Università di Pisa	A.A. 2017-2018

Data Mining II June 1th, 2018

2nd mid-term exam

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

a) Naive Bayes (**6 points**)

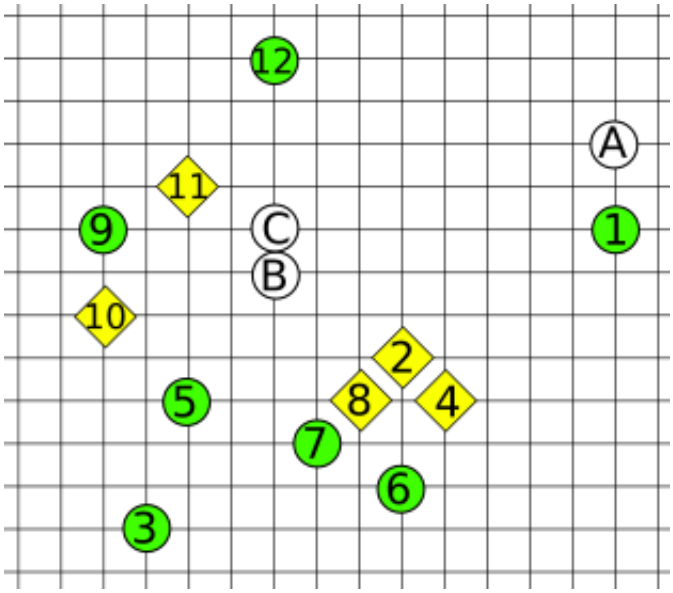
Given the training set on the left, build a Naive Bayes classification model and apply it to the test set on the right.

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

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

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.



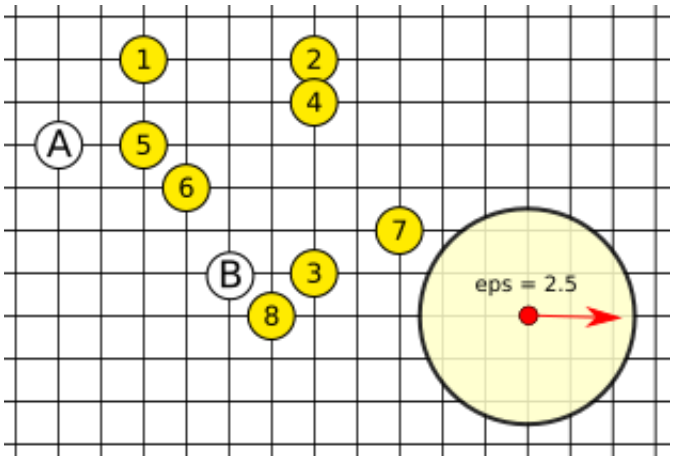
c) ANN (**1 point**)

An artificial neural network shows an accuracy of 90% over a given test set. When we build another neural network with exactly the same training set and parameters as before, but increasing by 1 the number of nodes in the last hidden layer, the new model has an accuracy of 88% on the same test set, i.e. slightly worse. Can it happen? 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 = 0.25? (Notice that in computing the density of a point, the point itself should not be counted)

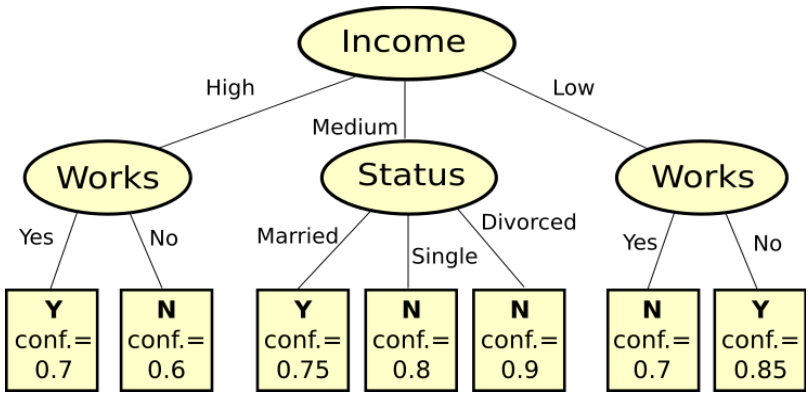


b) Density-based: LOF **(4 points)**
Compute the LOF score for points A and B by taking k=2, i.e. comparing each point with its 2-NNs (not counting the point itself). In order to simplify the calculations, the reachability-distance used by LOF can be replaced by the simple Euclidean distance.

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 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.



Income	Works	Status	class
High	Yes	Single	Y
Medium	No	Divorced	N
Medium	Yes	Married	Y
Low	No	Married	N
Medium	Yes	Single	Y
Medium	Yes	Married	Y
Low	No	Single	N
Low	Yes	Single	Y

b) AUC (**1 points**)
Compute the Area Under the Curve for the ROC in point a) above. What is the AUC of the optimal predictive model, and what that of a random classifier?