# BEFORE THIS TUTORIAL

https://aihubprojects.com/k-means-clustering-from-scratch-python/

# K-MODES CLUSTERING: MATHEMATICAL IMPLEMENTATION

Before entering the tutorial on k-modes, let's revisit the k-means clustering algorithm. K- means is an unsupervised partitional clustering algorithm that is based on grouping data into k – numbers of clusters by determining centroid using the Euclidean or Manhattan method for distance calculation. It updates or calculates the new centroid by calculating the mean of data points in that cluster. We repeat the process for a number of iterations till successive iterations cluster data items into the same group.

Similar to K-means, k-modes clustering is also an unsupervised algorithm used for clustering categorical variables. This algorithm groups data points in clusters based on the number of matching categories between centroid data points and observation data points. Unmatched datapoints are counted as +1 and similar data points are counted as zero. Centroids are updated based on the maximum frequency of categorical value i.e mode.

# THEORY ?? WE PREFER EXAMPLE

Let's kick off K-Means Clustering Scratch with a simple example. Suppose we have categorical data which should be clustered in 2 groups. Let's suppose, on column 1 (C1) we have categories (M, N, O), on column 2 (C2) we have categories (P, Q, R), and on column 3 (C3) we have categories (W, X, Y, Z). Categorical distribution in the dataset is shown in the table below:

| S.N. | C1 | C2 | C3 | D1 | D2 |
|---|---|---|---|---|---|
| 1 | M | Q | X | | |
| 2 | M | Q | Z | | |
| 3 | N | P | Z | | |
| 4 | O | Q | W | | |
| 5 | O | R | Y | | |
| 6 | N | P | X | | |
| 7 | N | P | W | | |
| 8 | N | R | Y | | |
| 9 | M | O | X | | |
| 10 | N | P | Y | | |

Step 1: It is already defined that k = 2 for this problem

Step-2: Since k = 2, we are randomly selecting two centroid as S.N = 2 & 6 i.e (M,Q,Z) & (N,P,X).

Step 3: In k-means, we calculate the distance of each point to each centroid using the euclidean distance calculation method. But it's different in the case of k-modes. We compare each data points to centroid to each observation data point. If any elements aren't equal, we add one (+1) in each unmatching case. But, if the elements aren't equal, we add zero i.e we do nothing.

| S.N. | C1 | C2 | C3 | D1 |
|---|---|---|---|---|
| 1 | M | Q | ***X*** | 0+0+1 = 1 |
| Centroid 1 | M | Q | ***Z*** | |

Here, we are comparing the first data points with centroid 1. Since the third column isn't matching, the total distance point is 1.

| S.N. | C1 | C2 | C3 | D2 |
|---|---|---|---|---|
| 1 | M | Q | X | 1+1+0 = 2 |
| Centroid 2 | N | P | X | |

Here, we are comparing the first data points with centroid 2. Since the first & the second column aren't matching, the total distance point is 2.

Similarly, we have calculated and completed the data points table. Data points are assigned to the closest centroid with a minimum difference value.

| S.N. | C1 | C2 | C3 | D1 | D2 |
|---|---|---|---|---|---|
| 1 | **M** | **Q** | **X** | 1 | 2 |
| 2 | **M** | **Q** | **Z** | 0 | 3 |
| 3 | N | P | Z | 2 | 1 |
| 4 | **O** | **Q** | **W** | 2 | 3 |
| 5 | **O** | **R** | **Y** | 3 | 3 |
| 6 | N | P | X | 3 | 0 |
| 7 | N | P | W | 3 | 1 |
| 8 | N | R | Y | 3 | 2 |
| 9 | **M** | **O** | **X** | 2 | 2 |
| 10 | N | P | Y | 3 | 1 |

From the table, we can see that data points 1, 2, 4, & 5 are assigned to cluster 1st, and the rest are assigned to the second cluster.

Step 4:- Here comes the actual use of modes in the algorithm. Maximum categorical occurrence (mode) is assigned to the updated centroid elements of each cluster.

| S.N. | C1 | C2 | C3 |
|---|---|---|---|
| 1 | **M** | **Q** | **X** |

| | | | |
|---|---|---|---|
| 2 | **M** | **Q** | **Z** |
| 4 | **O** | **Q** | **W** |
| 5 | **O** | **R** | **Y** |
| 9 | **M** | **O** | **X** |
| Centroid 1 | **M** | **Q** | **X** |

| | | | |
|---|---|---|---|
| 3 | N | P | Z |
| 6 | N | P | X |
| 7 | N | P | W |
| 8 | N | R | Y |
| 10 | N | P | Y |
| Centroid 2 | N | P | Y |

We have obtained the new cluster centroid as (M, Q, X) and (N, P, Y). Now we repeat the same process from step 3 till successive iterations cluster data items into the same group.

FINAL CLUSTER

After repeating step 3 & 4 several times we have obtained the final cluster as:-

| S.N. | C1 | C2 | C3 | D1(M, Q, X) | D2(N, P, Y) |
|---|---|---|---|---|---|
| 1 | **M** | **Q** | **X** | | |
| 2 | **M** | **Q** | **Z** | | |
| 3 | N | P | Z | | |
| 4 | **O** | **Q** | **W** | | |
| 5 | **O** | **R** | **Y** | | |
| 6 | N | P | X | | |
| 7 | N | P | W | | |
| 8 | N | R | Y | | |

| | | | | | |
|---|---|---|---|---|---|
| 9 | **M** | **O** | **X** | | |
| 10 | N | P | Y | | |