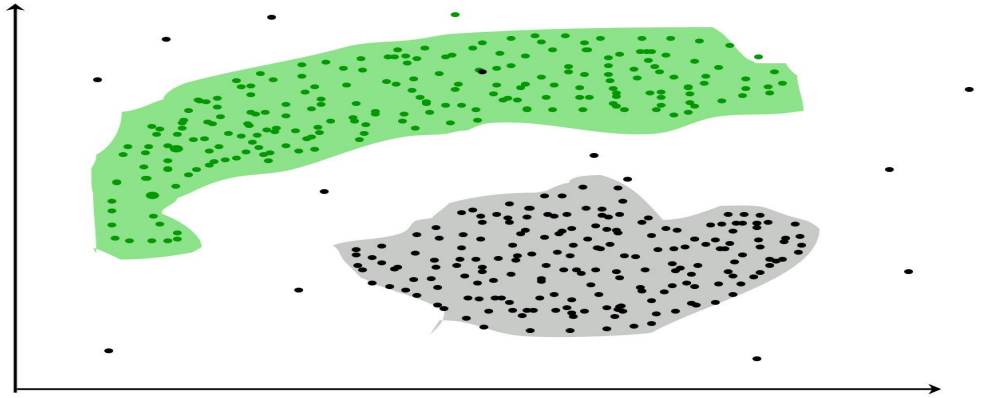
Introduction to Clustering

Introduction

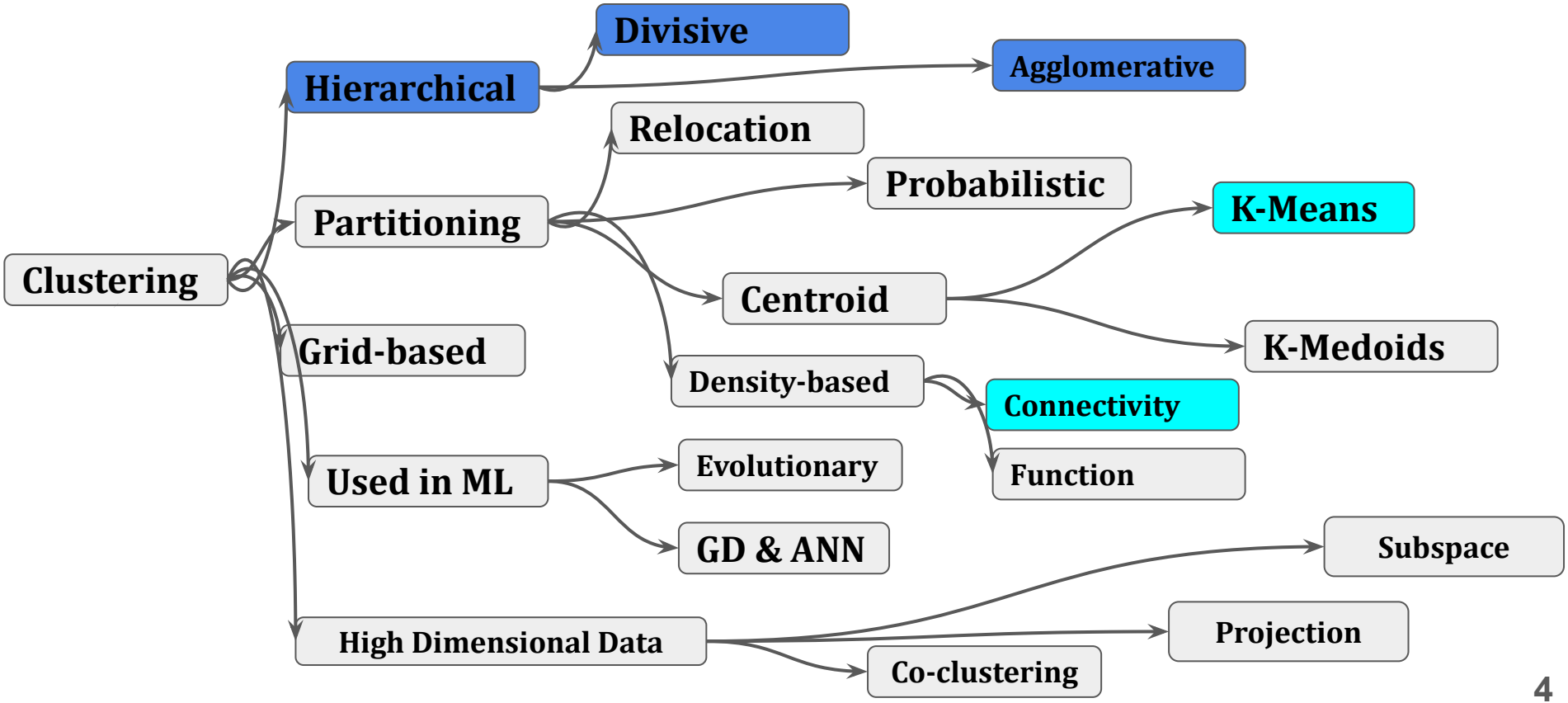
- Unsupervised Learning
- Process of grouping similar elements
- Intra-cluster similarity should be high
- Inter-cluster similarity should be low
- Grouping could be based on logical relationships or consumer preferences



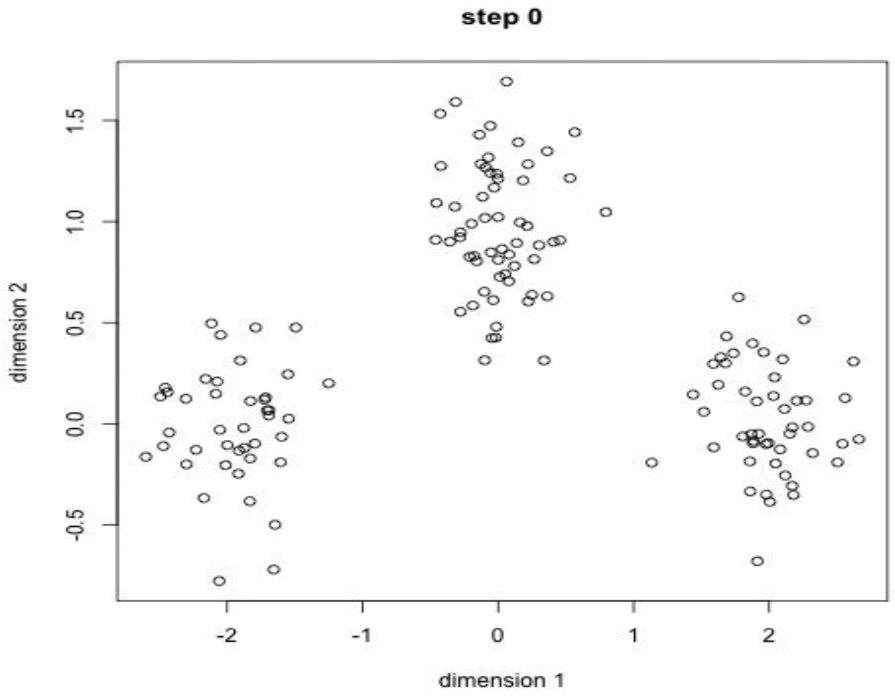
Purpose of Clustering

- High Dimensionality
- Ability to deal with noisy data
- Interpretability
- Scalability
- Ability to deal with different types of attributes
- Discovery of clusters with arbitrary shape

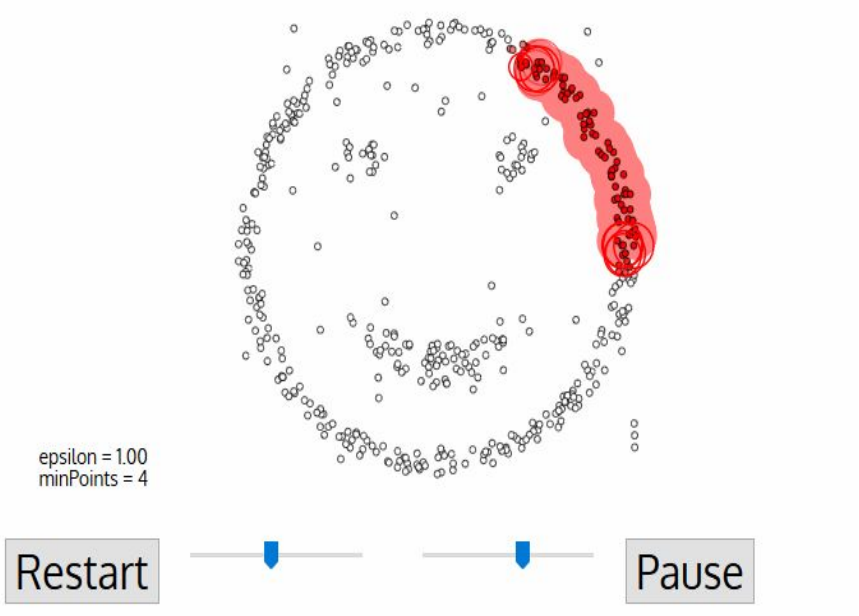
Types of Clustering



Example Algorithms



Centroid Based Algorithm



Density Based Algorithm

Centroid Based Clustering (K-Means)

- Select k number of classes and initialize their respective center points
- The center points are vectors of the same length as each data point vector
- Each data point is classified to be in the group whose center is closest to it
- Based on these classified points, recompute the group center
- Repeat these steps until the expected result
- Consider the following 2D points and find 3 cluster centroids for those

X	**1**	**-3**	**0**	**-7**	**0**	**6**	**2**	**-1**	**0**	**10**
Y	**2**	**6**	**3**	**-4**	**7**	**9**	**3**	**-1**	**-100**	**10**

Centroid Based Clustering (K-Means)

X	1	-3	0	-7	0	6	2	-1	0	10
Y	2	6	3	-4	7	9	3	-1	-100	10
Cluster										

Distance (C1) <1,2>										

Distance (C2) <-3, 6>										

Distance (C3) <0, 3>										

C1	(1,2)
C2	(-3,6)
C3	(0,3)

For simplicity, we will use squared distances

$$(e\_dist(x_1\ y_1)) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Centroid Based Clustering (K-Means)

X	1	-3	0	-7	0	6	2	-1	0	10
Y	2	6	3	-4	7	9	3	-1	-100	10
Cluster										

Distance (C1) <1,2>	0	5.66	1.41	10	5	8.6	1.41	3.6	102.005	12.04
Distance (C2) <-3, 6>	5.66	0	4.24	10.77	3.16	3.46	5.83	7.28	106.04	13.60
Distance (C3) <0, 3>	1.41	4.24	0	9.9	4	8.45	2	4.12	103	12.21

C1	(1,2)
C2	(-3,6)
C3	(0,3)

For simplicity, we will use squared distances

$$(e\_dist(x_1\ y_1)) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Centroid Based Clustering (K-Means)

X	1	-3	0	-7	0	6	2	-1	0	10
Y	2	6	3	-4	7	9	3	-1	-100	10
Cluster										

Distance (C1) <1,2>	0 = C1=1, 2	5.66	1.41	10	5	8.6	1.41=C1 =2, 3	3.6=C1= -1, -1	102.005 =C1=0, -100	12.04=C 1=10,10
Distance (C2) <-3, 6>	5.66	0 = C2=-3 , 6	4.24	10.77	3.16=C2 =0, 7	3.46=C2 =6,9	5.83	7.28	106.04	13.60
Distance (C3) <0, 3>	1.41	4.24	0=C3= 0, 3	9.9=C3 = -7, -4	4	8.45	2	4.12	103	12.21

C1	(1,2)
C2	(-3,6)
C3	(0,3)

For simplicity, we will use squared distances

$$(e\_dist(x_1\ y_1)) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Centroid Based Clustering (K-Means)

X	1	-3	0	-7	0	6	2	-1	0	10
Y	2	6	3	-4	7	9	3	-1	-100	10
Cluster										

Distance (C1) <1,2>	0 = C1=1, 2	5.66	1.41	10	5	8.6	1.41=C1 =2, 3	3.6=C1= -1, -1	102.005 =C1=0, -100	12.04=C 1=10,10
Distance (C2) <-3, 6>	5.66	0 = C2=-3 , 6	4.24	10.77	3.16=C2 =0, 7	3.46=C2 =6,9	5.83	7.28	106.04	13.60
Distance (C3) <0, 3>	1.41	4.24	0=C3= 0, 3	9.9=C3 = -7, -4	4	8.45	2	4.12	103	12.21

C1	(1,2), (-1, -1), (0, -100), (10, 10)(2, 3) = <(1+(-1)+0+10+2)/5, (2+(-1)+(-100)+10+3)/5> = <2.4, -17.2>
C2	(-3,6), (0, 7), (6, 9) = <(-3+0+6)/3, (6+7+9)/3> = <1, 7.3>
C3	(0,3), (-7, -4) = <(0+(-7))/2, (3+(-4))/2> = <-3.5, -0.5>

Centroid Based Clustering (K-Means)

X	1	-3	0	-7	0	6	2	-1	0	10
Y	2	6	3	-4	7	9	3	-1	-100	10
Cluster										

Distance (C1) <2.4, -17.2>	24.296347	28.78	25.373	20.797	29.3566	31.45	25.255	21.54	**77.540**	33.11
Distance (C2) <1, 7.3>	5.3	**4.206**	4.41	13.845	**1.044**	**5.28**	4.415	8.54	107.304	**9.396**
Distance (C3) <-3.5, -0.5>	**2.978**	5.5	**2.87**	**7.087**	6.55	11.32	**4.34**	**1.78**	100.67	14.94

C1	(0, -100)
C2	(-3,6), (0, 7), (6, 9), (10, 10)
C3	(1, 2), (0,3), (-7, -4), (2, 3), (-1, -1)

Centroid Based Clustering (K-Means)

X	1	-3	0	-7	0	6	2	-1	0	10
Y	2	6	3	-4	7	9	3	-1	-100	10
Cluster	C3	C2	C2	C3	C2	C2	C2	C3	C1	C2

Distance (C1) <2.4, -17.2>	19.25	23.82	20.34	16.2	24.32	26.5	20.2	16.55	82.84	28.24
Distance (C2) <1, 7.3>	5.3	4.2	4.41	13.84	1.044	5.3	4.42	8.54	107.3	9.4
Distance (C3) <-3.5, -0.5>	5.15	6.52	4.95	4.95	8.3	13.44	6.52	2.54	99.6	17.1

C1	(0, -100)= (0, -100)
C2	(-3,6), (0,3), (0, 7), (6, 9) , (2, 3), (10, 10) = (-3+0+0+6+2+10/6, 6+3+7+9+3+10/6) = (2.5, 6.33)
C3	(1, 2), (-7, -4), (-1, -1) = (-3.5, -1.5)

Centroid Based Clustering (K-Means)

X	1	-3	0	-7	0	6	2	-1	0	10
Y	2	6	3	-4	7	9	3	-1	-100	10
Cluster	C2	C2	C2	C3	C2	C2	C2	C3	C1	C2
Distance (C1) <0, -100>	102	106.04	103	96.25	107	109.17	103.02	99.01	0	110.45
Distance (C2) <2.5, 6.33>	4.58	5.5	4.16	14.03	2.58	4.4	3.36	8.12	106.35	8.34
Distance (C3) <-3.5, -1.5>	5.7	7.52	5.7	4.3	9.19	14.16	7.11	2.55	98.56	17.73

C1	(0, -100)
C2	(1, 2),(-3,6), (0,3), (0, 7), (6, 9) , (2, 3), (10, 10)
C3	(-7, -4), (-1, -1)

Centroid Based Clustering (K-Means)

X	1	-3	0	-7	0	6	2	-1	0	10
Y	2	6	3	-4	7	9	3	-1	-100	10
Cluster	C2	C2	C2	C3	C2	C2	C2	C3	C1	C2

Distance (C1) <0, -100>	102	106.04	103	96.25	107	109.17	103.02	99.01	0	110.45
Distance (C2) <2.5, 6.33>	4.58	5.5	4.16	14.03	2.58	4.4	3.36	8.12	106.35	8.34
Distance (C3) <-3.5, -1.5>	5.7	7.52	5.7	4.3	9.19	14.16	7.11	2.55	98.56	17.73

C1	(0, -100) = (0, -100)
C2	(1, 2),(-3,6), (0,3), (0, 7), (6, 9) , (2, 3), (10, 10) = ((1-3+0+0+6+2+10)/7, (2+6+3+7+9+3+10)/7)=(2.95, 5.72)
C3	(-7, -4), (-1, -1) = (-4, -2.5)

Centroid Based Clustering (K-Means)

X	1	-3	0	-7	0	6	2	-1	0	10
Y	2	6	3	-4	7	9	3	-1	-100	10
Cluster	C2	C2	C2	C3	C2	C2	C2	C3	C1	C2

Distance (C1) <0, -100>	102	106.04	103	96.25	107	109.17	103.02	99.01	0	110.45
Distance (C2) <2.95, 5.72>	4.2	5.96	4.01	13.91	3.22	4.48	2.88	7.8	105.76	8.25
Distance (C3) <-4, -2.5>	6.73	8.56	6.8	3.36	10.31	15.24	8.14	3.35	97.58	18.77

C1	(0, -100)
C2	(1, 2),(-3,6), (0,3), (0, 7), (6, 9) , (2, 3), (10, 10)
C3	(-7, -4), (-1, -1)

Thank You