Intro to Data Analysis 3

Gábor Békés

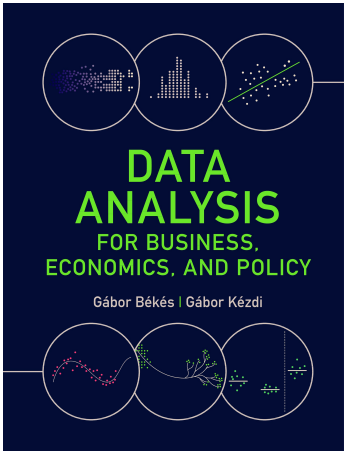
DA3 Prediction with Machine Learning

2025

Data Analysis 3

1. Framework for prediction (prediction error, loss function, RMSE, prediction with regression, overfitting, cross-validation)
2. Model building and selection (process, feature and label engineering, LASSO)
3. Regression trees (CART, stopping rules, pruning, search algorithms, regression vs CART)
4. Random forest (boosting, decorrelating trees, regression vs random forest) and GBM
5. Probability prediction and classification (threshold selection, ROC/AUC, classification with logit vs. random forest)
6. Forecasting from time series data (serial correlation, cross-validation in time series, ARIMA, vector autoregression)

Slideshow for the Békés-Kézdi Data Analysis textbook



- Cambridge University Press, 2021
- gabors-data-analysis.com
  - Download all data and code: gabors-data-analysis.com/data-and-code/
- Cover Chapter 13-18

There will be a bit of novelty

- towards the end

Course material

Material

- Lots of material, data and code to work through.
- Errata page: check, report if find typos or errors: gabors-data-analysis.com/errata/

Seminars and practice

- 12 lectures,
- 6 coding/practice seminars (Anja Hahn) on Fridays ONLINE (not compulsory)

Extras

- Q + A gabors-data-analysis.com/part3-qanda
- Beyond: gabors-data-analysis.com/beyond
- Reading gabors-data-analysis.com/readings
- Case studies github.com/gabors-data-analysis/da_case_studies
- Coding courses github.com/gabors-data-analysis/da-coding-rstats
  - Available for R and Python

Assessment

- In class Quizzes (10p)
  - Beginning of class: past lecture material, 2p each, max is 10p
- Assignments (90p)
  - 3 Assignments
  - One of 2 or 3 can be done in pairs.

- No exam.
- To pass, you will need to get at least 50% of the overall grade.