Examen

27 mars 2009 - Durée : 2h
Documents autorisés : Notes de cours, TD et TP (pas de livre)

Quelles que soient vos réponses, penser à les justifier, préciser l'annexe utilisée !!!
Une simple réponse oui/non ne suffit pas.

Un ensemble de sorties numériques pouvant vous être utiles pour répondre aux questions est disponible à la fin du sujet.

Partie A

On considère le célèbre jeu de données décrivant des iris, proposé par Fisher pour illustrer les méthodes d'analyse discriminante. Ce fichier contient la description de 150 observations de 3 espèces d'iris. Les fleurs sont observées par 4 variables : la longueur et la largeur des pétales, et la longueur et la largeur des sépales. Le problème posé est de construire une fonction permettant de prédire l'espèce à partir de la mesure de ces 4 quantités.
L'application de l'algorithme de construction d'arbres de décision a donné le résultat présenté en annexe. Pour obtenir ce résultat, le fichier des 150 iris a été découpé au hasard en deux parties : (1) un échantillon d'apprentissage de 105 individus qui a permis de construire l'arbre, (2) un échantillon de validation de 45 individus pour choisir le meilleur arbre élagué.
La figure en annexe contient, pour chacun des nœuds de l'arbre, les effectifs (et les pourcentages) dans chacune des classes pour l'échantillon d'apprentissage (colonne de gauche), et pour l'échantillon de validation (colonne de droite).

Voici un extrait de cette table :

Species_name	Petal_width	Petal_length	Sepal_width	Sepal_length
Verginica	20	49	28	56
Versicolor	16	51	27	60

Questions :

[1] Préciser quelles sont les types des variables mis en jeu
[2] Rappeler la différence entre échantillon de validation et échantillon de test.
[3] Calculer le taux d'erreur global de l'arbre sur l'échantillon d'apprentissage, ainsi que sur l'échantillon de validation.
[4] L'arbre présenté est-il complet pour l'échantillon d'apprentissage ? Pourquoi ?
[5] Classer les individus de l'extrait à l'aide de cet arbre. Ces deux individus sont-ils bien classés ? Si l'un n'est pas bien classé, cela vous étonne-t-il ? Pourquoi ?
[6] Procéder à l'élagage de l'arbre en suivant la méthode donnée dans le cours. Pour chacun des arbres de la séquence d'élagage, calculer le taux d'erreur global pour les échantillons d'apprentissage et de validation. Construire le graphique portant ces taux d'erreur en fonction du nombre de feuilles de l'arbre. Quel est le nombre de feuilles de l'arbre optimal ?

Partie B

On considère une étude mesurant les effets d'un traitement sur des personnes âgées atteintes de névralgie. Deux traitements et un placebo ont été comparés. La variable « Pain » indique si les personnes continuent à se plaindre (YES) ou non de douleurs (NO). Nous avons également mesuré l'âge et le genre des patients et la durée de la douleur avant que le traitement ne commence.

Treatment	Sex	Age	Duration	Pain
P	F	68	1	No
B	M	74	16	No
P	F	67	30	No
P	M	66	26	Yes
B	F	67	28	No
B	F	77	16	No
A	F	71	12	No
B	F	72	50	No
B	F	76	9	Yes
A	M	71	17	Yes
A	F	63	27	No
A	F	69	18	Yes
B	F	66	12	No
A	M	62	42	No
P	F	64	1	Yes
A	F	64	17	No

Questions :

[7] Préciser quelles sont les types des variables mis en jeu. L'expérimentation a porté sur combien de patients ?
[8] Vous êtes chargé d'analyser ce jeu de données, quelles méthodes pouvez-vous utiliser? (préciser pour chacune des méthodes quelle est la variable à expliquer et quelles sont les variables explicatives)
[9] Est-ce qu'il est pertinent de mettre une interaction Treatment*Sexe dans le modèle ?
[10] Dans l'annexe 5, quelle est la variable modélisée ? Quelle méthode a été utilisée ? Pourquoi ? Est-ce justifié ? Quelles sont les variables du modèle final ?
[11] Toujours sur les résultats de l'annexe 5, est-ce que les résultats de cette méthode sont satisfaisants. Si non, pourquoi ?
[12] Ecrire la règle de décision permettant de savoir si un patient continue à souffrir.
[13] L'individu n°1 est-il bien ou mal classé ?

Partie C

On s'intéresse ici à l'effet de la codéine (CODEINE) et de l'acupuncture (ACUPUNCTURE) sur les douleurs post-opératoire dentaires pour une population d'hommes. Deux traitements sont administrés. On mesure la douleur ressentie à l'issue de l'opération.

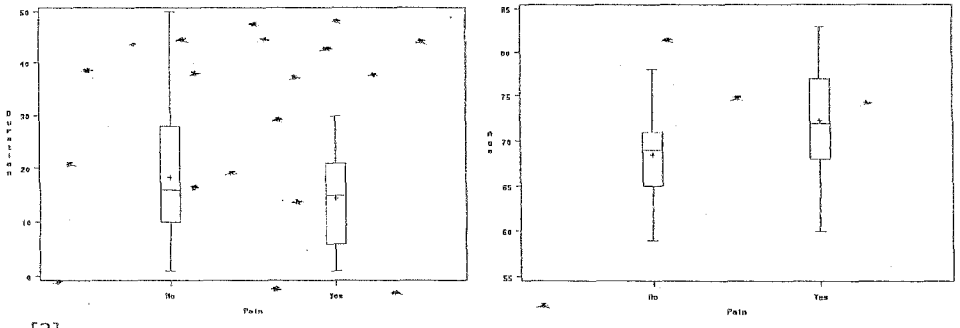
Les individus sont regroupés par niveau d'intensité de douleur (PainLevel).

PainLevel	Codeine	Acupuncture	Relief
1	1	1	0
1	2	1	0.5
1	1	2	0.6
1	2	2	1.2
2	1	1	0.3
2	2	1	0.6
2	1	2	0.7
2	2	2	1.3
3	1	1	0.4
3	2	1	0.8
3	1	2	0.8
3	2	2	1.6
4	1	1	0.4

[14] Combien d'hommes ont été opérés ?
[15] Quelle méthode est utilisée ?
[16] Est-ce que le modèle présenté à l'annexe 6 est pertinent ? Quelle que soit votre réponse, pourquoi ?
[17] Quels sont les facteurs qui ont un effet ? Pour ceux qui ont un effet, quel est-il ?

Annexes :

[1]



[2]

Treatment*Pain

Statistic	DF	Value	Prob
Chi-Square	2	13.7143	0.0011

Sexe*Pain

Statistic	DF	Value	Prob
Chi-Square	1	5.5543	0.0184

[3]

Treatment	Frequency	Percent	Cumulative Frequency	Cumulative Percent
A	20	33.33	20	33.33
B	20	33.33	40	66.67
P	20	33.33	60	100.00

Sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
F	30	50.00	30	50.00
M	30	50.00	60	100.00

Pain	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	35	58.33	35	58.33
Yes	25	41.67	60	100.00

[4]

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Treatment	2	11.9886	0.0025
Sex	1	5.3104	0.0212
Treatment*Sex	2	0.1412	0.9318
Age	1	7.2744	0.0070
Duration	1	0.0247	0.8752

[5]

The LOGISTIC Procedure

Model Information

Data Set	WORK.NEURALGIA
Response Variable	Pain
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	60
Number of Observations Used	60

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Treatment	2	11.9886	0.0025
Sex	1	5.3104	0.0212
Treatment*Sex	2	0.1412	0.9318
Age	1	7.2744	0.0070
Duration	1	0.0247	0.8752

Response Profile

Ordered Value	Pain	Total Frequency
1	No	35
2	Yes	25

Probability modeled is Pain='No'.

(...)

Step 0. Intercept entered:

(...)

Step 1. Effect Treatment entered:

(...)

NOTE: No effects for the model in Step 1 are removed.

Step 2. Effect Age entered:

(...)

NOTE: No effects for the model in Step 2 are removed.

Step 3. Effect Sex entered:

(...)

NOTE: No effects for the model in Step 3 are removed.

NOTE: No (additional) effects met the 0.05 significance level for entry into the model.

Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	Treatment		2	1	13.7143		0.0011
2	Age		1	2	10.6038		0.0011
3	Sex		1	3	5.9959		0.0143

(...)

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	15.8669	6.4056	6.1357	0.0132
Treatment	A	1	3.1790	1.0135	9.8375	0.0017
Treatment	B	1	3.7264	1.1339	10.8006	0.0010
Treatment	P	0	0	.	.	.
Sex	F	1	1.8235	0.7920	5.3013	0.0213
Sex	M	0	0	.	.	.
Age		1	-0.2650	0.0959	7.6314	0.0057

(...)

Association of Predicted Probabilities and Observed Responses

Percent Concordant	90.3	Somers' D	0.811
Percent Discordant	9.1	Gamma	0.816
Percent Tied	0.6	Tau-a	0.401
Pairs	875	c	0.906

[6]

The ANOVA Procedure

Dependent Variable: Relief

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	11.33500000	1.13350000	78.37	<.0001
Error	21	0.30375000	0.01446429		
Corrected Total	31	11.63875000			

R-Square	Coeff Var	Root MSE	Relief Mean
0.973902	10.40152	0.120268	1.156250

Source	DF	Anova SS	Mean Square	F Value	Pr > F
PainLevel	7	5.59875000	0.79982143	55.30	<.0001
Codeine	1	2.31125000	2.31125000	159.79	<.0001
Acupuncture	1	3.38000000	3.38000000	233.68	<.0001
Codeine*Acupuncture	1	0.04500000	0.04500000	3.11	0.0923

[7]

The ANOVA Procedure

Class Level Information

Class	Levels	Values
PainLevel	8	1 2 3 4 5 6 7 8
Codeine	2	1 2
Acupuncture	2	1 2

Number of Observations Read	32
Number of Observations Used	32

Dependent Variable: Relief

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	11.29000000	1.25444444	79.13	<.0001
Error	22	0.34875000	0.01585227		
Corrected Total	31	11.63875000			

R-Square	Coeff Var	Root MSE	Relief Mean
0.970035	10.88915	0.125906	1.156250

Source	DF	Anova SS	Mean Square	F Value	Pr > F
PainLevel	7	5.59875000	0.79982143	50.45	<.0001
Codeine	1	2.31125000	2.31125000	145.80	<.0001
Acupuncture	1	3.38000000	3.38000000	213.22	<.0001

Bonferroni (Dunn) t Tests for Relief

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	22
Error Mean Square	0.015852
Critical Value of t	3.55217
Minimum Significant Difference	0.3162

Means with the same letter are not significantly different.

Bon Grouping		Mean	N	Pain Level
	A	1.72500	4	8
	A			
	A	1.65000	4	7
	A			
B	A	1.55000	4	6
B				
B		1.25000	4	5
	C	0.90000	4	3
	C			
D	C	0.87500	4	4
D	C			
D	C	0.72500	4	2
D				
D		0.57500	4	1

Bonferroni (Dunn) t Tests for Relief

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	22
Error Mean Square	0.015852
Critical Value of t	2.07387
Minimum Significant Difference	0.0923

Means with the same letter are not significantly different.

Bon Grouping	Mean	N	Codeine
A	1.42500	16	2
B	0.88750	16	1

Bonferroni (Dunn) t Tests for Relief

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	22
Error Mean Square	0.015852
Critical Value of t	2.07387
Minimum Significant Difference	0.0923

Means with the same letter are not significantly different.

Bon Grouping	Mean	N	Acupuncture
A	1.48125	16	2
B	0.83125	16	1

[8] Arbre de décision sur le jeu de données « Iris de Fisher »