

# Machine Learning

## K-Modes Clustering

---

DR. BHARGAVI R

SCOPE

VIT CHENNAI

# K-Modes Clustering

---

- Unsupervised learning.
- Used for clustering categorical features/variables.
- It uses the total number of mismatches between the data points for measuring dissimilarity.

*Bhargavi R*

# Algorithm

---

1. **Initialize:** Randomly choose  $k$  observations as cluster centroids.
2. **Iterate:** Until the cluster assignments stop changing:
  - a. Calculate dissimilarities of each observation from all the clusters and assign each observation to the cluster whose centroid is closest
  - b. For each of the  $K$  clusters, compute the cluster centroid as the mode of the observations of the cluster. i.e  $k^{\text{th}}$  cluster centroid is the vector of the  $m$  feature modes for the observations in the  $k^{\text{th}}$  cluster.

# Example

---

- Styling suggestions based on personalities

Model	Hair Color	Eye Color	Skin Color
1	Blonde	Amber	Fair
2	Brunette	Gray	Brown
3	Red	Green	Brown
4	Black	Hazel	Brown
5	Brunette	Amber	Fair
6	Black	Gray	Brown
7	Red	Green	Fair
8	Black	Hazel	Fair

# Example

---

- Initialize Instances 1, 7, 8 as cluster centers

Model	Hair Color	Eye Color	Skin Color	Distance from C1	Distance from C2	Distance from C3	Cluster Assignment
1	Blonde	Amber	Fair	0	2	2	C1
2	Brunette	Gray	Brown	3	3	3	C1
3	Red	Green	Brown	3	1	3	C2
4	Black	Hazel	Brown	3	3	1	C3
5	Brunette	Amber	Fair	1	2	2	C1
6	Black	Gray	Brown	3	3	2	C3
7	Red	Green	Fair	2	0	2	C2
8	Black	Hazel	Fair	2	2	0	C3

# Example - Computing the distance

---

- Distance between (Blonde, Amber, Fair) and (Brunette, Gray, Brown) = 3 (since all the three values in the two vectors differ with respect to the corresponding value)
- Distance between (Blonde, Amber, Fair) and (Red, Green, Fair) = 2 (since two of the three values in the two vectors differ with respect to the corresponding value)

Bhargava R

# Example (cont...)

---

## Computing Centroids

- Centroid of C1 with observations 1, 2, and 5:
  - Mode of each variable for the observations 1, 2, 5 - (Brunette, Amber, Fair)
- Centroid of C2 with observations 3 and 7:
  - Mode of each variable for the observations 3, 7 - (Red, Green, Fair)
- Centroid of C3 with observations 4, 6 and 8:
  - Mode of each variable for the observations 4, 6, 8 - (Black, Hazel, Brown)
- Repeat the distance computations and reassignment of clusters with new centroids till there is no change in cluster assignments or max iterations.