Introduction to Data Science with R

for economists

Academic Level: MSc

Credit Value: 5 ECTS

Hours in class: 32 hours

Lecturer Sergey Peresvetov, peresvetov@econ.msu.ru

Aim of the course

This course is an introduction to data science applied to economics. The course covers computer programming and data analysis in R and MS Excel, econometrics (statistical analysis), financial economics, microeconomics, mathematical optimization, and probability models.

The emphasis of the course will be on making the transition from an economic model to an econometric model using real data. This involves: exploratory data analysis; specification of models to explain the data; estimation and evaluation of models; testing the economic implications of the model; forecasting from the model. The modeling process requires the use of economic theory, matrix algebra, optimization techniques, probability models, statistical analysis, and statistical software.

Plan of the course

Topics that will be covered in the class:

1. **Introduction to data science.** The data science process. Loading data into R. Exploring data. Managing data. Choosing and evaluating models. Mapping problems to machine learning tasks. Evaluating models. Validating models.
2. **Supervised methods.** Building single-variable models. Building models using many variables. Using cross-validation to estimate effects of overfitting. Using decision trees. Using nearest neighbor methods. Using Naive Bayes.
3. **Linear regression.** Building a linear regression model. Making predictions. Finding relations and extracting advice. Reading the model summary and characterizing coefficient quality. Linear regression takeaways.
4. **Logistic regression.** Building a logistic regression model. Making predictions. Finding relations and extracting advice from logistic models. Reading the model summary and characterizing coefficients. Logistic regression takeaways.
5. **Unsupervised methods.** Cluster analysis. Distances. Hierarchical clustering with hclust(). The k-means algorithm. Assigning new points to clusters. Clustering takeaways. Customer analytics with R.
6. **Advanced methods.** Using generalized additive models (GAMs). Using kernel methods. Using SVMs.

7. **Financial Methods.** Workflow of Portfolio optimization process. Work with time series data. Using finance packages: Rmetrics, PortfolioAnalytics, Quantmod. Mean-Variance Portfolios. Mean-CVaR Portfolios. Frontier Computation and Graphical Displays.
8. **Documentation and deployment**. Using knitr to produce milestone documentation. What is knitr, knitr technical details. Using knitr to document the data.

Assessment Methods / Grading

1. Homework and Computer labs 25%
2. 1 Midterm exam 25%
3. Final project 25%
4. Final Exam 25%

The homework, computer labs and project comprise the core of the course.

Recommended Reading

1. A Beginner's Guide to R by Alain Zuur, Elena Ieno and Erik Meesters, Springer-Verlag.
2. R Cookbook by Paul Teetor, O'Reilly.
3. Practical Data Science with R, Nina Zumel and John Mount, Manning Publications Co, 2014
4. Portfolio Optimization with R/Rmetrics Update 2015, Diethelm Würtz, Tobias Setz, Yohan Chalabi, William Chen, Andrew Ellis, Rmetrics Association and Finance Online Publishing, Zurich, Rmetrics eBooks 2009, NEW: Update 2015
5. Customer Analytics with R, Josep Curto, GitBook, 2017
6. R for Marketing Research and Analytics, Christopher N. Chapman, Elea McDonnell Feit, Springer International Publishing, 2015