Data Mining I

Corsi di Laurea Magistrale in Business Informatics, Informatica e Informatica Umanistica

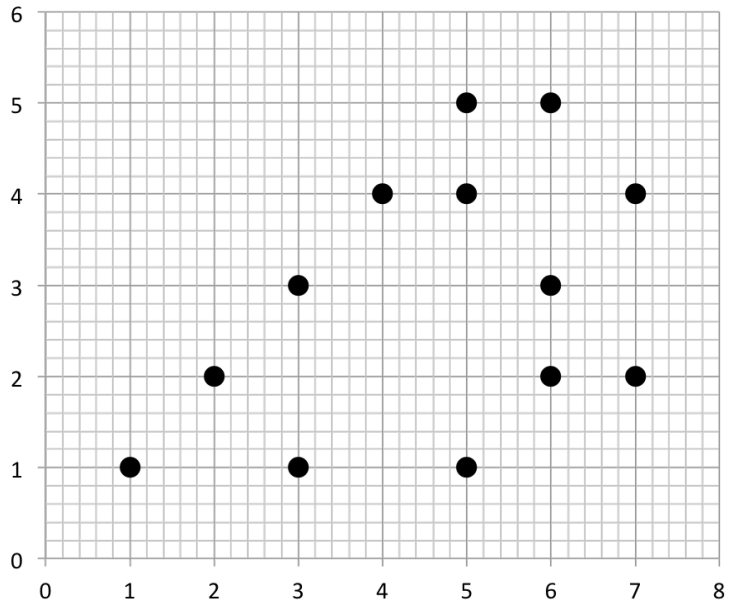
First part - Test 18.01.2016

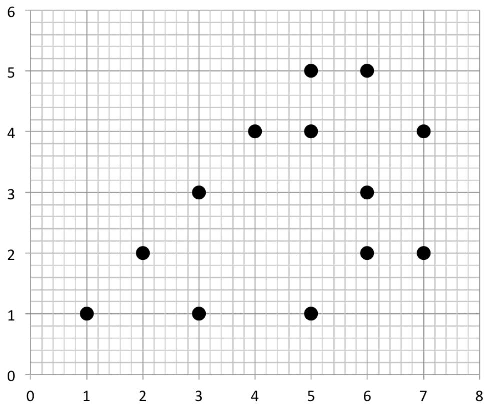
Docenti: Dino Pedreschi, Anna Monreale

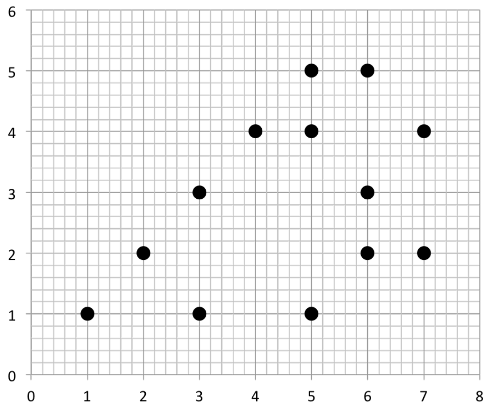
Exercise 1 (10 Points)

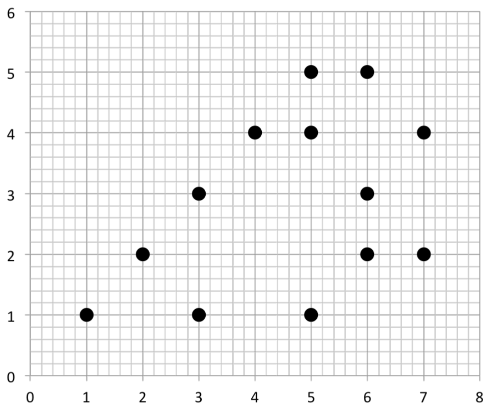
Apply **K-means** to the dataset in the below table and figure using K=3, and the centroids c1=P12, c2=P6 and c3=P13. Explain step by step the algorithm during the application providing for each itaration the centroids.

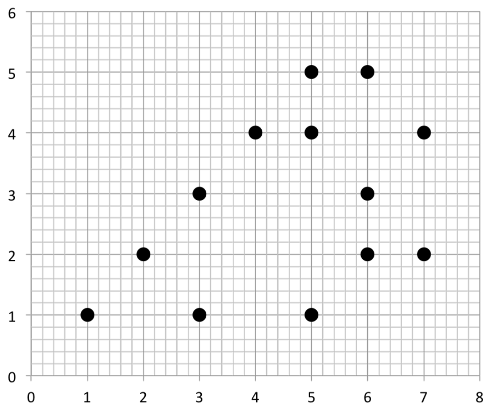
Points	X	Y
P1	3	3
P2	5	5
P3	6	5
P4	5	4
P5	3	1
P6	2	2
P7	6	2
P8	7	2
P9	5	1
P10	6	3
P11	4	4
P12	1	1
P13	7	4

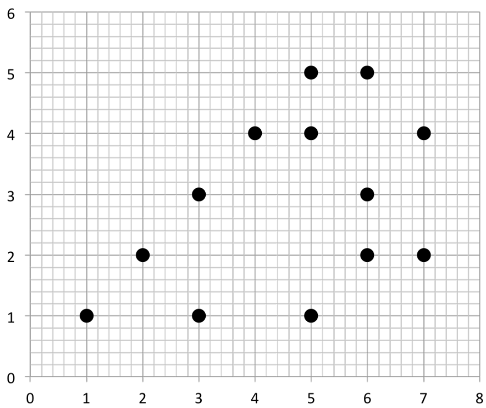


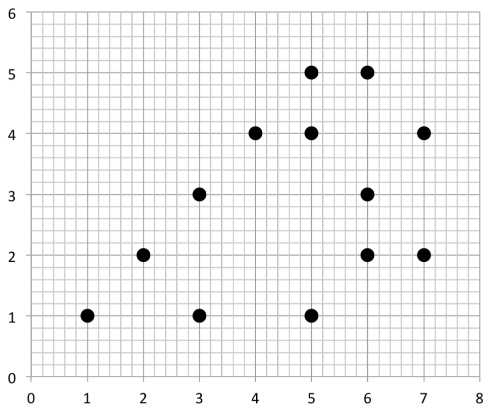

0 1 2 3 4 5 6
0 1 2 3 4 5 6 7 8


0 1 2 3 4 5 6
0 1 2 3 4 5 6 7 8


0 1 2 3 4 5 6
0 1 2 3 4 5 6 7 8


0 1 2 3 4 5 6
0 1 2 3 4 5 6 7 8

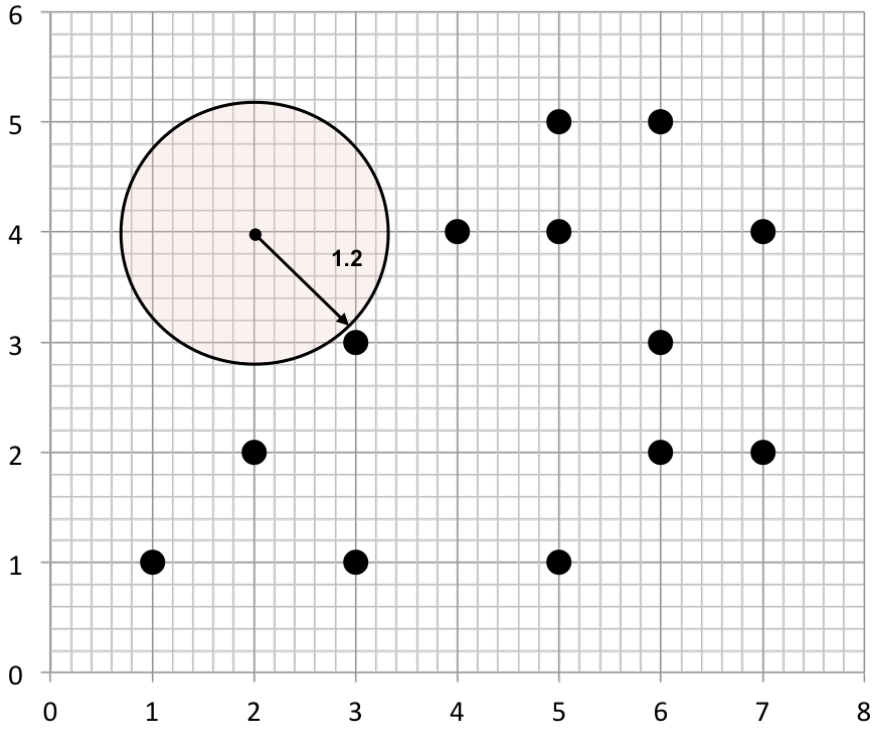

0 1 2 3 4 5 6
0 1 2 3 4 5 6 7 8


0 1 2 3 4 5 6
0 1 2 3 4 5 6 7 8

Exercise 2 (10 Points)

Apply **DBSCAN** to the dataset in the Exercise 1 with MinPts=3 (2 neighbors + the point we are considering as center for computing the density) and radius 1.2

1) Identify *core* points*, border* points and *noise* points.
2) Indicate the clusters obtained
3) Add some points in such a way to transform ech point in a core point



Exercise 3 (10 Points)

Compute the distance matrix by using the Euclideian distance and then apply the complete-linkage HAC and the following dataset and draw the corresponding dendogram.

Points	X	Y
P1	3	3
P2	5	5
P3	6	5
P4	2	1
P5	3	1
P6	4	2
P7	2	2

Exercise 4 (1 Points)

Given the following two vectors compute the cosine similarity

D1= 4 1 2 1 1
D2= 2 1 1 2 0

Exercise 5 (1 Points)

Given the following two binary vectors compute the Jaccard and Simple Matching Coefficient:

p= 011100

q= 110101