Università di Pisa	A.A. 2017-2018

Data Mining II June 1th, 2018

2nd mid-term exam – Solutions

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	

Solutions:

	Y	N
	4	
	A | Y	A | N
high	2	
medium	1	
low	1	
	B | Y	B | N
yes	1	
no	3	
	C | Y	C | N
yes	2	
no	2	

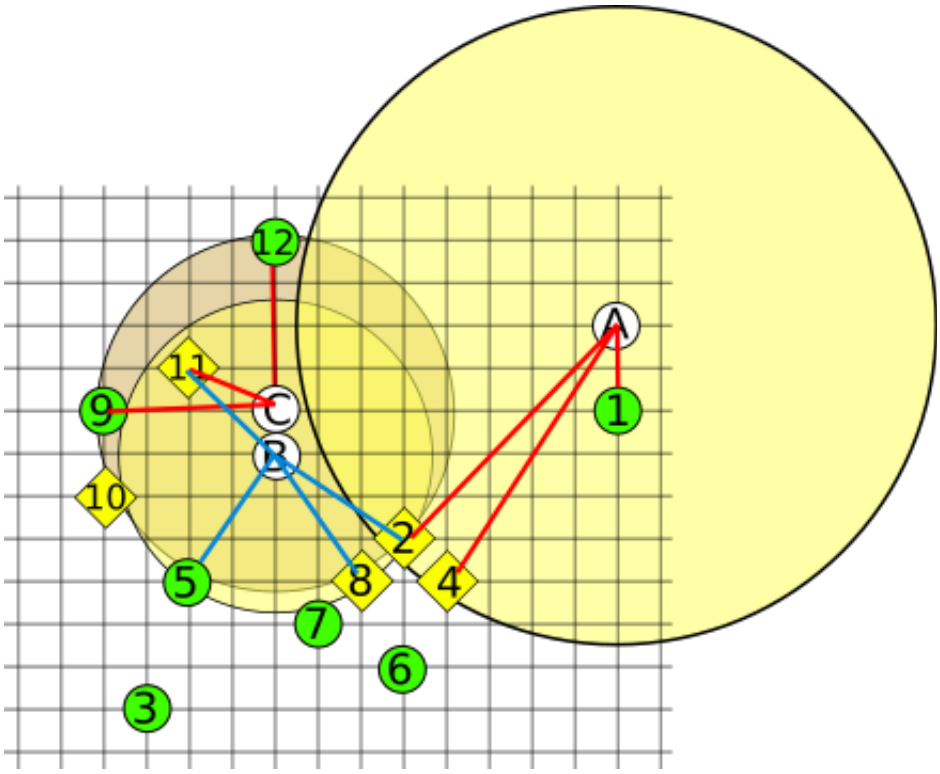
		Y	N
3		0.57	0.43
		A | Y	A | N
1	high	0.50	0.33
1	medium	0.25	0.33
1	low	0.25	0.33
		B | Y	B | N
2	yes	0.25	0.67
1	no	0.75	0.33
		C | Y	C | N
2	yes	0.50	0.67
1	no	0.50	0.33

	Class	Income	Married	Works	
		low	no	no	
Y	0.57	0.25	0.75	0.50	0.05
N	0.43	0.33	0.33	0.33	0.02
		high	yes	yes	
Y	0.57	0.50	0.25	0.50	0.04
N	0.43	0.33	0.67	0.67	0.06
		medium	yes	no	
Y	0.57	0.25	0.25	0.50	0.02
N	0.43	0.33	0.67	0.33	0.03

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.



KNN(A) = {1,2,4} → diamond
KNN(B) = {2,5,8,11} → diamond
KNN(C) = {9,11,12} → circle

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?

It can happen for several reasons. The main ones:

- The learning phase is inherently non-deterministic, since it starts from a random configuration (i.e. random weights) and from there it proceeds in a step-by-step local optimization. The situation is very similar to k-means clustering.
- Adding just one node increases the complexity of the model a little bit, therefore possibly increasing the risk of overfitting – especially if the starting condition was already an overfitting one.

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)

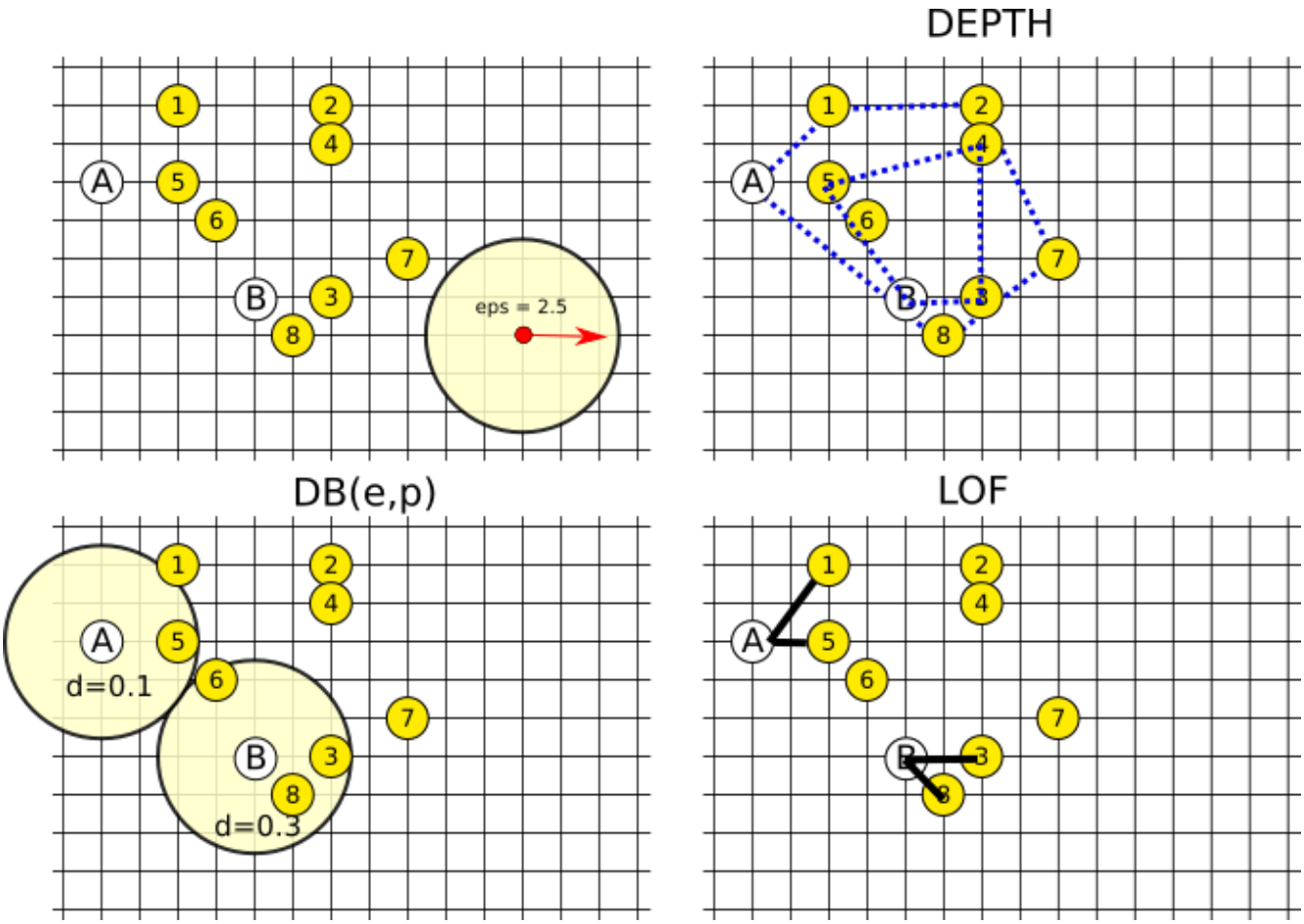
A → neighbours = {5} → d=0.1 => outlier B → neighbours = {3,6,8} → 0.3 => no outlier

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.

2-NN(A) = { 1,5 } LRD(A) = 1/ [(2 + √8)/2] = 0.414 LRD(1) = 1/ [(2 + √8)/2] = 0.414 LRD(5) = 1/ [**(√2 + 2 + 2)/3**] = 0.554 **LOF(A)** = ([LRD(1) + LRD(5)]/2) / LRD(A) = [(0.414 + 0.554) / 2] / 0.414 = **1.169 weak outlier**	**2-NN(B) = { 3,8 }** LRD(B) = 1/ [(2 + √2)/2] = 0.586 LRD(3) = 1/ [(2 + √2)/2] = 0.586 LRD(8) = 1/ [(√2 + √2)/2] = 0.707 **LOF(A)** = ([LRD(3) + LRD(8)]/2) / LRD(B) = [(0.586 + 0.707) / 2] / 0.586 = **1.103 weak outlier**

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

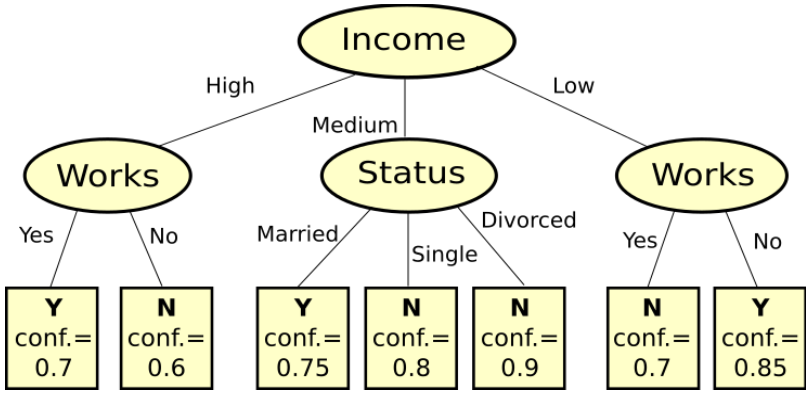
Depth 1 → A, 1, 2, 7, 8 Depth 2 → 3, 4, 5, B Depth 3 → 6



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

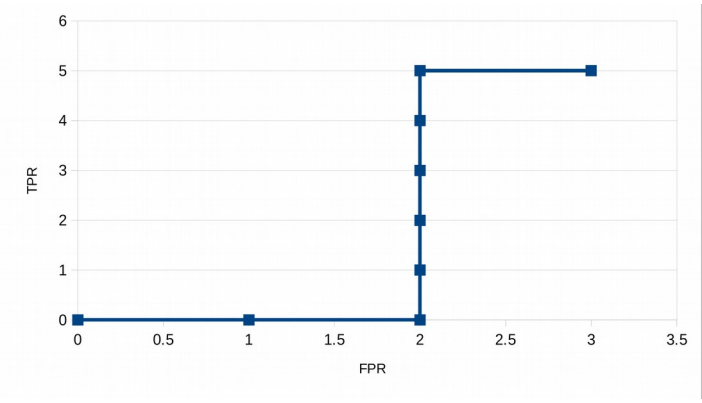
Solutions:

Income	Works	Status	class	predicted		score
High	Yes	Single	Y	Y	0.7	0.7
Medium	No	Divorced	N	N	0.9	0.1
Medium	Yes	Married	Y	Y	0.75	0.75
Low	No	Married	N	Y	0.85	0.85
Medium	Yes	Single	Y	N	0.8	0.2
Medium	Yes	Married	Y	Y	0.75	0.75
Low	No	Single	N	Y	0.85	0.85
Low	Yes	Single	Y	N	0.7	0.3

SORTED

Real Class
N
N
Y
Y
Y
Y
Y
N

	TPR	FPR	AUC partial
Score	0	0	
0.85	0	1	0
0.85	0	2	0
0.75	1	2	0
0.75	2	2	0
0.7	3	2	0
0.3	4	2	0
0.2	5	2	0
0.1	5	3	5
		AUC	**5**
		Normalized	**0.3333333333**



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?

AUC = 5 (= 0.333)
AUC optimal = 15 (=1.0)
AUC random = 7.5 (=0.5)