eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

TAREA 8

EJERCICIOS

1. El fichero `craneos.dat` ya fue analizado en una tarea previa.

   a) Haz un análisis discriminante. Asegúrate de entender el output de la función `discr` o lo que emplees en su lugar, haces uso de R en lugar de S-PLUS. En último caso, siempre podrías hacer discriminación a partir de los principios básicos (por ejemplo, utilizando la función `cancor`).

   b) Evalúa la matriz de clasificación intramuestral.

   c) Evalúa, mediante partición de la muestra, la tasa de error en la clasificación.

   d) Si en mitad del desierto encontrases un enterramiento con un cráneo cuyas medidas fueran $(130, 128, 103, 52)$, en ausencia de toda otra evidencia y sin posibilidad de utilizar métodos de fechado más sofisticados —como C_{14}—, ¿a qué periodo asignarías dicho cráneo? ¿Con mucha o poca convicción?

2. El fichero `cmc.dge` (que puedes leer con un `dget` para recuperar toda la información sobre nombres de variables y niveles de factores) contiene información sobre uso de anticonceptivos por parte de una muestra de mujeres indonesias encuestadas y diversas covariables. La información sobre procedencia de los datos se da como anexo.

   a) A la vista de dichas covariables, construye una regla para predecir si una mujer está utilizando anticonceptivos, y de qué tipo. Observa que muchas variables son factores.

   b) Discute las ventajas/inconvenientes relativos de los métodos que consideres. Discute las ventajas o inconvenientes de la especificación de los factores como ordenados.

AYUDAS, SUGERENCIAS Y COMPLEMENTOS

1. Cualquiera de los manuales utilizados en el curso trata, con mayor o menor desarrollo, el tema de análisis discriminante: [7], [4], [8] o [11], por ejemplo. Monografías especializadas son, entre otras, [9] (antiguo, pero todavía de útil y agradable lectura), [5], y [10]. Un libro moderno y muy bueno es [6]. [1] tiene también en su Capítulo 3 una presentación interesante del análisis discriminante, desde el punto de vista de las redes neuronales.

2. La referencia esencial sobre árboles binarios de regresión y clasificación continúa siendo [2]. Hay buena documentación *on line* en S-PLUS, y una descripción de las funciones correspondientes en [3].

3. En R "básico" no hay las funciones `tree` y asociadas para construir árboles[1]. Tampoco la función `discr`. Tienes, sin embargo, alternativas:

   a) Para construir árboles, tienes la librería `rpart`, que mejora bastante lo que S-PLUS ofrece como *standard*. Hay documentación disponible: [12], además de la ayuda *on-line*.

   b) Para hacer análisis discriminante, aunque lo podrías programar con facilidad haciendo uso de la función `cancor` o a partir de primeros principios, tienes también como alternativa la librería MASS, aneja al libro [13] y libremente disponible. Te interesarán las funciones `lda`, `qda` y `predict`.

4. Información sobre los datos: se reproduce como anexo. Las únicas alteraciones efectuadas han sido nombrar las variables y los niveles en la dataframe, para hacer los outputs más fáciles de interpretar.

Referencias

[1] C.M. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, 1996.

[2] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. Wadsworth, Belmont, California, 1984.

[3] J.M. Chambers and T.J. Hastie. *Statistical Models in S*. Wadsworth & Brooks/Cole, Pacific Grove, Ca., 1992.

[4] C.M. Cuadras. *Métodos de Análisis Multivariante*. Eunibar, Barcelona, 1981.

[5] D.J. Hand. *Discrimination and Classification*. Wiley, 1981.

[6] D.J. Hand. *Construction and Assessment of Classification Rules*. Wiley, 1997.

[7] R.A. Johnson and D.W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice Hall, 1992.

[8] W.J. Krzanowski. *Principles of Multivariate Analysis: A User's Perspective*. Oxford, 1988. Signatura: 519.23 KRZ.

[9] P.A. Lachenbruch. *Discriminant Analysis*. Hafner Press, New York, 1975.

[10] G.J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, 1992.

[11] G.A.F. Seber. *Multivariate Observations*. Wiley, New York, 1984.

[12] T.M. Therneau and E.J. Atkinson. An introduction to recursive partitioning unsing the RPART routines. Technical report, Mayo Foundation, 1997.

[13] W.N. Venables and B.D. Ripley. *Modern Applied Statistics with* S-PLUS. Springer-Verlag, New York, third edition, 1999.

[1] En realidad, hay un paquete añadido (`tree`) que replica con bastante exactitud lo disponible en S-PLUS. No obstante te conviene utilizar `rpart` en su lugar.

A. Datos cmc

1. Title: Contraceptive Method Choice

2. Sources:
   (a) Origin: This dataset is a subset of the 1987 National Indonesia Contraceptive Prevalence Survey
   (b) Creator: Tjen-Sien Lim (limt@stat.wisc.edu)
   (c) Donor: Tjen-Sien Lim (limt@stat.wisc.edu)
   (c) Date: June 7, 1997

3. Past Usage:
   Lim, T.-S., Loh, W.-Y. & Shih, Y.-S. (1999). A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-three Old and New Classification Algorithms. Machine Learning. Forthcoming. (ftp://ftp.stat.wisc.edu/pub/loh/treeprogs/quest1.7/mach1317.pdf or (http://www.stat.wisc.edu/~limt/mach1317.pdf)

4. Relevant Information:
   This dataset is a subset of the 1987 National Indonesia Contraceptive Prevalence Survey. The samples are married women who were either not pregnant or do not know if they were at the time of interview. The problem is to predict the current contraceptive method choice (no use, long-term methods, or short-term methods) of a woman based on her demographic and socio-economic characteristics.

5. Number of Instances: 1473

6. Number of Attributes: 10 (including the class attribute)

7. Attribute Information:

1.	Wife’s age	(numerical)	
2.	Wife’s education	(categorical)	1=low, 2, 3, 4=high
3.	Husband’s education	(categorical)	1=low, 2, 3, 4=high
4.	Number of children ever born	(numerical)	
5.	Wife’s religion	(binary)	0=Non-Islam, 1=Islam
6.	Wife’s now working?	(binary)	0=Yes, 1=No
7.	Husband’s occupation	(categorical)	1, 2, 3, 4
8.	Standard-of-living index	(categorical)	1=low, 2, 3, 4=high
9.	Media exposure	(binary)	0=Good, 1=Not good
10.	Contraceptive method used	(class attribute)	1=No-use 2=Long-term