Università di Pisa	A.A. 2016-2017

Data Mining II June 6th, 2017

mid-term exam

Exercise 1 - Classification **(13 points)**

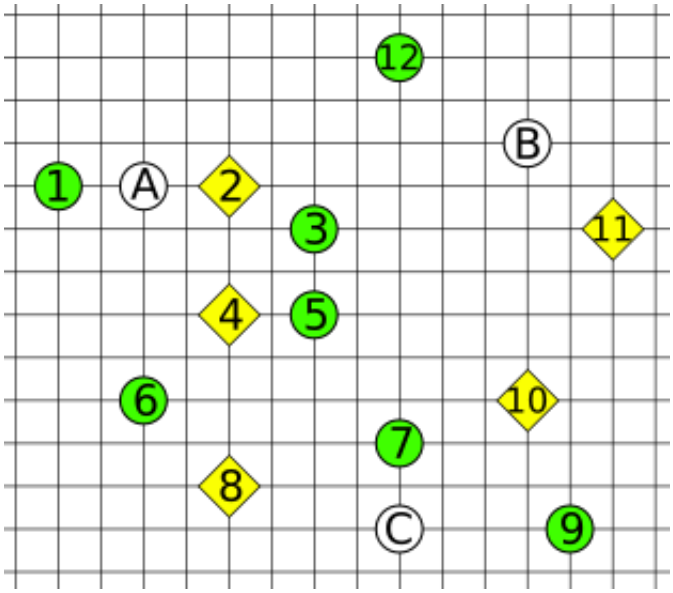
a) Naive Bayes (**6 points**)
Given the training set on the left, build a Naive Bayes classification model and apply it to the test set on the right.

A	B	C	class
high	no	green	Y
medium	no	red	Y
low	yes	green	N
high	no	red	N
low	yes	red	Y
high	no	green	Y
medium	yes	green	N

A	B	C	class
low	no	red	
high	yes	green	
medium	yes	red	

b) k-NN (**6 points**)

Given the training set below, composed of elements numbered from 1 to 12, and labelled as circles and diamonds, use it to classify the remaining 3 elements (letters A, B and C) using a k-NN classifier with k=3. For each point to classify, list the points of the dataset that belong to its k-NN set.



c) Ensembles (**1 point**)
For a binary classification problem (target values: Y and N) we are able to extract 1000 independent models, although each of them has a very poor classification, i.e. error = 50%. What is the accuracy that a bagging approach can achieve?

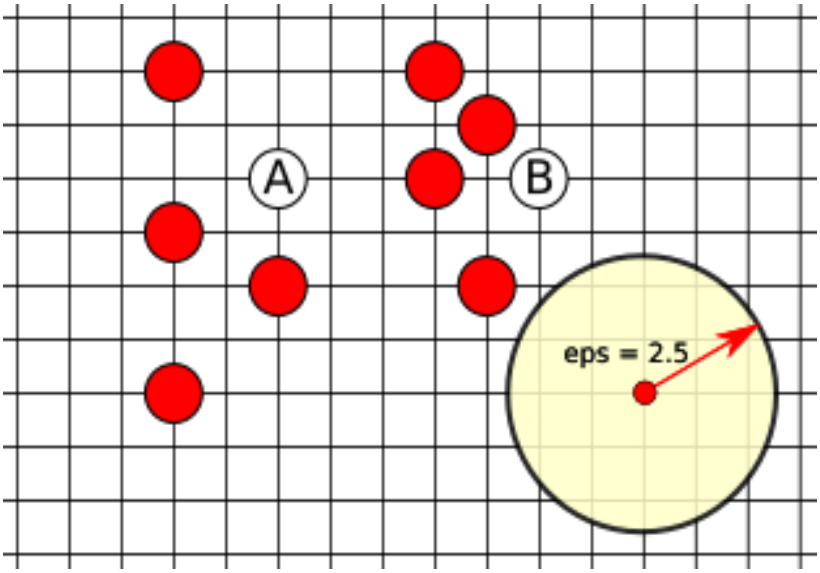
Exercise 2 - Outlier Detection **(12 points)**

Given the dataset of 10 points below, consider the outlier detection problem for points A and B, adopting the following three methods:

a) Distance-based: DB(ε,π) (**4 points**)
Are A and/or B outliers, if thresholds are forced to ε = 2.5 and π = 0.25 ?

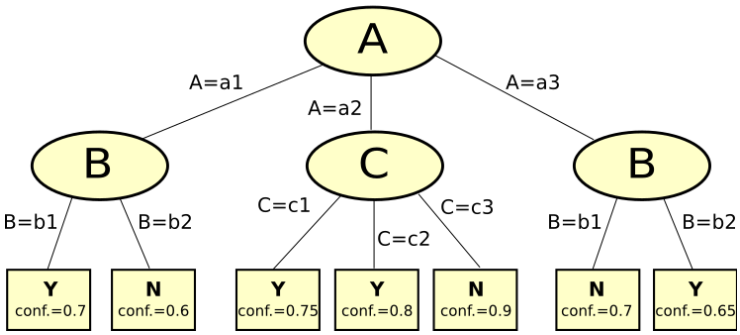
b) Density-based: LOF (**4 points**)
Compute the LOF score for points A and B by taking k=3, i.e. comparing each point with its 3 NNs (not counting the point itself). In order to simplify the calculations, the reachability-distance used by LOF can be replaced by the simple Euclidean distance.

c) Depth-based (**4 points**)
Compute the depth score of points A and B.



Exercise 3 - Validation **(7 points)**

a) ROC curve (**6 points**)
Given the following decision tree on left, where the leaves also show the confidence of each prediction, and given the test set on the right, build the corrisponding ROC curve.



A	B	C	class
a2	b1	c1	Y
a3	b1	c2	Y
a3	b2	c1	N
a2	b2	c3	N
a1	b1	c2	Y

b) Lift charts (**1 points**)
We have a test set containing 200 negative cases and 50 positive ones. Plot the lift charts we would obtain by applying to our test set (i) a perfect scoring/classification model, (ii) a random model, (iii) the worst possible model.