Exercise 3

Exercise 3 is due to 23.5.2017 (uploaded on CEWebS) and will be discussed on 24.5.2017.

1. Regression

The dataset `Toyota Corolla.csv` shows information about 1435 used Toyota cars. The following variables are available:

Price	Offer price in EURO
Age	in month
KM	Accumulate Km
Fuel Type	Fuel Type (petrol, Diesel, CNG)
HP	Horsepower
MetColor	Metallic Color (1 = yes, 0 = no)
Automatic	1= yes, 0 = no
Doors	Number of Doors
Weight	Weight in kilograms

Tasks:

a) Give a descriptive of the dataset.
b) Define a regression model which allows the prediction of the price in dependence of the other variables.
c) Split the data in a training set and test set and calculate the prediction error from the test set.

Note: This example is adapted from Leodolter: Data Mining and Business Analytics with R

2. Wholesale Customers

The dataset `WholesaleData.csv` shows for 440 customers of a wholesale distributor the following information:

FRESH	annual spending (m.u.) on fresh products (Continuous)
MILK	annual spending (m.u.) on milk products (Continuous)
GROCERY	annual spending (m.u.)on grocery products (Continuous)
FROZEN	annual spending (m.u.)on frozen products (Continuous)
DETERGENTS_PAPER	annual spending (m.u.) on detergents and paper products (Continuous)
DELICATESSEN	annual spending (m.u.)on and delicatessen products (Continuous)
CHANNEL	Horeca (Hotel/Restaurant/Café) or Retail channel (Nominal)
REGION	Lisbon, Oporto or Other (Nominal)

Tasks

a) Give a description of the data.
b) Find classification rules which allow the discrimination of the distribution channels.
c) Apply cluster analysis to find subgroups of the customers.

Note: This dataset was taken from the UCI Machine Learning Repository:
https://archive.ics.uci.edu/ml/datasets/Wholesale+customers

3. Credit Card Default

The dataset CreditDefault.xlsx informs about the defaults of 3000 credit card clients. The following variables are given:

X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
X2: Gender (1 = male; 2 = female).
X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
X4: Marital status (1 = married; 2 = single; 3 = others).
X5: Age (year).
X6 - X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .;X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -2,-1, 0 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.
X12-X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005.
X18-X23: Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005.

Y Credit default (1 = default, 0 = no default=

Tasks:

a) Using logistic regression to estimate the probability of credit default for the customers
b) Apply different classification methods for finding rules which determine the class membership.

Note: This example is adapted from the UCI Machine Learning Repository:
https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients