Data Mining 2

Module 2 - 2020/2021

Name _____ **Surname** _____ **ID:** _____ **Test id. AUTO**

Q1. Which one of the following statements can be considered true for a Naive Bayes Classifier?

1) the independence assumption always holds

2) robust to irrelevant attributes

3) suffers from isolated noise points

4) none of the others is true

5) is based on independece assumption

A1. _____

N.B.: this question can have more than one correct answer

Q2. Given the dataset and query record in the Figure which is the outcome of a Naive Bayes classifier?

	x1	x2	class
0	A	0	yes
1	A	1	yes
2	B	1	yes
3	A	1	no
4	A	0	no
5	B	0	no

q = ['A', 0]

A2. _____

Q3. Which is the class returned by a Naive Bayes Classifier which is classifying an instance with respect to a unique continuous attribute with value 20 knowking that the attribute has the following statistics?

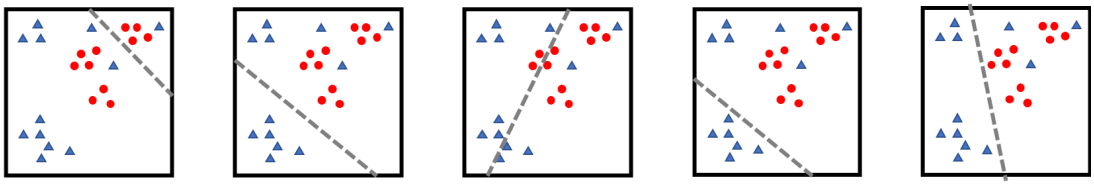
For Yes the mean is 15 and the standard deviation is 4. For No the mean is 36 and the standard deviation is 15.

A3. _____

Q4. In SVM in what consist the kernel trick?

A4. _____

Q5. Which hyperplane is better?



1) A

2) C

3) D

4) E

5) B

A5. _____

Q6. What happens in NN when the learnign rate parameter is close to 1?

1) The classification will be more likely 1

2) The new weight is mostly influenced by the value of the old weight

3) The new weight is mostly influenced by the current adjustment

4) The learning rate initially adapt in subsequent iterations

5) None of the others

A6. _____

Q7. Which is a correct usage of the validation set in training a NN?

1) Decide when to stop training by monitoring the error.

2) Handle missing attributes.

3) Update the weights.

4) Test the performance of the neural network.

5) None of the others.

A7. _____

Q8. Given 3 independent models for the same data with performance $e_1 = e_2 = e_3 = 0.4$, it is better to use a single model alone or to make a bagging with all the three models? Which is the error of the model ensemble? (example of answer: Single, 0.5.)

A8. _____

Q9. Which type of ensemble manipulate input features?

1) Engineering

2) Bagging

3) Random Forest

4) Boosting

5) None of the others

A9. _____

Q10. Is in boosting the generation of different samples independent?

1) No

2) No, unless the Gini Index is used

3) Yes

4) Yes, if the datsaet is imbalanced.

5) None of the others.

A10. _____

Q11. Given the dataset in the Figure run the first iteration of AdaBoost, find the best split, and fill the column 'norm weight' with the normalized new weights.

plan	sex	minutes	churn	weight	new weight	norm weight
travel	F	90	N			
travel	F	130	Y			
travel	M	70	N			
travel	M	80	N			
normal	M	90	Y			
normal	M	120	Y			
normal	F	100	Y			
normal	F	110	N			
travel	F	100	N			

Gain Function = Misclassification Error Z =

Split by

Error =

Alpha =

w^{j+1} = misclassified

w^{j+1} = classified correctly