Jessica Braganza
F171121S1
BE Comp 2

PAGE NO :
DATE :

ML Practical 4

Title : K-Means Clustering

1) Why does K-Means Clustering algorithm use only Euclidean distance metric?

Ans. 1. K-Means uses a vector quantization method used as a clustering method that does not explicitly use pairwise distance data points at all.

2. It repeatedly assigns points to the closest centroid by using Euclidean distance from data points to a centroid.

3. Euclidean distance is used because the sum of squared deviations from centroid is equal to the sum of pairwise squared Euclidean distances divided by the number of points.

4. The term "centroid" is itself from Euclidean geometry.

2) Explain K-Means algorithm with example.

Ans. Given $P_1 = [0.1, 0.6]$, $P_2 = [0.15, 0.71]$, $P_3 [0.08, 0.9]$, $P_4 [0.16, 0.85]$, $P_5 = [0.2, 0.3]$, $P_6 = [0.25, 0.5]$, $P_7 [0.24, 0.1]$, $P_8 [0.3, 0.2]$

M_1 (centroid) $\Rightarrow$ $[0.1, 0.6]$ C_1

M_2 (centroid) $\Rightarrow$ $[0.3, 0.2]$ C_2

} Clusters

x	y	$\sqrt{(x-0.1)^2 + (y-0.6)^2}$	$\sqrt{(x-0.3)^2 + (y-0.2)^2}$	Cluster
0.1	0.6	0	0.49	C_1
0.15	0.71	0.12	0.53	C_1
0.08	0.9	0.30	0.73	C_1
0.16	0.85	0.25	0.66	C_1
0.2	0.3	0.14	0.14	C_2
0.25	0.25	0.28	0.30	C_1
0.24	0.1	0.51	0.11	C_2
0.3	0.2	0.5	0	C_2

$c_1 = [P_1, P_2, P_3, P_4, P_6] = \{(0.1, 0.6), (0.08, 0.9), (0.16, 0.85), (0.25, 0.5)\}$

$c_2 = [P_5, P_7, P_8] = \{(0.2, 0.3), (0.24, 0.1), (0.3, 0.2)\}$

$$m_1 = \left(\frac{0.1 + 0.08 + 0.16 + 0.25}{4}, \frac{0.6 + 0.9 + 0.85 + 0.5}{4} \right)$$

$$= (0.15, 0.71)$$

$$m_2 = \left(\frac{0.2 + 0.24 + 0.3}{3}, \frac{0.3 + 0.1 + 0.2}{3} \right)$$

$$= (0.25, 0.2)$$

x	y	$\sqrt{(x-0.15)^2 + (y-0.71)^2}$	$\sqrt{(x-0.25)^2 + (y-0.2)^2}$	Cluster
0.1	0.6	0.12	0.43	c_1
0.15	0.71	0	0.52	c_1
0.08	0.9	0.2	0.72	c_1
0.16	0.85	0.14	0.66	c_1
0.2	0.3	0.41	0.11	c_2
0.25	0.5	0.23	0.3	c_1
0.24	0.1	0.62	0.10	c_2
0.3	0.2	0.53	0.05	c_2

$\therefore c_1 = \{P_1, P_2, P_3, P_4, P_6\}$ and $c_2 = \{P_5, P_7, P_8\}$

3) Write down any 2 applications of k-means in detail.

Ans. 1. Vector Quantization:

i) k-means originated from signal processing and still finds use in the domain. For eg. in computer graphics, color quantization is task of reducing the color palette of an image to a fixed no. of color k.

ii) A use case for this approach is image segmentation.

2. Feature learning:

i) k-means clustering has been used as a feature step for either supervised or un-supervised learning.

ii) Basic approach is to train a cluster first using training data.