Final project

The main goal of the final project is to apply methods taught in the course to a real–life data analysis problem, and to communicate your results to the course participants. The work on the project will include three stages: project proposal, presentation in class and final report.

Project proposal. You need to think about application area that interests you, find a dataset in this area, and formulate research questions. The dataset should be rich enough in order to allow application of a variety of different statistical learning methods. In the proposal you should provide explicit information on the source of the data set. Please begin to think about ideas for the final project as early as possible. We will schedule project hours in order to discuss the project ideas with the instructor. The data set and the proposal should be submitted for evaluation and approval until April 5th, 2019.

Your project proposal will be a PDF document that does not exceed 3 pages and has the following structure.

(a) *Motivation.* What is the problem that you want to solve? Describe research questions you want to address, and provide motivation.

(a) *Data set.* Describe the data set, variables, etc. Formulate research questions as statistical learning problems. Indicate explicitly the source of the data set.

(b) *Methods.* Which statistical learning algorithms do you plan to apply? Why do you think that the proposed methods are appropriate?

(c) *Experiments.* Which experiments do you intend to perform? How do you plan to evaluate algorithms performance?

Presentation in class. The project will be presented in class on June 7th, 2019. You will have 20 minutes in order to present the problem, your solution, and your findings. You can prepare presentation slides using R Markdown.

Final report. The final report on the project should be submitted in the form of a PDF file. You should use R Markdown in order to generate the file. The report should not exceed 6 typed pages including figures and bibliography. It should have the following structure.

- *Abstract.* One paragraph with brief description of the problem, solution and result.

- *Introduction.* Here you need to give motivation and to explain why the problem you are considering is interesting and important. You can discuss some background on the problem and refer to the relevant literature if necessary.

- *Data.* Describe your dataset. How many instances do you have? How was the dataset collected? What is the training and testing data? Depending on available space, you can include statistical summaries of some variables. Describe data preprocessing steps is applicable.

- *Methods.* Describe your algorithms. Make sure to include relevant mathematical formulas. For each algorithm discuss how it works. Provide necessary graphs, and summarize the obtained results.

- *Results, discussion and conclusion.* Compare results obtained by different algorithms. Include performance metrics (accuracy estimates, average errors, confusion matrices, ROC curves, etc.). Give interpretation to you quantitative results. Discuss figures and tables included in the text. Your figures should contain legends, axes labels, and figures and tables should be equipped with captions. Summarize your report and discuss the key points. What are the best–performing algorithms? Why do you think they performed better than other algorithms?

- *References.* This section may include a list of publications that are referred to in the text.

The final report should be submitted in the form of PDF file on June 14th, 2019. Along with the final report you need to submit a zip file containing your final R code for the project, the used data set and slides for class presentation. The results of your project should be reproducible!