Machine Learning and Statistics: A Common Ground

G. Alexi Rodríguez-Arelis

2023-06-06

This book aims to set a common ground between Machine Learning and Statistics regarding regression techniques, using `Python` and `R`, under two perspectives: inference and prediction.

Table of contents

Preface

Throughout my journey as a postdoctoral fellow in the Master of Data Science (MDS) at the University of British Columbia, I became aware of the fascinating overlap between Machine Learning and Statistics. Many Data Science students usually come across common Machine Learning/Statistics concepts or ideas that might only differ in names. For instance, simple terms such as weights in supervised learning (and their equivalent statistical counterpart as regression coefficients) might be misleading for students starting their Data Science formation. On the other hand, from an instructor's perspective in a Data Science program that subsets its courses in Machine Learning in `Python` and Statistics in `R`, regression courses in `R` also demand the inclusion of `Python`-related packages as alternative tools. In my MDS teaching experience, this is especially critical for students whose career plans lean towards industry where `Python` is more heavily used.

As a Data Science educator, I view this field as a substantial synergy between Machine Learning and Statistics. Nevertheless, I believe there are still many gaps to be addressed between both disciplines. Thus, closing these critical gaps is imperative in a domain with accelerated growth, such as Data Science. The MDS Stat-ML dictionary inspired me to write this book. It basically consists of common ground between foundational supervised learning models from Machine Learning and regression models commonly used in Statistics. I strive to explore common modelling approaches as a primary step while highlighting different terminology found in both fields. Furthermore, this discussion is not limited to a simple conceptual exploration. Hence, the second step is hands-on practice via the corresponding `Python` packages for Machine Learning and `R` for Statistics.

Audience

How this Book is Structured

1 Introduction

This is a book created from markdown and executable code.

See Knuth (1984) for additional discussion of literate programming.

```
1 + 1
```

```
[1] 2
```

```
lemurs <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/m
```

```
library(dplyr)
library(knitr)
lemur_data <- lemurs %>%
  filter(taxon == "ECOL",
         sex == "M",
         age_category == "adult") %>%
  select(c(age_at_wt_mo, weight_g)) %>%
  rename(Age = age_at_wt_mo,
         Weight = weight_g)
kable(head(lemur_data))
```

Age	Weight
129.90	2805
132.10	3001
140.32	2429
157.94	2597
164.58	2497
184.18	2225

```
lemur_data_py = r.lemur_data
lemur_data_py
```

```
         Age  Weight
0     129.90  2805.0
1     132.10  3001.0
2     140.32  2429.0
3     157.94  2597.0
4     164.58  2497.0
...      ...     ...
1302   59.77  2280.0
1303   61.08  2420.0
1304   61.15  2460.0
1305   61.25  2440.0
1306   61.68  2120.0

[1307 rows x 2 columns]
```

```python
import statsmodels.api as sm
y = lemur_data_py[["Weight"]]
x = lemur_data_py[["Age"]]
x = sm.add_constant(x)
mod = sm.OLS(y, x).fit()
lemur_data_py["Predicted"] = mod.predict(x)
lemur_data_py["Residuals"] = mod.resid
```

1.1 R

```r
fizz_buzz <- function(fbnums = 1:50) {
  output <- dplyr::case_when(
    fbnums %% 15 == 0 ~ "FizzBuzz",
    fbnums %% 3 == 0 ~ "Fizz",
    fbnums %% 5 == 0 ~ "Buzz",
    TRUE ~ as.character(fbnums)
  )
  print(output)
}

1 + 2
```

1.2 Python

```python
def fizz_buzz(num):
  if num % 15 == 0:
    print("FizzBuzz")
  elif num % 5 == 0:
    print("Buzz")
  elif num % 3 == 0:
    print("Fizz")
  else:
    print(num)
```

2 Summary

In summary, this book has no content whatsoever.

```
1 + 1
```

```
[1] 2
```

References

Knuth, Donald E. 1984. "Literate Programming." *Comput. J.* 27 (2): 97–111. https://doi.org/10.1093/comjnl/27.2.97.