

An Introduction to Machine Learning

Sudhakaran Prabakaran, Matthew Wayland and Christopher Penfold

2017-08-23

Contents

1	About the course	5
1.1	Overview	5
1.2	Registration	5
1.3	Prerequisites	5
1.4	Github	6
1.5	License	6
1.6	Contact	6
1.7	Colophon	6
2	Introduction	7
3	Linear models and matrix algebra	9
4	Linear and non linear logistic regression	11
5	Nearest neighbours	13
5.1	Example one	13
5.2	Example two	13
6	Decision trees and random forests	15
7	Support vector machines	17
8	Artificial neural networks	19
9	Dimensionality reduction	21
9.1	Linear Dimensionality Reduction	21
9.2	Nonlinear Dimensionality Reduction	21
10	Clustering	23
10.1	Types of cluster	23
10.2	Distance metrics	23
10.3	Clustering algorithms	23
10.4	Visualization	23
10.5	Examples	24
10.6	Summary	24
10.7	Exercises	24
10.8	Extended exercises	24
A	Resources	25
A.1	Python	25
A.2	Machine learning data set repository	25

B Solutions to exercises	27
B.1 Chapter 2 - Linear models and matrix algebra	27
B.2 Chapter 3 - Linear and non-linear logistic regression	27
B.3 Chapter 4 - Nearest neighbours	27
B.4 Chapter 5 - Decision trees and random forests	27
B.5 Chapter 6 - Support vector machines	27
B.6 Chapter 7 - Artificial neural networks	27
B.7 Chapter 8 - Dimensionality reduction	27
B.8 Chapter 9 - Clustering	27

Chapter 1

About the course

1.1 Overview

Machine learning gives computers the ability to learn without being explicitly programmed. It encompasses a broad range of approaches to data analysis with applicability across the biological sciences. Lectures will introduce commonly used algorithms and provide insight into their theoretical underpinnings. In the practicals students will apply these algorithms to real biological data-sets using the R language and environment.

During this course you will learn about:

- Some of the core mathematical concepts underpinning machine learning algorithms: matrices and linear algebra; Bayes' theorem.
- Classification (supervised learning): partitioning data into training and test sets; feature selection; logistic regression; support vector machines; artificial neural networks; decision trees; nearest neighbours, cross-validation.
- Exploratory data analysis (unsupervised learning): dimensionality reduction, anomaly detection, clustering.

After this course you should be able to:

- Understand the concepts of machine learning.
- Understand the strengths and limitations of the various machine learning algorithms presented in this course.
- Select appropriate machine learning methods for your data.
- Perform machine learning in R.

1.2 Registration

Bioinformatics Training: An Introduction to Machine Learning

1.3 Prerequisites

- Some familiarity with R would be helpful.
- For an introduction to R see An Introduction to Solving Biological Problems with R course.

1.4 Github

[bioinformatics-training/intro-machine-learning](#)

1.5 License

GPL-3

1.6 Contact

If you have any **comments**, **questions** or **suggestions** about the material, please contact the authors: Sudhakaran Prabakaran, Matt Wayland and Chris Penfold.

1.7 Colophon

This book was produced using the **bookdown** package (Xie, 2017), which was built on top of R Markdown and **knitr** (Xie, 2015).

Chapter 2

Introduction

You can label chapter and section titles using `{#label}` after them, e.g., we can reference Chapter 2. If you do not manually label them, there will be automatic labels anyway, e.g., Chapter ??.

Figures and tables with captions will be placed in `figure` and `table` environments, respectively.

```
par(mar = c(4, 4, .1, .1))  
plot(pressure, type = 'b', pch = 19)
```

Reference a figure by its code chunk label with the `fig:` prefix, e.g., see Figure 2.1. Similarly, you can reference tables generated from `knitr::kable()`, e.g., see Table 2.1.

```
knitr::kable(  
  head(iris, 20), caption = 'Here is a nice table!',  
  booktabs = TRUE  
)
```



Figure 2.1: Here is a nice figure!

Table 2.1: Here is a nice table!

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa
5.4	3.7	1.5	0.2	setosa
4.8	3.4	1.6	0.2	setosa
4.8	3.0	1.4	0.1	setosa
4.3	3.0	1.1	0.1	setosa
5.8	4.0	1.2	0.2	setosa
5.7	4.4	1.5	0.4	setosa
5.4	3.9	1.3	0.4	setosa
5.1	3.5	1.4	0.3	setosa
5.7	3.8	1.7	0.3	setosa
5.1	3.8	1.5	0.3	setosa

Chapter 3

Linear models and matrix algebra

Chapter 4

Linear and non linear logistic regression

Chapter 5

Nearest neighbours

5.1 Example one

5.2 Example two

Chapter 6

Decision trees and random forests

Chapter 7

Support vector machines

Chapter 8

Artificial neural networks

Chapter 9

Dimensionality reduction

9.1 Linear Dimensionality Reduction

9.1.1 Principle Component Analysis

9.1.2 Horeshoe effect

9.2 Nonlinear Dimensionality Reduction

9.2.1 t-SNE

9.2.2 Gaussian Process Latent Variable Models

9.2.3 GPLVMs with informative priors

Chapter 10

Clustering

10.1 Types of cluster

10.2 Distance metrics

Minkowski distance:

$$distance(x, y, p) = \left(\sum_{i=1}^n abs(x_i - y_i)^p \right)^{1/p} \quad (10.1)$$

10.3 Clustering algorithms

10.3.1 K-means

Pseudocode

10.3.2 DBSCAN

Density-based spatial clustering of applications with noise

10.4 Visualization

Dendrogram

Heatmap

10.5 Examples

10.5.1 Image segmentation

10.5.2 Quality control

10.6 Summary

10.6.1 Applications

10.6.2 Strengths

10.6.3 Limitations

10.7 Exercises

Exercise solutions: B.8

10.8 Extended exercises

Appendix A

Resources

A.1 Python

[scikit-learn](#)

A.2 Machine learning data set repository

[mldata.org](#)

This repository manages the following types of objects:

- Data Sets - Raw data as a collection of similarly structured objects.
- Material and Methods - Descriptions of the computational pipeline.
- Learning Tasks - Learning tasks defined on raw data.
- Challenges - Collections of tasks which have a particular theme.

Appendix B

Solutions to exercises

- B.1 Chapter 2 - Linear models and matrix algebra
- B.2 Chapter 3 - Linear and non-linear logistic regression
- B.3 Chapter 4 - Nearest neighbours
- B.4 Chapter 5 - Decision trees and random forests
- B.5 Chapter 6 - Support vector machines
- B.6 Chapter 7 - Artificial neural networks
- B.7 Chapter 8 - Dimensionality reduction
- B.8 Chapter 9 - Clustering

Bibliography

Xie, Y. (2015). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.

Xie, Y. (2017). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.4.