isep instituto superior de engenharia do porto

Department of Computer Engineering

Knowledge Discovery

Master's degree Computer Engineering

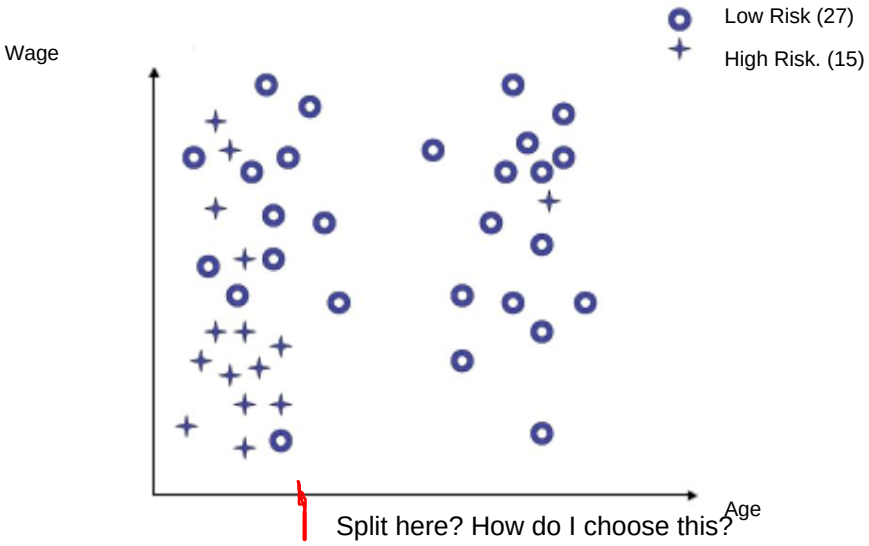
Normal Season Exam

July 7, 2022

Duration 2 hours

4 values

1. Consider the following data set:



a) Find the best partition, and the corresponding information gain. Consider at least two alternatives, one per axis, and compare the information gain of each one.

Information gain
= reduction in entropy

b) Multiple partitions of an attribute always result in a greater reduction in entropy than binary partitions. Say if you agree with the statement, justify it.

6 values

2. Consider the following data set regarding non-poisonous/poisonous mushrooms

	Heavy	Bad-Smelling	Mushroom	Spotted	Poisonous
A	0	0	0	0	
B	0	1	1	1	
W	1	1	0	0	
D	1	0	0	1	
AND	1	1	1	?	
F	0	0	1	?	
G	0	1	0	?	
H	1	0	1	?	

a) Without resorting to calculations, draw the decision tree with only the root attribute, relative to this dataset. Justify.

b) Present the confusion matrix. What is the model's Accuracy rate?

c) Should the presented model present different costs in relation to false predictions? Justify.

d) What is the most appropriate measure to evaluate the performance of this model? Present your value and compare it against the Accuracy of the model.

e) Using the Naive Bayes classifier on this data set, what would be the prediction for the instance E? Present the calculations. Why is the Naive Bayes classifier called Naïve?

f) Build a K-NN classifier based on Eucledean distance to predict the toxicity of the instance H. Use K = 1.

4 values

3. Suppose the Apriori algorithm produced the following frequent and respective itemsets supports:

Association Rules

L1	sup
{A}	0.4
{C}	0.3
{B}	0.6
{E}	0.5

L2	sup
{A, C}	0.2
{A,E}	0.2
{C,E}	0.2
{B,E}	0.4

L3	sup
{A, C, E}	0.1

L4 = Æ

a) Let b=0.5 be the minimum level of confidence, present all association rules with confidence greater than or equal to b.

b) Explain why the Apriori algorithm (which generates all frequent itemsets) always stops, i.e. is, because there is always a K such that Lk = Æ.

c) What are negatively correlated itemsets? What does this mean in practical terms?

d) Give an example of an A Þ B association rule that has support and confidence above the minimum requirements but where item sets A and B are negatively correlated.

3 values

4. Consider the following clustering algorithm called *Leader Clustering.* This algorithm takes two parameters: an integer value k and a real value t. It works analogously to K-means algorithm, that is, it starts by selecting k instances – Leaders and then assigns each training instance to the nearest leader, except if the distance from the instance to the leader is greater than the parameter t, then this instance becomes a new *leader.* after the processing of all training instances, the centers of each cluster are calculated, being these centers the new *leaders,* and the process is repeated until a stable partition is found.

a) Given a set of data, and the values of the parameters ket, will be the partition produced by Leader Clustering algorithm the same as the K-means algorithm? Assume that the k instances initials selected as leaders/centroids for both algorithms are the same.
Consider different cases depending on the parameter t, justify your answer.

b) Which of the two methods will be better at dealing with isolated values (outliers). Explain your response.

3 values

5. Consider the following array of documents characterized by the following Expressions Relevant (ERs) as attributes:

	ER1	ER2	ER3	ER4	ER5	ER6
Doc1	two	0	0	0	1	10
Doc2	two	0	14	0	1	10
Doc3	two	0	22	3	1	13
Doc4	3	0	9	3	1	13
Doc5	two	0	81	0	1	10
Doc6	two	1	30	0	1	10
Doc7	two	0	200	0	1	10
Doc8	1	1	100	0	1	10
Doc9	two	0	1	0	1	10

INSTITUTO POLITÉCNICO DO PORTO

a) Of the attributes presented above, are there any that are irrelevant? Why?

b) Which attribute would contribute most to document clustering?
considering the content residing in them? Why?

c) Between attributes ER4 and ER6, which one do you find more informative? Why?

d) With a view to calculating a matrix of similarities between documents, it would weigh accordingly Is the influence of each ER different in calculating these similarities? If yes, what is the principle basis of such a criterion?

END