MIS 637 B Final Exams
Data Analytics & Machine Learning

December 12, 2019
School of Business
Stevens Institute of Technology

Professor: M. Daneshmand

Student Name: SRUJAN JADHAV
CWID: 10445782

1. Describe the differences between clustering and classifications

ANSWER:

SR. No.	Clustering	Classification
1	It finds natural grouping of instances on given unlabeled data.	It learns a method to predict the instance class from pre labeled classified instances.
2	An unsupervised learning technique.	A supervised learning technique.
3	Works solely with unlabeled data.	Involves both labeled and unlabeled data.
4	Has a single phase. (Grouping)	Involves two phases. (Training and Testing)
5	Clustering does not poignantly employ training sets, which are groups of instances employed to generate the groupings.	Classification imperatively needs training sets to identify similar features.
6	Clustering groups objects with the aim to narrow down relations as well as learn novel information from hidden patterns.	Classification seeks to determine which explicit group a certain object belongs to.
7	Example of algorithms: K-Means, DBSCAN.	Example of algorithms: Decision trees, Random forest.

2. We have the following two-dimensional data points:
   a (3,2), b (3,3), c (4,3), d (5,3), e (1,2), f (4,2), g (1,1), h (2,1).
   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, plots, and final clusters**.

ANSWER:

FIRST PASS:

Given data: k = 2

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

Step 1: So, m1 = (4, 3) & m2 = (1, 1)
Calculate the distance of each point from m1 & m2 and divide them in to clusters C1 and C2.

Step 2:
Euclidean distance formula: $d_{Euclidean}$ (x, y) = √(x2-x1)^2 + (y2-y1)^2 where two points are (x1, y1) (x2, y2)

Points	Distance from Centre1 (4, 3)	Distance from Centre2 (1, 1)	Cluster
a (3,2)	1.414	2.236	C1
b (3,3)	1	2.828	C1
c (4,3)	0	3.605	C1
d (5,3)	1	4.472	C1
e (1,2)	3.162	1	C2
f (4,2)	1	3.162	C1
g (1,1)	3.605	0	C2
h (2,1)	2.828	1	C2

Points in cluster C1 are (a, b, c, d, f) and that in C2 are (e, g, h)

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

SSE = $1.414^2 + 1^2 + 0^2 + 1^2 + 1^2 + 1^2 + 0^2 + 1^2$

Hence, SSE= 6.999

d(m1,m2) = 3.605

$\frac{BCV}{WCV} = \frac{d(m_1, m_2)}{SSE}$ = 3.605/6.999 = 0.515

New Centroid: C1 = $\left(\frac{3+3+4+5+4}{5}, \frac{2+3+3+3+2}{5}\right)$ = (3.8, 2.6)

New Centroid: C2 = $\left(\frac{1+1+2}{3}, \frac{2+1+1}{3}\right)$ = (1.33, 1.33)

SECOND PASS:

Points	Distance from Centre1 (3.8, 2.6)	Distance from Centre2 (1.33, 1.33)	Cluster
a (3,2)	1	1.799	C1
b (3,3)	0.894	2.361	C1
c (4,3)	0.447	3.149	C1
d (5,3)	1.264	4.032	C1
e (1,2)	2.863	0.746	C2
f (4,2)	0.632	2.752	C1
g (1,1)	3.224	0.466	C2
h (2,1)	2.408	0.746	C2

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

SSE = $1^2 + 0.894^2 + 0.447^2 + 1.264^2 + 0.746^2 + 0.632^2 + 0.466^2 + 0.746^2$

SSE = 5.326

d(m1,m2) = 2.777

$\frac{BCV}{WCV} = \frac{d(m_1, m_2)}{SSE}$ = 2.777/5.326 = 0.521

New Centroid: C1 = $\left(\frac{3+3+4+5+4}{5}, \frac{2+3+3+3+2}{5}\right)$ = (3.8, 2.6)

New Centroid: C2 = $\left(\frac{1+1+2}{3}, \frac{2+1+1}{3}\right)$ = (1.33, 1.33)

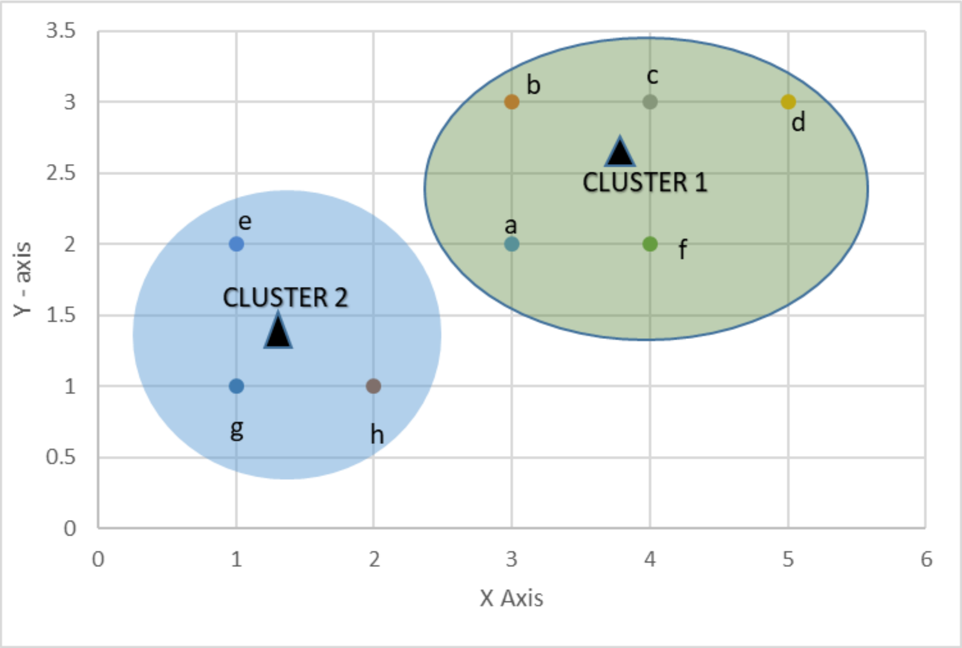
When the value of the centroid of two clusters doesn’t change then K-means algorithm terminates.

Final Clusters:

Cluster 1 = {a, b, c, d, f}

Cluster 2 = {e, g, h}

GRAPH PLOT:


3.5
3
2.5
2
1.5
1
0.5
0
Y - axis
0
1
2
3
4
5
6
X Axis
b
c
d
CLUSTER 1
a
f
e
CLUSTER 2
g
h