

Student Name | Nome do Aluno:

Number | Número:

ISEL - ADEETC

EXAM #1 | 1ST PHASE | 1ST SEMESTER | 2016 / 2017

SUBJECT: AMD – APRENDIZAGEM E MINERAÇÃO DE DADOS

COURSE: MESTRADO EM ENGENHARIA DE REDES DE COMUNICAÇÃO E MULTIMÉDIA

DURATION: 2'00"

3.FEV.2017

ATTENTION

ATENÇÃO

- All test sheets must be legibly identified with the name and number of the student.
Todas as folhas devem ser identificadas, de forma legível, com o nome e número do aluno.
- The questions can be answered with consultation from paper documentation.
As questões podem ser respondidas com consulta de documentação em papel.
- The questions should be answered directly in the test sheets.
As questões devem ser respondidas diretamente na folha do enunciado.
- The multiple-choice questions discount incorrect answers.
As questões de escolha múltipla descontam respostas incorretas.
- The interpretation of the questions is part of the evaluation.
A interpretação do enunciado faz parte da avaliação.
- The readability of the answer is part of the evaluation.
A legibilidade da resposta faz parte da avaliação.

GROUP . QUESTION	(ITEM) GRADE	TOTAL GROUP
I . 1	(a) 1; (b) 1.5; (c) 1.5; (d) 1.5	14
I . 2	(a) 1; (b) 1; (c) 1	
I . 3	(a) 1; (b) 1.5; (c) 1.5; (d) 1.5	
II . 1	(a) 1.5; (b) 1.5; (c) 1.5	6
II . 2	(a) 1.5	

GROUP I (14 VALUES)

1. Consider the evaluation process of data mining techniques.

Considere o processo de avaliação das técnicas de mineração de dados (*data mining*).

- (a) What is the assumption that originates the construction of differentiated datasets for training and for testing?

Que é o pressuposto que origina a construção de *datasets* diferenciados para treino e para teste?

- (b) Assume numeric attribute K and class C with the dataset: <1, X>, <3, X>, <4, X>, <2, Y>, <4, Y>, <6, Y>.

Mark each statement with an X either for true (T) or false (F).

Admita o atributo numérico K e classe C com o *dataset*: <1, X>, <3, X>, <4, X>, <2, Y>, <4, Y>, <6, Y>.

Assinale cada afirmação com X para indicação de verdadeiro (T) ou falso (F).

T	F	Statement Afirmação
		The stratified holdout with reservation of 1/3 should accept to reserve (from the dataset), as the testing set, the instances <4, X> and <4, Y>. O <i>stratified holdout</i> (reserva estratificada) com reserva de 1/3 deve aceitar que se reserve (do <i>dataset</i>), como conjunto de teste, as instâncias <4, X> e <4, Y>.
		The <i>leave-one-out</i> with a classifier that classifies always in accordance with the "majority rule" originates 100% of correct classifications (i.e., 0% error). O <i>leave-one-out</i> com um classificador que classifique sempre de acordo com a "regra da maioria" origina uma avaliação com 100% de classificações corretas (i.e., 0% erro).
		Using <i>bootstrap</i> it is to be expected that the testing set has only 2 instances (those not selected for the training) although it may have 3, 4, or 5 instances but never 6. Usando o <i>bootstrap</i> é de esperar um conjunto de teste com só 2 instâncias (aqueles não escolhidas para o treino) embora possa ter 3, 4 ou 5 instâncias mas nunca 6.
		With this <i>dataset</i> , the 2 × <i>stratified 3-fold cross validation</i> will have to repeat, for sure, 2 times the same training set throughout the full evaluation process. Com este <i>dataset</i> , o processo 2 × <i>stratified 3-fold cross validation</i> terá que repetir, de certeza, 2 vezes o mesmo conjunto de treino durante todo o processo de avaliação.
		During its execution the 2 × <i>stratified 3-fold cross validation</i> builds 6 classifiers. Durante a execução o 2 × <i>stratified 3-fold cross validation</i> constrói 6 classificadores.
		In <i>leave-one-out</i> with the repeating of the process we attenuate the evaluation bias. No <i>leave-one-out</i> o repetir do processo atenua o enviesamento (<i>bias</i>) da avaliação.
		To calculate the Kappa statistics it is useful to know the confusion matrix. Para calcular a estatística Kappa é útil conhecer a matriz de confusão.
		The ROC (<i>Receiver Operating Characteristics</i>) analysis is based on Kappa statistics. A análise ROC (<i>Receiver Operating Characteristics</i>) baseia-se na estatística Kappa.

Student Name | Nome do Aluno:

Number | Número:

- (c) Justify your (previous) answer. Justifique a sua resposta (anterior).

Statement: "The stratified holdout with reservation of 1/3 should accept to reserve (from the dataset), as the testing set, the instances <4, X> and <4, Y>."

Justification:

- (d) Justify your (previous) answer to the following statement.

Justifique a sua resposta (anterior) à afirmação que se segue.

Statement: "Using bootstrap it is to be expected that the testing set has only 2 instances (those not selected for the training) although it may have 3, 4, or 5 instances but never 6."

Justification:

2. Consider the relational model T1(A, L, U), T2(B, N, C, D) where primary keys are underlined.

Considere o modelo relacional T1(A, L, U), T2(B, N, C, D) onde as chaves primárias estão sublinhadas.

- (a) Complete the SQL statement to generate the dataset DS(A, B, C, D) that combines all the tuples of T1 with all those of T2 that satisfy the relation $L \leq N \leq U$.

Complete a diretiva SQL para gerar o *dataset* DS(A, B, C, D) que combina todos os tuplos de T1 com todos os de T2 que satisfaçam a relação $L \leq N \leq U$.

CREATE VIEW DS (

- (b) Represent DS considering the following tuples: <young, 10, 25>, <adult, 26, 50>, <a@x.y, 20, Lisboa, yes>, <b@x.y, 22, Porto, no>, <b@w.z, 40, Lisboa, yes>, <c@x.y, 50, Astana, yes>, <c@w.z, 22, Almaty, yes>.

Represente DS considerando os tuplos: <young, 10, 25>, <adult, 26, 50>, <a@x.y, 20, Lisboa, yes>, <b@x.y, 22, Porto, no>, <b@w.z, 40, Lisboa, yes>, <c@x.y, 50, Astana, yes>, <c@w.z, 22, Almaty, yes>.

- (c) You want to build a decision tree with ID3. Using DS (see item a) is there any attribute you must eliminate?

Pretende construir árvore de decisão com ID3. Se usar DS (alínea a) há algum atributo que deva eliminar?

Justification:

3. Consider the (simplistic) naïve Bayes model.

Considere o modelo (simplista) naïve de Bayes.

(a) Explain the assumption(s) that originates the naïve (simplistic) perspective of the model?

Explique o(s) pressuposto(s) que, esse modelo, dá origem à sua perspectiva naïve (simplista)?

The assumption(s):

The explanation of the assumption(s):

(b) Assume a dataset with the attributes X, Y, C with binary domain (i.e., 0 or 1) with C being the class. Knowing that $P(X=0 | C=0)=0.3$ e $P(Y=1 | C=0)=0.4$ e $P(C=0)=0.2$, mark with X for true (T) or false (F).

Admita um dataset com os atributos X, Y, C de domínio binário (i.e., 0 ou 1) sendo C a classe. Sabendo que $P(X=0 | C=0)=0.3$ e $P(Y=1 | C=0)=0.4$ e $P(C=0)=0.2$, assinale, com X, verdadeiro (T) ou falso (F).

T	F	Statement Afirmação
		The naïve Bayes value for the à-priori probability for the event C=0 is 0.3×0.4 O valor do naïve Bayes para a probabilidade à-priori do evento C=0 é 0.3×0.4
		The $\Pr(C=0 X=0, Y=1)$ is proportional to $0.3 \times 0.4 \times 0.2$ A $\Pr(C=0 X=0, Y=1)$ é proporcional a $0.3 \times 0.4 \times 0.2$
		If $\Pr(X=0, Y=1 C=0) = 0.24$, then X and Y are not conditionally independent given C Se $\Pr(X=0, Y=1 C=0) = 0.24$, X e Y não são condicionalmente independentes dado C
		If $\Pr(C=1 X=0, Y=1) \approx 0.5$, then the class value for $\langle X=0, Y=1 \rangle$ will be C=1 Se $\Pr(C=1 X=0, Y=1) \approx 0.5$, então o valor da classe quando $\langle X=0, Y=1 \rangle$ será C=1
		If $\Pr(X=1 C=0) = 0$, then applying Laplace estimator gets $P(X=1 C=0) = (0 + 1) = 1$ Se $\Pr(X=1 C=0) = 0$, então ao aplicar estimador Laplace $P(X=1 C=0) = (0 + 1) = 1$

(c) Justify your (previous) answer. Justifique a sua resposta (anterior).

Statement: If $P(X=0, Y=1 | C=0) = 0.24$, then X and Y are not conditionally independent given C

Justification (show all calculations):

(d) Justify your (previous) answer. Justifique a sua resposta (anterior).

Statement: If $P(C=1 | X=0, Y=1) \approx 0.5$, then the class value for $\langle X=0, Y=1 \rangle$ will be C=1

Justification (show all calculations):

GROUP II (6 VALUES)

1. The next figure presents values and classifications for 10 instances, with a real (continuous) attribute, R, a nominal attribute, N, and an attribute C that classifies in one of two class values: – or +.

A figura seguinte apresenta valores e classificações para 10 instâncias, com um atributo real (contínuo), R, um atributo nominal, N, e um atributo C que classifica num de dois valores de classe: – ou +.

C	–	–	+	–	+	–	+	+	–	+
R	– 0.1	0.7	1.0	1.0	2.0	2.5	3.2	3.2	4.1	4.9
N	a	a	b	a	a	a	b	a	b	b

- (a) Identify the intervals generated by the discretization process of the 1-R method applied to R attribute.

Identifique os intervalos gerados pelo processo de discretização do método 1-R aplicado ao atributo R.

R split point

- (b) What is the error of the 1R rules that consider the attribute N as follows: [if N=a then “–”] [if N=b then “+”] ?

Qual é o erro das regras 1R que consideram o atributo N como se segue: [if N=a then “–”] [if N=b then “+”] ?

The error of the 1R rules is: _____

Justification (show all the calculations):

- (c) Assume a classifier that applies the majority rule (i.e., zero-R) and, in case of equality, classifies as +. Apply it the leave-one-out validation and show the total number of errors and successes that we get.

Assuma um classificador que aplica a regra da maioria (i.e., zero-R) e, em caso de igualdade, classifica como +. Aplique-lhe o *leave-one-out validation* e mostre o total de erros e de sucessos que daí resulta.

The total number of errors is: _____ and the total number of successes is: _____

Justification (show all the calculations):

2. Assume the dataset with the Boolean (i.e., values 0 and 1) attributes X, Y, Z and W representing an identifier.

Admita o *dataset* com os atributos Booleano (i.e., valores 0 e 1) X, Y, Z e U a representar um identificador.

X	Y	Z	W
1	1	0	a
0	0	1	a
0	1	0	c
0	1	0	a
1	0	0	b
0	0	1	b
1	1	1	a

- (a) Assume that each attribute X, Y, Z is a product; e.g., *milk* (for X), *bread* (for Y) and *cheese* (for Z). Also assume that W represents the “client ID” (i.e., “a”, “b” and “c” are distinct clients). Transform the dataset for market-basket-analysis where each “client ID” is a transaction and 1 means “product is in client’s basket”.

Assuma que cada atributo X, Y, Z é um produto; e.g., *leite* (para X), *pão* (para Y) e *queijo* (para Z). Assuma também que W representa o “ID de cliente” (i.e., “a”, “b” e “c” são clientes distintos). Transforme o *dataset* para “market-basket-analysis” onde cada “ID de cliente” é uma transação e 1 significa que o “produto está no basket (cesto de compras) do cliente”.

Using the .tab or the .basket Orange format the dataset for market-basket-analysis is given by:

transaction	item	value
1	X	1
1	Y	1
1	Z	0
1	W	a
2	X	0
2	Y	0
2	Z	1
2	W	a
3	X	0
3	Y	1
3	Z	0
3	W	c
4	X	0
4	Y	1
4	Z	0
4	W	a
5	X	1
5	Y	0
5	Z	0
5	W	b
6	X	0
6	Y	0
6	Z	1
6	W	b
7	X	1
7	Y	1
7	Z	1
7	W	a