National Research University Higher School of Economics – St.Petersburg
Minor "Data Science"

Programming with Data and Reproducible Research
(or this year, due to a technical error "Анализ данных и технологии работы с данными")

Autumn 2016

Course Instructors: Alena Suvorova PhD, Ilya Musabirov MSc MA, Alexander Sirotkin PhD
Contacts: ds2016@piterdata.ninja

Course Summary

This course is a part of Data Science minor — 20 credit two-year 4 course sequence, dedicated to introduce fundamentals of Data Science to undergraduate non-STEM students. While each of the courses is focused on a specific skill — i.e. Programming, Data Analysis and Technologies, Data Mining, and Practice and Applications, all of them use iterative approach to introduce different aspects of Data Science.

Programming with Data is the first course in sequence, and its goal is to show students tool box of data analysis and reproducible research, based on R environment of statistical computing. We will cover basic of statistical computing in R, exploratory data analysis, data manipulation and graphics. By the end of the course students will be able to produce a dynamical data report employing all techniques they learned during the term.

How the course works

The course is structured around lectures and seminars/labs, with a large impact on independent work and class discussion. Seminars are the instructor-guided introductions to the tools or concepts with active discussions and independent programming tasks. Labs are provided in the form of partially completed code that can be used by students with different background: students with large programming experience can explore methods deeper, while students with no such experience just use code to solve their tasks. Each student is supposed to complete homework reading and programming assignment (homework project), complete online formative tasks, and to pass midterms and final exam.

Students are provided with web-based access to the programming environment that gives a transparent way to combining in-class and home-based activities. In addition, online exercises provide formative assessment, helping students to check their understanding of the material.
The course is accompanied with a web forum, so rising important, interesting or popular questions as well as helping others with their issues is essential part of the course.

Goals of the Course

- to introduce the idea and the tools of reproducible research;
- to show the data science applications in the wide range of research areas;
- to introduce the basics of statistical computing in R;
- to examine basic methods of exploratory data analysis;
- to explore the concept of supervised and unsupervised statistical learning;
- to explore classification vs regression problems;
- to introduce tree-based methods for regression and classification.

Target audience

The course is targeted at second-year undergraduate students. As part of the minor, the course is developed to be independent of student's major, it includes the basics of data analysis similar to the most research areas with examples from many of them (e.g. sociology, politics, economics, management, law). In this sense, the course may be interesting to students from various non-STEM faculties of the HSE, who want to learn about fundamentals of Data Science and explore appropriate tools for data analysis and reproducible research.

Requirements

Apart from the working knowledge of English language (including the ability to write and present in English) and the school-level mathematics and logics or better, the course does not have specific requirements, although basic knowledge of statistics is a plus. The course is structured around seminars and lab, with a large impact on independent work and class discussion.

Grading

- Online Homework Tests (Pass/Fail) 20%
- Midterm 1 (Practical part: Appendix 1) 15%
- Midterm 2 (Appendix 2) 15%
- Homework Project 20%
- Final Exam 30%

Provided that student is satisfied with his cumulative grade before the final, he/she may choose not to attend the final exam. In this case his/her cumulative grade becomes the final.
Active forum forum participation gives a bonus, added to a final grade, up to 20%

The resulting grade amounts to 5 ECTS credits.

Course overview

Week 1. INTRODUCTION. What is Data Science?

PART I. REPRODUCIBLE RESEARCH AND DATA SCIENCE
Week 2. Reproducible Research and Data Science
Week 3. Introduction to the tools: R, RStudio, RMarkdown. Reports and Citation

PART II. EXPLORATORY DATA ANALYSIS
Week 4. Making plots: grammar of graphics and ggplot2 package
Week 5. Data aggregation and manipulation: dplyr and lubridate packages
Week 6. Hypothesis testing: t-test, chi-square test, permutation test
Week 7. Midterm

PART III. INTRODUCTION TO STATISTICAL LEARNING
Week 8. Information systems and Data Science
Week 9. Statistical learning: Linear Regression Models, Decision Trees
Week 10. Linear Regression Models
Week 11. Regression and Classification Decision Trees
Week 12. Unsupervised learning: Introduction to Clustering

Readings

Several topics of the course are accompanied with readings. All texts are in English, and reading should be completed before class, in order to facilitate discussion.

READING 1 is devoted to the Introduction to Data Science and is discussed during first two seminars. Art of Data Science ch. 1--3

READING 2 is about basic concepts of statistical learning (Week 9).

- Chapters 2.1, 3.1-3.4 from James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112). New York: Springer.

Homework programming assignments

The course includes programming assignments.

ASSIGNMENT 1: Making reproducible report with citations
A student chooses any research topic of interest, then looks for appropriate articles in Google Scholar, export bibtex data for 3-5 of most popular or most recent articles and writes a report. Report in RMarkdown format should include an explanation why this topic is interesting and 2-3 paragraphs about the sources citing them in different ways.

ASSIGNMENT 2: Exploring dataset

Students are provided with a real dataset of game support and its description. The task to make a preliminary report on performance of the support service. Each report should include:

- Analytical questions (as precise as possible) which can provide useful insights to the stakeholders. They should connect the logic of business processes (user support in a fame company) to the data. Student is supposed to write down what is important to check for business and why, and how he/she are planning to check it using the data in hand.
- Attempts to get answers to these questions using visualization and aggregation.
- (after Week 6 seminar) Tests for dependences in dataset (with respect to main task)

Requirements:

- The report should look like the "real-life" one: without warning messages, with clear labels and captions, text should drive reader through the text, assumptions and results should be clearly stated.
- Simple summaries (means, medians, counts, etc.) and/or graphics (barplot, histogram, boxplot, etc.) will be enough.
- Report is in RMarkdown format.

ASSIGNMENT 3: Fitting regression tree

Students are provided with dataset that represents an extract from a commercial marketing database (the dataset can be changed; in this case, the task is the same with small correction of questions). The goal is to fit a regression tree to predict the annual income of a household from 13 demographic attributes and interpret the results. The report should include homework code, plots of trees and CP, detailed interpretation of results:

- What is the relation between the annual income and the other demographic predictors according to the best model
- What impact does parameter "control = rpart.control(cp = 10^-3)" have? Try to set cp = 0.02. Compare the results.
- Make cross-validation with help of functions plotcp and printcp. Look at values of errors and cp and prune the tree if it makes sense.
- How the model changed after pruning the tree? Why?
- Write down a row with your household information. Using the optimal tree and these data, predict your household income and draw tree.

Midterm and Final Exams

The course **midterm 1** covers Weeks 1-6 (Reproducible Research and Exploratory Data Analysis) and it is in the form of a classroom quiz. Students should answer the questions using methods of data manipulation, plotting and hypothesis testing. All groups have the same set of tasks (Appendix 1) but each of them has its own dataset, so each group has different answers.

The course **midterm 2** covers Weeks 8-12 (Introduction to Statistical Learning). It has both theoretical (assessing models, regression vs classification, interpreting trees) and programming part (one variant is shown in Appendix 2)

The **final** covers the whole course content, containing the tasks, similar to those in both midterms.

Main Topics Summary

WEEK 1 is the **Introduction** to the course; it introduces the area of Data Science, its main methods and techniques, focuses on breadth of topics and applications. The lecture suspects active discussions on following topics:

- who the data scientists are and what they do
- what skills are needed and what will be covered by the course
- Data Science vs Computational Social Science

The Week 1 also includes a set of presentations made by senior students about their project. These presentations, on the one hand, show possible applications of the skills, on the other hand, give the students an opportunity to join existing projects from the very beginning of the course.

WEEK 2 continues the discussion of Data Science applications on the base of Reading 1 assignment and introduces the concept of **Reproducible Research**, exploring what reproducibility is and why we need it. Reproducible research is helpful for the authors themselves: it gives an easier opportunity to reproduce figures in the revisions of a paper, to create earlier results again in a later stage of our research, etc. Another aim of making research reproducible is speeding up the research in the field in general: it is much easier to take up someone else's work if documented code, or data, or steps of the research are also available.

One of the steps towards the reproducible research is reproducible reports and well-documented code. Markup languages allow combining code and plain text to produce clearer reports. **WEEK 3** lab introduces the tool for making such reports: RMarkdown language. An introduction to RMarkdown is preceded by the introduction to the history of modern computer languages, explanation of the difference between translation, compilation and interpretation. This week also includes the introduction to the main tool of the course: R and RStudio.

The next part of the course is instrumentally-oriented and introduces the tools for **exploratory data analysis**, including

- data visualization: data types and graphics, grammar of graphics, ggplot2 package basic components (aesthetics, data, geoms, facets, ststs, scales, coordinate system) – **WEEK 4**
- data aggregation and manipulation: filtering, arranging, selecting, summarizing, and joining data (dplyr and tydir packages), working with dates (lubridate package) and strings (stringr package) – **WEEK 5**
- finding dependences in data: hypothesis testing (t-test, chi-square test, permutation test) – **WEEK 6**

This part is followed by large programming assignment (Assignment 2) and midterm, that summarized all the skills of Week 1-6.

The third part of the course introduces the concept of **statistical learning**. Statistical learning refers to a set of tools / methods / techniques for understanding data. These methods are classified as supervised and unsupervised. Supervised statistical learning involves building a statistical model for predicting an output based on inputs. In this case we have a dataset with known both output and inputs and trying to predict an output for new inputs. Problems of this nature occur in many fields as business, medicine, sociology, economics. In case of unsupervised statistical learning, there are inputs but no known output; the aim is to find relationships or structure from such data. Statistical learning includes many methods such as the linear and logistic regression, classification and regression trees, support vector machines, neural networks. The course discusses the application of the methods in **Information Systems** (**WEEK 8**), covers the fundamentals of statistical learning (**WEEK 9**), explores two supervised methods in details during seminar / lab sessions: **Linear Regression Models** (**WEEK 10**) and **Regression and Classification Decision Trees** (**WEEK 11**); and introduces **clustering** as one of the unsupervised methods during **WEEK 12**.

This part is also followed by large programming assignment (Assignment 3) and the final exam that summarized all the skills of Week 8-12.

Course Literature

- Roger D. Peng and Elizabeth Matsui. The Art of Data Science. A Guide for Anyone Who Works with Data. https://leanpub.com/artofdatascience
- Roger D. Peng. Exploratory Data Analysis with R. https://leanpub.com/exdata
- Roger D. Peng. R Programming for Data Science. https://leanpub.com/rprogramming
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. New York: Springer. http://www-bcf.usc.edu/~gareth/ISL/
- Chambers, J. (2008). Software for data analysis: programming with R. Springer Science & Business Media.
- Conway, D., & White, J. (2012). Machine learning for hackers. O'Reilly Media, Inc.
- Downey, A. B. (2012). Think complexity: Complexity science and computational modeling. " O'Reilly Media, Inc.".
- Downey, A. B. (2014). Think stats. " O'Reilly Media, Inc.".
- Gandrud, C. (2013). Reproducible Research with R and R Studio. CRC Press.
- Golder, S. A., & Macy, M. W. (2014). Digital footprints: Opportunities and challenges for online social research. Sociology, 40(1), 129.
- Kabacoff, R. (2015). R in action: data analysis and graphics with R. Manning Publications Co.
- Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., ... & Jebara, T. (2009). Life in the network: the coming age of computational social science. Science (New York, NY), 323(5915), 721.

- Loukides, M. (2010) What is data science? -O'Reilly Radar [WWWDocument]. URL: http://radar.oreilly.com/2010/06/what- is-data-science.html (accessed 12.5.14).
- Matloff, N. (2011). The art of R programming: A tour of statistical software design. No Starch Press.
- McCallum, Q. E. (2012). Bad Data Handbook: Cleaning Up The Data So You Can Get Back To Work. " O'Reilly Media, Inc.".
- Rodriguez, M., Helbing, D., & Zagheni, E. (2014). Migration of Professionals to the US. In Social Informatics (pp. 531-543). Springer International Publishing.
- Schutt, R., & O'Neil, C. (2013). Doing data science: Straight talk from the frontline. " O'Reilly Media, Inc.".
- Strohmaier, M., & Wagner, C. (2014). Computational social science for the world wide web. Intelligent Systems, IEEE, 29(5), 84-88.
- Wickham, H. (2009). ggplot2: elegant graphics for data analysis. Springer Science & Business Media.
- Yasseri, T., & Bright, J. (2014). Can electoral popularity be predicted using socially generated big data?. it-Information Technology, 56(5), 246-253.

Appendix 1. Midterm 1 Practical Part

Your Goal

This work is aimed to understand how effectively you deal with basics of data manipulation and visualization. Follow the instruction below.

Data Description

Current data provide extended information about users' transactions (order ID, time) and shop's inventory (amount, price per one). It also contains brief data about users (user ID, occupation, gender).

```
library(dplyr)
library(ggplot2)
library(readr)

orders = read_csv('~/materials/minor/midterm/orders.csv')
inventory = read_csv('~/materials/minor/midterm/inventory.csv')
users = read_csv('~/materials/minor/midterm/users.csv')
```

You can join datasets together by *user_id* and *inventory_id*.

Try to answer following question. If you feel you cannot answer one go to the next.

Task 1

1. Show top-10 people who use this shop more frequently.
2. What are three most expensive car?
3. What trademarks of phones is the most popular?

Task 2

1. What categories of products are preferred by people of different genders?
2. Who spends the most? What are their jobs?

Task 3 (pass with distinction)

1. Use a statistical test to find some associations between variables you explored in the previous task.
2. Try to find other interesting properties in the dataset.

Appendix 2. Midterm 2

1. Assessing classification models

You have a confusion table for your decision tree, where in rows are your model predictions, and in columns are true (observed) Y. The tree predicts if the person is a smoker or non-smoker.

	smoking	**non-smoking**
smoking	130	55
non-smoking	45	120

Use it to answer following questions (write down your calculations as well!):

A) How many smokers does the model predicts correctly?

B) How many errors does your model make?

C) How many non-smokers are in your data?

2. Regression vs. classification

For each of the following tasks answer if it is classification or regression problem. Choose three tasks and tell us how to re-formulate the task to the opposite problem (i.e. turn classification into regression and the other way around):

A) You are the weatherman. You have to predict the direction of the wind in a small French town called Buvil, knowing, for example, the proximity of the city (km) to the three sea winds.

B) You are manager in the bank. You have to decide, whether 21 year old student-sociologist return the loan or not.

C) You want to have something sweet, but you are afraid to become obese. You know what hundred of donuts consists on: cholesterol, sugar, colorant, and set of fats, proteins, carbohydrates. Also you have information about influence of donuts on obesity. You have to calculate probability of emergence of obesity based on your own donuts consumption.

D) Let's imagine that you are lector on "Data Science" minor. Student didn't come to test because he was ill, but he has no document with confirmation. Having information of students' attendance, howework grades, quantity of

questions and answers on QnA forum and absences and documents with confirmation, you have to define whether student lies or not.

E) The problem of determining the car license plate camera observations

3. Interpreting trees

You have a model built to classify US states according to their demographic data. Please, answer the following questions using this model.

Data Description

Demographic information of midwest counties

Variables:

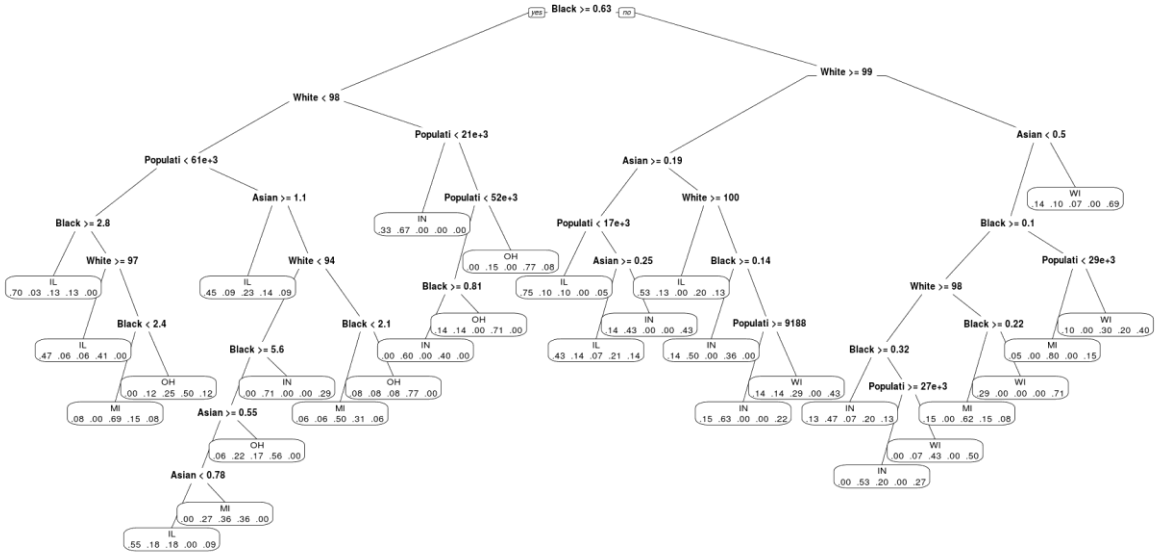
- Population - total population
- White - percentage of white population
- Black - percentage of black population
- Asian - percentage of asian population

Dependent variable: State

3.1 Interpreting the tree

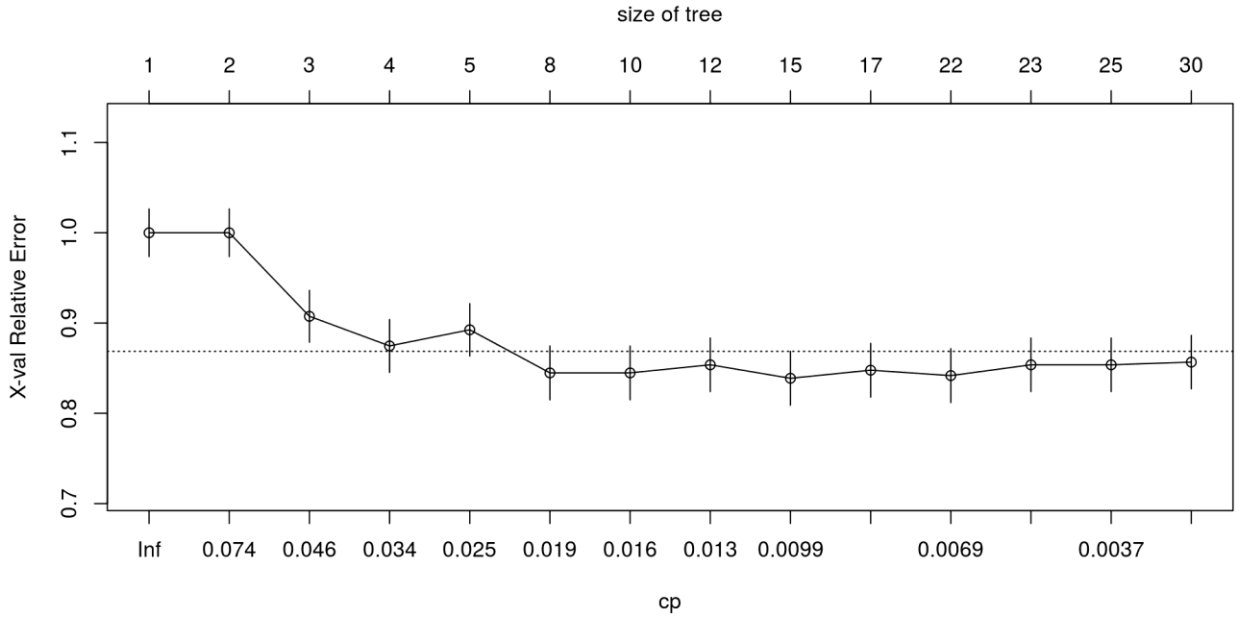
The following figure is built with prp with extra = 4.

- What does the numbers on tree nodes mean?
- Describe what observations are in the bottom right node.



3.2 Choosing the best tree

Which tree is the best according to minimum cross-validation error criterion? Provide a complexity parameter for the best tree. How many splits does it have?



CP	nsplit	rel error	xerror	xstd
0.09254	0	1	1	0.0264
0.0597	1	0.9075	1	0.0264
0.03582	2	0.8478	0.9075	0.02871
0.03284	3	0.8119	0.8746	0.02933
0.0194	4	0.7791	0.8925	0.02901
0.01791	7	0.7194	0.8448	0.02981
0.01493	9	0.6836	0.8448	0.02981
0.01095	11	0.6537	0.8537	0.02967
0.008955	14	0.6209	0.8388	0.0299
0.00796	16	0.603	0.8478	0.02977
0.00597	21	0.5612	0.8418	0.02985
0.004478	22	0.5552	0.8537	0.02967
0.002985	24	0.5463	0.8537	0.02967
1e-16	29	0.5313	0.8567	0.02963

3.3 Using tree for prediction

Once upon a time you met a nice person. He told you that he lived in a good county with a population 13500. He also told you that he rarely met an Asian since only 0.3% of the county population is Asian, but Black population is ten times bigger than Asian

and is approximately 3%. Assuming that there are no more other races rather than White in the county, could you predict in what state this man lived?
Provide you answer with a short explanation.

4. Building classification tree

Using the dataset HousetypeData that we upload for your in the next chunk show your skills in building, pruning and interpreting a classification tree.

```
HousetypeData <- read.csv("/principal/share/minor/midterm2/Housetype_Data.txt",
                          header = F,
                          col.names = c("type_of_home", "sex", "martial_status",
                                        "age", "edu","ocupation", "income",
                                        "living_time", "dual_incomes",
                                        "pers_in_house","pers_in_house_under18",
                                        "householder_status","ethnic_cl",
                                        "lang"))


HousetypeData <- within(HousetypeData, {
  sex <- factor(sex)
  martial_status <- factor(martial_status)
  edu <- factor(edu)
  ocupation <- factor(ocupation)
  living_time <- factor(living_time)
  dual_incomes <- factor(dual_incomes)
  householder_status <- factor(householder_status)
  type_of_home <- factor(type_of_home)
  ethnic_cl <- factor(ethnic_cl)
  lang <- factor(lang)
  age <- ordered(age)
  income <- ordered(income)
  pers_in_house <- ordered(pers_in_house)
  pers_in_house_under18 <- ordered(pers_in_house_under18)
})
```