Thống kê và vẽ đồ thị trong R

Lời mở đầu

Tác giả

Duc Nguyen | tuhocr.com

Nội dung cuốn sách này điểm qua hầu hết các chủ đề thống kê và vẽ đồ thị thường gặp, bao gồm các trích dẫn đến tài liệu toàn văn để thuận tiện cho người đọc dễ tra cứu. a

Trích dẫn

```
Duc Nguyen (2025). "Thống kê và vẽ đồ thị trong R". TUHOCR. https://thongkevavedothi.com

@Book{Nguyen2025,
  author    = {Duc Nguyen},
  publisher = {TUHOCR},
  title     = {Thống kê và vẽ đồ thị trong {R}},
  year      = {2025},
  url       = {https://thongkevavedothi.com},
}
```

1

Phân tích chuỗi thời gian

There are many packages in the `mlr3` ecosystem that you may want to use as you work through this book. All our packages can be installed from GitHub and R-universe[1]; the majority (but not all) packages can also be installed from CRAN. We recommend adding the mlr-org R-universe to your R options so you can install all packages with `install.packages()`, without having to worry which package repository it comes from. To do this, install `r ref_pkg("usethis")` and run the following:

1.1 aaa

bbb

[1] R-universe is an alternative package repository to CRAN. The bit of code below tells R to look at both R-universe and CRAN when trying to install packages. R will always install the latest version of a package.

Part I

Fundamentals

2

Data and Basic Modeling

Natalie Foss
University of Wyoming

Lars Kotthoff
University of Wyoming

In this chapter, we will introduce the mlr3 objects and corresponding R6 classes that implement the essential building blocks of machine learning. These building blocks include the data (and the methods for creating training and test sets), the machine learning algorithm (and its training and prediction process), the configuration of a machine learning algorithm through its hyperparameters, and evaluation measures to assess the quality of predictions.

In the simplest definition, machine learning (ML) is the process of learning models of relationships from data. Supervised learning is a subfield of ML in which datasets consist of labeled observations, which means that each data point consists of features, which are variables to make predictions from, and a target, which is the quantity that we are trying to predict. For example, predicting a car's miles per gallon (target) based on the car's properties (features) such as horsepower and the number of gears is a supervised learning problem, which we will return to several times in this book. In `mlr3`, we refer to datasets, and their associated metadata as tasks (Section 2.1). The term 'task' is used to refer to the prediction problem that we are trying to solve. Tasks are defined by the features used for prediction and the targets to predict, so there can be multiple tasks associated with any given dataset. For example, predicting miles per gallon (mpg) from horsepower is one task, predicting horsepower from mpg is another task, and predicting the number of gears from the car's model is yet another task.

Machine Learning/Supervised Learning

Supervised learning can be further divided into regression – which is the prediction of numeric target values, e.g. predicting a car's mpg – and classification – which is the prediction of categorical values/labels, e.g., predicting a car's model. **?@sec-special** also discusses other tasks, including cost-sensitive classification and unsupervised learning. For any supervised learning task, the goal is to build a model that captures the relationship between the features and target, often with the goal of training the model to learn relationships about the data so it can make predictions for new and previously unseen data. A model is formally a mapping from a feature vector to a prediction. A prediction can take many forms depending on the task; for example, in classification this can be a predicted label, or a set of predicted probabilities or scores. Models are induced by passing training data to machine learning algorithms, such as decision trees, support vector machines, neural networks, and many more. Machine learning algorithms are called learners in `mlr3` (**?@sec-learners**) as, given data, they learn models. Each learner has a parameterized space that potential models are drawn from and during the training process, these parameters are fitted to best match the data. For example, the parameters could be the coefficients used for individual features when training a linear regression model. During training, most machine learning algorithms are 'fitted'/'trained' by optimizing a loss-function that quantifies the mismatch between ground truth target values in the training data and the predictions of the model.

Regression
Classification

Model

Learners

For a model to be most useful, it should generalize beyond the training data to make 'good' predictions (**?@sec-predicting**) on new and previously 'unseen' (by the model) data. The simplest way to test this, is to split data into training data and test data – where the model is trained on the training data and then the separate test data is used to evaluate models in an unbiased way by assessing to what extent the model has learned the true relationships that underlie the data (**?@sec-performance**). This evaluation procedure estimates a model's generalization error, i.e., how well we expect the model to perform in general. There are many ways to evaluate models and to split data for estimating generalization error (**?@sec-resampling**).

Train/Test Data

Generalization Error

This brief overview of ML provides the basic knowledge required to use `mlr3` and is summarized in Figure 2.1. In the rest of this book, we will provide introductions to methodology when relevant. For texts about ML, including detailed methodology and underpinnings of different algorithms, we recommend Hastie, Friedman, and Tibshirani (2001), James et al. (2014), and Bishop (2006).

In the next few sections we will look at the building blocks of `mlr3` using regression as an example, we will then consider how to extend this to classification in **?@sec-classif**.

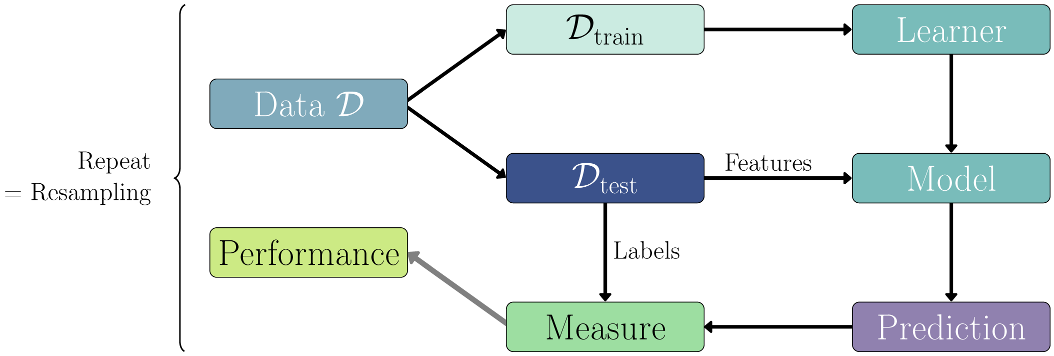

Figure 2.1: General overview of the machine learning process.

2.1 Tasks

Tasks are objects that contain the (usually tabular) data and additional metadata that define a machine learning problem. The metadata contain, for example, the name of the target feature for supervised machine learning problems. This information is extracted automatically when required, so the user does not have to specify the prediction target every time a model is trained.

3

References

Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning*. Springer.

Hastie, Trevor, Jerome Friedman, and Robert Tibshirani. 2001. *The Elements of Statistical Learning*. Springer New York. https://doi.org/10.1007/978-0-387-21606-5.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2014. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated. https://doi.org/10.1007/978-1-4614-7138-7.

Index