Practical work on Multivariate Analysis

MIRI - Data Mining and Business Intelligence

Any data matrix contains information about the generating phenomenon. The practice will consist in choosing a real problem and applying multivariate techniques to reveal the hidden information contained in the data set, as a prior step to modeling. The problem can be chosen from the machine learning repository http://archive.ics.uci.edu/ml/, or from the following list:

Problems: DIRECMARK, GENOME, INCOME, SAHEART, SPAM, VOWEL, ZIP, BCNSES, POTEC, BREAST I MICROARRAY, INSURANCE, SURVEY, CREDSCO, DRINK, FIBTELE, HEALTH, SIGMUND, .

Moreover, the students can provide their own dataset, previous approval of the professor.

The student must perform a multivariate approach of the data matrix (visualization, clustering and interpretation) plus a prediction model suitable for the undertaken problem. The student must write a complete report upon the solution envisaged.

Steps for conducting the practice

1. The student will choose a problem and read the corresponding documentation trying to understand what is the objective of the problem and the available data.
2. Pre-process of data. The student will perform a first summary of the data, and, eventually, detect errors, outliers and missing values and take the appropriate measures of correction. According to the problem and data, it may be necessary to perform a selection of variables (feature selection) and /or a derivation of new explanatory variables (feature extraction) if the problem needs it.

3. The student will choose the type of protocol for the validation (i.e. holdout or test sample to assess the quality of the final model). (Depending on the data size, it won't make sense to have a separate test data file).
4. The student will perform a multivariate exploratory analysis of the training data set, taking the test data as supplementary. The Multivariate exploration will consist on the visualisation of the information, detection of the hidden latent factors. The synthesis of the complexity by clustering and interpretation of the results.
5. Projection of the test individuals on the relevant dimensions and assignement to the detected clusters.
6. Then, according the undertaken problem, the student will choose a model for prediction of the response variable within the ones explained in the MVA course, will find its optimal parameters and will evaluate its performance (generalization error) according the established validation protocol.

The report should include:

1. A description of the problem and available data
2. The pre-process of data
3. The protocol of validation
4. The visualisation performed
5. The interpretation of the latent concepts.
6. The clustering performed
7. The interpretation of the found clusters.
8. Results obtained with the assignment of the test individuals.
9. The prediction model with its best parameterization
10. The final model and its generalization error
11. Scientific and personal conclusions

MARKDIRECT

Description: A on-line shopping company is interested in having a prediction model for optimizing the marketing campaign of one of its products. Response variables: "codi" and "bons". The remaining variables are explicative; the continuous ones had been normalized dividing by its maximum value.

VARIABLES			inici	len.
1 "edat"			1	8
2 "eciv"			10	1
	"casat"	1		
	"solter"	2		
3 "nens"			12	1
	"sense nens"	1		
	"amb nens"	2		
4 "tprof"			14	1
	"quadres"	1		
	"obrers"	2		
	"inactius"	3		
5 "antreb"			16	8
6 "prod_A"			25	1
	"A no"	1		
	"A si"	2		
7 "prod_B"			27	1
	"B no"	1		
	"B si"	2		
8 "prod_C"			29	1
	"C no"	1		
	"C si"	2		
9 "prod_D"			31	1
	"D no"	1		
	"D si"	2		
10 "prod_E"			33	1
	"E no"	1		
	"E si"	2		
11 "nprod"			35	1
	"menys de 3 prods"	1		
	"3 o mes prods"	2		
12 "data_B"			37	8
13 "interes"			46	1
	"centre A"	1		
	"centre B"	2		
	"centre C"	3		
14 "total comprat"			48	8
15 "temps sense comprar"			57	8
16 "targeta"			66	1
	"targeta no"	1		
	"targeta si"	2		
17 "sexe"			68	1
	"dona"	1		
	"home"	2		
18 "habitat"			70	1
	"habitat 1"	1		
	"habitat 2"	2		
	"habitat 3"	3		
	"habitat 4"	4		
19 "codi"			72	1
20 "bons"			74	1
	"dolents"	1		
	"bons"	2		

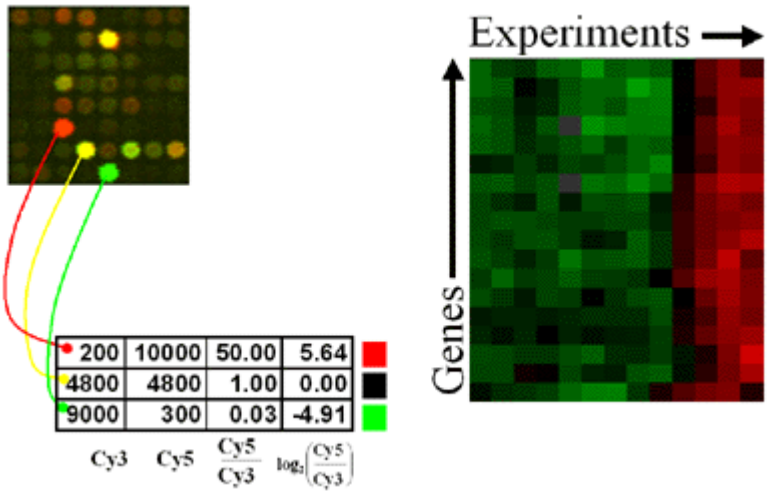
Genome. NCI microarray data

Source and reference:

http://genome-www.stanford.edu/nci60/

NCI microarray data

The data for one gene corresponds to one row, and each experiment is represented by a column. The ratio of induction/repression is such that the magnitude is indicated by the intensity of the colors displayed. If the color is black then the ratio of control to experimental cDNA is equal to 1, while the brightest colors (red and green) represent a ratio of 8 to 1. Ratios greater than 8 are displayed as the brightest color. In all cases red indicates an increase in mRNA abundance while green indicates a decrease in abundance in the experimental sample with respect to the control. Gray areas (when visible) indicate absent data, or data of low quality.



Systematic variation in gene expression patterns in human cancer cell lines.

Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M, Pergamenschikov A, Lee JC, Lashkari D, Shalon D, Myers TG, Weinstein JN, Botstein D, Brown PO.

Department of Biochemistry, Stanford University School of Medicine, Stanford, California, USA.

“We used cDNA microarrays to explore the variation in expression of approximately 8,000 unique genes among the 60 cell lines used in the National Cancer Institute's screen for anti-cancer drugs. Classification of the cell lines based solely on the observed patterns of gene expression revealed a correspondence to the ostensible origins of the tumours from which the cell lines were derived. The consistent relationship between the gene expression patterns and the tissue of origin allowed us to recognize outliers whose previous classification appeared incorrect. Specific features of the gene expression patterns appeared to be related to physiological properties of the cell lines, such as their doubling time in culture, drug metabolism or the interferon response. Comparison of gene expression patterns in the cell lines to those observed in normal breast tissue or in breast tumour specimens revealed features of the expression patterns in the

tumours that had recognizable counterparts in specific cell lines, reflecting the tumour, stromal and inflammatory components of the tumour tissue. These results provided a novel molecular characterization of this important group of human cell lines and their relationships to tumours in vivo.”

Se trata de ver hasta que punto la información proporcionada sobre los genes permiten validar los distintos tipos de tumores.

6830 genes (rows)
missing values have been imputed via SVD
60 cell columns, labels are below

CNS
CNS
CNS
RENAL
BREAST
CNS
CNS
BREAST
NSCLC
NSCLC
RENAL
RENAL
RENAL
RENAL
RENAL
RENAL
RENAL
BREAST
NSCLC
RENAL
UNKNOWN
OVARIAN
MELANOMA
PROSTATE
OVARIAN
OVARIAN
OVARIAN
OVARIAN
OVARIAN
PROSTATE
NSCLC
NSCLC
NSCLC
LEUKEMIA
K562B-repro
K562A-repro
LEUKEMIA
LEUKEMIA
LEUKEMIA
LEUKEMIA
LEUKEMIA
COLON
COLON
COLON
COLON
COLON
COLON
COLON

MCF7A-repro
BREAST
MCF7D-repro
BREAST
NSCLC
NSCLC
NSCLC
MELANOMA
BREAST
BREAST
MELANOMA
MELANOMA
MELANOMA
MELANOMA
MELANOMA
MELANOMA

Income Data

Marketing Database.

Source: Impact Resources, Inc., Columbus, OH (1987).
A total of N=9409 questionnaires containg 502 questions were filled out by shopping mall customers in the San Francisco Bay area.

The dataset income.data is an extract from this survey. It consists of 14 demographic attributes. The dataset is a good mixture of categorical and continuous variables with a lot of missing data. This is characteristic for data mining applications.

The goal is to predict the Anual Income of Household from the other 13 demographics attributes.

Attribute Information

1 ANNUAL INCOME OF HOUSEHOLD (PERSONAL INCOME IF SINGLE)
   1. Less than $10,000
   2. $10,000 to $14,999
   3. $15,000 to $19,999
   4. $20,000 to $24,999
   5. $25,000 to $29,999
   6. $30,000 to $39,999
   7. $40,000 to $49,999
   8. $50,000 to $74,999
   9. $75,000 or more

2 SEX
   1. Male
   2. Female

3 MARITAL STATUS
   1. Married
   2. Living together, not married
   3. Divorced or separated
   4. Widowed
   5. Single, never married

4 AGE
   1. 14 thru 17
   2. 18 thru 24
   3. 25 thru 34
   4. 35 thru 44
   5. 45 thru 54
   6. 55 thru 64
   7. 65 and Over

5 EDUCATION
   1. Grade 8 or less
   2. Grades 9 to 11
   3. Graduated high school
   4. 1 to 3 years of college
   5. College graduate
   6. Grad Study

6 OCCUPATION
   1. Professional/Managerial
   2. Sales Worker
   3. Factory Worker/Laborer/Driver

4. Clerical/Service Worker
5. Homemaker
6. Student, HS or College
7. Military
8. Retired
9. Unemployed

7 HOW LONG HAVE YOU LIVED IN THE SAN FRAN./OAKLAND/SAN JOSE AREA?
1. Less than one year
2. One to three years
3. Four to six years
4. Seven to ten years
5. More than ten years

8 DUAL INCOMES (IF MARRIED)
1. Not Married
2. Yes
3. No

9 PERSONS IN YOUR HOUSEHOLD
1. One
2. Two
3. Three
4. Four
5. Five
6. Six
7. Seven
8. Eight
9. Nine or more

10 PERSONS IN HOUSEHOLD UNDER 18
0. None
1. One
2. Two
3. Three
4. Four
5. Five
6. Six
7. Seven
8. Eight
9. Nine or more

11 HOUSEHOLDER STATUS
1. Own
2. Rent
3. Live with Parents/Family

12 TYPE OF HOME
1. House
2. Condominium
3. Apartment
4. Mobile Home
5. Other

13 ETHNIC CLASSIFICATION
1. American Indian
2. Asian
3. Black
4. East Indian
5. Hispanic
6. Pacific Islander

7. White
8. Other

14 WHAT LANGUAGE IS SPOKEN MOST OFTEN IN YOUR HOME?
1. English
2. Spanish
3. Other

Number of instances: 8993.
These are obtained from the original dataset with 9409 instances,by removing those observations with the response (Annual Income) missing.

The missing value flag is NA.

Coronary Heart Disease Survey

Medical database.

A retrospective sample of males in a heart-disease high-risk region of the Western Cape, South Africa. There are roughly two controls per case of CHD. Many of the CHD positive men have undergone blood pressure reduction treatment and other programs to reduce their risk factors after their CHD event. In some cases the measurements were made after these treatments. These data are taken from a larger dataset, described in Rousseauw et al, 1983, South African Medical Journal.

The goal is to predict the CHD from the other attributes.

Attribute Information

sbp	systolic blood pressure
tobacco	cumulative tobacco (kg)
ldl	low densiity lipoprotein cholesterol
adiposity	
famhist	family history of heart disease (Present, Absent)
typea	type-A behavior
obesity	
alcohol	current alcohol consumption
age	age at onset
chd	response, coronary heart disease

SPAM E-mail Database

Creator: George Forman, Hewlett-Packard Labs, 1501 Page Mill Rd., Palo Alto, CA 94304

Hewlett-Packard Internal-only Technical Report. External forthcoming.

Determine whether a given email is spam or not.
~7% misclassification error.

False positives (marking good mail as spam) are very undesirable.

If we insist on zero false positives in the training/testing set, 20-25% of the spam passed through the filter.

Relevant Information:

The "spam" concept is diverse: advertisements for products/web sites, make money fast schemes, chain letters, pornography...
Our collection of spam e-mails came from our postmaster and individuals who had filed spam.
Our collection of non-spam e-mails came from filed work and personal e-mails, and hence the word 'george' and the area code '650' are indicators of non-spam. These are useful when constructing a personalized spam filter. One would either have to blind such non-spam indicators or get a very wide collection of non-spam to generate a general purpose spam filter.

For background on spam:

Cranor, Lorrie F., LaMacchia, Brian A. Spam!
Communications of the ACM, 41(8):74-83, 1998.

Number of Instances: 4601 (1813 Spam = 39.4%)

Number of Attributes: 58 (57 continuous, 1 nominal class label)

Attribute Information:

The last column of 'spambase.data' denotes whether the e-mail was considered spam (1) or not (0), i.e. unsolicited commercial e-mail.

Most of the attributes indicate whether a particular word or character was frequently occuring in the e-mail.
The run-length attributes (55-57) measure the length of sequences of consecutive capital letters. For the statistical measures of each attribute, see the end of this file. Here are the definitions of the attributes:

48 continuous real [0,100] attributes of type word_freq_WORD = percentage of words in the e-mail that match WORD, i.e. 100 * (number of times the WORD appears in the e-mail) / total number of words in e-mail. A "word" in this case is any string of alphanumeric characters bounded by non-alphanumeric characters or end-of-string.

6 continuous real [0,100] attributes of type char_freq_CHAR = percentage of characters in the e-mail that match CHAR, i.e. 100 *

(number of CHAR occurences) / total characters in e-mail

1 continuous real [1,...] attribute of type capital_run_length_average = average length of uninterrupted sequences of capital letters

1 continuous integer [1,...] attribute of type capital_run_length_longest = length of longest uninterrupted sequence of capital letters

1 continuous integer [1,...] attribute of type capital_run_length_total = sum of length of uninterrupted sequences of capital letters = total number of capital letters in the e-mail

1 nominal {0,1} class attribute of type spam = denotes whether the e-mail was considered spam (1) or not (0), i.e. unsolicited commercial e-mail.

Missing Attribute Values: None

Statistics:

Spam	1813	(39.4%)
Non-Spam	2788	(60.6%)

This file: 'spambase.DOCUMENTATION' at the UCI Machine Learning Repository. http://www.ics.uci.edu/~mlearn/MLRepository.html

Vowel Recognition

SUMMARY: Speaker independent recognition of the eleven steady state vowels of British English using a specified training set of lpc derived log area ratios.

SOURCE: David Deterding (data and non-connectionist analysis)
Mahesan Niranjan (first connectionist analysis)
Tony Robinson (description, program, data, and results)

To contact Tony Robinson by electronic mail, use address "ajr@dsl.eng.cam.ac.uk"

MAINTAINER: neural-bench@cs.cmu.edu

PROBLEM DESCRIPTION:

The problem is specified by the accompanying data file, "vowel.data". This file is in the standard CMU Neural Network Benchmark format.

METHODOLOGY:

We have applied a variety of feed-forward networks to the task of recognition of vowel sounds from multiple speakers. Single speaker vowel recognition studies by Renals and Rohwer [RenalsRohwer89-ijcnn] show that feed-forward networks compare favourably with vector-quantised hidden Markov models. The vowel data used in this chapter was collected by Deterding [Deterding89], who recorded examples of the eleven steady state vowels of English spoken by fifteen speakers for a speaker normalisation study.

Report the number of test vowels classified correctly, (i.e. the number of occurrences when distance of the correct output to the actual output was the smallest of the set of distances from the actual output to all possible target outputs).

Though this is not the focus of Robinson's study, it would also be useful to report how long the training took (measured in pattern presentations or with a rough count of floating-point operations required) and what level of success was achieved on the training and testing data after various amounts of training. Of course, the network topology and algorithm used should be precisely described as well.

RESULTS:

Here is a summary of results obtained by Tony Robinson. A more complete explanation of this data is given in the exceprt from his thesis in the COMMENTS section below.

Classifier	no. of hidden units	no. correct	percent correct
Single-layer perceptron	-	154	33
Multi-layer perceptron	88	234	51
Multi-layer perceptron	22	206	45

Multi-layer perceptron	11	203	44
Nearest neighbour	-	260	56

The Speech Data

(An ascii approximation to) the International Phonetic Association (I.P.A.) symbol and the word in which the eleven vowel sounds were recorded is given in table 4.1. The word was uttered once by each of the fifteen speakers. Four male and four female speakers were used to train the networks, and the other four male and three female speakers were used for testing the performance.

vowel	word	vowel	word
i	heed	O	hod
I	hid	C:	hoard
E	head	U	hood
A	had	u:	who'd
a:	hard	3:	heard
Y	hud		

Table 4.1: Words used in Recording the Vowels

Front End Analysis

The speech signals were low pass filtered at 4.7kHz and then digitised to 12 bits with a 10kHz sampling rate. Twelfth order linear predictive analysis was carried out on six 512 sample Hamming windowed segments from the steady part of the vowel. The reflection coefficients were used to calculate 10 log area parameters, giving a 10 dimensional input space.

Each speaker thus yielded six frames of speech from eleven vowels. This gave 528 frames from the eight speakers used to train the networks and 462 frames from the seven speakers used to test the networks.

ZIP decoding

Normalized handwritten digits, automatically scanned from envelopes by the U.S. Postal Service. The original scanned digits are binary and of different sizes and orientations; the images here have been deslanted and size normalized, resulting in 16 x 16 grayscale images (Le Cun et al., 1990).

The data are in two gzipped files, and each line consists of the digit id (0-9) followed by the 256 grayscale values.

There are 7291 training observations and 2007 test observations, distributed as follows:

	0	1	2	3	4	5	6	7	8	9	Total
Train	1194	1005	731	658	652	556	664	645	542	644	7291
Test	359	264	198	166	200	160	170	147	166	177	2007

or as proportions:

	0	1	2	3	4	5	6	7	8	9
Train	0.16	0.14	0.1	0.09	0.09	0.08	0.09	0.09	0.07	0.09
Test	0.18	0.13	0.1	0.08	0.10	0.08	0.08	0.07	0.08	0.09

The test set is notoriously "difficult", and a 2.5% error rate is excellent. These data were kindly made available by the neural network group at AT&T research labs (thanks to Yann Le Cunn).

BCNSES. Evolution of the socioeconomic typology of Barcelona

Barcelona ha experimentado en los últimos años cambios notables. Se trata de realizar una síntesis de estos cambios a fin de poderlos cuantificar. Para ello se dispone de información sobre los 248 ZRPs (“Zones de Recerca Petites”) dando la repartición socioprofesional de sus habitantes, para el año 1988 y para el año 1996.
Es bien conocido que la posición social de las personas (y familias) es un factor explicativo de primer orden en múltiples comportamientos humanos, en política, en consumo, etc. Sin embargo la posición social no es fácil de medir. Una forma de definirlo es utilizando información secundaria (ya recogida) como son los datos padronales agregados por ZRPs (para evitar el problema de la confidencialidad de estos datos).

Se trata de obtener una tipología para las ZRPs de Barcelona a partir de la mínima información disponible con los datos de 1988, validarla y utilizar las reglas obtenidas para clasificar los datos de 1996 y evaluar los cambios producidos.

Barcelona has experienced in recent years remarkable changes. The objective is to detect and quantify these changes, using public information on the neighborhoods of Barcelona. Barcelona is divided into 248 neighborhoods, named ZRPs ("Zones of Recerca Petites"). The available data gives the distribution of the socioeconomic characteristics of its inhabitants for the year 1988 and for 1996.
It is well known that the social position of individuals (and families) is a prime explanatory factor in multiple human behavior; in politics, consumption, and so on. But the social position is not easy to measure. One way to define it is using secondary information as is the official data gathered by the municipality.

The practical work will consist of finding a typology for the ZRPs Barcelona from the data of 1988, then taking this typology as a response variable, find a prediction model using the same data of 1988, and apply this model to predict the data of 1996. Finally evaluate the found changes.

POTEC. Economic potential

This data was extracted from the census bureau database found at
http://www.census.gov/ftp/pub/DES/www/welcome.html
Donor: Ronny Kohavi and Barry Becker,
Data Mining and Visualization
Silicon Graphics.
e-mail: ronnyk@sgi.com for questions.
Split into train-test (2/3, 1/3 random).
32561 instances with some unknown values.
Duplicate or conflicting instances : 6
Class probabilities for adult.all file
Probability for the label '>50K' : 23.93%/24.78% (without unknowns)
Probability for the label '<=50K' : 76.07%/75.22% (without unknowns)

Extraction was done by Barry Becker from the 1994 Census database. A set of reasonably clean records was extracted using the following conditions:
((AAGE>16) && (AGI>100) && (AFNLWGT>1)&& (HRSWK>0))

Prediction task is to determine whether a person makes over 50K a year.

First cited in:
@inproceedings{kohavi-nbtree,
author={Ron Kohavi},
title={Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid},
booktitle={Proceedings of the Second International Conference on Knowledge Discovery and Data Mining},
year = 1996,
pages={to appear}}

Error Accuracy reported as follows, after removal of unknowns from train/test sets):
C4.5 : 84.46+-0.30
Naive-Bayes: 83.88+-0.30
NBTree : 85.90+-0.28

Following algorithms were later run with the following error rates, all after removal of unknowns and using the original train/test split.

	Algorithm	Error
1	C4.5	15.54
2	C4.5-auto	14.46
3	C4.5 rules	14.94
4	Voted ID3 (0.6)	15.64
5	Voted ID3 (0.8)	16.47
6	T2	16.84
7	1R	19.54
8	NBTree	14.10
9	CN2	16.00
10	HOODG	14.82
11	FSS Naive Bayes	14.05
12	IDTM (Decision table)	14.46
13	Naive-Bayes	16.12
14	Nearest-neighbor (1)	21.42
15	Nearest-neighbor (3)	20.35

 16 OC1   15.04
 17 Pebls   Crashed. Unknown why (bounds WERE
increased)

 Conversion of original data as follows:
 1. Discretized agrossincome into two ranges with threshold 50,000.
 2. Convert U.S. to US to avoid periods.
 3. Convert Unknown to "?"
 4. Run MLC++ GenCVFiles to generate data,test.

 Description of fnlwgt (final weight)

 The weights on the CPS files are controlled to independent estimates
of the civilian noninstitutional population of the US. These are
prepared monthly for us by Population Division here at the Census
Bureau. We use 3 sets of controls.
  These are:
    1. A single cell estimate of the population 16+ for each   state.
    2. Controls for Hispanic Origin by age and sex.
    3. Controls by Race, age and sex.

 We use all three sets of controls in our weighting program and "rake"
through them 6 times so that by the end we come back to all the
controls we used.

 The term estimate refers to population totals derived from CPS by
creating "weighted tallies" of any specified socio-economic
characteristics of the population.

 People with similar demographic characteristics should have
 similar weights. There is one important caveat to remember
 about this statement. That is that since the CPS sample is
 actually a collection of 51 state samples, each with its own
 probability of selection, the statement only applies within
 state.

>50K, <=50K.

age: continuous.
workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov,
Local-gov, State-gov, Without-pay, Never-worked.
fnlwgt: continuous.
education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-
acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th,
Doctorate, 5th-6th, Preschool.
education-num: continuous.
marital-status: Married-civ-spouse, Divorced, Never-married,
Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-
managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-
clerical, Farming-fishing, Transport-moving, Priv-house-serv,
Protective-serv, Armed-Forces.
relationship: Wife, Own-child, Husband, Not-in-family, Other-relative,
Unmarried.
race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
sex: Female, Male.
capital-gain: continuous.
capital-loss: continuous.
hours-per-week: continuous.

native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

BREAST. Diagnostic de càncer de mama.

Source Information
a) Creators:
Dr. William H. Wolberg, General Surgery Dept., University of Wisconsin, Clinical Sciences Center, Madison, WI 53792
wolberg@eagle.surgery.wisc.edu
W. Nick Street, Computer Sciences Dept., University of Wisconsin, 1210 West Dayton St., Madison, WI 53706
street@cs.wisc.edu 608-262-6619
Olvi L. Mangasarian, Computer Sciences Dept., University of Wisconsin, 1210 West Dayton St., Madison, WI 53706
olvi@cs.wisc.edu

b) Donor: Nick Street
c) Date: November 1995

See also:
http://www.cs.wisc.edu/~olvi/uwmp/mpml.html
http://www.cs.wisc.edu/~olvi/uwmp/cancer.html

Results:
- predicting field 2, diagnosis: B = benign, M = malignant
- sets are linearly separable using all 30 input features
- best predictive accuracy obtained using one separating plane in the 3-D space of Worst Area, Worst Smoothness and Mean Texture. Estimated accuracy 97.5% using repeated 10-fold crossvalidations. Classifier has correctly diagnosed 176 consecutive new patients as of November 1995.

4. Relevant information
Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.
A few of the images can be found at
http://www.cs.wisc.edu/~street/images/
Separating plane described above was obtained using Multisurface Method-Tree (MSM-T) [K. P. Bennett, "Decision Tree Construction Via Linear Programming." Proceedings of the 4th Midwest Artificial Intelligence and Cognitive Science Society, pp. 97-101, 1992], a classification method which uses linear programming to construct a decision tree. Relevant features were selected using an exhaustive search in the space of 1-4 features and 1-3 separating planes.
The actual linear program used to obtain the separating plane in the 3-dimensional space is that described in:
[K. P. Bennett and O. L. Mangasarian: "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets", Optimization Methods and Software 1, 1992, 23-34].
This database is also available through the UW CS ftp server:
ftp ftp.cs.wisc.edu
cd math-prog/cpo-dataset/machine-learn/WDBC/

5. Number of instances: 569

6. Number of attributes: 32 (ID, diagnosis, 30 real-valued input features)

7. Attribute information

1) ID number
2) Diagnosis (M = malignant, B = benign)

3-32)
Ten real-valued features are computed for each cell nucleus:
a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

All feature values are recoded with four significant digits.

8. Missing attribute values: none

9. Class distribution: 357 benign, 212 malignant

MICROARRAY. Predict disease classes using genetic microarray data

Data
Gene data is in genes-in-rows format, comma-separated values.
Take microarray.zip file, and unzip to extract 3 files:

- microarray_train.xls (training data, 1.7 MB)
- microarray_train_class.txt (training data classes)
- microrarray_test.xls (test data, 0.6MB)

Instructions

Training data: file microarray_train.xls, with with 7070 genes for 69 samples. A separate file microarray_train_class.txt has classes for each sample, in the order corresponding to the order of samples in microarray_train.xls. There are 5 classes, labelled EPD, JPA, MED, MGL, RHB.

Test data: file microrarray_test.csv, with 23 **unlabelled** samples and same genes. You can assume that the class distribution is similar.

Your goal is to learn the best model from the training data and use it to predict the label (class) for each sample in test data. You will also need to write a paper describing your effort.

Randomization experiments showed that one can get about 10-12 (from 23) correct answers with random guessing.

Below are suggested steps for doing this experiment, but you can vary and improve on the suggested approach, as long as you produce a prediction for the test set and describe your results.

Step 1. Data Exploration and Cleaning

Step 2. Selecting top genes (feature selection). Select the most discriminant genes of disease classes using the Fisher F. Craete new train and test data files with the retained genes.

Step 3. Find the best classifier/best gene set combination

Step 4. Generate predictions for the test set

Tingueu en compte que els fitxers venen amb els gens per fila i els cassos per columnes!

INSURANCE DATA

Description

Find a prediction rules for good and bad drivers (SPAD example)

1106 Belgian automobile insurance contracts.

Two groups, one with 0 claims (good), and other with some claims (bad).

List of explicative variables:

```
Accidents               2
      0 claim
      >1 claim
User                    2
      professional
      private
Age9                    9
Sex                     3
      male
      female
      company
Language                2
      french
      flemish
Postal_code12                 12
Bonus-maluscurrentyear11      11
Bonus-maulspreviousyear11     11
Age8                    8
Bonus-maluscurrentyear9       9
Bonus-maulspreviousyear9      9
Horsepower12                  12
Age3                    3
      1890-1949
      1950-1973
      ????
Bonus-maluspreviousyear2      2
      B-M 1
      others B-M
Contract_duration       2
      <86 contracts
      other contracts
Region                  2
      Brussels
      other region
Horsepower2             2
      10-39 HP
      >40 HP
Car_old                       2
      1933-1989 YVC
      1990-1991 YVC
Primes
```

CREDSCO DATA

Description

Find a prediction rules to imitate analysts when conceding or not a mortgage to clients.

Number of instances:
Missing value: 99999999
Variables (in Spanish). Response variable. Dictamen

```
   2 dictamen final
posi positivo
nega negativo
     motidict
   3 resc resultado del scoring
dene denegado
duda dudoso
apro aprovado
   2 formalizacion
fosi si
fono no
   2 morosidad
mosi si
mono no
   9 decision
apau aprovado automatico
amsa aprov. man sco-aprov
amsd aprov man sco duda
amsn aprov man sco negativo
dmsa deneg man sco aprov
dmsd deneg man sco duda
dmsn deneg man sco negat
deau denegado automatico
dere denegado por registros
     anti antiguedad en el trabajo
   6 vivienda
vial alquiler
viep escritura publica
vicp contrato privado
viic ignora contrato
vipa padres
viot otros
   4 cemp cargo empleados
ceac altos cargos
ceci cuadros intermedios
ceoe obrero especializado y administrativos
ceon obrero no especializado
     term plazo
  10 semp sector empleados
sefu funcionario
semi militares y policias
seco construcci¢n vidrio ceramica
sete textil
seme minero siderometalurgico
sequ quimica
sese servicios
seag agricultura
sere religiosos
```

```
seot otros
     edad
   5 eciv estado civil
solt soltero
casa casado
viud viudo
sepa separado
divo divorciado
  10 saut sector autonomos
satx taxi
satr transporte
saar arquitectos
saab abogado
same medico
sepl otras prof. liberales
saco comercio
sase servicios
saag agricultura
saot otros
   2 cliente
clno no
clsi si
   2 ebie estado del bien
nuev nuevo
viej viejo
  22 plaza
barc barcelona
bpro barcelona provincia
giro girona
llei lleida
tarr tarragona
vale valencia
cast castellon
alic alicante
murc murcia
alba albecete
jaen jaen
alme almeria
cord cordoba
sevi sevilla
cadi cadiz
gran granada
huel huelva
mala malaga
cace caceres
bada badajoz
madr madrid
plot otras
   2 registros
reno no
resi si
   4 titr tipo de trabajo
ttef empleado fijo
ttet empleado temporal
ttau autonomo
ttot otros
     puntos
     gastos
     ingresos
     cuota
     patrimonio
```

cpat cargas patrimoniales
importe
precio
vviv valor vivienda

DRINK_SURVEY DATA

Description

A drink manufacturer is interested in analysing its perception of drinking respect to other alcoholic alternatives and to build a prediction model of the likeliness to detect its drivers.

Instances:

Variables (questions in Spanish):

F3 **Anotar sexo (QUOTES)**

F3b **Me podría indicar cuántas personas viven en su hogar incluyéndose vd.?**

F4 **Por favor, ¿Podría decirme su edad exacta? (QUOTES)** F4

F6 **De qué nacionalidad es vd.?** F6

F7 **Anotar LUGAR DE RESIDENCIA (QUOTES)** F5

F7B **Anotar TAMAÑO DE HABITAT (QUOTES)** F5B

P4 **¿Con qué frecuencia suele beber …**

P8 ¿Por qué motivos consume usted.....? Y por qué más?

10 **Normalmente consumimos o compramos cosas porque tenemos necesidades que cubrir, por ejemplo vestirnos, saciar la sed, presumir....el ser humano tiene muchas necesidades que satisfacer y algunos productos nos ayudan a satisfacerlas. Ahora pensando en BEBIDAS CON ALCOHOL, hasta qué punto cada de las bebidas que le leo satisface sus necesidades? Para ello utilice una escala de 1 a 10, donde 1 es "No satisface para nada mis necesidades" y 10 es "Satisface totalmente mis necesidades".**

P16 **Me podría indicar qué bebidas consume habitualmente en cada una de las situaciones que le leo a continuación? Y cuál más?**

P21 **Y ahora cuando piensa en .....qué ideas le vienen a la cabeza**

P14 **Algunas veces las personas se ven imposibilitadas de tomar una bebida que ellos quieren por una razón u otra. Por favor, indíqueme de las razones que le voy a ir leyendo por cuáles no ha consumido alguna de estas bebidas**

P23 Hasta qué punto está usted de acuerdo con cada una de las frases que le leo a continuación? Para ello utilice una escala del 1 al 7 donde 1 significa "Completamente en desacuerdo" y 7 significa "Completamente de acuerdo".

c1 **Por favor, ¿me podría decir su estado civil?**

c2 **¿Puede decirme su nivel de estudios terminados?**

c3 **¿Es vd. el principal sustentador del hogar?**

c4 **Marque el nivel de estudios TERMINADOS del principal sustentador del hogar**

c5 **Marque su ocupación actual/ocupación actual del principal sustentador del hogar**

c6 **SOCIAL CLASS (QUOTES)**

FIBTELE_SURVEY DATA

Description

The purpose is to gain insight the alumni of two ICT schools according their perceptions about the performed career, to perform a clustering based on these perceptions, and use it to model the salary (three years after the graduation).

Variables:

Career	
Gender	
Age	
Studying	
Contract	
Salary	
Firmtype	
Accgrade	
Grade	
Startwork	
ima1	It's the best to study informatics
ima2	It is internationally recognized
ima3	It has a wide range of courses
ima4	The teachers are good
ima5	The facilities and equipment are good
ima6	Is leading research
ima7	It is highly regarded by companies
ima8	Can adapt to new needs and technologies
quaf1	Quality of the studies: the theoretical base
quaf2	Qualityof the studies: the technical competences
quaf3	Qualityof the studies: the applied training
qutr1	The ability to solve problems from
qutr2	Training in business management
qutr3	The written and oral communication skills
qutr4	Planning and time management acquired
qutr5	The ability to work in teams
val1	Allowed me to find a well-paid job
val2	I have prospects for improvement and promotion
val3	Allowed me to find a job that motivates me
val4	The training received is the basis on which I will build my career
sat1	I am satisfied with the training received
sat2	I am satisfied with my current situation
sat3	I think I'll have a good professional career
sat4	I think in the prestige of my work

DRINK_SURVEY DATA

Description,

The purpose is to model the main health risk factors for population : SBP (Systolic Blood Pressure), DBP (Diastolic Blood pressure), Glycemia and Cholesterol, from all available data. The risk factors can be defined according the following rules: hyperglycemia is based on having fasting plasma glucose greater than or equal to 126 mg/dL, or taking medication for diabetes. A person can be classified as having hypertension based on systolic blood pressure greater than 140 mmHg or having diastolic blood pressure greater than 90 mmHg (135 mmHg and 85 mmHg respectively for diabetic persons),or taking medication to control blood pressure. Cholesterolemia can be based on having more than 250 mg of cholesterol per deciliter of blood (mg/dL).

Variables:

gender
age
blood_pressure?
heart_attack?
heart_other?
diabetes?
cholesterol?
stroke?
heart_treat?
Blood_pressure_treat?
cholesterol_treat?
diabetes_treat?
blood_pressure_control?
cholesterol_control?
physical_exercise
strong_physical_exercise
alcohol
heavy_drinker
smoker
heavy_smoker
BMI
SBP
DBP
GLYCEMIA
CHOLESTEROL

SIGMUND ADMISSION_SURVEY DATA

Description,

The purpose is to model the admission to a large company using the answers given to a questionnaire.

Variables (in Spanish):

1. Gusto por el trabajo
2. Capacidad de trabajo
3. Voluntad y perseverancia
4. Ambici¢n
5. Sentido de la competitividad
6. Sentido de la eficacia
7. Autoridad natural
8. Capacidad para dirigir
9. Capacidad de persuasi¢n
10. Capacidad de negociaci¢n
11. Capacidad para asumir riesgos
12. Esp¡ritu de iniciativa
13. Capacidad para innovar
14. Sentido de la organizaci¢n y el m,todo
15. Capacidad de adaptaci¢n a las t,cnicas nuevas
16. Capacidad para trabajar en el extranjero
17. Presentaci¢n
18. Respeto por los usos y modos sociales
19. Tacto y delicadeza
20. Facilidad de contacto
21. Optimismo y alegria de vivir
22. Capacidad para trabajar en equipo
23. Tolerancia
24. Respeto por la jerarquia
25. Sensibilidad ala opinion de los dem s
26. Capacidad para escuchar
27. Capacidad para hablar en p£blico
28. Capacidad de adaptaci¢n a las situaciones nuevas
29. Discreci¢n
30. Equilibrio personal
31. Estabilidad de comportamiento
32. Espontaneidad
33. Resistencia a la frustraci¢n y al fracaso
34. Independencia
35. Confianza en si mismo
36. Sentido de la realidad
37. XXXX
38. PUNTS
39. ADMITIDO Si/No

SOCIOLOGICAL_SURVEY

Description: Establish a clustering of respondents according their opinions and then relate these clusters with the open questions about the marriage.

Questionnaire in French.