

1. Coursework details

You will undertake a project that will determine your final mark of the course by 30 per cent. The project will require you to analyse one or more real world datasets that will become available on the Moodle page of the course. The analysis will consist of three parts:

1. In the first part you will be given data collected from various individuals on several variables. The goal will be to use unsupervised learning techniques such as Principal Component Analysis or Cluster Analysis to summarise the information in the data by appropriate tables and/or plots.
2. In the second part you will be presented with a regression problem. The aim would be to compare various models and techniques for their estimation to allow meaningful interpretation and competitive predictive performance. The latter should be assessed by appropriate experiments based on training and test datasets. In addition to linear regression, Tree based methods, Non-linear models or other suitable techniques can be used if you think they can provide improvement.
3. Finally, in the third part you will be given a classification problem. The analysis will contain similar steps with the second part but you should be able to interpret the output from different models and compare their predictive performance taking into account that the response variable will be binary. In addition to appropriate regression or discriminant analysis, Tree-based methods, Non-linear models or other suitable techniques can be used if you think they will perform better.

In all of the three parts above, the output from these techniques should be described in non-technical language targeting people with a minimal quantitative background.

The results of the project should be presented in 10-page article in A4 format with Arial fonts (not Arial narrow) of size at least 11. All page margins should be at least 2 cm. The 10-page limit includes figures and tables but excludes the title page, table of contents and references. In addition to the 10-page article, which should be submitted both via a hard and a soft copy, your R code should also be submitted online via R script files which you can upload as a microsoft word document.

This coursework is due on the 1st of April

2. Coursework resources

In this section you will find resources related to your coursework such as the dataset to use for your report.

In the following sections you will find the datasets for each part of your project. You can import them all into R using the R script below:

```
#R code to import and prepare the EWCS dataset
ewcs=read.table("EWCS_2016.csv",sep=";",header=TRUE)
ewcs[,][ewcs[,] == -999] <- NA
kk=complete.cases(ewcs)
ewcs=ewcs[kk,]

#R code to import and prepare the student performance dataset
school1=read.table("student-mat.csv",sep=";",header=TRUE)
school2=read.table("student-por.csv",sep=";",header=TRUE)
schools=merge(school1,school2,by=c("school","sex","age","address","famsize","Pstatus","Medu","Fedu","Mjob","Fjob","reason","nursery","internet"))

#R code to import the bank marketing dataset

bank=read.table("bank.csv",sep=";",header=TRUE)
```

2.1. Part 1

Download the dataset for this part below

[EWCS_2016.csv](#)

As a reminder:

In the first part you will be given data collected from various individuals on several variables. The goal will be to use unsupervised learning techniques such as Principal Component Analysis or Cluster Analysis to summarise the information in the data by appropriate tables and/or plots.

Description for the EWCS dataset

Task:

Visualise and describe the data via unsupervised learning methods.

Source:

European Working Conditions Survey 2016.

Variable and value labels for EWCS 2016 dataset:

Q2a - Gender

Values 1: Male 2: Female

Q2b - Age

Values numeric

Q87a - I have felt cheerful and in good spirits [...which is the closest to how you have been feeling over the last two weeks]

Q87b - I have felt calm and relaxed [...which is the closest to how you have been feeling over the last two weeks]

Q87c - I have felt active and vigorous [...which is the closest to how you have been feeling over the last two weeks]

Q87d - I woke up feeling fresh and rested [...which is the closest to how you have been feeling over the last two weeks]

Q87e - My daily life has been filled with things that interest me [...which is the closest to how you have been feeling over the last two weeks]

Values for variables Q87a to Q87e

1. All of the time.
2. Most of the time
3. More than half of the time
4. Less than half of the time
5. Some of the time
6. At no time

Q90a - At my work I feel full of energy [Please tell me how often you feel this way...]

Q90b - I am enthusiastic about my job [Please tell me how often you feel this way...]

Q90c - Time flies when I am working [Please tell me how often you feel this way...]

Q90f - In my opinion, I am good at my job [Please tell me how often you feel this way...]

Values for variables Q90a to Q90f

1. Always.

2. Most of the time
3. Sometimes
4. Rarely
5. Never

You can download this information [here](#).

2.2. Part 2

You can download the datasets for this section below.

[student-mat.csv](#)

[student-por.csv](#)

As a reminder:

In the second part you will be presented with a regression problem. The aim would be to compare various models and techniques for their estimation to allow meaningful interpretation and competitive predictive performance. The latter should be assessed by appropriate experiments based on training and test datasets. In addition to linear regression, Tree based methods, Non-linear models or other suitable techniques can be used if you think they can provide improvement.

Description for Student Performance Dataset

Task:

Build a regression model for the variable G3 (final grade) without using the variables G1 and G2. Interpret the model and assess its predictive performance.

Source:

Paulo Cortez, University of Minho, Guimaraes, Portugal,

<http://www3.dsi.uminho.pt/pcortez>

Data Set Information:

This data approach student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features) and it was collected by using school reports and questionnaires.

Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por). In [Cortez and Silva, 2008], the two datasets were modeled under binary/five-level classification and regression tasks.

Important note: the target attribute G3 has a strong correlation with attributes G2 and G1. This occurs because G3 is the final year grade (issued at the 3rd period), while G1 and G2 correspond to the 1st and 2nd period grades. It is more difficult to predict G3 without G2 and G1, but such prediction is much more useful (see paper source for more details).

Attribute Information:

Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets:

- 1 school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
- 2 sex - student's sex (binary: 'F' - female or 'M' - male)
- 3 age - student's age (numeric: from 15 to 22)
- 4 address - student's home address type (binary: 'U' - urban or 'R' - rural)
- 5 famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
- 6 Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
- 7 Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
- 8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)

- 9 Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
- 10 Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
- 11 reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
- 12 guardian - student's guardian (nominal: 'mother', 'father' or 'other')
- 13 traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
- 14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
- 15 failures - number of past class failures (numeric: n if $1 \leq n < 3$, else 4)
- 16 schoolsup - extra educational support (binary: yes or no)
- 17 famsup - family educational support (binary: yes or no)
- 18 paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
- 19 activities - extra-curricular activities (binary: yes or no)
- 20 nursery - attended nursery school (binary: yes or no)
- 21 higher - wants to take higher education (binary: yes or no)
- 22 internet - Internet access at home (binary: yes or no)
- 23 romantic - with a romantic relationship (binary: yes or no)
- 24 famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
- 25 freetime - free time after school (numeric: from 1 - very low to 5 - very high)
- 26 goout - going out with friends (numeric: from 1 - very low to 5 - very high)
- 27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
- 28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
- 29 health - current health status (numeric: from 1 - very bad to 5 - very good)
- 30 absences - number of school absences (numeric: from 0 to 93)
- # these grades are related with the course subject, Math or Portuguese:
- 31 G1 - first period grade (numeric: from 0 to 20)
- 31 G2 - second period grade (numeric: from 0 to 20)
- 32 G3 - final grade (numeric: from 0 to 20, output target)

Relevant Papers:

P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th Future Business Technology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

You can download this information [here](#).

2.3. Part 3

You can download the dataset for this part below.

[bank.csv](#)

As a reminder:

Finally, in the third part you will be given a classification problem. The analysis will contain similar steps with the 2nd part but you should be able to interpret the output from different models and compare their predictive performance taking into account that the response variable will be binary. In addition to appropriate regression or discriminant analysis, Tree-based methods, Non-linear models or other suitable techniques can be used if you think they will perform better.

Description for Bank Marketing Dataset

Task:

Build a Classification model to predict if the client will subscribe (yes/no) a term deposit (variable y). Interpret the model and assess its predictive performance.

Source:

[Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

Data Set Information:

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

Attribute Information:

Input variables:

bank client data: 1 - age (numeric)
 2 - job : type of job (categorical: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')
 3 - marital : marital status (categorical: 'divorced', 'married', 'single', 'unknown'; note: 'divorced' means divorced or widowed)
 4 - education (categorical: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')
 5 - default: has credit in default? (categorical: 'no', 'yes', 'unknown')
 6 - housing: has housing loan? (categorical: 'no', 'yes', 'unknown')
 7 - loan: has personal loan? (categorical: 'no', 'yes', 'unknown')
 # related with the last contact of the current campaign:
 8 - contact: contact communication type (categorical: 'cellular', 'telephone')
 9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
 10 - day_of_week: last contact day of the week (categorical: 'mon', 'tue', 'wed', 'thu', 'fri')
 11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be

discarded if the intention is to have a realistic predictive model.

other attributes:

12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)

14 - previous: number of contacts performed before this campaign and for this client (numeric)

15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure', 'nonexistent', 'success')

social and economic context attributes

16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)

17 - cons.price.idx: consumer price index - monthly indicator (numeric)

18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)

19 - euribor3m: euribor 3 month rate - daily indicator (numeric)

20 - nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):

21 - y - has the client subscribed a term deposit? (binary: 'yes','no')

Relevant Papers:

S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimaraes, Portugal, October, 2011. EUROSIS. [bank.zip]

You can download this information [here](#).