_0 + 1) \times \dots \times (N_0 + N - 1)$$

**Step 2: Eliminate random variables**

$$\begin{aligned} (z_1, \dots, z_N) &\sim \text{CRP}(N_0, N) \\ x_i &\sim f(\theta_{z_i}) \end{aligned} , \text{ where } \theta \text{ are the hyper parameters of the predictive distribution } f(x|\theta)$$

This would be based on the assumption that the data indicates random  $K+$  number of predictive distributions describing each cluster ( $N_0$ ), as well as the rate of  $K+$  increasing with  $N$  that is controlled by ( $N_0$ ).

**Step 3: Update the Algorithm**

To derive the probability of X, the log of  $N_k^{-i}$  can be subtracted from the distance of the cluster centroids, which is determined by the Euclidean metric  $\frac{1}{2} ||\cdot||_2^2$ .

$$p(X, z|N_0) = p(z_1, \dots, z_N) \prod_{i=1}^N \prod_{k=1}^K f(x_i | \theta_k^{-i})^{\delta(z_i, k)} , \text{ where } \delta(x, y) = 1, \text{ if } x = y \text{ and } 0 \text{ otherwise.}$$

Additionally, the composite variance  $\sigma_k^{-i} + \hat{\sigma}^2$  as shown in the distance calculation implies the smaller the composite becomes, the less significant the number of data points in the cluster  $N_k^{-i}$  becomes to the assignment, and vice versa. This further exhibits the composite variance to be the controlling factor that equalises geometry and density.

**Step 4: Compute Existing & New Cluster respectively**

$$\begin{aligned} d_{i,k} &= -\ln f(x_i | \theta_k^{-i}) \\ d_{i,K+1} &= -\ln f(x_i | \theta_0) \end{aligned} , \text{ } x_i \text{ is obtained by computing } z_i = \arg \min_{k \in 1, \dots, K+1} [d_{i,k} - \ln N_k^{-i}].$$

Thus, using the MAP-DP method, clusters density are accounted for, which exhibits a consistent way of inferring missing values from DP mixtures and predicting unknown data. As the assumption of equal cluster density is removed, this model resolves the issue of K-means inability to detect small clusters.

# Bibliography

- Tan, S., Chan, S., Lee, P., & Li, L. (2019). *Association and Clustering Study Guide*. Singapore: Singapore University of Social Sciences: Educational Technology & Production.
- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics: Volume 20*, 53-65.
- Raykov, Y. P., Boukouvalas, A., Baig, F., & Little, M. A. (2016). What to Do When K-Means Clustering Fails: A Simple yet Principled Alternative Algorithm. *PLoS One*. v. 11(9).
- Li, J., Le, T. D., Liu, L., Liu, J., Zhou, J., & Sun, B. (2013). Mining Causal Association Rules. *Proceedings of the 2013 IEEE 13th International Conference on Data Mining Workshops* (pp. 114-123). ICDMW.