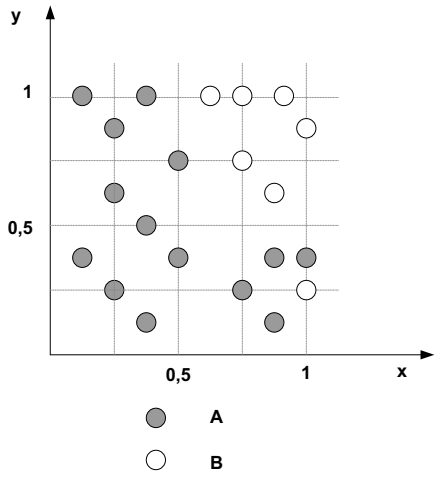
4 valores

1. Considere o conjunto de treino representado na figura abaixo, onde os valores de x e y estão restritos ao intervalo [0.0, 1.0]. Neste conjunto, embora os atributos sejam originalmente contínuos, todos os valores foram discretizados usando o valor 0.5 como limite.



a) Determine a incerteza associada a este conjunto de treino

b) Apresente a árvore de decisão gerada com este conjunto de treino, usando o critério de Ganho de Informação

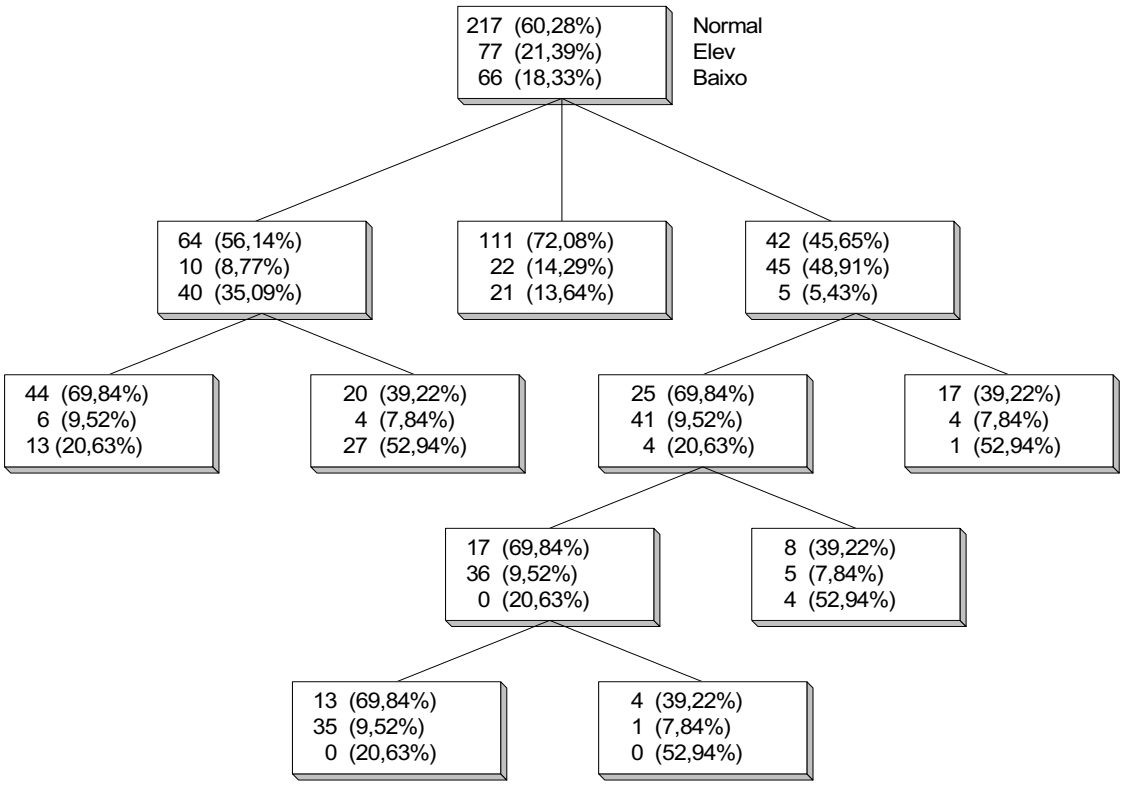
c) Quais as principais vantagens e desvantagens dos algoritmos de classificação baseados em árvores de decisão.

d) Indique vantagens e desvantagens da discretização de atributos quando aplicados na construção de um modelo de classificação

e) Apresente sucintamente dois processos para estimar o erro de um classificador num conjunto de dados

5 valores

2. Considere a seguinte árvore de decisão obtida com o algoritmo C5.0 aplicado a um conjunto de dados



a) Construa a matriz de confusão relativa a esta árvore

b) Calcule a taxa Precisão do modelo, o que pode concluir?

c) Qual o significado do valor no cruzamento da linha identificada com o texto **"Baixo"** com a coluna identificada com o texto **"^Elev"** ?

d) Qual das previsões: **^Normal, ^Elev, ^Baixo** apresenta melhor taxa Positivos Verdadeiros face aos Positivos Falsos, justifique apresentando os cálculos. O que pode concluir?

e) Apresente a forma final da árvore de decisão acima, após a aplicação do algoritmo de poda dado nas aulas que considera a seguinte regra para cálculo da estimativa de erros:

estimativa (erros, #exemplos) = (2 * erros + 1) / (1 + #exemplos)

Justifique a resposta apresentando os cálculos

5 valores

3. Considere o seguinte conjunto de pontos a 1-dimensão: {6, 12, 18, 24, 30, 42, 48}

a) Para cada um dos seguintes conjuntos iniciais de centróides:

i. 18 e 45

ii. 15 e 40

b) Crie dois clusters atribuindo cada ponto ao centróide mais próximo e em seguida calcule a soma do quadrado dos erros para cada um dos conjuntos de clusters. Mostre ambos os clusters e a soma do quadrado dos erros.

c) Ambos os conjuntos de centróides apresentam soluções estáveis, isto é, se aplicasse o algoritmo k-means neste conjunto de pontos com os mesmos centróides iniciais obteria os mesmos clusters? Justifique a resposta sem recorrer a cálculos.

d) Quais os clusters produzidos pelo algoritmo single-link, distância mínima?

e) A que definição de clustering correspondem as partições obtidas com o algoritmo K-means (considere a partição com menor soma do quadrado dos erros) e com o algoritmo single-link?

f) Qual a característica do algoritmo K-means que explica o comportamento anterior?

INSTITUTO POLITÉCNICO DO PORTO

4 valores

4. Considere a seguinte tabela de contingência de vendas dos três itens A, B e C.

			A	
			0	1
C = 0	**B**	1	20	15
		0	0	25
C = 1	**B**	1	15	10
		0	5	10

a) Indique o Suporte, a Confiança e o Interesse das regras:

i. B C → A

ii. B → A C

iii. O que pode concluir face aos itens destas regras?

b) Assuma que a regra **{a, b} → {c, d} se encontra no conjunto final** de regras apresentadas pelo algoritmo Apriori e a regra **{c, d} → {a, b} não se encontra**.

Para cada uma das seguintes regras:

i. {a, b, c} →{d}

ii. {a} →{b, c, d}

iii. {b, c, d}→{a}

iv. {c} →{a, b, d}

Indique, justificando se:

a) a regra encontra-se no conjunto final

b) a regra não se encontra no conjunto final

c) há possibilidade da regra aparecer no conjunto final de regras

2 valores

5. Considere o seguinte problema de mineração de dados. Dada uma base de dados constituída pelas tabelas:

Cliente (IdCli, Cidade, CodPostal)
Loja (IdLoja, Nome, Cidade, CodPostal)
Compra (IdCli, IdLoja, Item, Data)

Pretende-se conhecer a evolução das compras dos clientes que moram na cidade com o CodPostal A, que efectuaram as compras na loja B que fica na cidade com o CodPostal C.

a) Dos algoritmos estudados, indique o mais adequado para resolver este problema

b) Defina os parâmetros necessários ao algoritmo bem como os dados que usaria para extrair os padrões pretendidos, e os resultados que serão fornecidos pelo algoritmo

c) Será necessária alguma adaptação ao algoritmo indicado? Justifique.

INSTITUTO POLITÉCNICO DO PORTO

FIM