K Modes Clustering

- K modes clustering is an unsupervised machine learning algorithm that is used to cluster categorical variables.
- In Kmeans, we use euclidean distance to cluster continuous data. Centroids are updated by means.
- In Kmodes, it uses the dissimilarities (total mismatch) between data points. Lesser the dissimilarities the more our data points are closer. It uses Mode

Example →

Person	hair colour	eye colour	skin colour
P1	blonde	amber	fair
P2	brunette	gray	brown
P3	red	green	brown
P4	black	hazel	brown
P5	brunette	amber	fair
P6	black	gray	brown
P7	red	green	fair
P8	black	hazel	fair

Step 1 →

Suppose K=3, pick 3 observations at random and use them as centroids.
Cluster 1 → P1 (blonde, amber, fair), Cluster 2 → P7 (red, green, fair), Cluster 3 → P8 (black, hazel, fair)

Step 2 → Calculate the dissimilarities (no of mismatch) and assign each observation to its closest cluster.

Eg for P1, Cluster 1 → 0 (dissimilarity), Cluster 2 → 2, Cluster 3 → 2.
After this calculate all dissimilarities and assign the observation to its closest cluster that has the least dissimilarity.

	Cluster 1(P1)	Cluster 2(P2)	Cluster 3(P3)	Cluster
P1	0	2	2	cluster 1
P2	3	3	3	Cluster 1
P3	3	1	3	Cluster 2
P4	3	3	1	Cluster 3
P5	1	2	2	Cluster 1
P6	3	3	2	cluster 3
P7	2	0	2	Cluster 2
P8	2	2	0	Cluster 3

Obs. P1, P2, P5 assigned to cluster 1. P3, P7 assigned to Cluster 2. P4, P6, P8 assign to cluster 3.

Note: If a point have all equal number, randomly give any cluster. Example - P2.

Step 3 - Define new Modes. Mode is most observed value.
Cluster 1 observation (P1, P2, P5) has brunette as most observed as hair Colour, amber as most observed eye colour, and fair as most observed skin.

Note: If we observe same occurance of value, take mode randomly. In our case of Cluster 2 (P3, P7) have one occurance brown and fair in skin colour, randomly give any value, we chose brown as Mode.

New Centroids →

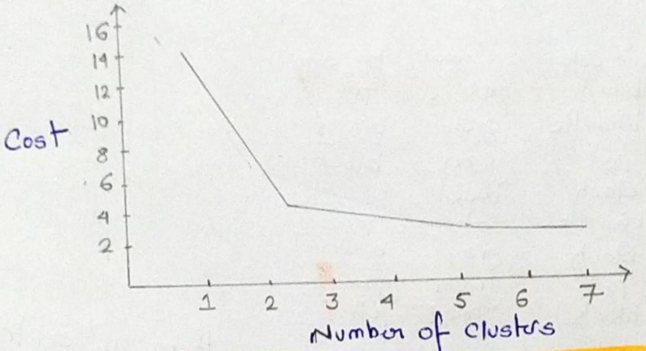
	hair colour	eye colour	skin colour
Cluster 1	brunette	amber	fair
Cluster 2	red	green	fair
Cluster 3	black	hazel	brown

Repeat step 2-3

After obtaining the new leaders, again calculate the dissimilarity between the observations and the newly obtained leaders.

We will see again the reassignment of clusters, will do until there is no change in the assignment of observations

How to choose number of clusters?



ELBOW METHOD FOR OPTIMAL K

Cost is the sum of all the dissimilarities between the clusters.