Introdução *Data Mining* 3 a 6 de Julho de 2017

Exercícios Práticos

1. Introdução

1.1. Discuta o slide 14 "Data Mining Vs. Estatística" à luz da produção estatística no INE - Refira casos em que se aplica Estatística e quando se pode aplicar o Data Mining.

1.2. Refira exemplos da possível aplicação de Data Mining no INE (refira-se ao slide 17 Principais tarefas do Data Mining).

1.2.1. Em que projetos estatísticos faz sentido fazer descoberta de padrões?

1.2.2. Com que finalidade?

2. Preparação dos dados

2.1. Discretizar o atributo 'Humidade', do ficheiro **weather.numeric.csv**, usando os *softwares* RapidMiner e R. Considerar 5 intervalos.

2.1.1. Discretização com Intervalos Iguais.

2.1.2. Discretização com Intervalos com Frequências Iguais.

2.2. Obter os pontos de corte do atributo 'Humidade', do ficheiro de dados weather.numeric", e calcular o ganho de informação para um ponto de corte à escolha.

3. Avaliação de Modelos

3.1. Considere, para 25 exemplos, as classes '**+**' e '**-**', reais e previstas pelos modelos 1 e 2 (M1 e M2).

Real	+	+	-	-	-	+	+	-	-	-	+	+	+	-	+	-	-	+	+	+	-	-	-	-	+
Prevista M1	+	+	+	-	-	+	-	+	-	+	+	-	-	+	+	+	-	-	+	+	-	-	+	-	+
Prevista M2	+	-	+	-	+	+	+	-	-	-	+	+	-	-	+	-	+	+	+	-	+	-	-	-	+

3.1.1. Construa a matriz de confusão para cada um dos modelos (M1 e M2).

3.1.2. Calcule as medidas: taxa de acerto, precisão, *recall* e medida-f1, para cada um dos modelos (M1 e M2). Qual dos modelos obteve o melhor desempenho?

4. Modelos Exploratórios e Preditivos

Clustering

4.1. Com base na matriz que se apresenta de seguida, fazer os vários passos da agregação com base no método de *Furthest Neighbour* (*Complete Linkage)* e desenhar o respetivo dendograma.

	a	**b**	**c**	**d**
a	0	5	3	9
b	5	0	7	4
c	3	7	0	6
d	9	4	6	0

4.2. Com base na seguinte matriz de dados (Sharma, 1996), que descreve 6 indivíduos e 4 variáveis, foram calculados os erros ESS para implementação do método de *Ward*.

Subject Id	Income ($ thous.)	Education (years)
S1	5	5
S2	6	6
S3	15	14
S4	16	15
S5	25	20
S6	30	19

	Members in Cluster					
Cluster Solution	1	2	3	4	5	ESS
(a) All Possible Five-Cluster Solutions						
1	S1,S2	S3	S4	S5	S6	1.0
2	S1,S3	S2	S4	S5	S6	90.5
3	S1,S4	S2	S3	S5	S6	110.5
4	S1,S5	S2	S3	S4	S6	312.5
5	S1,S6	S2	S3	S4	S5	410.5
6	S2,S3	S1	S4	S5	S6	72.5
7	S2,S4	S1	S3	S5	S6	90.5
8	S2,S5	S1	S3	S4	S6	278.5
9	S2,S6	S1	S3	S4	S5	372.5
10	S3,S4	S1	S2	S5	S6	1.0
11	S3,S5	S1	S2	S4	S6	68.0
12	S3,S6	S1	S2	S4	S5	125.0
13	S4,S5	S1	S2	S3	S6	53.0
14	S4,S6	S1	S2	S3	S5	106.0
15	S5,S6	S1	S2	S3	S4	13.0
(b) All Possible Four-Cluster Solutions						
1	S1,S2,S3	S4	S5	S6		109.333
2	S1,S2,S4	S3	S5	S6		134.667
3	S1,S2,S5	S3	S4	S6		394.667
4	S1,S2,S6	S3	S4	S5		522.667
5	S1,S2	S3,S4	S5	S6		2.000
6	S1,S2	S3,S5	S4	S6		69.000
7	S1,S2	S3,S6	S4	S5		126.000
8	S1,S2	S4,S5	S3	S6		54.000
9	S1,S2	S4,S6	S3	S5		107.000
10	S1,S2	S5,S6	S3	S4		14.000

Desenhe o dendograma correspondente a este processo de *clustering*.

4.3. Aplicação no R: Ward Hierarchical Clustering

```
# http://www.statmethods.net/advstats/cluster.html
f<-read.csv("fogos.csv", header=TRUE, sep=";")
d <- dist(f, method = "euclidean") # distance matrix
fit <- hclust(d, method="ward")
plot(fit) # display dendogram
groups <- cutree(fit, k=5) # cut tree into 5 clusters
# draw dendogram with red borders around the 5 clusters
rect.hclust(fit, k=5, border="red")
#avaliar clusters (apenas para 3 variáveis)
f$groups<-groups
f$groups<-as.factor(f$groups)
attach(f)
aggregate(f,by=list(f$groups),FUN=mean)
a<-aov(X1980~groups, data=f)
TukeyHSD(a)
#para ordenar por grupos, ver ex. 4.13 com o o ficheiro das profissões
f<-read.csv("exemplo cluster.csv", header=TRUE, sep=";")
#para estandardizar as variáveis (uma vez que estão expressas e unidades diferentes),
fazer:
prestig<-(prestig-mean(prestig))/sd(prestig)
educ<-(educ-mean(educ))/sd(educ)
rendim<-(rendim-mean(rendim))/sd(rendim)
educ<-(educ-mean(educ))/sd(educ)

d <- dist(f, method = "euclidean") # distance matrix
fit <- hclust(d, method="ward")
plot(fit) # display dendogram
groups <- cutree(fit, k=5) # cut tree into 5 clusters
f$groups<-groups
f$groups<-as.factor(f$groups)
aggregate(f,by=list(f$groups),FUN=mean)
attach(f)
a<-aov(prestig~groups, data=f)
```

```
summary (a)
TukeyHSD(a)
```

4.4. Aplicação no R: comparar com o método single linkage

```
# Single-linkage Hierarchical Clustering
# http://www.statmethods.net/advstats/cluster.html
fit <- hclust(d, method="single")
plot(fit) # display dendogram
groups <- cutree(fit, k=5) # cut tree into 5 clusters
# draw dendogram with red borders around the 5 clusters
rect.hclust(fit, k=5, border="red")
```

4.5. Aplicaçao no R: K-Means Cluster Analysis

```
f<-read.csv("exemplo cluster.csv", header=TRUE, sep=";")
fit <- kmeans(f[,2:5], 5) # 5 cluster solution
fit #ver resultados obtidos
# get cluster means
aggregate(f,by=list(fit$cluster),FUN=mean)
# append cluster assignment
mydata <- data.frame(f, fit$cluster)
#o fit$cluster contém o cluster membership
#permite-nos obter a dimensão de cada um dos 5 grupos formados
table(mydata$fit.cluster)
#ordenar por cluster
ordem <- order(f$fit.cluster)
ordem
f[ordem,]
```

4.6. Determinar o número de clusters para o kmeans

```
f<-mydata
mydata<-mydata[,2:5]
wss <- (nrow(mydata)-1)*sum(apply(mydata,2,var))
for (i in 2:15) wss[i] <- sum(kmeans(mydata,  centers=i)$withinss)
```

```
plot(1:15, wss, type="b", xlab="Number of Clusters", ylab="Within groups sum of squares")
```

Regras de Associação

4.7. Considere as transações da tabela seguinte:

ID	p1	p2	p3	p4
1	1	1	0	0
2	0	1	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

4.7.1. Indique os *itemsets* frequentes, para suporte 40%.

4.7.2. Qual a confiança da regra de associação {p1, p2}→{p3}?

4.7.3. E qual o lift da regra {p1}→{p2}?

4.8. Abrir o ficheiro de dados **supermarket sample.xls**.

- Obter regras de associação com um suporte mínimo igual a 50% e confiança superior a 20%. Altere estes valores e compare os resultados.

K-NN

4.9. Considere o seguinte conjunto de treino:

ID	a1	a2	a3	classe
1	1	1	0	X
2	0	1	1	Z
3	0	0	1	X
4	1	1	1	X
5	0	1	0	X
6	1	0	1	Z

Classifique o exemplo x = {a1=1, a2=0, a3=1} usando um classificador 1-vizinho mais próximo. Classifique-o usando um classificador 5-vizinhos mais próximos.

Naive Bayes

4.10. Considere o seguinte conjunto de treino:

ID	a1	a2	a3	classe
1	1	1	0	X
2	0	1	1	Z
3	0	0	1	X
4	1	1	1	X
5	0	1	0	X
6	1	0	1	Z

Classifique o exemplo x = {a1=0, a2=1, a3=1} usando um classificador Naive Bayes.

4.11. Considere os dados do ficheiro "golf.csv".

Outlook	Temperature	Humidity	Wind	Play
sunny	85	85	FALSE	no
sunny	80	90	TRUE	no
overcast	83	78	FALSE	yes
rain	70	96	FALSE	yes
rain	68	80	FALSE	yes
rain	65	70	TRUE	no
overcast	64	65	TRUE	yes
sunny	72	95	FALSE	no
sunny	69	70	FALSE	yes
rain	75	80	FALSE	yes
sunny	75	70	TRUE	yes
overcast	72	90	TRUE	yes
overcast	81	75	FALSE	yes
rain	71	80	TRUE	no

Classifique o exemplo x ={Outlook = "sunny", Temperature = 66, Humidity = 90, Wind = TRUE} usando um classificador Naive Bayes.

Árvores de Decisão e Regras de Decisão

4.12. Fazer o diagrama em árvore da disjunção (i.e. OR).

A	B	V
0	0	0
0	1	1
1	0	1
1	1	1

4.13. Considere o conjunto de dados apresentado na tabela seguinte em que **DMC** é a duração média de uma chamada, **FUM** é a faturação do último mês, e **CT** é uma variável que indica se o contrato terminou. A partir destes três atributos pretende-se determinar o valor do atributo **Ab** (abandonou?)

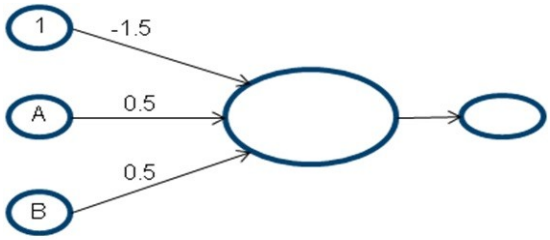
DMC	FUM	CT	Ab
alta	alta	sim	sim
alta	baixa	não	não
baixa	alta	não	não
baixa	baixa	sim	sim
baixa	alta	sim	não

Extraia as Regras de Decisão utilizando o algoritmo oneR.

Redes Neuronais

4.14. Considere os seguintes conjunto de dados da tabela e o *perceptron*.

A	B	∧
0	0	0
0	1	0
1	0	0
1	1	1



4.14.1. Utilizando a taxa de aprendizagem de $\eta = 0.25$ e a função de ativação,

$$f(x_1, x_2, ..., x_n) = \begin{cases} 1, & \text{sse } \sum x_i w_i > 0 \\ 0, & \text{c.c.} \end{cases},$$

descreva uma iteração do algoritmo de correção de erro do *perceptron* para o conjunto de testes da tabela (função AND).

4.14.2. O algoritmo convergiu?

5. Modelos de Procura Evolutivos

Algoritmos Genéticos

5.1. Crie 10 instâncias para o problema "*Knapsack*". Considere uma taxa de seleção de 50% (as melhores 5 instâncias contribuem com 2 cópias cada). Como *crossover*, escolha 3 pares diferentes das instâncias selecionadas e troque os últimos 3 genes. Finalmente, insira uma mutação aleatória em cada instância. Repita o processo 2 vezes e indique a solução final.

ITEM	SURVIVAL POINTS	WEIGHT
pocketknife	10.00	1.00
beans	20.00	5.00
potatoes	15.00	10.00
unions	2.00	1.00
sleeping bag	30.00	7.00
rope	10.00	5.00
compass	30.00	1.00

5.2. Use o script "knapsack.r" para obter uma solução ótima. Indique qual é.