Data Mining I

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

First Part Test del 06.09.2017

Docenti: Dino Pedreschi, Anna Monreale

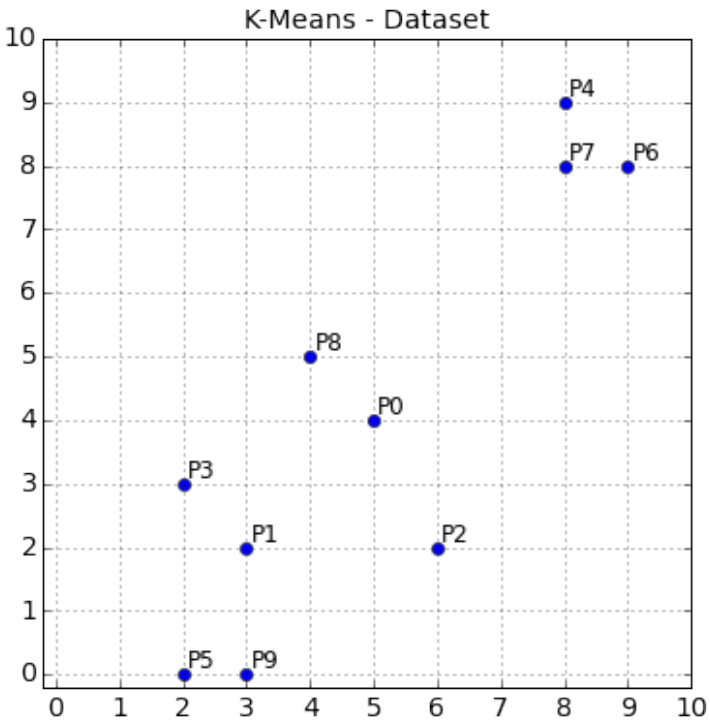
Exercise 1 (16 Points)

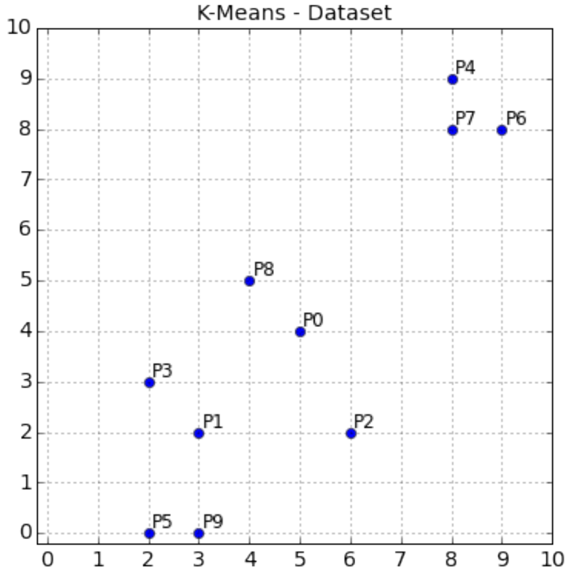
a) Apply **K-means** to the dataset in the below table and figure using K=2, and the centroids c1=P2 and c2=P9. Explain what happens in any iteration **(10 points).**

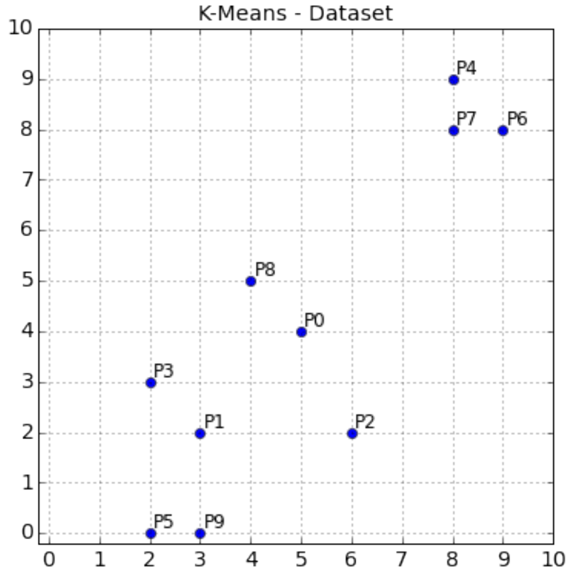
b) Discuss the reason of the k-means termination **(4 points)**.

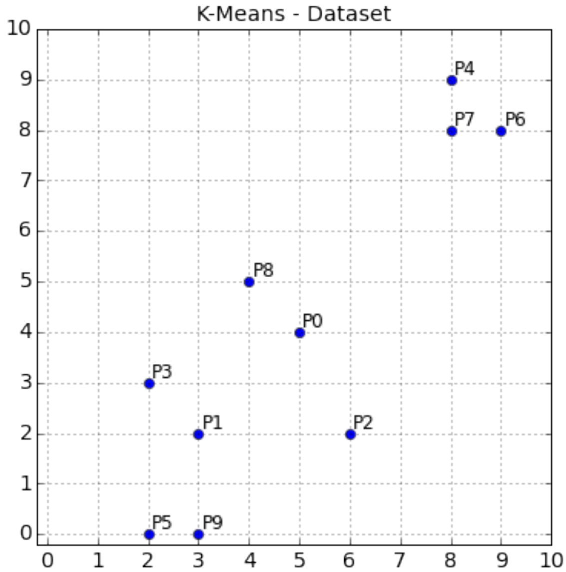
c) Identify another couple of initial centroids leading to the same clustering obtained in a**) (2 points)**.

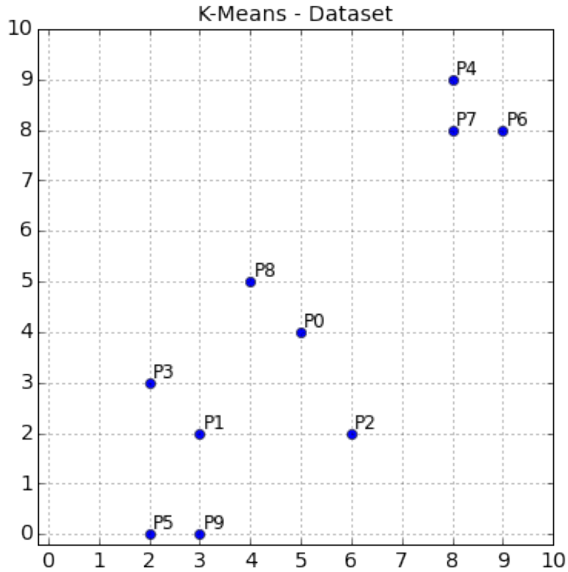
Points	X	Y
P0	5	4
P1	3	2
P2	6	2
P3	2	3
P4	8	9
P5	2	0
P6	9	8
P7	8	8
P8	4	5
P9	3	0

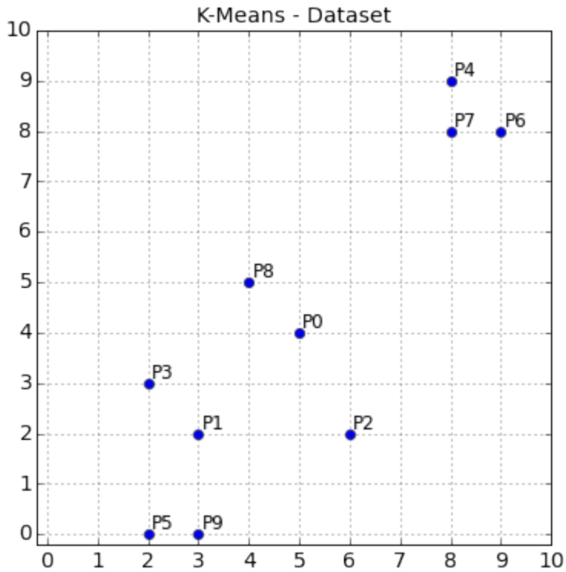


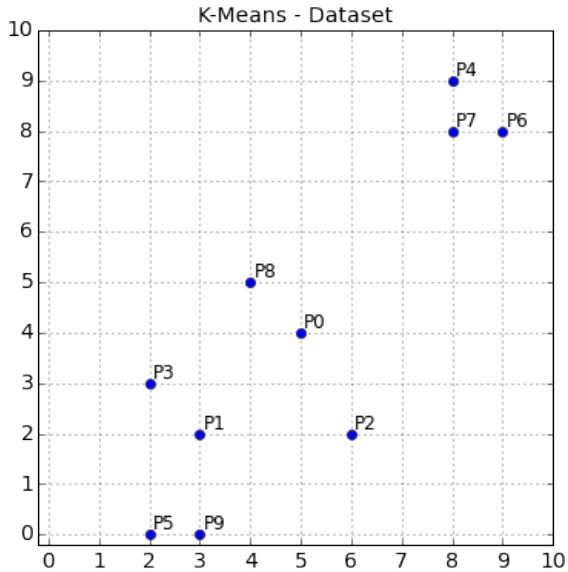

K-Means - Dataset
P4
P7
P6
P8
P0
P3
P1
P2
P5
P9


K-Means - Dataset
P4
P7
P6
P8
P0
P3
P1
P2
P5
P9


K-Means - Dataset
P4
P7
P6
P8
P0
P3
P1
P2
P5
P9


K-Means - Dataset
P4
P7
P6
P8
P0
P3
P1
P2
P5
P9

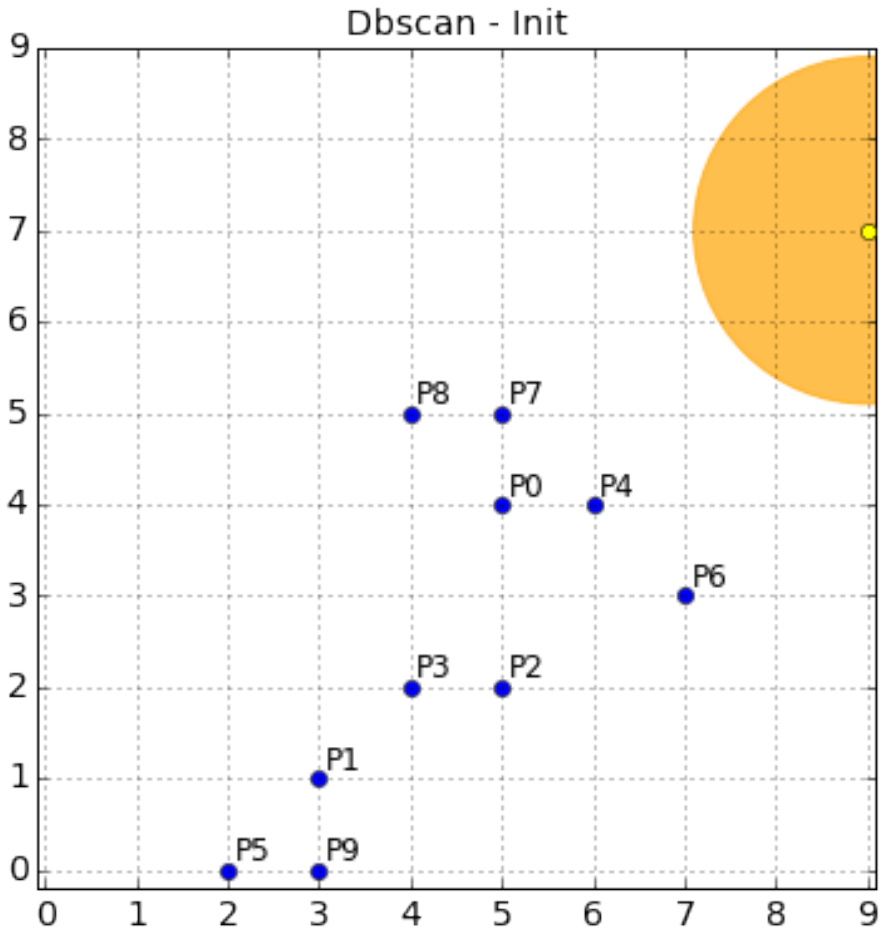

K-Means - Dataset
P4
P7
P6
P8
P0
P3
P1
P2
P5
P9


K-Means - Dataset
P4
P7
P6
P8
P0
P3
P1
P2
P5
P9

Exercise 2 (15 Points)

a) Apply **Dbscan** on the data of previous exercise with radius eps=1.9 and minPts=3 (2 neighbor plus the point itself) and for each point specify if it is a core point, border point or noise **(10 points).**
b) Indicate the composition of the clusters obtained **(2 points).**
c) Which is the minimum eps to obtain a unique cluster? How many core and border points are presents in this new clustering? **(3 points).**

Points	X	Y
P0	5	4
P1	3	1
P2	5	2
P3	4	2
P4	6	4
P5	2	0
P6	7	3
P7	5	5
P8	4	5
P9	3	0



Data Mining I

Corsi di Laurea in Business Informatics, Informatica e Informatica Umanistica

First Part Test del 06.09.2017

Docenti: Dino Pedreschi, Anna Monreale

Exercise 1 (14 Points)

Consider the following transactions

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

A) Extract the frequent itemsets by *Apriori* using *min sup=30%*, showing and discussing the different steps of the algorithm **(7 points)**

B) Extract the association rules using minimum confidence equal to 70% **(5 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. (**2 points**)

Exercise 2 (17 Points)

Training Data

	State	Contract	Sex	PreviousCompany	Churn
0	Italy	Classic	F	Tim	YES
1	German	Travel	M	Wind	NO
2	France	Travel	M	Tim	YES
3	Italy	Young	F	Tim	NO
4	German	Travel	F	Tim	NO
5	Italy	Classic	M	Tim	YES
6	Italy	Classic	M	Wind	NO
7	German	Young	F	Wind	NO
8	German	Young	M	Tim	NO
9	France	Travel	F	Tim	NO
10	Italy	Young	M	Tim	YES

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 or the number of records in a leaf is equals or higher than 2. **(12 points)**

B) Provide the confusion matrix and evaluate the accuracy, precision, recall and f1-measure of the tree with respect to the following test **AND** training set above. You must explicitly provide the formulas of accuracy, precision, recall and f1-measure **(4 Points)**

C) Compute the Entropy on the attribute Contract to estimate its variability without considering the class Churn (**1 Point**).

Test Data

	State	Contract	Sex	PreviousCompany	Churn
0	Italy	Classic	F	Tim	YES
1	German	Travel	M	Wind	YES
2	Italy	Travel	M	Wind	YES
3	German	Young	M	Tim	NO