K-Modes Clustering

Prepared By: Kimia Aksir and Fakhreldin Saeed

Clustering

- **Clustering** is a method of grouping a set of objects in a way that ensures the objects that are similar to each other remain in the same group/cluster.
- Or, in other words, **Clustering** is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups.

When to use Clustering?

- When there is Unlabelled data, Clustering can be one of the best ways of analysing data to discover hidden patterns.
- Clustering is especially useful for exploring data you know nothing about.

When to use Clustering? Continued…

- Suppose that a bank wants to give credit card(s) to the customers. How do they decide the “offer”, they want to give their customers?
- Does that make sense to look at each customers details separately and make the decision?
- So what does the bank do?

Different types of Clustering

- **K-Modes Clustering**
- K-Means Clustering
- DBSCAN Clustering
- Gaussian Mixture Model
- Mean-Shift Clustering
- OPTICS algorithms

Clustering: K-Modes Clustering

- K-Modes is a clustering algorithm that is specifically designed for clustering categorical data, which is data that consists of non-numeric values, such as colours, types, or categories.
- It is an extension of the more well-known K-Means algorithm, which is designed for clustering numeric data.

Clustering: K-Modes Clustering Continued…

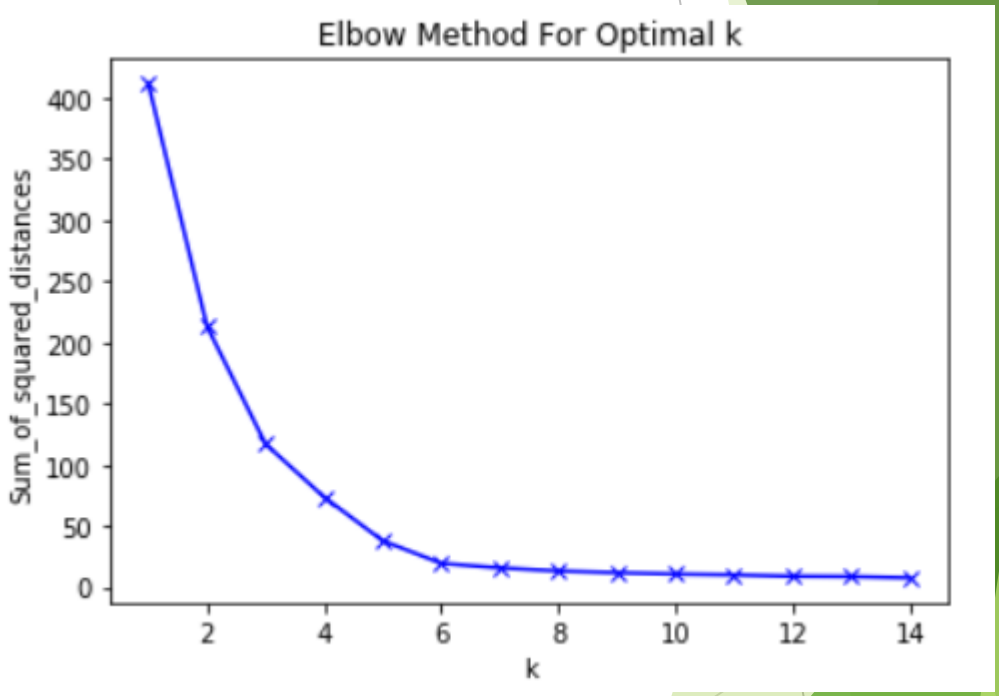
- The K-Modes algorithm works by identifying clusters of data points that have similar categorical values.
- It does this by assigning each data point to the cluster that has the most similar categorical values, based on a distance metric that measures the dissimilarity between different categories.
- It then updates the **centroids** of each cluster based on the data points that belong to it and repeats the process until the clusters stabilize.

How to choose “K” in K-Modes Clustering

- Choosing the appropriate value of "K" (the number of clusters) in K-Modes clustering can be a challenging task. Here are a few methods that can be used to determine the optimal value of K:
  - Elbow Method
  - Silhouette Method
  - Domain Knowledge
  - Trial and Error

Choosing “K” value in K-Modes Clustering: Elbow Method

- The elbow method involves plotting the within-cluster sum of squares (WCSS) against different values of K, and selecting the value of K where the decrease in WCSS starts to level off.
- This method is similar to the method used for K-Means clustering, and can be a good starting point for determining the optimal value of K.



Centroid

- A centroid is a representative object or data point that defines the centre of a cluster
- Unlike K-Means clustering, which uses the mean (or average) of the numeric data points to define the centroid, K-Modes uses the mode (or most common value) of the categorical data points.
- The centroid of a cluster in K-Modes clustering is defined by a vector of the most frequent values for each categorical feature in the cluster.

Distances with Centroid(s) in K-Modes Clustering

- The distance between a data point and a centroid is used to determine which cluster the data point belongs to.
- Since K-Modes clustering deals with **categorical** data, a distance metric that is appropriate for categorical data is used to measure the **dissimilarity** between a data point and a centroid.
- The most commonly used distance metric in K-Modes clustering is the simple matching distance.
- Other distance metrics that can be used in K-Modes clustering include the **Jaccard** distance and the **Hamming** distance.

How does the K-Modes algorithm work?

1. Pick **K** observations **randomly** and use them as leaders/clusters
2. Calculate the **dissimilarities** and assign each observation to its **closest** cluster
3. Define new modes for the clusters
4. Repeat 2–3 steps until there is no re-assignment required

An Example of K-Modes Clustering

Our Data

Individual	Q1	Q2	Q3	Q4	Q5
1	A	B	A	B	C
2	A	A	A	B	B
3	C	A	B	B	A
4	A	B	B	A	C
5	C	C	C	B	A
6	A	A	A	A	B
7	A	C	A	C	C
8	C	A	B	B	C
9	A	A	B	C	A
10	A	B	B	A	C

An Example of K-Modes Clustering Continued…

- Step 1: Pick K observations at **random** and use them as leaders.
- Our **K =3**

Cluster	Q1	Q2	Q3	Q4	Q5
1(1)	**A**	**B**	**A**	**B**	**C**
2(5)	**C**	**C**	**C**	**B**	**A**
3(10)	**A**	**B**	**B**	**A**	**C**

Individual	Q1	Q2	Q3	Q4	Q5
1	A	B	A	B	C
2	A	A	A	B	B
3	C	A	B	B	A
4	A	B	B	A	C
5	C	C	C	B	A
6	A	A	A	A	B
7	A	C	A	C	C
8	C	A	B	B	C
9	A	A	B	C	A
10	A	B	B	A	C

An Example of K-Modes Clustering Continued…

- Step 2 part one is to Calculate the **dissimilarities**.

Cluster	Q1	Q2	Q3	Q4	Q5
1(1)	A	B	A	B	C
2(5)	C	C	C	B	A
3(10)	A	B	B	A	C

Individual	Q1	Q2	Q3	Q4	Q5	C1	C2	C3
1	A	B	A	B	C	**0**	**4**	**2**
2	A	A	A	B	B	**2**	**4**	**4**
3	C	A	B	B	A	**4**	**2**	**4**
4	A	B	B	A	C	**2**	**5**	**0**
5	C	C	C	B	A	**4**	**0**	**5**
6	A	A	A	A	B	**3**	**5**	**4**
7	A	C	A	C	C	**2**	**4**	**3**
8	C	A	B	B	C	**3**	**3**	**3**
9	A	A	B	C	A	**4**	**4**	**3**
10	A	B	B	A	C	**2**	**5**	**0**

An Example of K-Modes Clustering Continued…

- Step 2 part two is to **assign** each observation to its **closest** cluster.

Cluster	Q1	Q2	Q3	Q4	Q5
1(1)	A	B	A	B	C
2(5)	C	C	C	B	A
3(10)	A	B	B	A	C

Individual	Q1	Q2	Q3	Q4	Q5	C1	C2	C3
1	A	B	A	B	C	0	4	2
2	A	A	A	B	B	2	4	4
3	C	A	B	B	A	4	2	4
4	A	B	B	A	C	2	5	0
5	C	C	C	B	A	4	0	5
6	A	A	A	A	B	3	5	4
7	A	C	A	C	C	2	4	3
8	C	A	B	B	C	3	3	3
9	A	A	B	C	A	4	4	3
10	A	B	B	A	C	2	5	0

Cluster	Individual
1	1,2,6,7,8
2	3,5
3	4,9,10

An Example of K-Modes Clustering Continued…

- Step 3 : Define new modes for the clusters (Centroid)

Individual	Q1	Q2	Q3	Q4	Q5	C1	C2	C3
1	A	B	A	B	C	0	4	2
2	A	A	A	B	B	2	4	4
3	C	A	B	B	A	4	2	4
4	A	B	B	A	C	2	5	0
5	C	C	C	B	A	4	0	5
6	A	A	A	A	B	3	5	4
7	A	C	A	C	C	2	4	3
8	C	A	B	B	C	3	3	3
9	A	A	B	C	A	4	4	3
10	A	B	B	A	C	2	5	0
Cluster 1	**A**	**A**	**A**	**B**	**C**			
Cluster 2	**C**	**A**	**B**	**B**	**A**			
Cluster 3	**A**	**B**	**B**	**A**	**C**			

An Example of K-Modes Clustering Continued…

- Step 4 : Repeat 2–3 steps until there is no re-assignment required

Cluster	Q1	Q2	Q3	Q4	Q5
Cluster 1	A	A	A	B	C
Cluster 2	C	A	B	B	A
Cluster 3	A	B	B	A	C

Individual	Q1	Q2	Q3	Q4	Q5	C1	C2	C3
1	A	B	A	B	C	1	4	3
2	A	A	A	B	B	1	3	4
3	C	A	B	B	A	3	0	4
4	A	B	B	A	C	3	4	0
5	C	C	C	B	A	4	2	5
6	A	A	A	A	B	2	4	3
7	A	C	A	C	C	2	5	3
8	C	A	B	B	C	2	1	3
9	A	A	B	C	A	3	2	3
10	A	B	B	A	C	3	4	0

Where to stop?

- In K-Modes clustering, the iterative clustering process is continued until the centroids no longer move or until a pre-defined stopping criterion is met.
- The most common stopping criterion in K-Modes clustering is a maximum number of iterations or a minimum decrease in the cost function.
- The cost function in K-Modes clustering is a measure of the dissimilarity between the data points and the centroids.

Advantages of K-Modes Clustering

- K-Modes has several advantages over other clustering algorithms for categorical data, including:
  - Simplicity
  - Scalability
  - ability to handle large datasets.
- It is also less sensitive to outliers than other clustering algorithms, which makes it useful in real-world applications where there may be noisy or incomplete data.

Disadvantages of K-Modes Clustering

1. Sensitivity to Initialization: Depending on the initial values chosen, the algorithm can converge to different local optima, which can affect the quality of the resulting clusters.
2. Scalability: Although K-Modes is designed to handle large datasets, it can still be computationally expensive for very large datasets. This can make it challenging to use in applications where there are very large amounts of categorical data.
3. Binary Distance Metric: The K-Modes algorithm uses a binary distance metric to measure the dissimilarity between different categorical values. This means that it treats all categories as equally dissimilar, which may not be appropriate in all cases
4. Difficulty with High-Dimensional Data: Like many clustering algorithms, K-Modes can struggle with high-dimensional data. This is because the number of possible combinations of categories grows exponentially with the number of dimensions, which can make it difficult to identify meaningful clusters.

Applications of Clustering

- Recommendation Engines
- Customer Segmentation
- Document Clustering
- Market Segmentation
- Image Segmentation
- Anomaly Detection

and many more….