Final assignment

After the five chapters of the course there is a final assignment that will be given for the last week of the course. The idea is to go into one of the topics of the course in a bit greater detail. The available topics are the particular statistical methods that we have gone through in chapters 2-5.

You may select any one of those methods for your assignment, go into it in a some way greater detail, and add your report of it to the total package of your weekly submissions.

The requirements differ depending on your group. (Recall that the groups are based on the desired level of studies: Master's/Doctoral studies is one group and Bachelor's studies is the other group).

**Group 1 (Doctoral/Master’s level of studies)**

**Choose one:**

* Utilize one of the methods introduced during this course to analyze your own dataset.
* Choose one of the datasets utilized during the course. Combine two statistical methods introduced during the course to analyze the data (for example do both dimensionality reduction and regression analysis). Ideas for analysis can be seen below. Do not use a data-method combination already introduced during the course.

Deadline of the final assigments is on Wednesday 8th March 2017 at 23:55. After that, they will be graded by the teachers of this course.

Useful resources

Below you will find links to open source datasets and ideas for the final assignment.

Open source datasets

UCI Machine Learning repository https://archive.ics.uci.edu/ml/index.html

Kaggle https://www.kaggle.com/datasets

Avoindata https://www.avoindata.fi/en (mainly in Finnish)

Helsinki Region Infoshare http://www.hri.fi/en/

You can also use your own data! Check out the Group details.

Ideas for Analysis

Note: We cannot guarantee that you will get interesting results. Your focus should be on utilizing and explaining the method of your choice and describing the results of your analysis, whatever they may be.

Data: learning2014

- Create a binary variable -> Logistic Regression

- Create a categorical variable -> LDA

- PCA (Advanced: combine with regression)

- K-means (Advanced: combine with LDA)

Data: student alcohol consumption

- MCA (on the categorical variables, perhaps categorize more variables?)

- Linear regression (for the grades variable, for example)

- LDA (on categorized grades for example)

Data: Boston

- Linear regression (for the crime variable, for example)

- Logistic regression (for the crime variable, for example)

- PCA (Advanced: combine with regression)

Data: human

- Linear regression on one of the "Health and knowledge" variables

- Logistic regression on one of the "Health and knowledge" variables

- LDA analysis on one of the "Health and knowledge" variables

- K-means

Also you can use these datasets from the Factominer-package to do Multiple Correspondence Analysis:

Hobbies

Poison

Children

To get started with your final assignment, check out the **General instructions**page to see the assignment for your group and instructions on how to get started as well as general guidelines for the assignment. Also see the **Useful resources**page for ideas for analysis and links to data sources.

To submit your final assignment, include all analysis and results in the index.Rmd file using RMarkdown syntax and publish the results to the GitHub page of your IODS-final repository. Then enter the final assignment submission page (click on the link 'Final assignment' above) and provide a link to your web page.

**The index file and your Github webpage should include:**

* Your full name, date and email address at the beginning of the page. Use the yaml header of the RMarkdown document to set these.
* Brief description of the "research question" you are exploring, possibly including your hypothesis **(max 2 points)**
* A link to your data wrangling script. See the general instructions. **(max 5/10 points)**
* Description of your data and its variables. Where is the data from, what does it relate to, what do the variables represent, what has been done to the data before analysis? **(max 2 points)**
* Visually clear and interesting explorations of the variables of interest in the data, from the point of view of your research question. Include interpretations of the distributions and relationships of the variables. Use captions to draw the reader’s focus on the interesting parts of your tables and graphics. **(max 8 points)**
* Brief description of the method you are using in your own words **(max 3 points)**
* Presentation of the results of your analysis including visualizations and summaries and a thorough interpretation of the results including a validation analysis of the method. **(max 16 points)**
* Conclusions and discussion **(max 2 points)**
* An ‘abstract’ at the beginning of the page with a summary of your analysis **(max 2 points)**

**The total maximum of Final Assignment is 40 points.**

Optional / advanced:

* Look at the RMarkdown html document options and finalize the appearance of your page: <http://rmarkdown.rstudio.com/html_document_format.html>.
* Rewrite your repository' s readme file now that you have a better understanding of your project.

**Enter the final assignment submission page (click on the link 'Final assignment' above) and provide a link to your web page.**