Haonan Hu

4/6/2021

2863545

1. Assume you are working with a data set of samples each with 3 features: a, b, and c, represented as (a, b, c) in feature space. If the centroid of a k- means cluster is located at (1,4,8), what is the distance from the sample (3,8,2) to the centroid?

   D = sqrt((3-1)^2), sqrt((8-4)^2), sqrt((2-8)^2) ~=7.48

2. Assume you are trying to cluster six samples with k=2 and after the first iteration, the first cluster contains the samples with the features: (2,3,4), (1,6,3), and (2,1,5) and the second cluster contains the samples: (6,3,5), (9,8,7), and (7,2,6). What would the k-means clustering algorithm generate as centroids for the next iteration?

	Cluster1		
	x	y	z
	2	3	4
	1	6	3
	2	1	5
Mean = new centroids	5/3	10/3	4

	Cluster2		
	x	y	z
	6	3	5
	9	8	7
	7	2	6
Mean = new centroids	22/3	13/3	6

3. Using the same samples, clusters, and centroids as in Problem 2, what would the Reconstruction Error be for this set of clusters?

	Cluster1				Cluster2		
	x	y	z		x	y	z
x^t	2	3	4		6	3	5
	1	6	3		9	8	7
	2	1	5		7	2	6
m_i	5/3	10/3	4		22/3	13/3	6
$(x^t - m_i)^2$	1/9	1/9	0		16/9	16/9	1
	4/9	64/9	1		25/9	121/9	1
	1/9	49/9	1		1/9	49/9	0
$\sum_t (x^t - m_i)^2$	46/3				82/3		
Reconstruction Error	46/3 + 82/3 ~=42.67						