Name : Pawan Suresh Bathe Roll No.: CS18M519

CS6745: Mining Massive DataSets

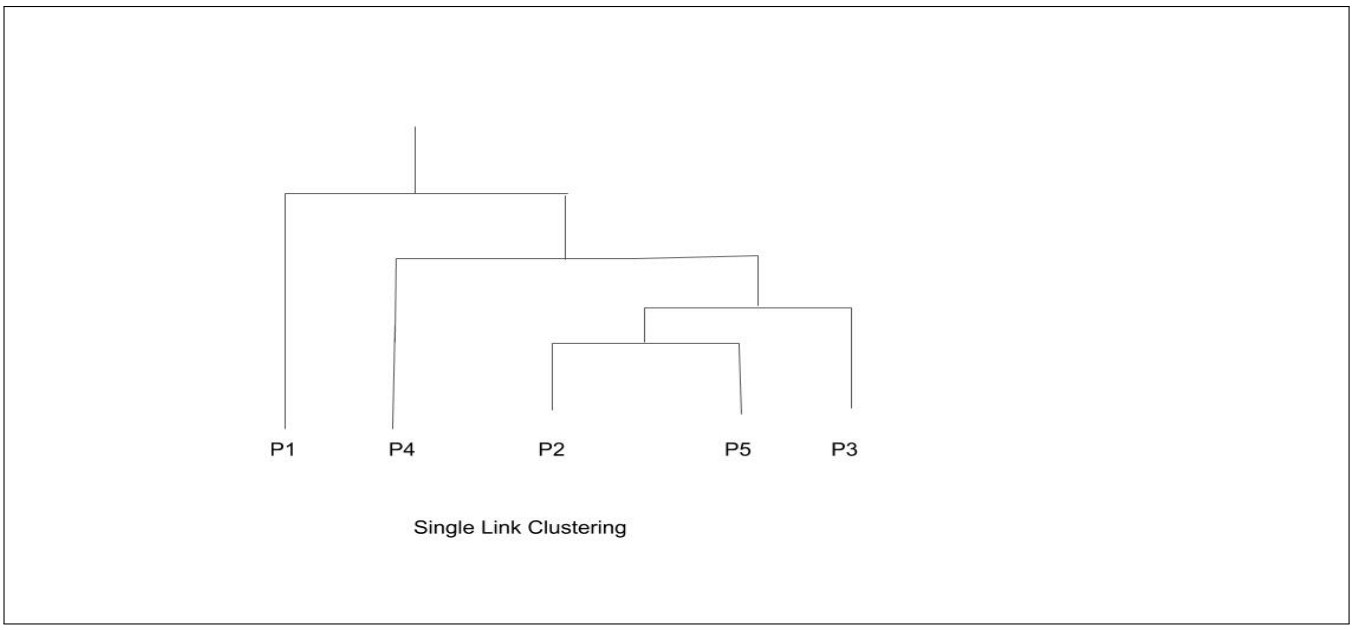
Tutorial 7

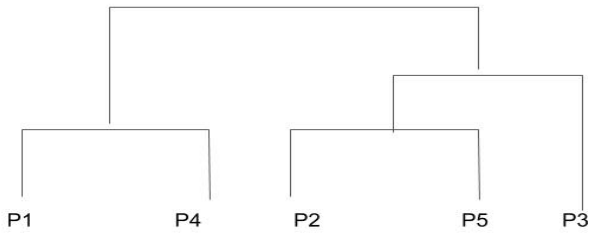
October 21, 2019

- Write your name and roll number in the space provided
- Be neat, and use the space judiciously.
- **Rough sheets won't be evaluated.**

1. (4 marks) For the similarity matrix given below show the hierarchy of clusters created by single link and complete link clustering algorithms. The final results should be shown as dendrograms.

	P1	P2	P3	P4	P5
P1	1	0.1	0.41	0.55	0.35
P2	0.1	1	0.64	0.47	0.98
P3	0.41	0.64	1	0.44	0.85
P4	0.55	0.47	0.44	1	0.76
P5	0.35	0.98	0.85	0.76	1





Complete Link Clustering

2. (2 marks) When can k-means fail? Suggest some alternative.

K-means fails when:

1. Clusters are not spherical or it also fails when clusters are from circles with same centre but with different radius with different circles with same centre representing different clusters.

2. K-means also fails when we unevenly sized cluster data.

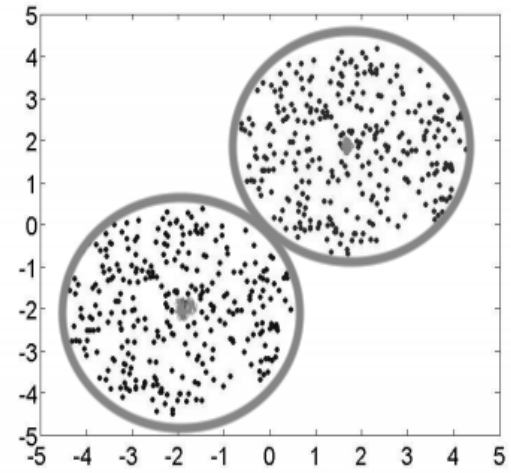
3. K-means becomes difficult with more numbers of dimension

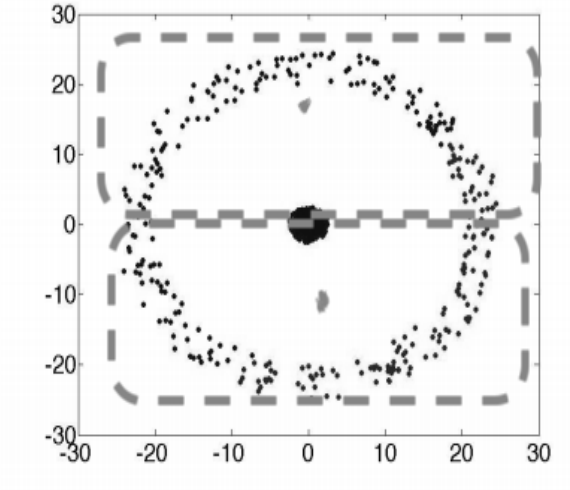
In such cases, Alternative to this ,we can use some hierarchical clustering methods like single link and complete link clusterings.
Also, to deal with high number of dimensions, we can reduce dimensionality by using feature selection like PCA.

3. (a) (2 marks) How well do you think k-means clustering works for each dataset in the figure below. Explain.

1. K-means perform good on dataset represented by first figure.
2. K-means doesn't perform very well on dataset represented by second figure.

(b) (2 marks) Demarcate possible clusters when k = 2.


5
4
3
2
1
0
-1
-2
-3
-4
-5
-5 -4 -3 -2 -1 0 1 2 3 4 5


30
20
10
0
-10
-20
-30
-30 -20 -10 0 10 20 30