MIS 637 A Final Exam
Data Analytics and Machine Learning

May 03, 2023
12:00 Noon to 2:30 PM
School of Business
Stevens Institute of Technology

Professor M. Daneshmand

Student Name:……Poorvi Rajendra Raut…………………………

1. Describe the differences between clustering and classifications.

2. We have the following two-dimensional data points:

   a (4,2), b (2,3), c (4,3), d (3,5), e (1,2), f (2,1), g (1,1), h (3,2).

   Identify the cluster by applying the k-means algorithm, with k=2. Show that the ratio of the between-cluster variation to the within-cluster variation increases with each pass of the algorithm. Please show your work and how the algorithm works: **passes, steps, formulas, calculations, tables, and final clusters.**

Solution 1:
Clustering and classification are two fundamental techniques of Data Analytics Machine learning, but both differ in objectives and methodology.

Differences between clustering and classification is as follows:

Classification	Clustering
1. Classification is a supervised learning mechanism. It involves assigning class or label to new data point based on features or characteristics	1.Clustering is an unsupervised learning mechanism. It involves grouping similar data points together based on similarity of features and characteristics.
2. As classification has labels so there is need of training and testing data for verifying the model created.	2.There is no need of training and testing data in clustering as the model learns on its own without prior knowledge.
3. Classification is more complex as there are many levels of	3.In clustering only grouping of similar data is done so it is much

classification for example in decision tree classification, we require to make decision at every level and calculate the metrics.	less complex than classification.
4. The classification algorithms use various techniques to learn the decision boundaries between classes such as decision trees, logistic regression, support vector machine (SVM) and neural networks.	4.Clustering algorithms require forming of clusters based on similarity of data points which is usually determined by distance metric like Euclidian distance. Clustering algorithms include K-means, Hierarchical clustering and density-based clustering.

Solution 2:
Given data points:
a (4,2), b (2,3), c (4,3), d (3,5), e (1,2), f (2,1), g (1,1), h (3,2).

Step 1: k = 2 specifies number of clusters to partition

Step 2: Randomly assign k=2 cluster centers for example m1= (1,1) and m2= (4,3)

First Iteration:

Step 3: For each record find nearest cluster center by calculating the Euclidean distance between the points and cluster centers and determine the closest values to m1 and m2 and divide in clusters of k=2

Euclidean distance between points x (x1, x2) and y (y1, y2)= sqrt((x1-y1)^2 + (x2-y2)^2)

Point	a	b	c	d	e	f	g	h
Distance from m1	3.16	2.23	3.60	4.47	1	1	0	2.23
Distance from m2	1	2	0	2.236	3.16	2.82	3.60	1.41
Cluster Membership	C2	C2	C2	C2	C1	C1	C1	C2

cluster m1 contains: {e,f,g} and cluster m2 contains {a,b,c,d,h }

cluster membership is assigned and now calculate SSE

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} d(p, m_i)^2$$

1^2+1^2+0+1^2+2^2+0+(2.236)^2+(1.41)^2 = 13.98

Recall clusters constructed where between-cluster variation (BCV) large, as compared to within-cluster variation (WCV)

BCV/WCV
= d(m1,m2)/SSE
= 3.6/13.98 = 0.25

Ratio BCV/WCV expected to increase for successive iterations.

Step 4: For *k* clusters, find cluster centroid, update location. Calculate the new cluster centers as the mean of the data points assigned to each cluster.

Cluster 1: Mean = ((1+2+1)/3, (2+1+1)/3))= (1.33,1.33)
Cluster 2: Mean = ((4+2+4+3+3)/5, (2+3+3+5+2)/5)) = (3.2,3)

Step 5: Repeats Steps 3 – 4 until convergence or termination

Second Iteration: Repeat steps 3 and 4
Again m1= (1.33,1.33) m2= (3.2,3). calculating the Euclidean distance between the points and cluster centers and determine the closest to new values to m1 and m2 and divide in clusters of k=2

Point	a	b	c	d	e	f	g	h
Distance from m1	2.75	1.79	3.14	4.03	0.74	0.74	0.466	1.799
Distance from m2	1.28	1.2	0.8	2.00	2.41	2.33	2.97	1.019
Cluster Membership	C2	C2	C2	C2	C1	C1	C1	C2

cluster m1 contains: {e,f,g} and cluster m2 contains {a,b,c,d,h }

cluster membership is assigned and now calculate SSE

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} d(p, m_i)^2$$

(0.74)^2+(0.74)^2+(0.466)^2+(1.28)^2+(1.2)^2+(0.8)^2+2^2+(1.019)^2
=10.06
Recall clusters constructed where between-cluster variation (BCV) large, as compared to within-cluster variation (WCV)

BCV/WCV
= d (m1, m2)/SSE
0.3
Ratio BCV/WCV increases as compared to previous iteration 0.25 to 0.3.

Step 4: For *k* clusters, find cluster centroid, update location. Calculate the new cluster centers as the mean of the data points assigned to each cluster.

Cluster 1: Mean = ((1+2+1)/3, (2+1+1)/3))= (1.33,1.33)
Cluster 2: Mean = ((4+2+4+3+3)/5, (2+3+3+5+2)/5)) = (3.2,3)

Step 5: Repeat steps 3 and 4 until convergence or termination. Since the mean values of clusters /centroids remain unchanged, the algorithm terminates.