Statistics Reminder-Introduction to R

Haute école de gestion de Genève
Informatique de Gestion
Alexandros Kalousis

Problem description

The director of the HR is trying to understand the reasons for which employes are quiting their jobs. In order to do so the HR department has created a database with the description of a number of employes over the last year. Your objective is to understand what are the main factors that determine whether an employ will stay or leave.

TP1: Exploratory analysis

- In this first TP you need to perform an exploratory and qualitative data analysis in order to better understand the data and the different variables
- You are going to use R to do the analysis and develop the necessary scripts.
- Based on the results of the analysis we will have to write a small report, in which you will deliver what is requested bellow, and describe what according to your analysis are the most important factors that determine whether an employ leaves or stays.

Deliverables I

D1: R script: generic, i.e. *runnable on any dataset*

- single main control function

```
runAnalysis(datasetName,
            DiscreteAttributesIndeces,
            AttrsToRemoveIndeces,
            ClassAttributeIndex)

#datasetName               : file name of the training set
#DiscreteAttributesIndeces: vector with discrete attr indeces
#AttrsToRemoveIndeces      : vector with attrs to remove
#ClassAttributeIndex       : index of class attribute
```

  - we type runAnalysis('train.csv',...) and get all the results
- fully functional - we test and crashes on errors = 0 points

best coding practices - statement of purpose and authorship, does not repeat (functions, loops), user/technical documentation (comments), independent of data (no hard-coded values, re-usable for other datasets), clear structure, etc.

Deliverables II

D2: Report

- explain the problem and discuss the results
- summary of main findings - factual, clear, and concise
- based on your code results - if report $\neq$ code = 0 points

standard *best report writing* practices - statement of purpose and authorship, clear structure, reader-oriented, distinguish facts from opinions, etc.

Preliminary analysis

Before you begin the analysis, you need to understand your dataset. For this you need to clearly answer in your report the following questions:

- How many instances and how many variables (attributes) are there?
- What is the target variable and is it quantitative or qualitative?
- The other attributes, are they quantitative or qualitative?
- Shall any variables be dropped from the analysis? Why?
- Are there any missing data?

Exploratory analysis

For each **qualitative** attribute $\mathbf{f}$:

- compute the probability distribution $P(\mathbf{f})$
- compute the conditional probability of the target variable, y, given the attribute values $P(y|\mathbf{f})$

TP1: Part 1 III

Choose some qualitative variable, $\mathbf{f}$, and the target variable y and show examples of how the following probability rules hold for them:

- $P(f, y) = P(f|y)PyY) = P(y|f)P(f)$, known as the multiplication rule.
- $P(y|f) = P(f|y)P(y)/P(f)$, known as the bayes rule.

with the help of the probability matrices that you have established above and the contigency matrix.

Using your examples explain the meaning of each one of the terms that you see above, i.e. $P(f, y), P(f|y), P(y), ...$ etc.

TP1: Part 1 IV

For each **quantitative** attribute $\mathbf{f}$:

- compute the mean $\mu(\mathbf{f})$ and variance $\sigma^2(\mathbf{f})$
- compute the mean $\mu(\mathbf{f}|target)$ and variance $\sigma^2(\mathbf{f}|target)$ conditioned on the target variable
- Order the variables in terms of their importance by computing the class conditional mean difference scaled by the standard deviation.

TP1: Part 1 V

Choose one quantitative variable, $\mathbf{f}$, and

- write down the normal distribution for the following cases: $P(\mathbf{f}), P(\mathbf{f}|y)$, explain what changes between the different distributions.
- plot these normal distributions, compare them to the respective histograms that we will create bellow, discuss the similarities and differences.

TP1: Part 1 VI

For each **qualitative** attribute $\mathbf{f}$:

- visualise the probability distribution $P(\mathbf{f})$ and the conditional probabilities $P(y|\mathbf{f})$ by bar charts

For each **quantitative** attribute $\mathbf{f}$:

- visualise the distribution $p(\mathbf{f})$ and the conditional distributions $p(\mathbf{f}|target)$ by histograms.

Which attributes seem the most and the least useful to predict the target variable?

Pick two continuous variables you think are the most useful and visualise their effect on the target in a scatter plot (nuage de points).

Hint: Use different colors (or symbols) to distinguish the target variable classes for the various combinations of the two continuous variables.