Business Analytics
Exercise Sheet 1

Josif Grabocka (josif@ismll.de)
Information Systems and Machine Learning Lab (ISMLL)
Universität Hildesheim

30 April 2013
Sumbission Until 5 May 2013 23:59

Question 1: R - Input/Output (1 point)

Download the dataset named 'sal-adv-inc.csv' from the course's website.
Load the CSV dataset in R and print the min, max and mean value of each column.

Question 2: R - Graphics (1 point)

- Create a multiple plot consisting of three rows and one column of scatter plots, respectively of the columns SALES vs ADVT, SALES vs INCOME and ADVT vs INCOME.
- Appropriately name the axis and the title of subplots.

Question 3: R - Kendall Tau Rank Correlation (4 points)

Assume your data is in form of tuples $D = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$
A pair is defined as $P_{i,j} = ((x_i, y_i), (x_j, y_j))\ \forall i, j \in \{1, ..., n\}$
The accordance or dis-accordance of a pair is defined in Equation 1.

$$\begin{array}{ll} \text{if} & (x_i \geq x_j \wedge y_i \geq y_j) \vee (x_i < x_j \wedge y_i < y_j) \ \rightarrow \ P_{i,j} \text{ is accordant} \\ \text{else} & P_{i,j} \text{ is dis-accordant} \end{array} \tag{1}$$

Finally the Kendall Tau Rank Correlation of two vectors $x, y \in R^n$ is defined in Equation 2.

$$\tau(x, y) = \frac{|\{(i,j)|P_{i,j} \text{ is accordant}\}| - |\{(i,j)|P_{i,j} \text{ is dis-accordant}\}|}{\frac{1}{2}n(n-1)} \tag{2}$$

- Implement an R function named kendalltau(x,y) wich receives two vectors x,y and computes the aforementioned correlation.
- Compute and print the Kendall Tau correlation of the ADVT and INCOME columns of your dataset. Comment briefly the results? Does investing on advertisement (ADV) pays of in form of income (INCOME)?

Question 4: R - Data Normalization (4 points)

The mean and standard deviation are commonly known vector metrics. Normalization of a vector intends to transform the data into having zero mean and a standard deviation of 1. More concretely the mean, standard deviation and normalization are defined in Equation 3.

$$\begin{aligned} \text{Let:} \quad & x \in R^n \\ \overline{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ \sigma_x &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} \\ x_i^{norm} &\leftarrow \frac{x_i - \overline{x}}{\sigma_x}, \quad \forall i \in \{1, ..., n\} \end{aligned} \tag{3}$$

- Create an R function which normalizes every column of a matrix/data sheet.
- Apply the normalization to the dataset of Question 1 and print the result.

Submission

- Electronically to josif@ismll.de
- Email title must be *BA2013-NAME-Tutorial-NO*, e.g.: *BA2013- JosifGrabocka-Tutorial-1*
- Report file must be a PDF with file name like the email title.
- Source codes and other materials must be a ZIP with file name like the email title.