Seminar course on Data visualization and analysis with R (23932)

Dr. Shailendra Gupta

www.sbi.uni-rostock.de

Course description

Data scientists are in high demand to generate knowledge out of the data. This course suits all students who want to understand data science and machine learning. We cover the basics of R programming language, data preparation, data visualization, and machine learning. Throughout the course, you learn various statistical concepts for data exploration in R using multiple datasets. You will be able to understand methods for static as well as interactive data visualization in R.

You will learn a variety of machine learning algorithms and data science topics, including significance testing, linear regression, logistic regression, and more advanced topics such as decision trees, random forests, and support vector machines for machine learning.

Course outline and learning outcomes

Week 1 (09.04.2024)

- **Introduction to exploratory data analysis tools**
  Learning outcomes: You will learn basic exploratory data analysis tools such as histogram, the Q-Q plot, scatter plots, box plots, log transforms, p values, significance testing, and other statistical methods to summarize the data.
- **Getting started with R and RStudio**
  We will provide a brief description of the setting of R environment and use the R programming language.
  Learning outcomes: you will learn to install R and R packages, customize RStudio for data science, set projects in RStudio, get data from GitHub, reading .txt and .csv files in RStudio.

Week 2 (16.04.2024)

- **Data types in R**
  This section will detail various data types like vectors, matrices, lists, arrays, factors, and data frames in the R programming language.
  Learning outcomes: you will learn various data types in R using interactive exercises. you will also learn to perform basic operations on the data.
- **R programming basics**

Week 2 will start with the programming basics and best practices to keep the code tidy. You will perform various exercises to use logical operators, conditional statements, loops, functions, math functions with R, and regular expressions on various data types.
Learning outcomes: you will get familiar with R programming syntax and learn basic operations on the data.

Week 3 (23.04.2024)

- **Data preparation and transformation**

  We will guide you to Importing Data into the R environment, handling data with missing values, using data filters for missing data, and methods for replacing missing data, including the factual analysis method, median imputation method, and deriving values method.
  L*earning outcomes:* you will learn to import and prepare data for the analysis in R. The learning will be achieved through various exercises and real experimental data.

Week 4 (30.04.2024)

- **p-values and significance testing**

  Using various examples, we will show you how statistical inferences, such as p-values, null distribution, and confidence intervals, support scientific statements.
  Learning outcomes: you will learn about random sampling from a population, calculation of significance testing, and p-values using R.

Week 5 (07.05.2024)

- **Data visualization using GGPlot**

  We will guide you to prepare data for visualization, GGPlot introduction, setting aesthetic properties, GGPlot Geoms, working with multiple charts layering and text, and creating multiple charts with facets.

  Learning outcomes: you will understand the ggplot2 package for creating publication-quality graphs from the data. You will also learn to interpret the data from the graph.

Week 6 (14.05.2024)

- **Interactive data visualization using Plotly, and Data Tables, Working with RMarkdown document**

  In this section, we will guide you to make interactive plots deployed on the web browser using Plotly and Highcharter. We will also provide a detailed description of the R interface to the JavaScript library Data Tables to display R data objects as tables on HTML pages. We will guide you to work with RMarkdown to save, execute and share R code.

  Learning outcomes: You will learn to create interactive plots ready for web deployment. With Data Tables, you will learn to visualize your data using interactive tables with features, such as filtering, pagination, and sorting.

Week 7 (28.05.2024)

- **Interactive data visualization Shiny webapp and Flexdashboards,**

  We will work in small groups to create interactive browser-based R applications using Shiny webapp and Flexdashboards.

  Learning outcomes: you will learn to create interactive reports and web-based applications using real experimental data.

Week 8 (04.06.2024)

- **Machine learning with R: Linear and logistic regression**

  Revisiting some of the very basics of machine learning, we will discuss the structure of the R package caret with a focus on how to use it. We will learn about regression and classification problems, and the performance metrics that are relevant for these. We will build simple regression and classification models using datasets available at Kaggle.

  Learning outcomes: You will be able to produce the basic structure used for supervised and unsupervised learning and know how to implement these using the caret package in R. You will be able to correctly evaluate regression and classification models in different settings. You will be able to train linear regression and logistic regression using the R package caret.

Week 9 (11.06.2024)

- **Machine learning with R: K nearest neighbors, K-means clustering**

  We will provide the introduction to K nearest neighbors (KNN) classification algorithm along with the implementation of KNN in R. We will build simple regression and classification models using datasets available at Kaggle. As an example of unsupervised learning, we will introduce the k-means clustering algorithm.

  Learning outcomes: You will be able to train k nearest neighbor models and perform k-means clustering using the R package caret. You will be able to able to use KNN in R for classification problems. You will be able to differentiate between various unsupervised algorithms and list their strengths and weaknesses. You will also learn to visualize K-means clustering results using GGPlot2.

Week 10 (18.06.2024)

- **Machine learning with R: Support Vector Machines (SVM), Feature selection**

  In this section, we extend the set of machine learning approaches by introducing Support Vector Machines. We focus on how they differ from the kinds of models introduced earlier and learn how to use them using caret. Furthermore, we will introduce the very basics of feature selection and a few algorithms with different characteristics.

Learning outcomes: You will understand the supervised learning method SVM and be able to build and test model in R. You will be able to chose and implement a feature selection method for high-dimensional problems.

Week 11 (25.6.2024)

- **Machine learning with R: Decision trees and Random Forest, Feature Representation**

  Decision trees and Random Forests are introduced as the final class of machine learning models for this lecture. We will discuss the underlying algorithms and then their implementation in R caret. Furthermore, we will discuss how to represent non-numeric data types (such as text) in order to be used in machine learning.

  Learning outcomes: You will understand the theory underlying decision tree and random forest methods and be able to use these methods in R for classification problems. You will have a good understanding of the differences between the different classes of ML models. Furthermore, you will have a basic understanding of how to use methods for the numerical representation of non-numeric features.

Week 12 (02.06.2024)

- **Assignment of project topics**

  We will assign project topics from Kaggle or from real projects from our department. You will work in a group of 3-4 students and apply various data analysis and visualization methods on the selected dataset. You will prepare an interactive report on your project.

Week 12 (09.07.2024)

- Discussion and intermediate review of the project.

Week 13 (16.07.2024)

- Presentation and evaluation of project report.

Course language

All the presentation lecture notes, slides, and datasets for this course will be available in English.