IA – Eduardo P2

Parte 2) Classificação

Exercício 1 (Árvores de Classificação)**:**
a) obter uma árvore de decisão para a base de dados abaixo, considerando que COMPRAR é o atributo meta (variável dependente):

SEXO	PAIS	IDADE	COMPRAR
M	França	25	Sim
M	Inglaterra	25	Sim
F	França	25	Sim
F	Inglaterra	25	Sim
F	França	55	Não
M	Alemanha	55	Não
M	Alemanha	55	Não
F	Alemanha	55	Não
F	França	55	Não
M	França	55	Não

b) Repita o exercício, agora eliminando o atributo IDADE da base de dados. Qual árvore é melhor, considerando “explicar o passado” e “predizer o futuro”?

Exercício 2:
Para responder às perguntas 1-5, considere a seguinte base de dados: (Parte2 Pg 54)

A_1	A_2	A_3	A_4	Classe
S	H	H	W	N
S	H	H	S	N
O	H	H	W	Y
R	M	H	W	Y
R	C	N	W	Y
R	C	N	S	N
O	C	N	S	Y
S	M	H	W	N
S	C	N	W	Y
R	M	N	W	Y
S	M	N	S	Y
O	M	H	S	Y
O	H	N	W	Y
R	M	H	S	N

1) Obter a árvore de classificação pelo ganho de informação;

2) Quais são as regras de classificação obtidas por meio desta árvore?

3) Qual é a acurácia deste conjunto de regras para o conjunto de treinamento? Esta acurácia é uma estimativa adequada para a capacidade de generalização do classificador (ao se considerar dados não vistos)?

4) Supondo-se que não se pode dispor adicionalmente de mais dados, descreva um procedimento que permita estimar melhor a acurácia do classificador em questão para dados novos (e.g., ainda não observados e que serão classificados pela árvore de decisão). Dica: ver validação cruzada.

Parte 3) Regressão

Exercício 3 (Árvores de Regressão):
- Realizar o teste para autônomo, complementar árvore

Exercício 4 (K-NN para regressão):
- Estimar Y para o objeto #9 com k=1,2,3,4,5 (com e sem ponderação)
- Diferença ao de classificação: estimar **média** para cada k

Exemplo	a_1	a_2	a_3	Y
1	0	250	36	10
2	10	150	34	15
3	2	90	10	5
4	6	78	8	20
5	4	20	1	30
6	1	170	70	40
7	8	160	41	25
8	10	180	38	35
9	6	200	45	?

d(9,1) = 65
d(9,2) = 65
d(9,3) = 149
d(9,4) = 159
d(9,5) = 226
d(9,6) = 60
d(9,7) = 46
d(9,8) = 31

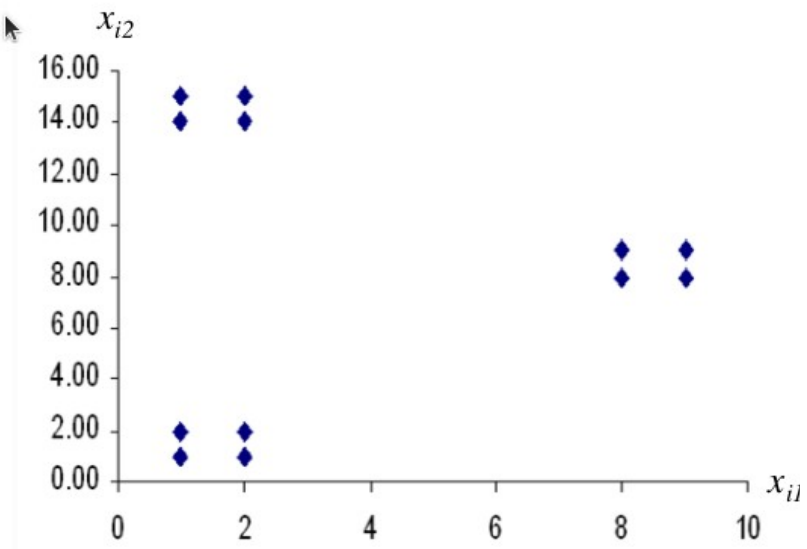
a) Sem ponderação
k = 1 : exemplo 8 -> Y = 38
k = 2 : exemplo 8,7 -> Y = (35 + 25)/2 = 30
k = 3 : exemplo 8,7,6 -> Y = (35 + 25 + 5)/3 = 21,666
k = 4 : exemplo 8,7,6,1 -> Y = (35 + 25 + 5 + 10)/4 = 18,75
k = 5 : exemplo 8,7,6,1,2 -> Y = (35 + 25 + 5 + 10 + 15)/5 = 18

b) Com ponderação (Inverso da distância manhattan)
k = 1 : exemplo 8 -> Y = 38
k = 2 : exemplo 8,7 -> Y = (35 + 25)/2 = 30
k = 3 : exemplo 8,7,6 -> Y = (35 + 25 + 5)/3 = 21,666
k = 4 : exemplo 8,7,6,1 -> Y = (35 + 25 + 5 + 10)/4 = 18,75
k = 5 : exemplo 8,7,6,1,2 -> Y = (35+25+5+10+15)/5 = 18

Inversos: voto do 8 = 1/31 = 0,0322
voto do 7 = 1/46 = 0,0217
voto do 6 = 1/60 = 0,0166
voto do 1 = 1/65 = 0,0153
voto do 2 = 1/65 = 0,0153

Parte 4) Clustering
Exercício 5 (K-means):

Objeto $\mathbf{x}_i$	x_{i1}	x_{i2}
1	1	2
2	2	1
3	1	1
4	2	2
5	**8**	**9**
6	**9**	**8**
7	**9**	**9**
8	**8**	**8**
9	1	15
10	2	15
11	1	14
12	2	14



Executar k-means com $k = 3$ nos dados acima a partir dos protótipos [6 6], [4 6] e [5 10]. Quais foram as partições e os centróides obtidos?

Objeto/centróide	[6,6]	[4,6]	[5,10]
1	$5^2 + 4^2 = 41$	$3^2 + 4^2 = 25$	$4^2 + 8^2 = 80$
2	$4^2 + 5^2 = 41$	$2^2 + 5^2 = 29$	$3^2 + 9^2 = 90$
3	$5^2 + 5^2 = 50$	$3^2 + 5^2 = 34$	$4^2 + 9^2 = 97$
4	$4^2 + 4^2 = 32$	$2^2 + 4^2 = 18$	$3^2 + 8^2 = 73$
5	$(-2)^2 + (-3)^2 = 13$	$(-4)^2 + (-3)^2 = 25$	$(-3)^2 + (1)^2 = 10$
6	$(-3)^2 + (-2)^2 = 13$	$(-5)^2 + (-2)^2 = 29$	$(-4)^2 + (2)^2 = 18$
7	$(-3)^2 + (-3)^2 = 18$	$(-5)^2 + (-3)^2 = 34$	$(-4)^2 + (1)^2 = 17$
8	$(-2)^2 + (-2)^2 = 8$	$(-4)^2 + (-2)^2 = 18$	$(-3)^2 + (2)^2 = 13$
9	$(5)^2 + (-9)^2 = 106$	$(3)^2 + (-9)^2 = 90$	$(4)^2 + (-5)^2 = 41$
10	$(4)^2 + (-9)^2 = 97$	$(2)^2 + (-9)^2 = 85$	$(3)^2 + (-5)^2 = 34$
11	$(5)^2 + (-8)^2 = 89$	$(3)^2 + (-8)^2 = 73$	$(4)^2 + (-4)^2 = 32$
12	$(4)^2 + (-8)^2 = 80$	$(2)^2 + (-8)^2 = 66$	$(3)^2 + (-4)^2 = 25$

P1[6,6] = {6,8} - c1x = (9 + 8)/ 2 = 8,5 | c1y = (8 + 8)/ 2 = 8
P1[4,6] = {1,2,3,4} - c2x = (1 + 2 + 1 + 2)/ 4 = 1,5 | c2y = (2+1+1+2)/4 = 1,5
P1[6,6] = {5,7,9,10,11,12} -c3x=(8+9+1+2+1+2)/6=3,83 |c3y=(9+9+15+15+14+14)/6= 12,6

Exercício 6 (Silhueta Simplificada):
Calcule o valor para as silhuetas para a partição *correta* acima e também para uma partição formada por dois clusters à sua escolha.

C1 = (1.5, 1.5) ;
C2 = (8.5, 8.5) ;
C3 = (1.5, 14.5);

C1 e C2 = 49 + 49 = 98;
C1 e C3 = 169;
C2 e C3 = 49 + 36 = 85;

C1 → C2 ; C2 → C3 ; C3 → C2

1) Dissimilaridade entre clusters
- Distância E^2 entre centroides → saber cluster mais próximo

2) para **k = 3**
P/ cluster 1:
i1 = (1,2), bi1 =$7.5^2 + 6.5^2$ = 98,5 , ai1 = $0.5^2 + 0.5^2$ = 0,5
s(i1) = 98,5 – 0,5 / 98,5 = 98/ 98,5 = 0,9949 = s(i2)

i3 = (1,1), bi3 =$7.5^2 + 7.5^2$ = 112,5 , ai3 = $0.5^2 + 0.5^2$ = 0,5
s(i3) = 112,5 – 0,5 / 112,5 = 112/ 112,5 = 0,9955

i4 = (2,2), bi4 =$6.5^2 + 6.5^2$ = 84,5 , ai4 = $0.5^2 + 0.5^2$ = 0,5
s(i4) = 84,5 – 0,5 / 84,5 = 84/ 84,5 = 0,9940

P/ cluster 2: {…}
P/ cluster 3: {…}

SSWC = 0,9949 + 0,9949 + 0,9955 + 0,9940 + … + s(i)12 / 12 =