Departamento de Engenharia Informática
Faculdade de Ciências e Tecnologia
Universidade de Coimbra

Tópicos de Ciência de Dados

6 de Janeiro, 2023
Exame Época Normal
Duração: 2:30h

Observe que:
Deverá realizar as partes A e B em folhas separadas.
Exame com consulta
Qualquer tentativa de fraude conduzirá à anulação da prova para todos os intervenientes.

Parte A:

1 – Considere o seguinte conjunto de variáveis categóricas adquiridas de um conjunto de 11 doentes:

Hipertensão	S	S	S	S	S	N	S	S	N	N	S
Tabagismo	S	S	N	N	S	N	S	S	S	S	S
Colesterol	S	N	S	N	N	N	N	N	N	S	N
Antecedentes familiares	S	S	S	S	S	N	N	N	N	N	N
Risco Cardiovascular	A	A	A	A	B	B	B	A	A	A	A

Assumindo que S=”Sim”, N=”Não” e que A=”Alto”, B=”Baixo”, considere que lhe é pedido que projete um preditor do risco cardiovascular usando os atributos “Hipertensão”, “Tabagismo”, “Colesterol” e “Antecedentes Familiares”. Nesse contexto, responda às seguintes questões:

a) Quais são as vantagens decorrentes da redução da dimensionalidade do vetor de atributos? Que metodologias dispõe para o fazer?
b) Apresente o ranking de atributos usando a metodologia Goodman Kruskal Lambda. Deverá apresentar todos os cálculos necessários.

2 – Considere o conjunto de pontos (x,y) das classes C1 e C2, respetivamente:

Tabela 1: Pontos das classes C1 e C2.

Classe C1 (x,y)	Classe C2 (x,y)
P1=(4, 5) P2=(3, 6) P3=(4, 8) P4=(8, 10) P5=(6, 1) P6=(18, 5) P7=(8, 16) P8=(3, 0) P9=(18, 18) P10=(19, 14) P11=(8, 9) P12=(2, 11) P13=(5, 4) P14=(8, 9) P15=(11, 19) P16=(5, 10)	P26=(6, 2) P27=(13, 14) P28=(25, 31) P29=(26, 28) P30=(32, 29)

P17=(12, 10) P18=(14, 4) P19=(4, 9) P20=(2, 12) P21=(28, 32) P22=(22, 33) P23=(22, 30) P24=(34, 16) P25=(15, 20)	

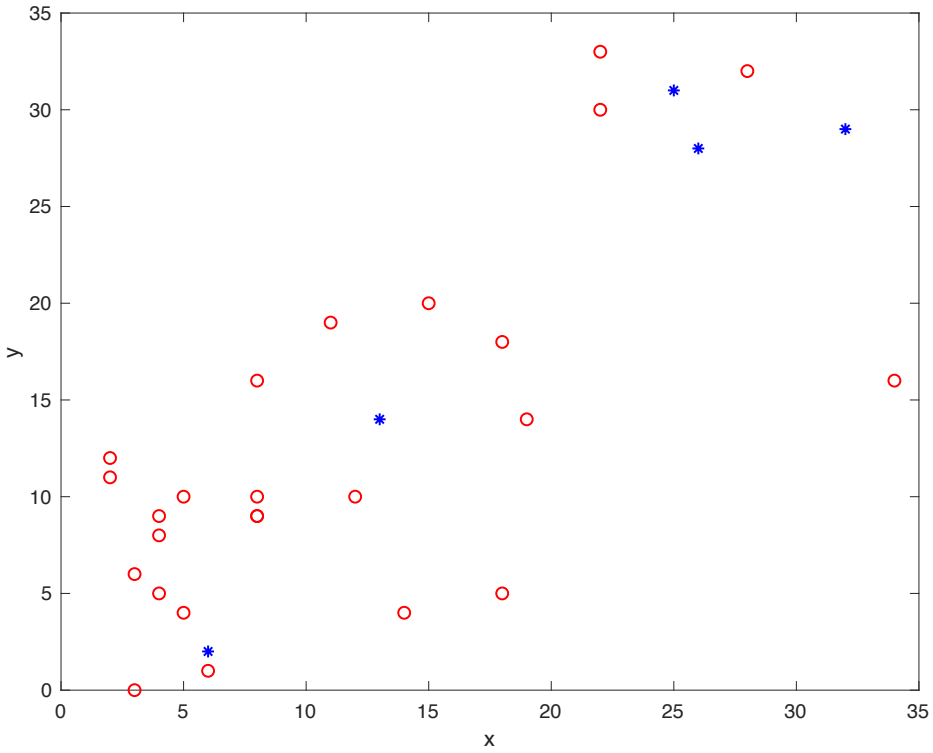

Figura 1: 'o' pontos da classe C1, '*' pontos da classe C2.

a) Usando a metodologia *Adaptive Synthetic Sampling Approach* com K=5, determine o conjunto de dados decorrentes do processamento dos pontos P24 e P28. Considere que beta = 1. Apresente todos os cálculos necessários.
b) Determine a projeção que maximiza a separabilidade das duas classes C1 e C2, usando os dados da tabela 1. Apresente todas as matrizes e indique todos os cálculos necessários.

3 – A Transformada de Fourier do sinal x(t) é dada por

$$X(\omega) = \begin{cases} 0, & \omega < -20\pi \ \vee \ \omega > 20\pi \\ |20\pi - \omega|, & -20\pi < \ \omega < 20\pi \end{cases}$$

Considere agora o sinal y[n], amostrado de x(t) com uma frequência de amostragem 15Hz. Represente graficamente, em Hz, o resultado da aplicação da Discrete Time Fourier Transform (DTFT) ao sinal y[n], para o domínio -20Hz a 20Hz.

Part B

It is required to construct a computational model to predict houses categories in a city, to be classified as Luxury (Lux) or not (N). A data set of 80 samples is available were attributes like Area (m^2), Construction Year (Old or New) and Neighborhood category (H: high, M: medium, L: low) were collected. The Table 2 represents a sub-sample of the training set.

Table 2

Class	Area [m^2]	Construction Age	Neighborhood
Lux	520	Old	H
N	300	Old	M
Lux	900	New	H
Lux	470	New	H
N	518	New	L
Lux	1080	New	M
N	340	New	M
N	500	Old	L
N	400	Old	M
Lux	1200	Old	H

On this sense two methods will be tested. OneR and a ID3 decision tree. On this sense implement the following steps.

a) Discretize the continuous attribute using the entropy-based binning algorithm. Considered as possible thresholds the midpoint between successive points of different classes. Express the procedure and the resultant attributes. Use the arrange below.

Class	N	N	N	L	N	N	N	L	L	L
Area	300	340	400	470	500	518	520	900	1080	1200

b) Apply the OneR approach. Indicates the procedure for selecting the best rule.

c) Considering the discrete attributes (obtained in a) define the first node of a decision tree based in the ID3 algorithm. Describe the procedure and results, include a design of the partial tree.

d) Describe the next step that should be conducted in the implementation of a decision tree (not calculi are needed here).

e) Considering the full data set, indicate which data partition approaches are suitable for model development and evaluation.

f) Considering the information contained in Table 1, indicates two approach that could be used to depict the relation between the house's areas and the house's price (not given data). Also indicate one metric that could be used in model development.

g) Considering the approach in c), describe the learning paradigms it belongs. Justify.