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	

Model probabilities

	Y	N
	0.57	0.43
	A | Y	A | N
high	0.50	0.33
medium	0.25	0.33
low	0.25	0.33
	B | Y	B | N
yes	0.25	0.67
no	0.75	0.33
	C | Y	C | N
green	0.50	0.67
red	0.50	0.33

Predictions

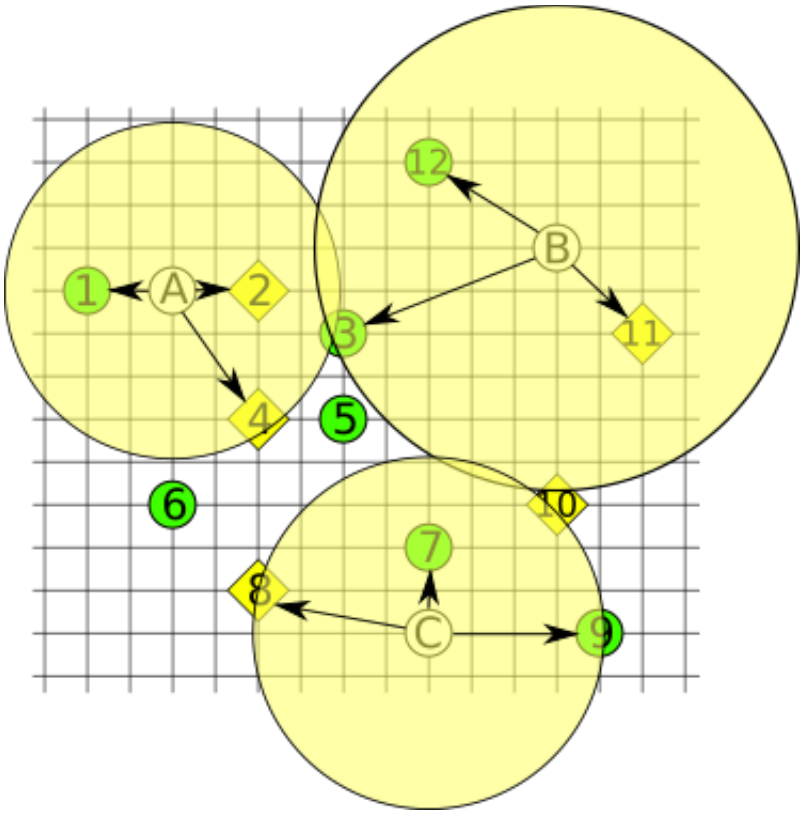
		low	no	red	
Y	0.57	0.25	0.75	0.50	0.05
N	0.43	0.33	0.33	0.33	0.02
		high	yes	green	
Y	0.57	0.50	0.25	0.50	0.04
N	0.43	0.33	0.67	0.67	0.06
		medium	yes	red	
Y	0.57	0.25	0.25	0.50	0.02
N	0.43	0.33	0.67	0.33	0.03

==> Output: Y, N, N

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.

KNNs:



Results:

kNN(A) = {1,2,4} ==> A = diamond
kNN(B) = {11,12,3} ==> B = circle
kNN(C) = {7,8,9} ==> C = circle

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?

Answer: feeding the formula for the global error of the ensemble model with epsilon = 0.5, we get that the overall error is again 0.5. Indeed, each of the 2^1000 possible configurations of outcomes of the 1000 base models (e.g. < M1=right, M2=right, M3=wrong, … M1000=right>) have the same probability 0.5^1000. Since for each configuration where the "correct" models win there is one (for instance, the complementary one) where they loose, it happens that the probability of having a successful configuration is the same as having a failure. More intuitively: a model with 0.5 corresponds to throwing a coin. Throwing it 1000 times does not help to predict anything.

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 ?

Answer: A is an outlier, B is not. See figure.

b) Density-based: LOF (**4 points**)

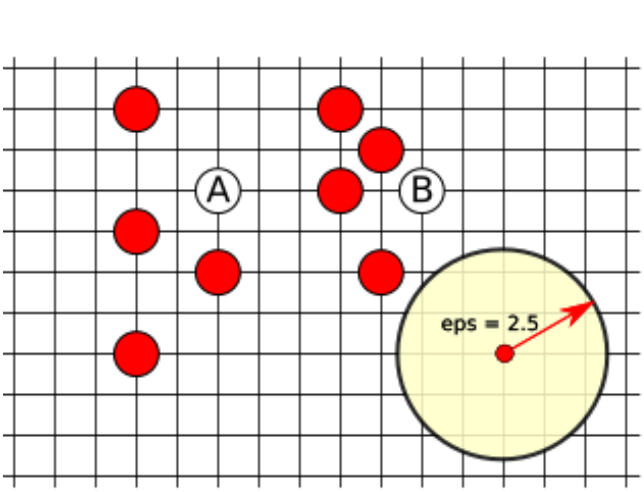
Compute the LOF score for points A and B by taking k=3, i.e. comparing each point with its 3 NNs (not counting the point itself). In order to simplify the calculations, the reachability-distance used by LOF can be replaced by the simple Euclidean distance.

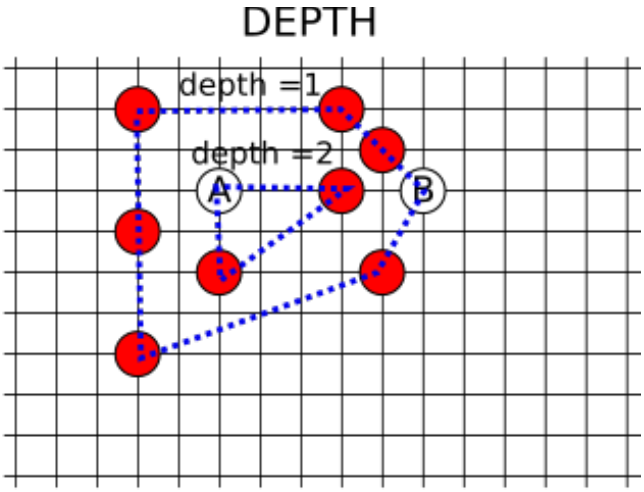
Answer: see 3-Nns for A and B in the figure. LOF(A) is lower than LOF(B).

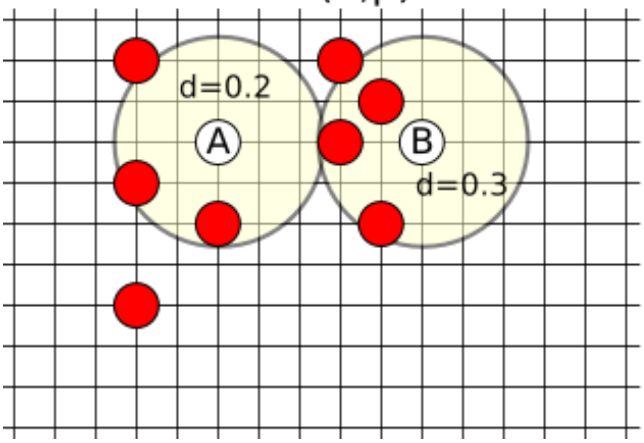
c) Depth-based (**4 points**)

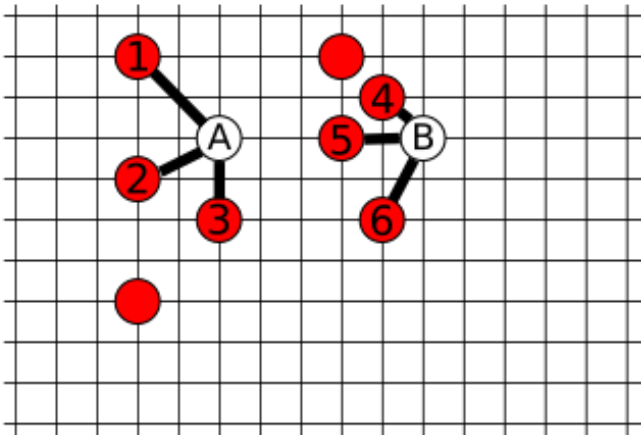
Compute the depth score of points A and B.

Answer: A → 2, B → 1. See figure.



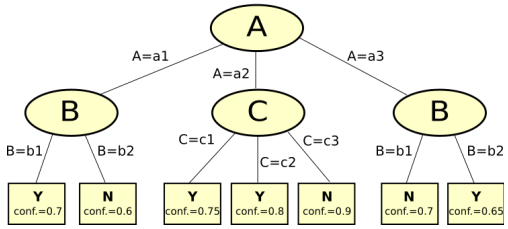






Exercise 3 - Validation **(7 points)**

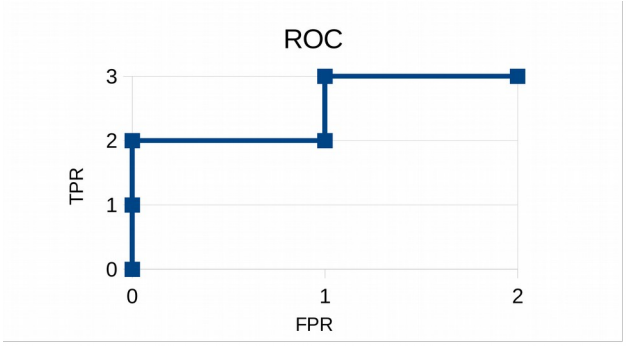
a) ROC curve (**6 points**)
Given the following decision tree on left, where the leaves also show the confidence of each prediction, and given the test set on the right, build the corrisponding ROC curve.



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

Answer:

A	B	C	class	predicted	p(Y)	FPR	TPR
a2	b1	c1	Y	Y 0.75	0.75	0	0
a1	b1	c2	Y	Y 0.7	0.7	0	1
a3	b2	c1	N	Y 0.65	0.65	1	2
a3	b1	c2	Y	N 0.7	0.3	1	3
a2	b2	c3	N	N 0.9	0.1	2	3



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.

Answer: