

# Applied probability lab session 1

## 1 Data Cleaning

The `forestfires.txt` dataset is a collection of observations from the Monteshinho park in Portugal, that were used to build a predictive model of forest fire areas. The columns of the data frame are:

- `X` - x-axis spatial coordinate within the Montesinho park map: 1 to 9
- `Y` - y-axis spatial coordinate within the Montesinho park map: 2 to 9
- `month` - month of the year: “jan” to “dec”
- `day` - day of the week: “mon” to “sun”
- `temp` - temperature in degrees Celsius
- `RH` - relative humidity in %: 15.0 to 100
- `wind` - wind speed in km/h
- `rain` - outside rain in mm/m2
- `area` - the burned area of the forest (in ha)

1. Your dataset is given in the course website in Moodle, under `<groupXX>_forestfires.csv`. Read the dataset into R. The dataset is in a comma separated format, and has a header line.
2. Print the dataset out on the screen using the `print` command. Scroll through the dataset. With a small dataset such as this one, scrolling through the dataset can be useful to familiarize yourself with the data and spot problems. Can you see anything in this dataset that might pose a problem for future analysis? Can you see any entries that appear to be errors? If so, how would you propose fixing those errors? (Don't copy-past the entire dataframe).
3. Now summarize the dataset using the `summary` command. Does anything strike you as unusual in the data summaries? Investigate further and propose fixes for any errors or irregularities.
4. Create a scatterplot matrix of the variables and once again look for anomalies in the plots. Can you see any plots which might suggest outlying or extreme observations? Investigate these with individual scatterplots and/or boxplots. Are the outliers reasonable, or do they signify an error?
5. Provide a cleaned version of the dataset as an appendix at the end of assignment file, attempting to preserve as much data as possible. Justify each change you make.
6. Provide a statistical description of the data: each variable and each pair of variables. Comment.

Please make sure that your answers are as specific as possible. For example, do not write “I deleted some of the records because some entries appear to be errors.” Instead write record no... was deleted because field... appears to be incorrect - it contains a value which is ...” Unspecific answers will not be fully graded. Include all the R commands, plots, summaries and any other output of R you consider relevant. Submit your work and the full script file through Moodle.

## 2 Market Basket Analysis

The 8 items under study are: Hammer, Nails, Screws, Screwdriver, Wrench, Level, Drill, and Brush. The probabilities to put an item as the first item in the basket are (respectively): 0.02, 0.2, 0.2, 0.2, 0.2, 0.05, 0.01, 0.02, 0.3. Your matrix of conditional probabilities is given in the course website in Moodle under `<groupXX>_CondProbs.txt`. The cell  $M_{ij}$  in the matrix is the probability to buy item  $j$  given that the first item in the basket is  $i$ . We assume that the events “item  $i$  is in the basket” and “item  $j$  is in the basket” are conditionally independent given the event “item  $k$  was the first item in the basket.”

In this exercise, a rule will be called significant if it satisfies the following conditions:

- `support > 0.15`
- `confidence > 0.5`
- `lift > 1.1`

### Theoretical Calculations

1. Provide the formulas for computing the 1-item support, 2-item support, 2-item confidence and 2-item lift from the probabilities you were given.
2. Read the conditional probability matrix into R using the command `read.table`. The command `read.table` creates a data frame. In order to use mathematical operations, convert the data frame into a (numerical) matrix using the command:  
`cond.probs<-data.matrix(cond.probs)`
3. Compute the 1-item support, 2-item support, 2-item confidence and 2-item lift from the probabilities you were given.
4. Provide a list of significant 2-item rules.
5. Discuss shortly the rules you expect to find significant in the simulation.

### Data simulation and basket analysis of the simulated dataset

In this section you will simulate 5000 baskets using the probabilities you are given above. The dataset will be the matrix containing 5000 rows and eight columns. The rows correspond to baskets (transactions), while the eight columns represent the items ( “Hammer,” “Nails,” “Screws,” etc). Every row contains eight numbers, 0 or 1: 1 means that the corresponding item is in the basket, and 0 if it is not. The header of the dataset should contain the item names.

1. Set the random number generator seed to 34567 using the command `set.seed()`.
2. For each observation (transaction), draw the first item in the basket according to the given probabilities. Then using the conditional probabilities and the assumption of conditional independence, draw other items that will be bought in the transaction. Some useful commands: `sample()`, `runif()` . Propose a code without using “for loops”.
3. In this question we are interested in all two-item rules of the form  $A \Rightarrow B$ . Using matrix and vector operations (without using “for loops”):
  - (a) Compute the support for all single items.
  - (b) Compute the support for all two-item rules.
  - (c) Using only your support measurements, compute the confidence of all two-item rules.

- (d) Using only your support and confidence measurements, compute the lift of all two-item rules.
  - (e) List all the significant two-item rules. You may use a “for loop” to print the rules.
4. Discuss shortly the two-item rules you have found. Do they match the theoretical calculations you have performed in the previous section?

Include in your work all the R commands and any output of R that you think relevant. Submit your work and the full script file through Moodle.