

797ML Handbook

Steve Linberg

2022-04-03

Contents

1	About	9
2	Simple Linear Regression	11
2.1	TL;DR	11
2.2	What it does	11
2.3	When to do it	11
2.4	How to do it	12
2.5	How to interpret the output	12
2.6	Where to learn more	12
3	Multiple Linear Regression	13
3.1	TL;DR	13
3.2	What it does	13
3.3	When to do it	13
3.4	How to do it	14
3.5	How to interpret the output	14
3.6	Where to learn more	14
4	Logistic Regression	15
4.1	TL;DR	15
4.2	What it does	15
4.3	When to do it	15
4.4	How to do it	15
4.5	How to interpret the output	15
4.6	Where to learn more	15
5	Multiple Logistic Regression	17
5.1	TL;DR	17
5.2	What it does	17
5.3	When to do it	17
5.4	How to do it	17
5.5	How to interpret the output	17
5.6	Where to learn more	17

6	Linear Discriminant Analysis	19
6.1	TL;DR	19
6.2	What it does	19
6.3	When to do it	19
6.4	How to do it	19
6.5	How to interpret the output	19
6.6	Where to learn more	19
7	Quadratic Discriminant Analysis	21
7.1	TL;DR	21
7.2	What it does	21
7.3	When to do it	21
7.4	How to do it	21
7.5	How to interpret the output	21
7.6	Where to learn more	21
8	Naive Bayes	23
8.1	TL;DR	23
8.2	What it does	23
8.3	When to do it	23
8.4	How to do it	23
8.5	How to interpret the output	23
8.6	Where to learn more	23
9	K-Nearest Neighbors	25
9.1	TL;DR	25
9.2	What it does	25
9.3	When to do it	25
9.4	How to do it	25
9.5	How to interpret the output	25
9.6	Where to learn more	25
10	Poisson Regression	27
10.1	TL;DR	27
10.2	What it does	27
10.3	When to do it	27
10.4	How to do it	27
10.5	How to interpret the output	27
10.6	Where to learn more	27
11	Cross-Validation	29
11.1	TL;DR	29
11.2	What it does	29
11.3	When to do it	29
11.4	How to do it	29
11.5	How to interpret the output	29

11.6 Where to learn more	29
12 Bootstrap	31
12.1 TL;DR	31
12.2 What it does	31
12.3 When to do it	31
12.4 How to do it	31
12.5 How to interpret the output	31
12.6 Where to learn more	31
13 Best Subset Selection	33
13.1 TL;DR	33
13.2 What it does	33
13.3 When to do it	33
13.4 How to do it	33
13.5 How to interpret the output	33
13.6 Where to learn more	33
14 Stepwise Selection	35
14.1 TL;DR	35
14.2 What it does	35
14.3 When to do it	35
14.4 How to do it	35
14.5 How to interpret the output	35
14.6 Where to learn more	35
15 Ridge Regression	37
15.1 TL;DR	37
15.2 What it does	37
15.3 When to do it	37
15.4 How to do it	37
15.5 How to interpret the output	37
15.6 Where to learn more	37
16 Lasso	39
16.1 TL;DR	39
16.2 What it does	39
16.3 When to do it	39
16.4 How to do it	39
16.5 How to interpret the output	39
16.6 Where to learn more	39
17 Principal Component Regression	41
17.1 TL;DR	41
17.2 What it does	41
17.3 When to do it	41
17.4 How to do it	41

17.5	How to interpret the output	41
17.6	Where to learn more	41
18	Bagging	43
18.1	TL;DR	43
18.2	What it does	43
18.3	When to do it	43
18.4	How to do it	43
18.5	How to interpret the output	43
18.6	Where to learn more	43
19	Random Forests	45
19.1	TL;DR	45
19.2	What it does	45
19.3	When to do it	45
19.4	How to do it	45
19.5	How to interpret the output	45
19.6	Where to learn more	45
20	Boosting	47
20.1	TL;DR	47
20.2	What it does	47
20.3	When to do it	47
20.4	How to do it	47
20.5	How to interpret the output	47
20.6	Where to learn more	47
21	Bayesian Additive Regression Trees	49
21.1	TL;DR	49
21.2	What it does	49
21.3	When to do it	49
21.4	How to do it	49
21.5	How to interpret the output	49
21.6	Where to learn more	49
22	Support Vector Machines	51
22.1	TL;DR	51
22.2	What it does	51
22.3	When to do it	51
22.4	How to do it	51
22.5	How to interpret the output	51
22.6	Where to learn more	51
23	Principal Component Analysis	53
23.1	TL;DR	53
23.2	What it does	53
23.3	When to do it	53

23.4	How to do it	53
23.5	How to interpret the output	53
23.6	Where to learn more	53
24	K-Means Clustering	55
24.1	TL;DR	55
24.2	What it does	55
24.3	When to do it	55
24.4	How to do it	55
24.5	How to interpret the output	55
24.6	Where to learn more	55
25	Hierarchical Clustering	57
25.1	TL;DR	57
25.2	What it does	57
25.3	When to do it	57
25.4	How to do it	57
25.5	How to interpret the output	57
25.6	Where to learn more	57

Chapter 1

About

This book is being written as part of a final project for 797ML at UMass Amherst, spring 2022. It contains a simple reference and breakdown for a couple of dozen core methods used in machine learning.

The intent is twofold:

1. Serve as a reference for the basics of the material covered in the class, using language and examples that are as simple as possible to explain the core concepts and how to do them;
2. Force myself to learn these techniques better by carrying out the above.

Chapter 2

Simple Linear Regression

2.1 TL;DR

What it does Looks to see how well a single predictor variable predicts an outcome, like *how well do years of education predict salary?*

When to do it When you want to see if pretty much the simplest possible model provides enough of an explanation of variance for your purposes

How to do it With the `lm()` function, among other ways

How to assess it Look for a significant p -value for the predictor, and a reasonable R^2

2.2 What it does

Simple linear regression is where it all begins; among the simplest of all of the regression techniques in analysis, which attempts to estimate a slope and an intercept line for a set of observations using a single predictor variable X and an output variable Y . It uses ordinary least squares (OLS) to build its model, looking for the line through the mean of X and Y that has the smallest sum of squares between the predicted and observed values.

2.3 When to do it

It is a simple first step for looking at data to see if there is an easy single-variable model that does a reasonable job predicting outcomes using one predictor variable. Sometimes, it can be good enough! It has the advantage of being easy to execute, to understand and to communicate, and the value of these factors should not be underestimated. Communicating with non-specialists is an important aspect of a data scientist's job.

Linear regression requires a dataset with a continuous outcome variable; it is easiest and most effective if the predictor variable is also numeric, whether continuous or discrete. It is possible to do linear regression with non-numeric predictors, such as true/false or ordered responses, by converting the predictors to a numeric scale.

2.4 How to do it

2.5 How to interpret the output

2.6 Where to learn more

Chapter 3

Multiple Linear Regression

3.1 TL;DR

What it does Looks to see how well multiple predictor variables predict an outcome, like *how well do years of education and age predict salary?*

When to do it When a simple linear regression doesn't provide a good enough explanation of variance, and you want to see if adding additional variables provides a better one

How to do it With the `lm()` function, utilizing more than one predictor

How to assess it Look for significant p -values for the predictors, and a reasonable adjusted- R^2

3.2 What it does

Multiple linear regression is the first natural extension of simple linear regression. It allows for more than one predictor variable to be specified. It is also possible to combine predictors in interactions, to find out if combinations of predictors have different effects than simply adding them to the model. XXX explain/demo

3.3 When to do it

Use multiple linear regression when a simple linear regression doesn't provide a good enough explanation of the variance you're observing, and you want to see if adding more predictors provides a better fit. Typically, this would be in response to either a low R^2 that leaves a lot of unexplained variance, or even just a visual conclusion drawn from seeing a plot of a linear model with an unsatisfactory regression line.

3.4 How to do it

The `lm()` function, using more than one predictor in the formula.

3.5 How to interpret the output

3.6 Where to learn more

Chapter 4

Logistic Regression

4.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

4.2 What it does

4.3 When to do it

4.4 How to do it

4.5 How to interpret the output

4.6 Where to learn more

Chapter 5

Multiple Logistic Regression

5.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

5.2 What it does

5.3 When to do it

5.4 How to do it

5.5 How to interpret the output

5.6 Where to learn more

Chapter 6

Linear Discriminant Analysis

6.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

6.2 What it does

6.3 When to do it

6.4 How to do it

6.5 How to interpret the output

6.6 Where to learn more

Chapter 7

Quadratic Discriminant Analysis

7.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

7.2 What it does

7.3 When to do it

7.4 How to do it

7.5 How to interpret the output

7.6 Where to learn more

Chapter 8

Naive Bayes

8.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

8.2 What it does

8.3 When to do it

8.4 How to do it

8.5 How to interpret the output

8.6 Where to learn more

Chapter 9

K-Nearest Neighbors

9.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

9.2 What it does

9.3 When to do it

9.4 How to do it

9.5 How to interpret the output

9.6 Where to learn more

Chapter 10

Poisson Regression

10.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

10.2 What it does

10.3 When to do it

10.4 How to do it

10.5 How to interpret the output

10.6 Where to learn more

Chapter 11

Cross-Validation

11.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

11.2 What it does

11.3 When to do it

11.4 How to do it

11.5 How to interpret the output

11.6 Where to learn more

Chapter 12

Bootstrap

12.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

12.2 What it does

12.3 When to do it

12.4 How to do it

12.5 How to interpret the output

12.6 Where to learn more

Chapter 13

Best Subset Selection

13.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

13.2 What it does

13.3 When to do it

13.4 How to do it

13.5 How to interpret the output

13.6 Where to learn more

Chapter 14

Stepwise Selection

14.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

14.2 What it does

14.3 When to do it

14.4 How to do it

14.5 How to interpret the output

14.6 Where to learn more

Chapter 15

Ridge Regression

15.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

15.2 What it does

15.3 When to do it

15.4 How to do it

15.5 How to interpret the output

15.6 Where to learn more

Chapter 16

Lasso

16.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

16.2 What it does

16.3 When to do it

16.4 How to do it

16.5 How to interpret the output

16.6 Where to learn more

Chapter 17

Principal Component Regression

17.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

17.2 What it does

17.3 When to do it

17.4 How to do it

17.5 How to interpret the output

17.6 Where to learn more

Chapter 18

Bagging

18.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

18.2 What it does

18.3 When to do it

18.4 How to do it

18.5 How to interpret the output

18.6 Where to learn more

Chapter 19

Random Forests

19.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

19.2 What it does

19.3 When to do it

19.4 How to do it

19.5 How to interpret the output

19.6 Where to learn more

Chapter 20

Boosting

20.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

20.2 What it does

20.3 When to do it

20.4 How to do it

20.5 How to interpret the output

20.6 Where to learn more

Chapter 21

Bayesian Additive Regression Trees

21.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

21.2 What it does

21.3 When to do it

21.4 How to do it

21.5 How to interpret the output

21.6 Where to learn more

Chapter 22

Support Vector Machines

22.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

22.2 What it does

22.3 When to do it

22.4 How to do it

22.5 How to interpret the output

22.6 Where to learn more

Chapter 23

Principal Component Analysis

23.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

23.2 What it does

23.3 When to do it

23.4 How to do it

23.5 How to interpret the output

23.6 Where to learn more

Chapter 24

K-Means Clustering

24.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

24.2 What it does

24.3 When to do it

24.4 How to do it

24.5 How to interpret the output

24.6 Where to learn more

Chapter 25

Hierarchical Clustering

25.1 TL;DR

What it does :

When to do it :

How to do it :

How to assess it :

25.2 What it does

25.3 When to do it

25.4 How to do it

25.5 How to interpret the output

25.6 Where to learn more