Introduction to Machine Learning in R

Program: MA & PhD Workshop, Fall 2024
Room: TBD
Time: Tuesday / Thursday, 15:00–18:00

Contact Information

Instructor: Patrick Kraft, PhD
Office: 18.2.A19
Email: patrickwilli.kraft@uc3m.es
Office Hours: Mondays, 15:00–17:00, or by appointment.

I Overview

Course Description

This four-day intensive course introduces participants to machine learning concepts and practical applications in R, focusing on supervised and unsupervised learning, along with advanced text analysis techniques. Participants will learn to implement various machine learning methods, evaluate models, and explore text analysis tools such as topic modeling and word embeddings. Through lectures and hands-on labs, students will develop the skills needed to apply machine learning to diverse datasets.

Target Audience

Social scientists, data analysts, and anyone with basic statistical and R programming knowledge who are interested in machine learning and text analysis.

Prerequisites

- Basic knowledge of R (data manipulation, basic plotting)
- Understanding of linear regression and fundamental statistical concepts

II Schedule

Day 1: Introduction to Machine Learning in R (03/12/2024)

Lecture (1 hour):

- Overview of Machine Learning: Key concepts and applications
- Introduction to R packages for machine learning (`caret`, `tidymodels`, etc.)
- Data preprocessing and feature engineering
- Cross validation and model evaluation

Lab (2 hours):

- Setting up your R environment
- Loading and exploring datasets
- Data cleaning and feature engineering in R

Day 2: Supervised and Unsupervised Learning (05/12/2024)

Lecture (1 hour):

- Supervised Learning: Regression Trees and k-Nearest Neighbors (kNN)
- Regularization and overfitting: Pruning and kNN weighting
- Unsupervised Learning: Clustering techniques (k-Means and Hierarchical Clustering)
- Dimensionality Reduction: PCA

Lab (2 hours):

- Implementing regression trees and kNN in R
- Evaluating supervised models using cross-validation and performance metrics
- Clustering and dimensionality reduction using PCA

Day 3: Text Analysis I – Fundamentals and Topic Modeling (10/12/2024)

Lecture (1 hour):

- Introduction to text data in R
- Preprocessing text: Tokenization, stop-word removal, and stemming
- Topic Modeling: Latent Dirichlet Allocation (LDA)

Lab (2 hours):

- Text preprocessing using the `tidytext`, `tm`, and `stm` packages
- Implementing LDA for topic modeling
- Visualizing and interpreting topics from a text corpus

Day 4: Text Analysis II – Advanced Methods and Model Evaluation (12/12/2024)

Lecture (1 hour):

- Word embeddings: Word2Vec and GloVe
- Sentiment analysis and text classification
- Model evaluation for text data: Precision, Recall, AUC-ROC
- Ethical considerations: Bias, fairness, and ethical AI

Lab (2 hours):

- Implementing word embeddings in R using `text2vec`
- Building a sentiment analysis pipeline
- Evaluating model performance on text classification tasks

III Course Outcomes

By the end of this course, participants will:

1. Understand key machine learning concepts, including supervised and unsupervised learning.
2. Be proficient in using R for data preprocessing, model building, and evaluation.
3. Apply machine learning techniques, such as regression trees, kNN, and dimensionality reduction.
4. Implement text analysis methods, including topic modeling and word embeddings, for real-world data.
5. Understand ethical considerations in machine learning and how to address bias.

IV Resources

Here's a list of online books and free resources that complement the course content, organized by topic:

General Machine Learning

1. **An Introduction to Statistical Learning by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani**
   A beginner-friendly resource covering foundational machine learning methods with R code examples.
   https://www.statlearning.com/
2. **Hands-On Machine Learning with R by Brad Boehmke and Brandon Greenwell**
   A practical guide to implementing machine learning techniques in R.
   https://bradleyboehmke.github.io/HOML/

Supervised and Unsupervised Learning

1. **R for Data Science by Hadley Wickham and Garrett Grolemund**
   Covers data manipulation, visualization, and modeling in R, providing a strong foundation for working with machine learning models.
   https://r4ds.had.co.nz/
2. **Tidy Modeling with R by Max Kuhn and Julia Silge**
   A practical guide to using 'tidymodels,' a collection of software for model building in R.
   https://www.tmwr.org/
3. **Applied Machine Learning Using mlr3 in R by Bernd Bischl, Raphael Sonabend, Lars Kotthoff, and Michel Lang**
   Covers supervised and unsupervised learning, as well as model tuning and evaluation using R.
   https://mlr3book.mlr-org.com/

Text Analysis and NLP

1. **Text Mining with R by Julia Silge and David Robinson**
   An excellent resource for text analysis using the `tidytext` package, covering preprocessing, topic modeling, and sentiment analysis.
   https://www.tidytextmining.com/
2. **Supervised Machine Learning for Text Analysis in R by Emil Hvitfeldt and Julia Silge**
   Building on *Text Mining with R*, this book shows you how to learn and make predictions from text data with supervised models using tidy principles.
   https://smltar.com/
3. **Quanteda Tutorials by Kohei Watanabe and Stefan Müller**
   A step-by-step introduction to quantitative text analysis using the `quanteda` package in R.
   https://tutorials.quanteda.io/

Deep Learning (Optional/Advanced)

1. **TensorFlow for R Tutorial**
   A great resource for those interested in exploring neural networks in R using `keras` and `tensorflow`.
   https://tensorflow.rstudio.com/tutorials/
2. **Fast.ai's Free Course: Practical Deep Learning for Coders**
   While focused on Python, it provides an accessible introduction to deep learning concepts, which could be useful for participants looking to explore this area further.
   https://course.fast.ai/

Supplementary Resources

1. **DataCamp's Free Tutorials**
   Covers various machine learning and R topics with interactive coding exercises.
   https://www.datacamp.com/community/tutorials
2. **The Comprehensive R Archive Network (CRAN) Task Views**
   A curated collection of R packages and resources for different machine learning tasks.
   https://cran.r-project.org/web/views/MachineLearning.html