University of Roehampton London

| CMP020L012A | Data Analytics | 2022/23 |
|---|---|---|

# Week-11
# K-Mode Clustering
# Solution

## Answer the following Multiple-Choice Questions

1. What is K-modes clustering?
   a) A Supervised Learning method
   b) An Unsupervised Learning Method
   c) None of the above
   d) Both of the above

2. Which of the following metric is used for measuring K-Mode Clustering?
   a) Hamming Distance
   b) Euclidean distance
   c) Similarity Cosine
   d) None of the above

3. What initialization method is chosen in K-Mode Clustering?
   a) K-means Initialization
   b) Density based Initialization.
   c) Random Initialization
   d) None of the above

4. K-means clustering is good for continuous data whereas K-mode clustering is good for categorical data
   a) True
   b) False

5. What is the role of parameter K in the K-mode Clustering?
   a) determines the differences to be used
   b) determines the distance to be used
   c) determines the initialization method to be used
   d) determines the number of clusters to be formed

6. What is the advantage of K-Mode clustering over K-Means clustering?
    a) K-Mode Clustering can handle categorical data.
    b) K-Mode is faster than K-Means Clustering
    c) K-Mode Clustering is more accurate than K-Means Clustering
    d) All of the above

7. There can be multiple Centroids in one cluster.
    a) True
    b) False

University of Roehampton London

# Answer the following Question

Assuming you are given with a dataset, contains 16 records in it.

| Record | Gender | Age Range | Education Level |
|---|---|---|---|
| 1 | Female | 20-30 | Bachelor's Degree |
| 2 | Male | 30-40 | High School Diploma |
| 3 | Female | 20-30 | Master's Degree |
| 4 | Female | 20-30 | Bachelor's Degree |
| 5 | Male | 30-40 | High School Diploma |
| 6 | Male | 40-50 | Master's Degree |
| 7 | Female | 30-40 | High School Diploma |
| 8 | Male | 20-30 | Master's Degree |
| 9 | Female | 30-40 | Bachelor's Degree |
| 10 | Male | 20-30 | High School Diploma |
| 11 | Female | 40-50 | Master's Degree |
| 12 | Male | 30-40 | Bachelor's Degree |
| 13 | Female | 20-30 | High School Diploma |
| 14 | Male | 40-50 | Master's Degree |
| 15 | Female | 30-40 | Bachelor's Degree |
| 16 | Female | 20-30 | Master's Degree |

Manually group these data into three clusters using the k-modes clustering algorithm.

**Answer:**

K-modes is a clustering algorithm that works on categorical data. We have 16 records in our dataset, and we want to cluster them into 3 groups using k-modes clustering.

To perform k-modes clustering manually, we can use the following steps:

1. Choose k = 3 as the number of clusters.
2. Randomly assign each record to one of the three clusters.
3. Calculate the mode for each cluster for each categorical variable.
4. For each record, calculate the distance to each cluster by counting the number of variables for which it differs from the mode of that cluster.
5. Assign each record to the cluster for which it has the smallest distance.
6. Repeat steps 3-5 until the clusters no longer change.

Using this method, we can manually cluster the 16 records as follows:

**Initial clustering:**

Cluster 1: [1, 4, 9, 12]

Cluster 2: [3, 8, 11, 14, 16]

Cluster 3: [2, 5, 6, 7, 10, 13, 15]

**Mode for each cluster:**

Cluster 1: Female, 20-30, Bachelor's Degree

Cluster 2: Female, 30-40, Bachelor's Degree

Cluster 3: Male, 30-40, Master's Degree

**Distance from each record to each cluster:**

| Record | Cluster 1 | Cluster 2 | Cluster 3 |
|---|---|---|---|
| **1** | **0** | **1** | **3** |
| **2** | **3** | **2** | **0** |
| **3** | **3** | **0** | **3** |
| **4** | **0** | **1** | **3** |
| **5** | **3** | **2** | **2** |
| **6** | **3** | **3** | **1** |
| **7** | **3** | **0** | **3** |
| **8** | **3** | **0** | **2** |
| **9** | **0** | **1** | **3** |
| **10** | **3** | **2** | **2** |
| **11** | **3** | **0** | **1** |
| **12** | **0** | **1** | **3** |
| **13** | **3** | **3** | **1** |
| **14** | **3** | **0** | **3** |
| **15** | **3** | **3** | **1** |
| **16** | **3** | **0** | **3** |

**Assign each record to the cluster for which it has the smallest distance.**

Cluster 1: [1, 4, 9, 12]

Cluster 2: [3, 5,6,7,10, 11, 14, 16]

Cluster 3: [2, 6, 13, 15]

**Repeat steps 3-5 until the clusters no longer change.**