Aprendizaje estadístico(1^o cuatrimestre 2020)

Trabajo práctico final

Cada grupo deberá analizar el set de datos que corresponda según su número de grupo, y realizar el análisis que crea pertinente acorde a todo lo visto en la materia. En un plazo de 30 días deberán entregar un informe a jeminagarcia@gmail.com en un archivo con formato pdf.

Breves explicaciones de los datos:

1. El presente trabajo requiere la carga (usando el comando *load*) del dataset *BostonHousing*. Una descripción del mismo se puede hallar en `https://www.kaggle.com/c/boston-housing`. Basado en el dataset "Boston Housing", se le pide encontrar un modelo para predecir y explicar la variable "medv" (valor mediano de las porpiedades) en función del resto de las covariables, usando todo lo aprendido en el curso. Justificar claramente su elección.

2. El archivo *Vidrios.txt* se tiene datos de 9 variables predictivas para poder clasificar el tipo de vidrio, el cual consta de 7 categorías. Las categorías 4 y 6 no son tenidas en cuenta por falta de datos.

   Detalle:
   The study of classification of types of glass was motivated by criminological investigation. At the scene of the crime, the glass left can be used as evidence...if it is correctly identified!
   Attribute Information:
   1. Id number: 1 to 214
   2. RI: refractive index
   3. Na: Sodium (unit measurement: weight percent in corresponding oxide, as are attributes 4-10)
   4. Mg: Magnesium
   5. Al: Aluminum
   6. Si: Silicon
   7. K: Potassium
   8. Ca: Calcium
   9. Ba: Barium
   10. Fe: Iron
   11. Type of glass: (class attribute)
   9. Class Distribution: (out of 214 total instances)

3. El archivo *Sat.txt* se tiene datos de valores de espectros de pixels en una imagen de satélite, para predecir la clase del suelo. Use los métodos que conoce y compárelos mediante CV. Luego aplíquelos a la muestra de test sat.tst y analice los resultados obtenidos.

   Detalle

   PURPOSE The database consists of the multi-spectral values of pixels in 3x3 neighbourhoods in a satellite image, and the classification associated with the central pixel in each neighbourhood. The

aim is to predict this classification, given the multi-spectral values. In the sample database, the class of a pixel is coded as a number. This database was generated from Landsat Multi-Spectral Scanner image data.

DESCRIPTION One frame of Landsat MSS imagery consists of four digital images of the same scene in different spectral bands. Two of these are in the visible region (corresponding approximately to green and red regions of the visible spectrum) and two are in the (near) infra-red. Each pixel is a 8-bit binary word, with 0 corresponding to black and 255 to white. The spatial resolution of a pixel is about 80m x 80m. Each image contains 2340 x 3380 such pixels.

The database is a (tiny) sub-area of a scene, consisting of 82 x 100 pixels. Each line of data corresponds to a 3x3 square neighbourhood of pixels completely contained within the 82x100 sub-area. Each line contains the pixel values in the four spectral bands (converted to ASCII) of each of the 9 pixels in the 3x3 neighbourhood and a number indicating the classification label of the central pixel. The number is a code for the following classes:

Number Class

1 red soil 2 cotton crop 3 grey soil 4 damp grey soil 5 soil with vegetation stubble 6 mixture class (all types present) 7 very damp grey soil

NB. There are no examples with class 6 in this dataset.

The data is given in random order and certain lines of data have been removed so you cannot reconstruct the original image from this dataset.

In each line of data the four spectral values for the top-left pixel are given first followed by the four spectral values for the top-middle pixel and then those for the top-right pixel, and so on with the pixels read out in sequence left-to-right and top-to-bottom. Thus, the four spectral values for the central pixel are given by attributes 17,18,19 and 20. If you like you can use only these four attributes, while ignoring the others. This avoids the problem which arises when a 3x3 neighbourhood straddles a boundary.

NUMBER OF EXAMPLES training set 4435 test set 2000

NUMBER OF ATTRIBUTES 36 (= 4 spectral bands x 9 pixels in neighbourhood)

ATTRIBUTES The attributes are numerical, in the range 0 to 255.

CLASS There are 6 decision classes: 1,2,3,4,5 and 7.

NB. There are no examples with class 6 in this dataset- they have all been removed because of doubts about the validity of this class.

4. En este problema, para cada muestra de tejido hay un microarray que contiene las expresiones de 2000 genes. Buscamos clasificar en una de dos clases: Normal o con tumor, usando 2000 variables explicativas correspondientes a los genes, a partir de una muestra de tamaño 62. El objetivo es, además de encontrar una regla de clasificación adecuada, hallar los genes más relevantes para clasificar. Los datos de este problema corresponden al artículo 'Broad patterns of gene expression revealed by

clustering of tumor and normal colon tissues probed by oligonucleotide arrays' U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, Proc. Natl. Acad. Sci. USA, Vol. 96, Issue 12, 6745-6750, June 8, 1999. Colon_X tiene la traspuesta de X: cada bloque de 62 números es una columna de X, y cada microarray es una fila. Use los métodos que le resulten adecuados, incluyendo "Nearest shrunken centroids", que fue inventado especialmente para estos casos. Los microarrays tienen mucha variabilidad; dos microarrays con la misma muestra de tejido pueden dar muy distintos. Si para cada uno de los 62 tejidos calculamos mediana y MAD y los graficamos, se ve que ambas varían enormemente, y que tienen una relación lineal. Lo que se acostumbra en estos casos es tomar logaritmo de todo.

5. Los siguientes datos corresponden a un trabajo para determinar la composición de un conjunto de vasijas de vidrio de un yacimiento arqueológico. Como el análisis espectrométrico es más barato que el análisis químico, se procuró calibrar el primero para que reemplace al segundo. Con este objetivo se tomó una muestra de 180 vasijas , a las que se realizó una espectrometria de rayos X sobre 1920 frecuencias, y también un análisis de laboratorio para determinar el contenido de 13 compuestos químicos, a saber:

   Cada fila del archivo Vessel_X es el espectro de una vasija, limitado a las frecuencias 100 a 400, pues las demás tienen valores casi nulos. Cada fila del archivo Vessel_Y tiene los contenidos de los 13 compuestos en esa vasija. Vamos a comparar distintos métodos. Ahora se trata de predecir el compuesto 1 (óxido de sodio). Utilizando todo lo aprendido, encontrar un modelo para poder predecir.