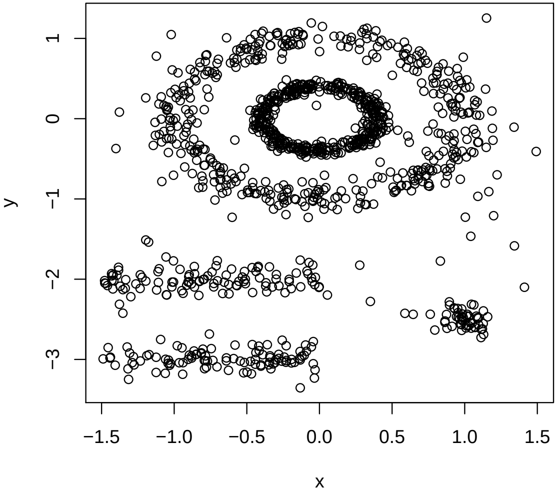
Density-Based Spatial Clustering Application with Noise

Mengta Chung, PhD

Department of Management Sciences
Tamkang University

Goal for Today



DBSCAN

- Density-Based Spatial Clustering Application with Noise (DBSCAN)
- DBSCAN was proposed by Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu in 1996.

  Ester, Martin; Kriegel, Hans-Peter; Sander, Jörg; Xu, Xiaowei (1996). Simoudis, Evangelos; Han, Jiawei; Fayyad, Usama M. (eds.). A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp. 226–231.

Advantage of DBSCAN over K-means

- We do not need to predetermine the number of clusters.
- DBSCAN can identify outliers (observations not belonging to any cluster).
- Except for outliers, every observation will be assigned to a cluster eventually, even when observations are scattered far away.

Two Parameters in DBSCAN

- eps (radius around each point)
- minPts (minimum # of points)

Example

#	A	B
1	1	2
2	3	4
3	2.5	4
4	1.5	2.5
5	3	5
6	2.8	4.5
7	2.5	4.5
8	1.2	2.5
9	1	3
10	1	5
11	1	2.5
12	5	6
13	3.6	4
14	2.1	2.5

Example

Consider the point $(1, 2)$

Use eps = 0.6 and minPts = 4

(1). Calculate the Euclidean distance

(2). eps < 0.6, only $(1.2, 2.5)$ and $(1, 2.5)$

(3). 2 < minPts = 4

$(1, 2)$ will not develop a cluster

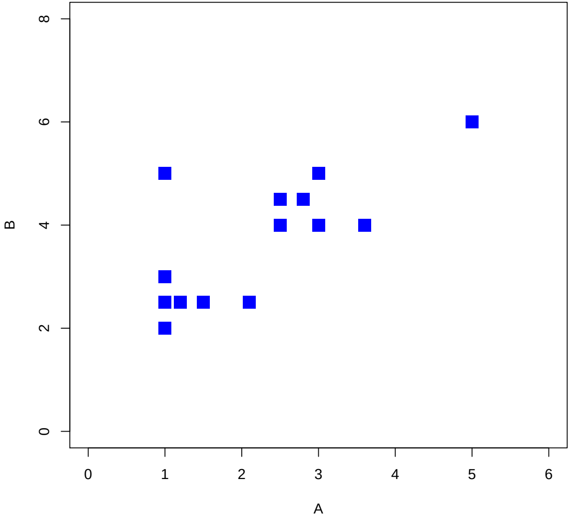
A	**B**	**dist from** $(1, 2)$
1	2	0
3	4	2.828
2.5	4	2.500
1.5	2.5	0.707
3	5	3.606
2.8	4.5	3.081
2.5	4.5	2.915
1.2	2.5	0.539 < 0.6
1	3	1
1	5	3
1	2.5	0.500 < 0.6
5	6	5.657
3.6	4	3.280
2.1	2.5	1.208

Example

#	Point	Neighbors				Cluster
1	(1, 2)	(1.2, 2.5)	(1, 2.5)			2
2	(3, 4)	(2.5, 4)	(2.8, 4.5)	(3.6, 4)		1
3	(2.5, 4)	(3, 4)	(2.8, 4.5)	(2.5, 4.5)		1
4	(1.5, 2.5)	(1.2, 2.5)	(1, 2.5)	(2.1, 2.5)		2
5	(3, 5)	(2.8, 4.5)				1
6	**(2.8, 4.5)**	(3, 4)	(2.5, 4)	(3, 5)	(2.5, 4.5)	C1
7	(2.5, 4.5)	(2.5, 4)	(2.8, 4.5)			1
8	**(1.2, 2.5)**	(1, 2)	(1.5, 2.5)	(1, 3)	(1, 2.5)	C2
9	(1, 3)	(1.2, 2.5)	(1, 2.5)			2
10	(1, 5)					
11	**(1, 2.5)**	(1, 2)	(1.5, 2.5)	(1.2, 2.5)	(1, 3)	C3 = C2
12	(5, 6)					
13	(3.6, 4)	(3, 4)				1
14	(2.1, 2.5)	(1.5, 2.5)				2

- (2.8, 4.5), (1.2, 2.5), (1, 2.5) meet the criteria of eps = 0.6 and minPts = 4 (C1, C2, C3)
- Cluster 2 = Cluster 3
- (1, 2) is a point in cluster 2 => C2
- (3, 4) is a point in cluster 1 => C1
- (3.6, 4) doesn't belong to any clusters, but (3, 4) is in its neighborhood, so (3.6, 4) => C1
- (2.1, 2.5) doesn't belong to any clusters, but (1.5, 2.5) is in its neighborhood, so (2.1, 2.5) => C2
- (1, 5), (5, 6) are outliers

Plot



Lab

Analyze multishapes.csv using DBSCAN in Python