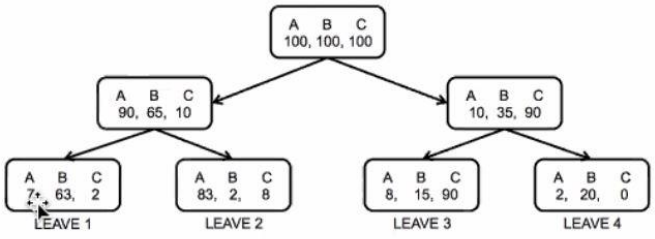
c) Records in a database;
d) None of the above;

2. 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:

a) 13%;
b) 15%;
c) 20%;
d) None of the above;



3. 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:
   a) Has a higher number of components than the original space;
   b) Has the same number of components as the original space;
   c) Has a lower number of components than the original space;
   d) 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;

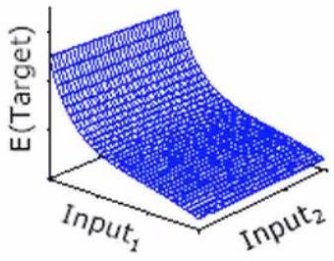
4. I've trained a SOM, represented by the two neurons in the table, now I need to classify the three individuals shown below (x1, x2, x3). We can say that:

   a) The individual x1 has the worst representation in the SOM;
   b) The individual x2 has the worst representation in the SOM;
   c) The individual x3 has the worst representation in the SOM;
   d) There is not enough information to answer the question;

Neurons		Individuals		
N1	N2	x1	x2	x3
0.1	0.9	1	0	1
0.2	0.9	1	0	1
0.8	0.1	0	1	1
0.9	0.2	0	1	1

5. 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?

   a) I should choose input 1;
   b) I should choose input 2;
   c) Both variables are important, none should be discarded;
   d) The information provided is not sufficient to make a decision;



6. 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:

   a) The analysis was correctly done;
   b) The customers from cell 554 should have a monetary value below the one in the 555 cell, thus the analysis is incorrect;
   c) The customers from cell 554 should have a Freq value below the one of the 555 cell, thus the analysis is incorrect;
   d) The table does not allow any conclusion on the correctness of the analysis;

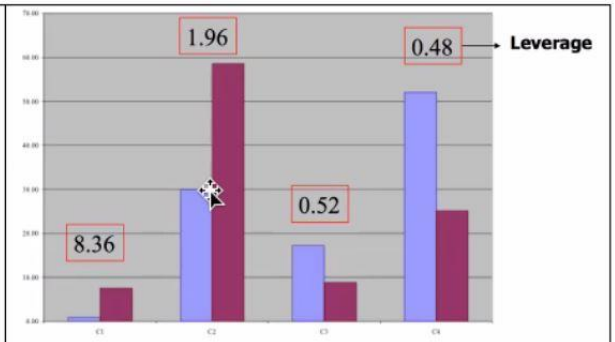
Custid	Recency	Freq	Monetary	RFM
10341	52	151	2648	555
1650	52	150	2644	555
4790	52	147	2635	555
9210	52	162	3648	554
3310	52	160	3621	554
5289	52	158	3607	554

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

   d) The table does not allow any conclusion on the correctness of the analysis;

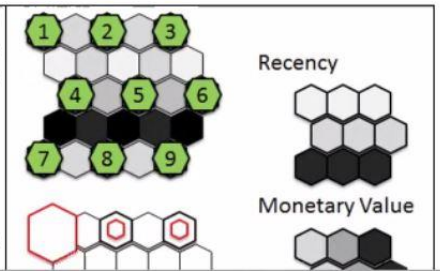
7. During the classes we talked about the concept of leverage, thus we can say:

   a) Cluster C4 is the most valuable cluster in the database;
   b) Cluster C1 is the most valuable cluster in the database;
   c) Cluster C2 is the cluster with the most valuable customers;
   d) Cluster C1 is the cluster with the most valuable customers;



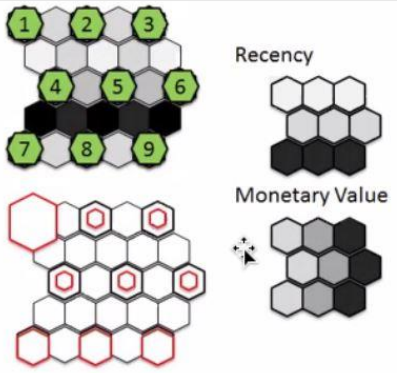
8. I've done a segmentation based on a SOM with 9 neurons, the results are shown in the figure (black=highest, white=lowest). Taking into account the information provided by the figure (Matrix U, 2 component planes, and a Hit Map) we can say that:

   a) There are 2 main clusters in the dataset;
   b) Neurons 1 and 4 represent the older clients;
   c) Neurons 1, 2 and 3 represent customers that have not buy for a long time;



8. 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:

   a) There are 2 main clusters in the dataset;
   b) Neurons 1 and 4 represent the older clients;
   c) Neurons 1, 2 and 3 represent customers that have not buy for a long time;
   d) The information provided is not sufficient to make a decision.



9. I want to normalize the following dataset:

   a) Using MinMax, the value of VarX for the record with id 2 is 0;
   b) Using MinMax, the value of VarX for the record with id 2 is 1;
   c) Using Z-score, the value of VarX for the record with id 2 is 0;
   d) Using Z-score, the value of VarX for the record with id 2 is 1;

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

10. Which of the following algorithms is most sensitive to outliers?

    a) K-means clustering algorithm;
    b) K-medoids clustering algorithm;
    c) DBSCAN;
    d) None of these algorithms is especially sensitive to outliers;

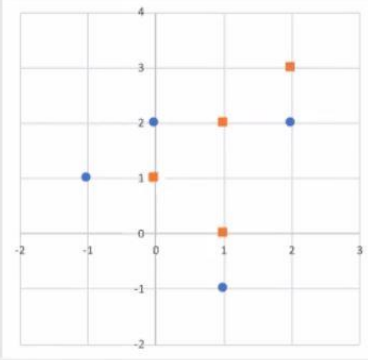
11. Suppose, you have given the following data (table and scatter plot below) where x and y are the 2 input variables and Class is the dependent variable. Suppose now that you want to predict the class of a new data point x=1 and y=1 using Euclidian distance in 3-NN. In which class this data point belongs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

10. Which of the following algorithms is most sensitive to outliers?

    a) K-means clustering algorithm;
    b) K-medoids clustering algorithm;
    c) DBSCAN;
    d) None of these algorithms is especially sensitive to outliers;

11. 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



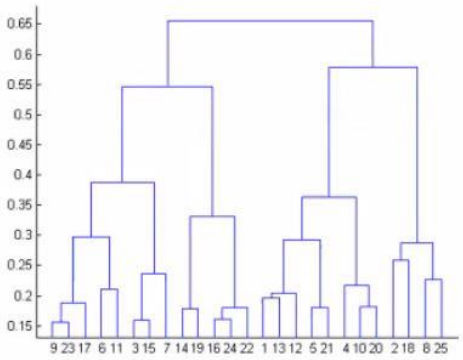
a) The class represented by the orange square;
b) The class represented by the blue circle;
c) Belongs to both classes;
d) The information provided is not sufficient to make a decision;

12. 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?

a) N1 will be updated to [0.76; 0.24; 0.8; 0.96];
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) None of the options is correct;

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

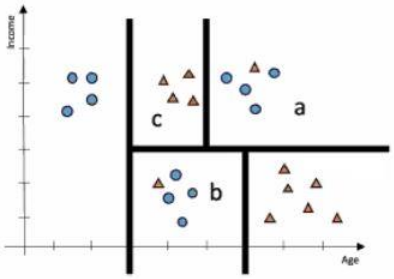
13. What is the most appropriate number of clusters for the dataset represented by the following dendrogram:

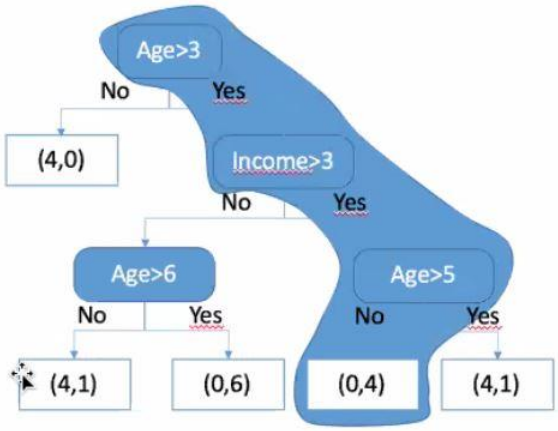


a) 4
b) 6
c) 7
d) 25

14. 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:

14. 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:





a) a);
b) b);
c) c);
d) None of the above;

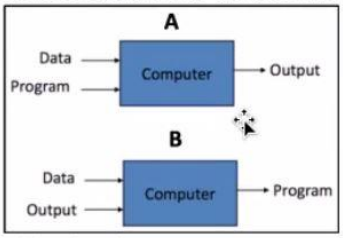
15. In which of the following cases will K-Means clustering fail to give good results?

a) A dataset with outliers;
b) A dataset with spherical-shaped clusters;
c) Both of the above;
d) None of the above;

NOVA IMS Information Management School

True or false Questions

1. 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:

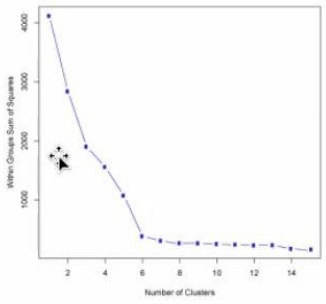


We can say that **A** represents machine learning;

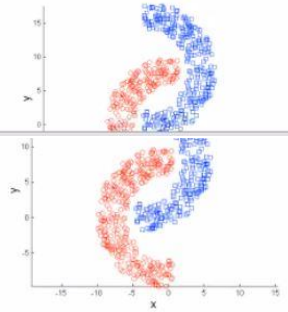
2. Given the figure the optimal number of clusters is 3.

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

2. Given the figure the optimal number of clusters is 3.



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

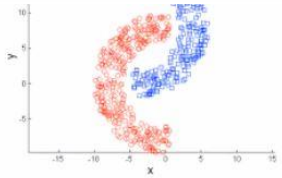


Original Points

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

5. 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

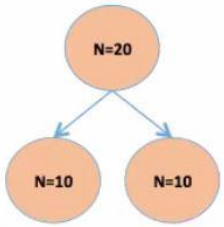


Original Points

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

5. 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

6. Given the classification tree on the figure, the proposed partition should be rejected because it increases the entropy.	Ent.=0.931  Ent.=0.971 Ent.=0.469

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

School

Data Mining – 2020-2021 1st Call Exam

Questions Point Values (100 points Total): M.C. 5 (-1.25); T.F. 2,5 (-1.25); Essay 10

Duration: 45 minutes

NAME:	Nº.

Answer Matrix:

I-1	a	b	c	d
I-2	a	b	c	d
I-3	a	b	c	d
I-4	a	b	c	d
I-5	a	b	c	d
I-6	a	b	c	d
I-7	a	b	c	d
I-8	a	b	c	d
I-9	a	b	c	d
I-10	a	b	c	d
I-11	a	b	c	d
I-12	a	b	c	d
I-13	a	b	c	d
I-14	a	b	c	d
I-15	a	b	c	d
II-1	T	F		
II-2	T	F		
II-3	T	F		
II-4	T	F		
II-5	T	F		
II-6	T	F		

RESPOSTAS

1	Distance between Neuros – (b)
2	?
3	B
4	C
5	A
6	The last 3 customers should have RFM 555 – i think the answer is B
7	D
8	A
9	A
10	A
11	A
12	B
13	A – 4 clusters
14	C
15	A
True or False Questions	
1	F
2	F
3	F
4	F
5	F
6	T