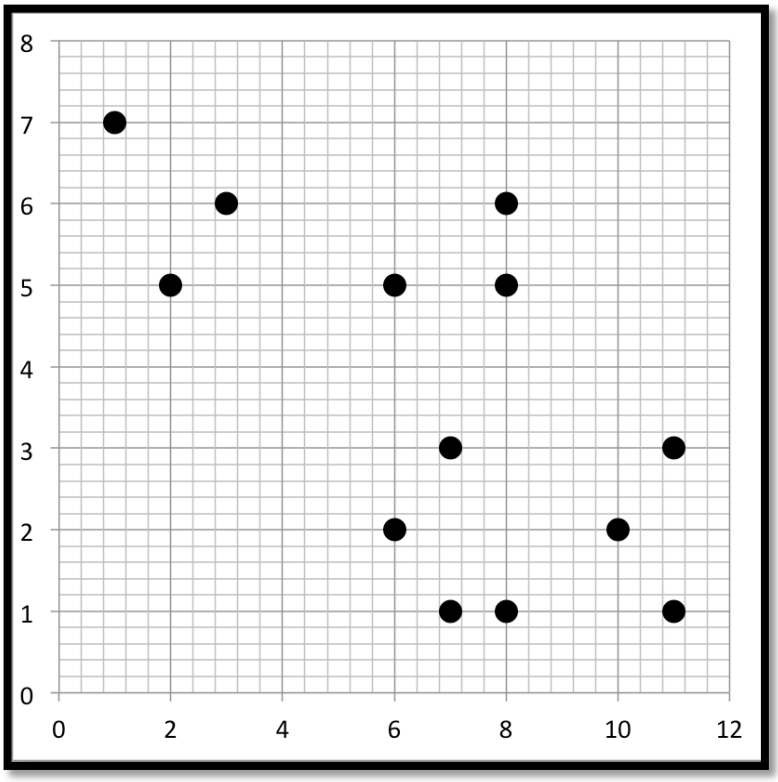
Data Mining I
Corsi di Laurea Magistrale in Business Informatics, Informatica e Informatica Umanistica

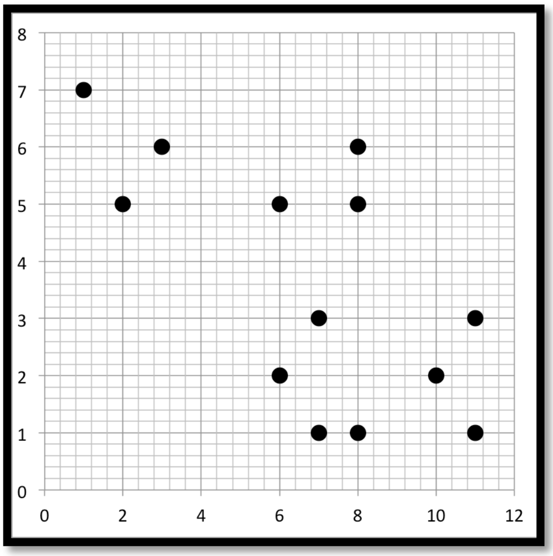
First Part Test del 19.01.2017
Docenti: Dino Pedreschi, Anna Monreale

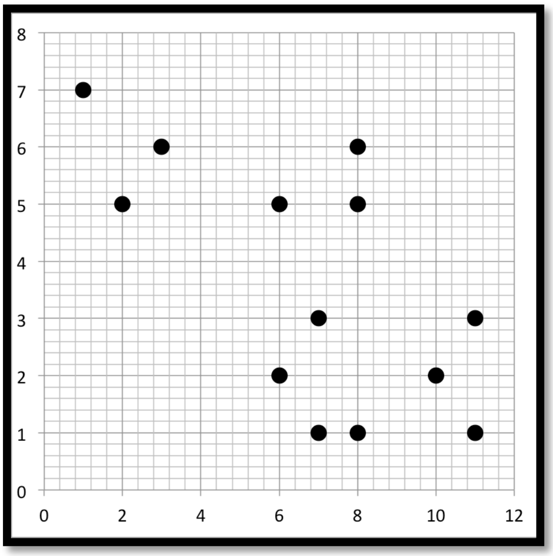
Exercise 1 (15 Points)

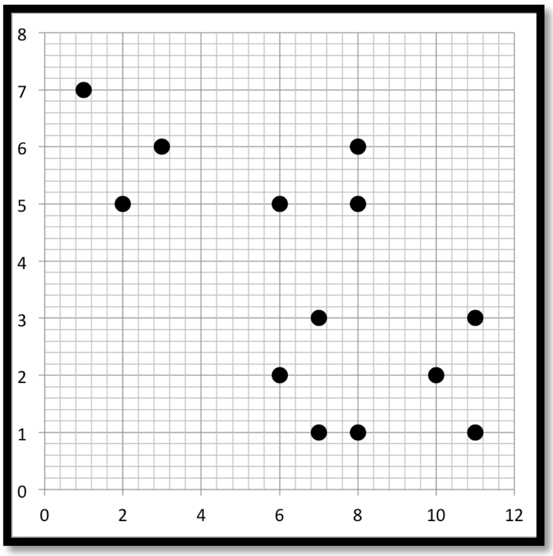
a) Apply **K-means** to the dataset in the below table and figure using K=2, and the centroids c1=P5 and c2=P8. Explain what happens in any iteration.
b) Discuss the reason of the k-means termination

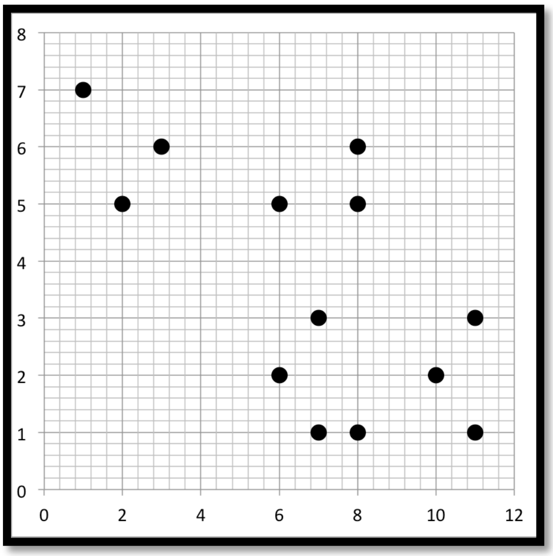
Points	X	Y
P1	1	7
P2	2	5
P3	3	6
P4	10	2
P5	11	1
P6	11	3
P7	6	2
P8	7	1
P9	7	3
P10	8	1
P11	6	5
P12	8	6
P13	8	5

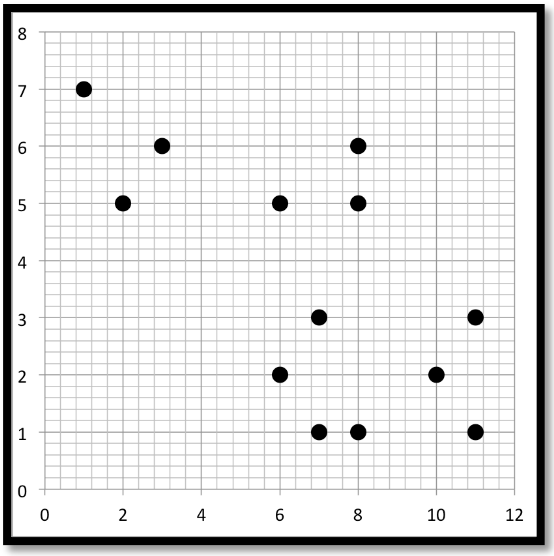


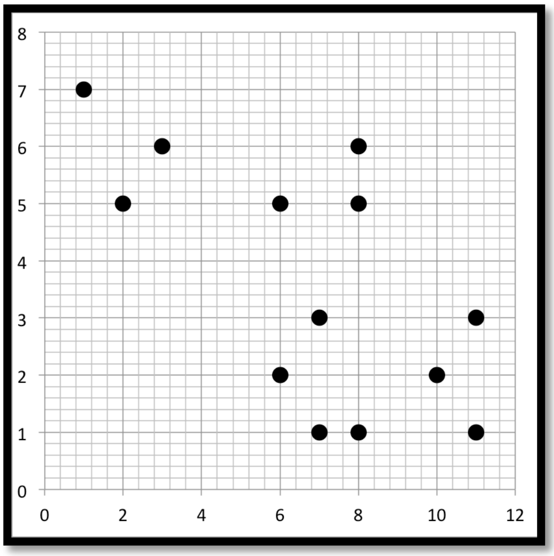

8
7
6
5
4
3
2
1
0
0 2 4 6 8 10 12


8
7
6
5
4
3
2
1
0
0 2 4 6 8 10 12


8
7
6
5
4
3
2
1
0
0 2 4 6 8 10 12


8
7
6
5
4
3
2
1
0
0 2 4 6 8 10 12


8
7
6
5
4
3
2
1
0
0 2 4 6 8 10 12


8
7
6
5
4
3
2
1
0
0 2 4 6 8 10 12

Exercise 2 (16 Points)

Consider the following points and use the Manhattan distance to solve the following exercises:

1. Apply the single-linkage HAC on the dataset and draw the corresponding dendogram.
2. Apply the complete-linkage HAC on the dataset and draw the corresponding dendogram.

P1	1	7
P2	6	2
P3	7	1
P4	7	3
P5	8	1
P6	6	5
P7	8	6
P8	8	5

Data Mining I

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

Second Part -Test 19.01.2017
Docenti: Dino Pedreschi, Anna Monreale

Exercise 1 (14 Points)

Consider the following transactions

Transaction ID	Itemsets
1	{A,E}
2	{E,D}
3	{B,E,D}
4	{A,C,D,E}
5	{A,B,E,F}
6	{A,B}
7	{B,C,F}
8	{E,B,F}
9	{A,D,F}
10	{A,D,C,F}

A) Extract the frequent itemsets by *Apriori* using *min sup=20*%, showing and discussing the different steps of the algorithm **(7 points)**

B) Extract the association rules using minimum confidence equal to 70% **(3 points)**

A) Compute the lift for the rules extracted in the previous point and explain the lift measure highlighting when it is particularly useful and explaining the relation among the variables in case of lift value greater than 1, lower than 1 and equal 0. (**3 points**)

C) How many frequent subsets with maximum length 3 does the frequent pattern {a, b, c, d, e, u,f,h} contain? (**1 points**)

Exercise 2 (17 Points)

Consider the following dataset

Training Data

State	Total Calls	Sex	Service Calls	CHURN
Italy	<100	F	> 10	YES
German	<100	M	<= 10	NO
Italy	>= 100	M	> 10	YES
Italy	<100	F	<= 10	NO
German	< 100	F	> 10	NO
German	>=100	M	> 10	YES
German	< 100	M	<= 10	YES
German	>= 100	F	<= 10	YES
German	>=100	M	> 10	NO
German	< 30	F	<= 10	NO
German	>=100	F	> 10	YES
Italy	<100	F	> 10	NO

A) Use the above training dataset for building a decision tree based on misclassification rate for the variable "CHURN", expanding the nodes of the tree until no split provides a gain. **(12 points)**

B) Provide the confusion matrix and evaluate the accuracy, precision and recall of the tree with respect to the following test and training set **(4 Points)**

C) Given the training set compute the Gini Index on the attribute Sex without considering the class (**1 Point**)

Test Data

State	Total Minutes	Sex	Service Calls	CHURN
German	90	M	<= 10	YES
Italy	70	F	<= 10	NO
German	166	F	<= 10	YES
Italy	90	M	> 10	YES