NOVA IMS | moodle

202021 - Data Mining - S1

Started on	Monday, 25 January 2021, 3:04 PM
State	Finished
Completed on	Monday, 25 January 2021, 3:38 PM
Time taken	33 mins 13 secs

Question **1**

Complete

Marked out of 10.00

Flag question

In the context of relevant analysis for CRM, we talked about cell-based segmentation. The situation is the following: I manage a telco company, below you can see a segmentation of my customers based on the evolution of the number of minutes used during a year. Q means quartile, dot (.) represents customers that are not part of the database in one of the periods and the letters are references, to identify the cells, that you can use in your explanation. What's your analysis of the evolution of my business? Please, be detailed and, among other things, explain which cells you consider to be more relevant for the analysis.

Contigency Table		Period 06.2013/06.2014					Totals
		.	<Q1	[Q1,Q2[	[Q2,Q3[	>=Q3	
Period 12.2012 / 12.2013	.	0 a	30603 b	17050 c	8815 d	6427 e	62895 13.5%
	<Q1	28734 f	55411 g	13600 h	2772 i	178 j	100695 21.6%
	[Q1, Q2[	14834 k	15506 l	52421 m	16838 n	1097 o	100696 21.6%
	[Q2, Q3[	6540 p	3450 q	19346 r	59756 s	11608 t	100700 21.6%
	>=Q3	3765 u	760 v	1835 w	14733 x	79604 y	100697 21.6%
Totals		53873 11.6%	105730 22.7%	104252 22.4%	102914 22.1%	98914 21.2%	465683 100%

79% of the >=Q3 clients shifted quartile between periods.

78,3% of the <Q1 clients also shifted their quartile between periods.
11,6% of the clients didn't buy anything on the 2nd period, even if they bought on the previous period
On the other hand, 6427 (cell e) didn't buy anything on the 1st period, but are now over the 3rd quartile.
All cells havd some importance in this context, depe Google Chrome tion one wants to answer.

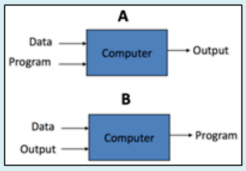
Question **2**

Correct

Marked out of 2.50

Flag question

During the classes we discussed the differences between traditional programming and machine learning, to highlight the differences between the two we used the figure below:



We can say that **A** represents machine learning;

Select one:

- a. I don't want to answer this question;
- b. False ✔
- c. True

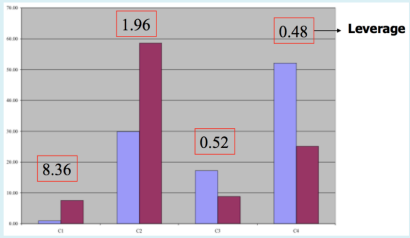
Question **3**

Incorrect

Marked out of 5.00

Flag question

During the classes we talked about the concept of leverage, thus we can say:



Select one:

- a. Cluster C2 is the cluster with the most valuable customers;
- b. I don't want to answer this question; ✖
- c. Cluster C1 is the cluster with the most valuable customers;
- d. Cluster C4 is the most valuable cluster in the database;
- e. Cluster C1 is the most valuable cluster in the database;

Question 4

Correct

Marked out of 5.00

Flag question

During the training of a SOM, using a learning rate of 0.6, neighborhood function of 0, and input pattern x1, and the initial weights of neurons N, shown in the table; which of the following is true?

x1	N1	N2
1	0.2	0.8
1	0.6	0.4
0	0.7	0.7
0	0.9	0.3

Select one:

- a. I don't want to answer this question;
- b. N2 will be updated to [0.92; 0.76; 0.28; 0.12]; ✔
- c. N2 will be updated to [0.68; 0.84; 0.28; 0.36];
- d. N1 will be updated to [0.76; 0.24; 0.8; 0.96];
- e. None of the options is correct;

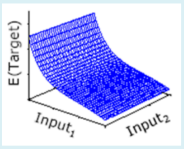
Question 5

Correct

Marked out of 5.00

Flag question

Given that we are trying to create a predictive model that allows us to make good predictions for variable E(Target), and assuming that we can only use one of the two input variables presented in the graphic, which one should we choose to include in the model?



Select one:

- a. Both variables are important, none should be discarded;
- b. I should choose input 2;
- c. I should choose input 1; ✔
- d. I don't want to answer this question;
- e. The information provided is not sufficient to make a decision;

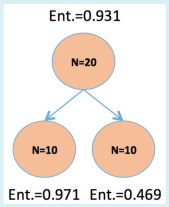
Question 6

Correct

Marked out of 2.50

Flag question

Given the classification tree on the figure, the proposed partition should be rejected because it increases the entropy.



Select one:

- a. False ✔
- b. True
- c. I don't want to answer this question;

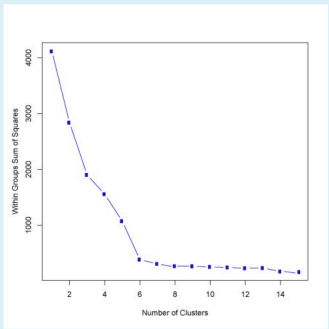
Question 7

Correct

Marked out of 2.50

Flag question

Given the figure the optimal number of clusters is 3.



Select one:

- a. I don't want to answer this question;
- b. False ✔
- c. True

Question 8

Incorrect

Marked out of 2.50

Flag question

I am using a hierarchical_clustering algorithm, taking into account the distance matrix shown, the first pair to be grouped should be BA-TO

	BA	FI	MI	NA	RM	TO
BA	0	662	877	255	412	996
FI	662	0	295	468	268	400
MI	877	295	0	754	564	138
NA	255	468	754	0	219	869
RM	412	268	564	219	0	669
TO	996	400	138	869	669	0

Select one:

- a. True ✖
- b. I don't want to answer this question;
- c. False

Question 9

Incorrect

Marked out of 5.00

Flag question

I want to normalize the following dataset:

Id	VarX
1	2.1
2	1.3
3	3.5
4	2.8
5	6.3
6	5.4

Select one:

- a. Using MinMax, the value of VarX for the record with id 2 is 1;
- b. I don't want to answer this question; ✖
- c. Using Z-score, the value of VarX for the record with id 2 is 0;
- d. Using MinMax, the value of VarX for the record with id 2 is 0;
- e. Using Z-score, the value of VarX for the record with id 2 is 1;

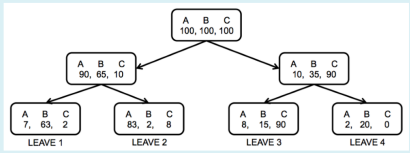
Question 10

Incorrect

Marked out of 5.00

Flag question

I've developed a classification tree (below) based on the "DDT" method, studied in class, and using as discriminant measure: f(A) = ΣCi/n where A is an attribute, Ci is the number of samples correctly classified by the majority class, n is the total number of examples. We can say that the error rate of this tree is:



Select one:

- a. 20%;
- b. 15%;
- c. None of the above;
- d. I don't want to answer this question; ✖
- e. 13%;

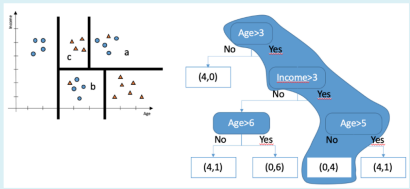
Question 11

Incorrect

Marked out of 5.00

Flag question

I've developed a predictive model using classification_trees. This model allows me to classify my customers based on their propensity to buy a certain product I'm promoting. Given the tree below we can say that an individual following the path highlighted on the tree (right side) belongs to group:



Select one:

- i. I don't want to answer this question;
- ii. None of the above;
- iii. (a); ✖
- iv. (c);
- v. (b);

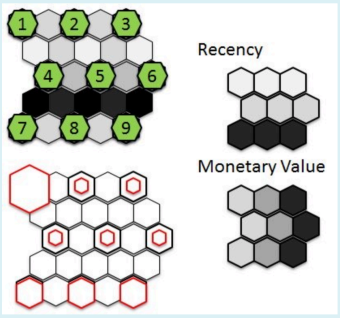
Question 12

Incorrect

Marked out of 5.00

Flag question

I've done a segmentation based on a SOM with 9 neurons, the results are shown in the figure (black=highest, white=lowest). Taking into account the information provided by the figure (Matrix U, 2 component planes, and a Hit Map) we can say that:



Select one:

- a. Neurons 1 and 4 represent the older clients;
- b. There are 2 main clusters in the dataset;
- c. I don't want to answer this answer;
- d. The information provided is not sufficient to make a decision.
- e. Neurons 1, 2 and 3 represent customers that have not bought for a long time; ✖

Question 13

Correct

Marked out of 5.00

Flag question

I've done an RFM analysis, after doing the analysis I've ordered the individuals based on the RFM cell to which they belong, next by days since the last buy, frequency of purchase, and finally the monetary value (figure). Now I'm a bit confused with the results, specifically I don't know if the analysis was correctly done. Based on the table we can conclude that:

Custid	Recency	Freq	Monetary	RFM
10341	52	151	2648	555
1650	52	150	2644	555
4790	52	147	2635	555
9210	52	162	3648	554
3310	52	160	3621	554
5289	52	158	3607	554

Select one:

- a. The customers from cell 554 should have a Freq value below the one of the 555 cell, thus the analysis is incorrect;
- b. The customers from cell 554 should have a monetary value below the one in the 555 cell, thus the analysis is incorrect; ✔
- c. I don't want to answer this question;
- d. The table does not allow any conclusion on the correctness of the analysis;
- e. The analysis was correctly done;

Question 14

Incorrect

Marked out of 5.00

Flag question

I've trained a SOM, represented by the two neurons in the table, now I need to classify the three individuals shown below (x1, x2, x3). We can say that:

Neurons

N1	N2
0.1	0.9
0.2	0.9
0.8	0.1
0.9	0.2

Individuals

x1	x2	x3
1	0	1
1	0	1
0	1	1
0	1	1

Select one:

- a. The individual x2 has the worst representation in the SOM;
- b. The individual x1 has the worst representation in the SOM;
- c. The individual x3 has the worst representation in the SOM;
- d. I don't want to answer this question; ✖
- e. There is not enough information to answer this question;

Question 15

Incorrect

Marked out of 5.00

Flag question

In which of the following cases will K-Means clustering fail to give good results?

Select one:

- a. A dataset with spherical-shaped clusters;
- b. I don't want to answer this question;
- c. None of the above;
- d. A dataset with outliers;
- e. Both of the above; ✖

Question **16**

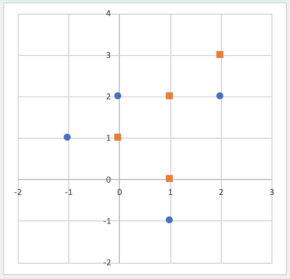
Correct

Marked out of 5.00

Flag question

Suppose, you have given the following data (table and scatter plot below) where x and y are the 2 input variables and Class is the dependent variable. Suppose now that you want to predict the class of a new data point x=1 and y=1 using Euclidian distance in 3-NN. In which class this data point belongs to?

x	y	Class
-1	1	0
0	2	0
1	-1	0
2	2	0
0	1	1
1	0	1
1	2	1
2	3	1



Select one:

- a. Belongs to both classes;
- b. The information provided is not sufficient to make a decision;
- c. The class represented by the orange square; ✔
- d. The class represented by the blue circle;
- e. I don't want to answer this question;

Question **17**

Incorrect

Marked out of 2.50

Flag question

t-SNE can be used as a dimension reduction algorithm just like PCA.

Select one:

- a. I don't want to answer this question;
- b. True ✖
- c. False

Question **18**

Correct

Marked out of 5.00

Flag question

The U-Matrix represents:

Select one:

- a. Distances between neurons; ✔
- b. Neurons;
- c. Records in a database;
- d. I don't want to answer this question;
- e. None of the above;

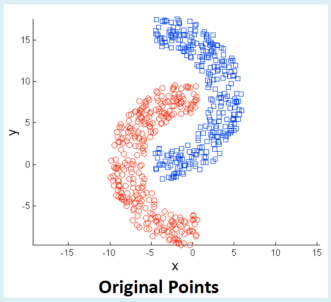
Question **19**

Correct

Marked out of 2.50

Flag question

We can say that k-means would be a good option to cluster this dataset.



Select one:

- a. True
- b. False ✔
- c. I don't want to answer this question;

Question **20**

Correct

Marked out of 5.00

Flag question

We use principal component analysis to transform a data set from a given basis (space), into an equivalent set, in a new space. This new space:

Select one:

- a. Has the same number of components as the original space; ✔
- b. Has a lower number of components than the original space;
- c. Has a higher number of components than the original space, but if the intrinsic dimension of the data is smaller, we can disregard these extra components, and stick with a number of components equal to the original space;
- d. Has a higher number of components than the original space;
- e. I don't want to answer this question;

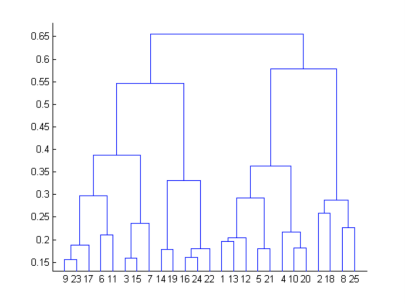
Question **21**

Correct

Marked out of 5.00

Flag question

What is the most appropriate number of clusters for the dataset represented by the following dendrogram:



Select one:

- a. 25
- b. I don't want to answer this question;
- c. 7
- d. 6
- e. 4 ✔

Question **22**

Correct

Marked out of 5.00

Flag question

Which of the following algorithms is most sensitive to outliers?

Select one:

- a. I don't want to answer this question;
- b. None of these algorithms is especially sensitive to outliers;
- c. K-means clustering algorithm; ✔
- d. K-medoids clustering algorithm;
- e. DBSCAN;