Methods for Data Science (M345 A50)

Coursework 1 – Machine Learning

Deadline: Friday 16 November 2018, 5pm.

General instructions

The goal of this project is to analyse a dataset using the tools from Part 1 of our course. Note that coursework projects are different from exams. They are more open-ended and may require going beyond what we did in lectures. Initiative and creativity are important as is the ability to pull together the course content, draw new links between subjects and back up your analysis with relevant computations. **The quality of presentation and communication are very important**, so use good combinations of tables and figures to present your results.

You can find detailed submission instructions at the end of this document.

Dataset

The file 'druguse.Rdata' contains data on personality traits and use of various substances and illegal drugs for 1,885 individuals. The file is available in BB and the data is already as an R data frame.

The dataset includes **5 predictors** on the individual's background (`agegroup, gender, education, country, ethnicity`), plus **7 predictors** with scores for personality traits (`neuroticism, extraversion, opentoexperience, agreeableness, conscientiousness, impulsiveness, sensation`), and **4 predictors** on consumption of legal substances (`caffeine, chocolate, nicotine, alcohol`).

The next **14 columns** include consumption of various illegal substances. These columns *might* be predictors, depending on the question being asked (e.g. can personality and cannabis use predict heroin consumption?). In Question 3, you will need to think by yourself about a specific question and choose the right methods to answer it.

The consumption of all substances (legal and illegal) ranges from 0 to 6, meaning: never used (0), used over a decade ago (1), used in the last decade (2), used in last year (3), used in last month (4), used in last week (5), used in last day (6).

The column `any` counts whether the subjects have reported use of any of the illegal substances at all. The last two columns:

- `severity` is a score of the severity of drug consumption.
- `UseLevel` reports the 'high' or 'low' consumption of drugs. This was created from the rest of the data by adding up the reports for the other drugs, and counting subjects as 'high' users if their total was above a certain threshold.

Question 1

1.1 (Rmd only) Make two coloured histograms (e.g. using `geom_bar` in `ggplot`): **(a)** one with country on the x axis and set the fill aesthetic to UseLevel. This shows how many individuals are from each country, and within each country, how many have high vs low use levels. **(ii)** repeat but with gender on the x axis.

1.2 Perform exploratory data analysis. Create various visualisations (e.g. scatter plots, jitter plots, coloured bar plots, tile plots, etc) that illustrate the relationship between substance use and the predictor variables. **(i)** One of these should relate the categorical outcome `UseLevel` to one or more predictors, e.g. a scatter plot of extraversion and impulsiveness, coloured by `UseLevel`. **(ii)** One should relate the quantitative outcome, severity to one or more predictors. Feel free to use multiple panels/plots to present a visual exploration of the data. Present up to 6 plots in total. Discuss any relevant patterns you observe in the data.

Question 2

2.1 (Rmd only) Use logistic regression to build a classifier that predicts if an individual's substance use level will be 'high' or 'low' based on the predictors in the first 16 columns of the data. Hint: create another data frame with only the appropriate columns (1:16) and the `UseLevel` column. Use only the first 1400 observations to train the model and leave the rest as a validation data set. Are smokers more likely or less likely to have a high use level? What about chocolate eaters? Discuss.

2.2 (Rmd only) Make predictions on your validation data using the `predict` function. Make classifications by predicting 1 or TRUE ('high') when the probability of outcome is bigger than 0.5 and 0 of FALSE ('low') if it is less. Make a `table` showing how correct your model is.

2.3 What is the accuracy of your logistic regression model? Use ROC analysis to examine the quality of the classifier, explain and discuss.

2.4 Use K-fold cross-validation with *K=10* to estimate the accuracy you would see if you had a new dataset.

Question 3

3.1 Use another method or combination of methods to solve the same problem (namely: predict the UseLevel based on some or all of the predictors in the first 16 columns). Note: You may need to make a copy of the data frame with non-numerical columns converted to numerical values, because some of the methods we used in the course do not work for R factors or character predictors (e.g. `knn`).

3.2 What is the accuracy? What have you done to estimate how your method performs on observations that are not in the training set? Explain.

Question 4

4.1 (Rmd only) Create a variable that is 'yes' or 'no', representing whether the patient reports that they ever used heroin. Use a random forest to predict whether someone has ever used heroin. As before, use the first 16 columns as predictors, but this time include all the illicit drugs as well. Do not include the summaries of overall use (`any`, `severity` and `UseLevel`).

As above, train on the first 1400 rows. Report the accuracy when you test your model on the remaining rows. Explain and discuss.

4.2 Based on your own interests and your EDA, find another classification or regression question that you can ask with this dataset. E.g. you could try to predict reported use of other drugs, or you could also create new outcomes such as whether someone uses both crack *and* cocaine, either crack *or* cocaine, etc. Another example is to predict alcohol consumption based on the personality traits. There are many interesting options. Be creative! Note that predicting `UseLevel` from the consumption of all illicit drugs is not interesting, because `UseLevel` was constructed directly from them (see page 1).

Build the appropriate machine learning model(s) to perform this task.

Explain why you chose a specific method. How does it perform? Show the results with a table or plot and estimate the accuracy.

Question 5

Based on your answers to Questions 1-3 and further exploration, can you say which predictors seem important in predicting substance use? Why or why not? Discuss.

Mastery Question

(Rmd only) Choose and study one of these two papers:

[1] Pan, Sinno Jialin and Qiang Yang. "A Survey on Transfer Learning." *IEEE Transactions on Knowledge and Data Engineering* 22 (2010): 1345-1359.

[2] Acion L, Kelmansky D, van der Laan M, Sahker E, Jones D, et al. (2017) Use of a machine learning framework to predict substance use disorder treatment success. PLOS ONE 12(4): e0175383.

The PDFs are available from these links:

https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0175383&type=printable

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5288526

Note: You can download the PDFs from within the Imperial's network.

Discuss your chosen paper and state what you think are the 2-3 strengths and weaknesses of the machine learning methods presented ('transfer learning' in [1], or 'super learning' in [2]). Suggest a future research direction.

Submission instructions

You will hand in three documents, wrapped into **a single .zip file**:

1) Your code in .Rmd format.
2) The html that you create by knitting your .Rmd file.
3) A poster in pdf format.

You are also required to comply with these specific requirements:

- Name your files as 'SurnameCID.zip', e.g. Smith1234567.zip. **Do not submit multiple files**. This will slow down the marking and reduce our ability to give you detailed feedback on your work.
- Your .Rmd file must produce all plots that appear in your poster.
- Your html must make it clear where the answers to each question are. Use clear headings 'Question 1(a)', etc.
- Your poster should include the phrase "*The contents of this work and the associated code are my own unless otherwise stated*".

Note on Rmd files: Your .Rmd file should not be an unstructured long file that has everything you ever tried in it. The .Rmd is not a diary or lab notebook. Be succinct, concise and comment your code in enough detail to make it clear what each code block is doing. As we discussed in class, it is important you demonstrate you understand what the code is doing. A good strategy is first doing the project in your own R scripts, and then preparing an .Rmd file that summarizes your final answers.

R markdown

You may use any R libraries you want. But do list them at the **top** of your .Rmd file in the setup section by using `require(library_name)`. This way, markers can ensure that they have installed the necessary libraries when running your code.

- **Do not** refer to files that need to be downloaded or read from your hard drive, other than 'druguse.Rdata'. Do not forget the basic rule: make sure that your files can be opened and run on a computer different from yours. Failure to do this will significantly delay the marking and you may lose marks.
- You may define your own functions.
- Use text and comments to make it clear what your code is doing. Failure to explain your analyses or code may cause you to lose marks.
- Where a question has been labelled "**Rmd only**", you **do not** need to include the result in your poster.

Poster

Your poster should tell the story of your data science project. Imagine showing it to a potential employer or tweeting about it. Your reader should be able to understand the data you had, the question you asked, how you answered it, and what you found from the analysis.

Avoid using lots of text paragraphs - the poster should look good and be accessible to read. Of course, you will need to use *some* text to explain what you did and why, but try to keep the text concise and clear. Bullet points are typically more effective than long text paragraphs.

You can create your poster with any software of choice, but you are encouraged to use LaTeX. There are poster templates at www.overleaf.com which you can use without even having a LaTeX installation on your computer.

Keep in mind:

- The poster should be written clearly and should communicate your project's story with correct spelling and grammar. Plots must have titles and axis labels with legible font size. Unlabelled plots may cause you to lose marks.
- Because of its format, it is impossible to include every detail in a poster. This is OK: a key to good communication is having the 1-minute, 3-minute, 10-minute versions of your story. Think about how to describe your project in three sentences: what you asked, what you did, something you found.
- The poster does not need to have your answers to the "Rmd only" questions.
- All figures in your poster must be produced in your .Rmd file.

Marking scheme

Clarity/legibility of Rmd file (20), Html file (50), Poster (30). Total: 100 marks.

Marks may be deducted if the code cannot be run in the markers' computer, e.g. due to external files being required, etc.