Enunciados para trabajo práctico final

Cada grupo o persona deberá analizar el set de datos que elija, y realizar el análisis que crea pertinente acorde a todo lo visto en la materia. Deberán entregar un informe a jmgarcia@fi.uba.ar en un archivo con formato pdf,y luego en una semana a convenir se realizarán las presentaciones.

Breves explicaciones de los datos:

1. El archivo *heart.csv* contiene 14 variables. Se busca predecir la presencia de una enfermedad al corazon (heart disease) en el paciente. Las variables son:

   a) age
   b) sex
   c) chest pain type (4 values)
   d) resting blood pressure
   e) serum cholestoral in mg/dl
   f) fasting blood sugar > 120 mg/dl
   g) resting electrocardiographic results (values 0,1,2)
   h) maximum heart rate achieved
   i) exercise induced angina
   j) oldpeak = ST depression induced by exercise relative to rest
   k) the slope of the peak exercise ST segment
   l) number of major vessels (0-3) colored by flourosopy
   m) thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
   n) target: 0= no heart disease, 1= heart disease

2. Set de datos *diabetes.csv*.
   Context
   This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

   Content
   The datasets consists of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

   Inspiration
   Can you build a machine learning model to accurately predict whether or not the patients in the dataset have diabetes or not?

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.

   Cuidado que son dos los archivos, colon_X son todos los datos de las covariables, pero después está colon_T donde tienen la variable respuesta
   . Los datos del paper: The matrix Colon_X contains the expression of the 2000 genes with highest minimal intensity across the 62 tissues. The genes are placed in order of descending minimal intensity. Each entry in Colon_X is a gene intensity derived from the 20 feature pairs that correspond to the gene on the chip, derived using the filtering process described in the 'materials and methods' section. The data is otherwise unprocessed (for example it has not been normalized by the mean intensity of each experiment).

   The identity of the 62 tissues is given in file Colon_tissues. The numbers correspond to patients, a positive sign to a normal tissue, and a negative sign to a tumor tissue.

   O sea, para cada muestra de tejido hay un microarray que contiene las "expresiones" de 2000 genes. Tenemos un problema de clasificación con p=2000 (genes), n=62(tejidos) y 2 clases ("+"=normal, "-"=tumor). 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.

6. Rendimiento academico de los estudiantes en matemática y portugues. Archivos *student-mat.csv* y *student-por.csv*. Se busca predecir la nota final de los estudiantes.

   Context:
   The data were obtained in a survey of students math and portuguese language courses in secondary school. It contains a lot of interesting social, gender and study information about students.

   Content:
   Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course)

   Variables:

   a) school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)

   b) sex - student's sex (binary: 'F' - female or 'M' - male)

   c) age - student's age (numeric: from 15 to 22)

   d) address - student's home address type (binary: 'U' - urban or 'R' - rural)

   e) famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)

   f) Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)

   g) Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)

   h) Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)

   i) Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')

   j) Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other') reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')

   k) guardian - student's guardian (nominal: 'mother', 'father' or 'other')

   l) traveltime - home to school travel time (numeric: 1 - 1 hour)

   m) studytime - weekly study time (numeric: 1 - 10 hours)

   n) failures - number of past class failures (numeric: n if 1¡=n¡3, else 4)

   ñ) schoolsup - extra educational support (binary: yes or no)

   o) famsup - family educational support (binary: yes or no) paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)

   p) activities - extra-curricular activities (binary: yes or no)

   q) nursery - attended nursery school (binary: yes or no)

   r) higher - wants to take higher education (binary: yes or no)

s) internet - Internet access at home (binary: yes or no)

t) romantic - with a romantic relationship (binary: yes or no)

u) famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)

v) freetime - free time after school (numeric: from 1 - very low to 5 - very high)

w) goout - going out with friends (numeric: from 1 - very low to 5 - very high)

x) Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)

y) Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)

z) health - current health status (numeric: from 1 - very bad to 5 - very good)

These grades are related with the course subject, Math or Portuguese:

G1 - first period grade (numeric: from 0 to 20)
G2 - second period grade (numeric: from 0 to 20)
G3 - final grade (numeric: from 0 to 20, output target)

Additional note: there are several (382) students that belong to both datasets . These students can be identified by searching for identical attributes that characterize each student, as shown in the annexed R file.

7. En el archivo *weatherAUS.csv* se encuentran datos acerca de vasios factores climáticos en Australia, entre otras cosas. El objetivo es decidir si puede crearse algún modelo ocn estas variables para poder predecir si va a llover mañana.

   Context
   Predict whether or not it will rain tomorrow by training a binary classification model on target RainTomorrow

   Content
   This dataset contains daily weather observations from numerous Australian weather stations.

   The target variable RainTomorrow means: Did it rain the next day? Yes or No.

   Note: You should exclude the variable Risk-MM when training a binary classification model. Not excluding it will leak the answers to your model and reduce its predictability.