

Practical Data Science with R

Instructor

Matthew Renze

Twitter: @matthewrenze

Email: matthew@matthewrenze.com

Web: <http://www.matthewrenze.com>

Course Description

R is a very popular open-source programming language for machine learning. Its interactive programming environment and powerful data analysis capabilities make R an ideal tool for machine learning.

This workshop will provide an introduction to the R programming language using RStudio. In addition, we will demonstrate how we can use R to train a series of machine learning models. We'll cover supervised and unsupervised learning in the form of classification, regression, and clustering. Finally, we'll learn how to deploy these models to production to make predictions given new data.

Prerequisites

Please bring your own Windows laptop and complete Lab 0 to install all of the necessary software before the workshop begins.

Module Descriptions

1. **Introduction** – introduce machine learning and the R programming language
2. **Classification** – learn how to predict categorical outcomes
3. **Regression** – learn how predict numeric outcomes
4. **Clustering** – learn how to predict groups of data based on similarity
5. **R in Practice** – learn how to optimize and deploy models into production

Learning Objectives

When students are finished with this workshop, they should understand the following:

Introduction

- What machine learning is, why it is important, and how the machine learning process works
- What R is and why it has become so popular for machine learning
- How to create data types, data structures, subset data tables, and find help on R topics

Classification

- What classification is, how it works, and applications for classification
- What are decision trees and neural networks
- How to train, test, and make predictions with a decision tree classifier and a neural network

Regression

- What regression is, how it works, and applications for regression
- What are linear regression, non-linear regression, and multiple regression
- How to train, test, and make predictions with linear regression and a neural network

Clustering

- What clustering is, how it works, and applications for clustering
- What are k-Means and hierarchical clustering
- How to group similar data points using k-means and hierarchical clustering algorithms

ML in Practice

- What are overfitting, underfitting, and regularization
- How to evaluate machine learning model performance
- How to tune hyperparameters to optimize machine learning models
- What are the options for deploying machine learning into production
- How to deploy machine learning models into production with R

Course Outline

Introduction

Lecture

- What is machine learning?
- Why is machine learning important?
- How does the machine learning process work?
- What is R?
- Why is R so popular for machine learning?
- R language basics

Lab

- Installation and setup
- Hello World
- Working with data types
- Working with data structures
- Working with data frames
- Miscellaneous topics

Classification

Lecture

- Classification
- Decision tree classifier
- Neural network classifier
- Training and test data sets

Lab

- Creating a training and test data set
- Predicting categories with a decision tree
- Predicting categories with a neural network

Regression

Lecture

- Regression
- Simple linear regression
- Non-linear regression
- Multiple variable regression

Lab

- Predicting values with simple linear regression
- Predicting values with a neural network

Clustering

Lecture

- Clustering
- k-Means clustering
- Hierarchical clustering

Lab

- Grouping data with k-means clustering
- Grouping data with hierarchical clustering

Machine Learning in Practice

Lecture

- Overfitting, underfitting, and regularization
- Evaluating model performance
- Tuning hyperparameters to optimize model
- Deploying ML in production with R