Università di Pisa	A.A. 2018-2019

Data Mining II June 6th, 2019

2nd mid-term exam

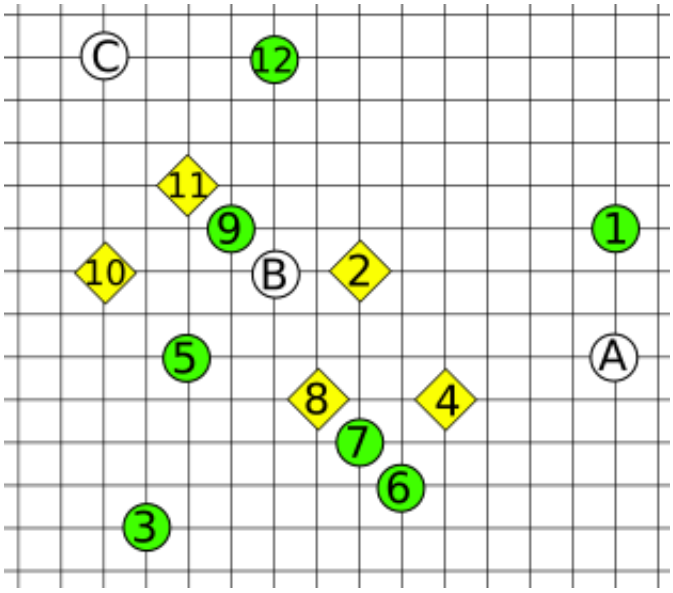
Exercise 1 - Classification **(13 points)**

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.

Income	Loan	Works	class
high	no	yes	Y
medium	no	yes	N
low	yes	no	N
medium	no	no	Y
high	yes	no	N
low	yes	no	Y
low	no	yes	Y

Income	Loan	Works	class
low	no	no	
high	no	yes	
medium	yes	yes	

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.

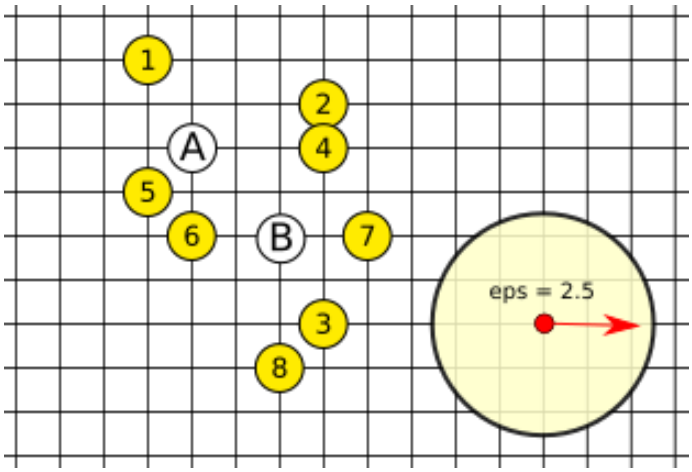


c) Ensemble mathods (**1 point**)
When using AdaBoosting, the number of weak classifiers to use is a parameter of the method. What happens to the accuracy of the strong model produced when that number becomes very large? Why?

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

Given the dataset of 10 points below (all positioned at an intersection of the regular grid depicted), 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 = 0.35? (Notice that in computing the density of a point, the point itself should not be counted)



b) Density-based: LOF (**4 points**) Compute the LOF score for points A and B by taking k=2, i.e. comparing each point with its 2-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 test set with the predictions (and associated confidence) returned by our model, build the corrisponding ROC curve.

Record	Real Class	Predicted	
row 1	Y	Y	0.74
row 2	Y	Y	0.88
row 3	N	N	0.77
row 4	Y	N	0.66
row 5	Y	Y	0.92
row 6	N	N	0.99
row 7	Y	N	0.82
row 8	N	N	0.93
row 9	N	Y	0.98
row 10	N	Y	0.95

b) AUC (**1 points**)
Assume to have a trivial classifier that always predicts "Y" with 100% confidence. What is its AUC (area under the ROC curve) on the test set above?