WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

Data Science and Machine Learning

Introduction and Organization

Katharina Fenz, Thomas Mitterling,
Lukas Schmoigl, Maximilian Thomasberger,
September 20, 2021

Outline

Part I: Course Outline and Organization

- Content
- Class Design
- Prerequisites

Specialization

1. Introduction and Organization (08.10)
2. Supervised Learning and Cross Validation (15.10+22.10)
3. Spatial Methods and Spatial Data Visualization (29.10)
4. Random Forest and Boosted Regression Trees (05.11)
5. tbd (12.11)
6. First Exam (19.11)
7. (Spatial) Bayesian Model Averaging (26.11)
8. Unsupervised Learning and Clustering Algorithms (03.12+10.12)
9. Text Mining and Text Classi cation (17.12)
10. tbd (14.01)
11. Project Discussion (21.01)
12. Second Exam (28.01)

Seminar

1. Data Wrangling (13.10+20.10)
2. GitHub (27.10)
3. Exploratory Data Analysis and Data Visualization (03.11+10.11)
4. Practice: Spatial Methods and Spatial Data Visualization (17.11)
5. Practice: Random Forest and Boosted Regression Trees (24.11)
6. Practice: (Spatial) Bayesian Model Averaging (01.12)
7. Practice: Unsupervised Learning and Clustering Algorithms (15.12)
8. Practice: Text Mining and Text Classi cation (22.12)
9. Presentations (12.01+19.01+26.01)

Class Design

Specialization (1827): Fri, 10:00-13:00

- Introduction into data science and machine learning methods
- Two exams (scheduled: 19.11 and 28.01)
- Competition
  - Each group (2-3 students) gets the same data set with a full set of predictors and the outcome to be predicted as well as a hold-out data set with predictors only
  - Train and validate a model with the full data set and use it to predict in the hold-out
  - Group with the lowest RMSE to the real outcomes in the hold-out wins maximum points a other points will be scaled accordingly
- Grading
  - Exams (35%+35%)
  - Competition assignment (20%)
  - Class participation (10%)

Class Design

Seminar (1915): Wed, 12:00-14:00

- Computer Exercises
  - Application of methods from lecture using R
- Paper
  - Topic and data of your choice applying any method(s) learned in class
  - Including good documentation of code and appealing illustration with data visualizations
  - Groups of 2-3 students
  - Proposal with research question, methods and data (November)
  - Presentation and discussion of the paper in class (January)
  - Final paper (February)
- Grading
  - Paper and presentation (70%)
  - Assignments (20%)
  - Class participation (10%)

Prerequisites

- Coding
  - Examples and and assignments will be given using R
  - Basic programming skills in R or a similar language are required
- Mathematics and Statistics
  - Basic understanding of probability, statistics, linear algebra and calculus is necessary
  - However this is an applied class: no formal proofs and no advanced mathematics
  - Ideally you bring a good understanding of statistical modelling in practice

Outline

Part II: Introduction to Course Content

- Data Science
  - Data Wrangling
  - Exploratory Data Analysis and Data Visualization
- Machine Learning
  - What is Machine Learning?
  - What can you do with Machine Learning?
- Spatial Methods

Data Wrangling

Source: https://r4ds.had.co.nz/

- Data Wrangling is the process of
  - Reading in the data
  - Cleaning the data
  - Joining the data
  - Aggregating the data
  - Reshaping the data
- In order to be able to apply machine learning methods and visualizations.

Exploratory Data Analysis and Data Visualization

- Exploratory data analysis is the process of asking and answering questions about your data and generating hypotheses.
  - Assessing the quality of your data
  - Plotting & Summarizing your data
  - Iterative process throughout a data-science project
  - Focus on ggplot2 and "grammar of graphics"
- Data visualization is the process of creating plots for the communication of your ideas and results.
  - Making your plots self-explanatory and beautiful
  - Best and worst practices
  - Focus on ggplot2 and "grammar of graphics"

Curve Fitting Methods...

Source: https://xkcd.com/

...and the Messages they Send

Source: https://xkcd.com/

Aim of Machine Learning

- In a classic understanding the aim of machine learning is prediction: when we use Machine Learning Methods we are usually interested in the best possible prediction of some outcome variable given a set of input variables - we are not interested in estimat causal e ects
- In that sense those methods are often a "black box"
- However Machine Learning can be used when a problem of causal nature is transform into prediction problem (e.g. Causal Random Forest)
- Most methods can be used for both regression and classi cation problems

Branches of Machine Learning

- Supervised Learning:
  - Data set already contains classified or measured outcome variables that take on predefined values
  - Parameters are optimized in a test and training data set in such a way that the predictions are as "accurate" as possible without overfitting the data in a test data set
  - e.g. Random Forests, Support Vector Machine, Linear Models with regularization term, Neural Networks, Recommender Systems, etc.
- Unsupervised Learning:
  - There is no predetermined classification or measure of some outcome variable
  - Rather the algorithm comes up with a systematization of the input data (the data "speaks to us")
  - e.g. Principal Component Analysis, Clustering Algorithms, etc.
- Reinforcement Learning:
  - The algorithm is rewarded by ending up in a "good spot" and subsequently chooses the best path through the game
  - e.g. robotic movement, gaming, etc.

What can you do with Machine Learning?

- Tackle model uncertainty and incorporate spatial spillover effects in cross-country regressions
- Disaggregate population numbers to a more granular level
- Classification in remote sensing
- Automatic text classification (e.g.: spam filters, "Robocop" by APA)
- Imputing information in incomplete data sets