

ISEL - ADEETC

EXAM #1 | 1ST PHASE | 1ST SEMESTER | 2019 / 2020

SUBJECT: AMD – APRENDIZAGEM E MINERAÇÃO DE DADOS

COURSE: MESTRADO EM ENGENHARIA DE REDES DE COMUNICAÇÃO E MULTIMÉDIA

DURATION: 1'45"

20.JAN.2020

## 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.

QUESTION	(ITEM) GRADE	TOTAL
1	(a) 1; (b) 1; (c) 1.5; (d) 1.5; (e) 1.5; (f) 1.5	8
2	(a) 2=[0.25, 0.25, 0.25, 0.25, 0.25, 0.25, 0.5]; (b) 2; (c) 2; (d) 2	8
3	(a) 1.5; (b) 1; (c) 1.5	4

1. Considere o método APRIORI e o dataset que identifica **ITEMS\_BOUGHT** no contexto dos restantes atributos.

<b>customerID</b>	<b>paymentID</b>	<b>DATE</b>	<b>ITEMS_BOUGHT</b>
1	1001	05/01/2020	{i1, i4, i5}
1	1031	08/01/2020	{i1, i2, i3, i5}
2	1050	08/01/2020	{i1, i2, i4, i5}
2	1011	08/01/2020	{i1, i3, i4, i5}
3	1023	08/01/2020	{i2, i3, i5}
3	1051	08/01/2020	{i2, i4, i5}
4	1088	08/01/2020	{i3, i4}
4	1015	10/01/2020	{i1, i2, i3}
5	1020	10/01/2020	{i1, i4, i5}
5	1004	10/01/2020	{i1, i2, i5}

- (a) Calcule o **suporte** de {i5} considerando **paymentID** como o “market basket”

**suporte** sendo **paymentID** o “market basket” (apresente todos os cálculos):

- (b) Calcule o **suporte** de {i2, i4}=>{i5} considerando **paymentID** como o “market basket”

**suporte** sendo **paymentID** o “market basket” (apresente todos os cálculos):

- (c) Sendo **paymentID** o “market basket”, **use o resultado** das 2 alíneas anteriores para calcular a **confiança** de {i2, i4}=>{i5} e de {i5}=>{i2, i4}

**confiança** de {i2, i4}=>{i5}:

**confiança** de {i5}=>{i2, i4}:

atenção: apresente todos os cálculos e em cada cálculo tem que usar o(s) resultado(s) das alíneas anteriores

- (d) Calcule o **suporte** de {i2, i4}=>{i5} considerando **customerID** como o “market basket”; cada item é tratado como variável binária (i.e., 1 se ocorre pelo menos num **paymentID** de cada **clientID**; e 0 c.c.)

**suporte** sendo **customerID** o “market basket” (apresente todos os cálculos):

- (e) Diga, justificando, quais os “frequent” 3-itemset gerados (pelo APRIORI) a partir dos “frequent” 2-itemset: {i1,i2}, {i1,i3}, {i2,i3}, {i2,i5}

**Justificação** (apresente todos os cálculos):

- (f) Apresente o dataset no formato .basket do Orange considerando **DATE** como o “market basket”

**Justificação** (apresente todos os cálculos):

2. Admita o dataset com atributos R, G, B e output Booleano C.

R	G	B	C
0	1	0	0
1	0	1	1
0	1	0	0
0	0	0	1
1	0	0	0
1	1	1	1
0	0	0	0

(a) Pretende prever C usando naive Bayes. Preencha a tabela apresentando todos os cálculos.

$P( R=0   C=0 ) =$	
$P( R=1   C=0 ) =$	
$P( G=1   C=0 ) =$	
$P( G=1   C=1 ) =$	
$P( C=0 ) =$	
$P( C=1 ) =$	
verosimilhança para: $P(C=1   <R=0, G=1, B=0>) =$	

(b) Utilizando o estimador de Laplace o valor de  $P(B=1 | C=0)$  será: \_\_\_\_\_

Apresentação dos cálculos

(c) NÃO utilize o estimador de Laplace e diga como será classificada a instância  $<R=0, G=1, B=1>$ : \_\_\_\_\_

Apresentação dos cálculos

(d) O naive Bayes admite como pressuposto que “todos os atributos são condicionalmente independentes dada a classe”. Justifique se este pressuposto é, ou não, satisfeito neste dataset.

riscar a opção incorreta: o pressuposto [ é | não-é ] satisfeito

Apresentar todos os cálculos

sugestão: considere  $P(R=1 | C=0)$  e  $P(G=1 | C=0)$

3. Considere o *dataset*:  $\langle R, G, 0 \rangle$ ,  $\langle R, B, 0 \rangle$ ,  $\langle R, B, 1 \rangle$ ,  $\langle K, G, 1 \rangle$  onde o último atributo é a classe (com valores “0” e “1”). Indique se cada afirmação seguinte é verdadeira ou falsa.

(a) Preencha a tabela seguinte indicando, com cruz, se cada afirmação é verdadeiro ou falso (V ou F).

V	F	Afirmiação
		O processo <i>stratified holdout</i> com reserva de 1/2 deve aceitar reservar (do <i>dataset</i> ), como conjunto de teste, as instâncias $\langle R, B, 0 \rangle$ e $\langle R, G, 0 \rangle$ .
		A aplicação (ao <i>dataset</i> ) do processo <i>leave-one-out</i> classificando segundo a regra da maioria origina uma avaliação com 0% de classificações corretas (i.e., 100% erro).
		Usando o <i>bootstrap</i> é de esperar que o conjunto de teste apenas tenha 1 instância (diferente das usadas no treino) embora possa ter 2 ou até 3 instâncias mas nunca 4.
		Com este <i>dataset</i> , o processo $1 \times$ <i>stratified 2-fold cross validation</i> terá que repetir, de certeza, 2 vezes o mesmo conjunto de treino durante todo o processo de avaliação.
		Este <i>dataset</i> tem 2 classes logo só existe uma forma de calcular os verdadeiros e falsos positivos; com 3 ou mais classes estes conceitos têm várias formas de calcular.
		Este <i>dataset</i> tem 2 classes pelo não é possível ter as métricas <i>F1-score</i> , <i>precision</i> e <i>recall</i> todas com valor 1.
		A estatística “kappa” (com a matriz de confusão) usa-se para se conseguir estender a avaliação binária (com 2 classes) para a avaliação multi-classe.
		Na avaliação de um classificador podem usar-se vários métodos para gerar os <i>dataset</i> de treino e teste, mas a métrica de avaliação tem que ser diferente para cada método.

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

Afirmiação: A aplicação (ao *dataset*) do processo *leave-one-out* classificando segundo a regra da maioria origina uma avaliação com 0% de classificações corretas (i.e., 100% erro)

Justificação:

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

Afirmiação: Usando o *bootstrap* é de esperar que o conjunto de teste apenas tenha 1 instância (diferente das usadas no treino) embora possa ter 2 ou até 3 instâncias mas nunca 4.

Justificação: