Exercises 3 and 5

Exercises 3 and 5 are due to 18.6.2018 (uploaded on CEWebS) and will be discussed on 20.5.2018.

For each exercise a maximum of 4.5 points will be assigned.

If you have further questions you can contact me by email.

1. Wholesale Customers 1

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

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

Tasks

a) Use descriptive statistics for business and data understanding and summarize the findings.

b) Find classification rules which allow the discrimination of the distribution channels and.

c) Find classification rules for discrimination of the regions.

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

2. Wholesale Customers 2

a) Apply for the dataset of exercise 1 different cluster analysis methods and select an appropriate solution for clustering.

b) Visualize the cluster solutions with different techniques and interpret the solutions.

c) Compare the cluster solution with the classification rules of exercise 1 b) and 1c). Are the results comparable?

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) Use descriptive statistics for business and data understanding and summarize the findings.

b) Split the dataset in a training and a test set (70%, 30%) and learn a classification rule from the training set using logistic regression. Find an appropriate threshold for the classification. Evaluate the solution for the test set.

c) Apply at least two other classification methods and compare the solutions with logistic regression from a numerical point of view (accuracy, sensitivity, specificity) and from a practitioners point of view (interpretation and understanding of the results).

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

4. Temporal analysis

The data Timeseries.xlsx show the monthly access to a server from users in two different countries at different locations for a period of three years (36 month). In Country 1 there are 6 locations and in country 2 there are 8 locations. Additionally, the total accesses from the two countries are shown.

The following questions are of interest:

a) Are there differences in the patterns at different locations over the period of three years?

b) Are the access patterns of locations representative for the overall traffic in the two countries?

c) Are there substantial differences in the access patterns between the two countries?

For analyzing question a) split the time series from each location into 3 annual time series of length 12 month. Use for the resulting 18 time series of country 1 and the 24 time series of country 2 time warping for finding the warping distances. Apply cluster analysis for the 18 resp. 24 time series based on the warping distances. What is your conclusion with respect to the annual patterns?

Formulate an algorithmic procedure for answering question b), compute the solution and answer the question.

For answering question c) find appropriate response features which can characterize the 14 time series and apply appropriate classification methods for classification the time series into the groups defined by the countries.