IAA Predictive Modeling with R and Python

Don Hale

2025-10-22

Table of contents

About **3**

1 Introduction **4**

2 Introduction to Machine Learning **5**
2.1 What is Machine Learning? 5
2.2 Major Types of Machine Learning 5
2.2.1 1. Supervised Learning 6
2.2.2 2. Unsupervised Learning 6
2.2.3 3. Semi-Supervised Learning 6
2.2.4 4. Reinforcement Learning 6
2.3 Objectives of Machine Learning 6
2.4 Bias-Variance Tradeoff 7
2.5 The Machine Learning Process 7
2.6 Summary 8

3 resampling-model-selection **9**

4 generalized-additive-models **10**

5 tree-based-models **11**

6 neural-network-models **12**

7 naive-bayes-models **13**

8 model-agnostic-interpretability **14**

9 support-vector-machines **15**

About

This book is a companion to the Machine Learning Class. It contains both R and Python Code

To learn more about Quarto books visit https://quarto.org/docs/books.

1 Introduction

2 Introduction to Machine Learning

Machine Learning (ML) is a branch of artificial intelligence (AI) that focuses on building systems that **learn patterns from data** and make predictions or decisions without being explicitly programmed to perform specific tasks. Instead of writing rules by hand, a machine learning algorithm "learns" from examples in the data.

2.1 What is Machine Learning?

Formally, machine learning is the process of using data to **train a model** to make accurate predictions or decisions. The model identifies patterns in the training data and generalizes these patterns to new, unseen data.

ML can be thought of as:

- **Predictive modeling** – learning a function that maps inputs (features) to outputs (targets).
- **Automated decision-making** – systems that improve their performance over time without explicit reprogramming.

2.2 Major Types of Machine Learning

Machine learning is commonly categorized into several types based on the **nature of the task** and **availability of labeled data**.

2.2.1 1. Supervised Learning

- Uses **labeled data**, meaning that each training example has an input and a known output.
- The goal is to **learn a function** that maps inputs to outputs accurately.
- Typical tasks:
  - **Regression**: Predict a continuous variable (e.g., house prices).
  - **Classification**: Predict a categorical variable (e.g., spam vs. non-spam email).

2.2.2 2. Unsupervised Learning

- Works with **unlabeled data**, where the output is unknown.
- The goal is to **discover patterns or structure** in the data.
- Typical tasks:
  - **Clustering**: Group similar observations together (e.g., customer segmentation).
  - **Dimensionality reduction**: Reduce the number of features while retaining important information (e.g., PCA).

2.2.3 3. Semi-Supervised Learning

- Combines a small amount of labeled data with a large amount of unlabeled data.
- Useful when labeling data is expensive or time-consuming.

2.2.4 4. Reinforcement Learning

- Focuses on **learning through trial and error**.
- An agent learns to take actions in an environment to **maximize cumulative reward**.
- Examples: Robotics, game AI, recommendation systems.

2.3 Objectives of Machine Learning

The main objective of machine learning is to **build models that generalize well** from historical data to make accurate predictions or decisions on **new, unseen data**. Key considerations include:

1. **Accuracy** – How well does the model predict outcomes?

2. **Generalization** – Does the model perform well on unseen data, or is it overfitting the training data?
3. **Interpretability** – Can humans understand how the model makes predictions?

2.4 Bias-Variance Tradeoff

A fundamental concept in machine learning is the **bias-variance tradeoff**, which explains the balance between:

- **Bias**: Error due to overly simplistic models that **underfit** the data.
- **Variance**: Error due to overly complex models that **overfit** the training data.

The goal is to find a model that **minimizes the total prediction error**:

[Total Error = Bias^2 + Variance + Irreducible Error]

- **Underfitting** → High bias, low variance
- **Overfitting** → Low bias, high variance

The optimal model balances bias and variance for the best predictive performance.

2.5 The Machine Learning Process

A typical workflow in machine learning consists of the following steps:

1. **Data Collection**
   - Gather data from experiments, databases, or external sources.
2. **Data Preprocessing**
   - Handle missing values, outliers, and feature engineering.
3. **Model Selection**
   - Choose an appropriate algorithm (e.g., linear regression, random forest, neural networks).
4. **Model Training**

- Fit the model to the training data.

5. **Model Evaluation**
   - Assess performance using metrics such as RMSE, accuracy, precision, recall, or AUC.
6. **Model Tuning**
   - Adjust hyperparameters to improve performance.
7. **Model Deployment**
   - Use the trained model to make predictions on new data.
8. **Monitoring and Maintenance**
   - Continuously track model performance and retrain if necessary.

2.6 Summary

Machine learning allows us to build predictive models that **learn from data**. Understanding the different types of ML, their objectives, and the tradeoffs between bias and variance is critical to applying ML effectively. The remainder of this book explores practical algorithms, their implementation in **R and Python**, and strategies for model selection and interpretation.

3 resampling-model-selection

4 generalized-additive-models

5 tree-based-models

6 neural-network-models

7 naive-bayes-models

8 model-agnostic-interpretability

9 support-vector-machines