

Marek Gagolewski

Machine Learning Classics with R

Contents

List of Tables	xi
List of Figures	xiii
Preface	xix
1 Introduction	1
1.1 What is Machine Learning?	1
1.1.1 Data Sources	1
1.1.2 Main Types of Machine Learning Problems	2
1.2 Input Data, \mathbf{X}	3
1.2.1 Abstract Formalism	3
1.2.2 Concrete Example	4
1.3 Unsupervised Learning	7
1.3.1 Dimensionality Reduction	7
1.3.2 Anomaly Detection	9
1.3.3 Clustering	10
1.4 Supervised Learning	11
1.4.1 Desired Outputs, \mathbf{y}	11
1.4.2 Types of Supervised Learning Problems	12
1.4.3 One Dataset – Many Problems	13
2 Agglomerative Hierarchical Clustering	17
2.1 Dataset Partitions	18
2.1.1 Label Vectors	18
2.1.2 K-Partitions	18
2.1.3 “Interesting” Partitions	20
2.2 Euclidean Distance	21
2.3 Hierarchical Clustering at a Glance	23
2.4 Cluster Dendograms	27
2.5 Linkage Functions	28
2.6 Exercises	32
2.7 Remarks	36
3 Classification with Nearest Neighbours	39
3.1 Introduction	39
3.2 K-Nearest Neighbours Classifier	41

3.3	Example in R	44
3.4	Classifier Assessment	45
3.5	Classifier Selection	49
3.6	Implementing a K-NN Classifier (*)	52
3.6.1	Main Routine	52
3.6.2	Mode	53
3.6.3	NN Search Methods	54
3.7	Remarks	57
4	Feature Engineering	59
4.0.1	Feature Engineering	60
4.0.2	Different Metrics (*)	62
4.1	Exercises	64
4.1.1	Wine Quality – Best K-NN Parameters via Cross-Validation (*)	64
4.2	Remarks	65
5	Classification with Decision Trees	67
5.1	Introduction	67
5.1.1	Classification Task	67
5.1.2	Data	69
5.2	Decision Trees	70
5.2.1	Introduction	70
5.2.2	Example in R	71
5.2.3	A Note on Decision Tree Learning	76
5.3	Exercises	76
5.3.1	EdStats – Where Girls Are Better at Maths Than Boys?	76
5.3.2	EdStats and World Factbook – Joining Forces	80
5.3.3	Wine Quality – Random Forest and XGBoost (*)	83
5.4	Outro	85
6	Simple Linear Regression	87
6.1	Simple Regression	87
6.1.1	Introduction	87
6.1.2	Side Note: K-NN Regression	87
6.1.3	Search Space and Objective	91
6.2	Simple Linear Regression	94
6.2.1	Introduction	94
6.2.2	Solution in R	94
6.2.3	Analytic Solution	97
6.2.4	Derivation of the Solution (**)	98
6.3	Exercises	100
6.3.1	The Anscombe Quartet	100
6.3.2	Median House Value in Boston	104
6.4	Outro	106
7	Multiple Regression	107

7.1	Introduction	107
7.1.1	Formalism	107
7.1.2	Simple Linear Regression - Recap	108
7.2	Multiple Linear Regression	109
7.2.1	Problem Formulation	109
7.2.2	Fitting a Linear Model in R	109
7.3	Finding the Best Model	111
7.3.1	Model Diagnostics	111
7.3.2	Variable Selection	121
7.3.3	Variable Transformation	129
7.3.4	Predictive vs. Descriptive Power	130
7.4	Exercises	134
7.4.1	Anscombe's Quartet Revisited	134
7.4.2	Countries of the World – Simple models involving the GDP per capita	137
7.4.3	Countries of the World – Most correlated variables (*)	141
7.4.4	Countries of the World – A non-linear model based on the GDP per capita	144
7.4.5	Countries of the World – A multiple regression model for the per capita GDP	150
7.4.6	Median House Value in Boston (Continued)	153
7.5	Outro	154
7.5.1	Remarks	154
7.5.2	Other Methods for Regression	154
7.5.3	Derivation of the Solution (**)	155
7.5.4	Solution in Matrix Form (***)	156
7.5.5	Pearson's r in Matrix Form (***)	158
8	Classification with Linear Models	161
8.1	Introduction	161
8.1.1	Classification Task	161
8.1.2	Data	162
8.2	Binary Logistic Regression	164
8.2.1	Motivation	164
8.2.2	Logistic Model	165
8.2.3	Example in R	166
8.2.4	Loss Function: Cross-entropy	168
8.3	Exercises	170
8.3.1	EdStats – Fitting of Binary Logistic Regression Models	170
8.3.2	EdStats – Variable Selection in Binary Logistic Regression (*) .	173
8.3.3	Currency Exchange Rates Growth/Fall	179
8.4	Outro	180
8.4.1	Remarks	180
9	Continuous Optimisation with Iterative Algorithms	183

9.1	Introduction	183
9.1.1	Optimisation Problems	183
9.1.2	Types of Minima and Maxima	183
9.1.3	Example Objective over a 2D Domain	187
9.1.4	Example Optimisation Problems in Machine Learning	189
9.2	Iterative Methods	189
9.2.1	Introduction	189
9.2.2	Example in R	190
9.2.3	Convergence to Local Optima	191
9.2.4	Random Restarts	193
9.3	Gradient Descent	194
9.3.1	Function Gradient (*)	194
9.3.2	Three Facts on the Gradient	195
9.3.3	Gradient Descent Algorithm (GD)	197
9.3.4	Example: MNIST (*)	201
9.3.5	Stochastic Gradient Descent (SGD) (*)	204
9.4	A Note on Convex Optimisation (*)	208
9.5	Outro	209
9.5.1	Remarks	209
10	Clustering with K-Means	213
10.1	Within-Cluster Sum of Squares	213
10.2	K-means Clustering	214
10.2.1	Example in R	214
10.2.2	Problem Statement	216
10.2.3	Algorithms for the K-means Problem	218
10.2.4	K-means Revisited	221
10.2.5	optim() vs. kmeans()	223
10.3	Exercises	227
10.3.1	Clustering of the World Factbook	227
10.3.2	Unbalance Dataset – K-Means Needs Multiple Starts	233
10.3.3	Clustering of Typical 2D Benchmark Datasets	237
10.3.4	Wine Quality – volatile.acidity and sulphates	241
10.3.5	Wine Quality – chlorides and total.sulfur.dioxide	241
10.4	Outro	242
10.4.1	Remarks	242
11	Discrete Optimisation	245
11.1	Introduction	245
11.1.1	Recap	245
11.2	Outro	246
11.2.1	Remarks	246
12	Feature Selection	249
12.1	Introduction	249
12.1.1	Recap	249

12.2	Outro	250
12.2.1	Remarks	250
13	Shallow and Deep Neural Networks	253
13.1	Introduction	253
13.1.1	Binary Logistic Regression: Recap	253
13.1.2	Data	254
13.2	Multinomial Logistic Regression	257
13.2.1	A Note on Data Representation	257
13.2.2	Extending Logistic Regression	257
13.2.3	Softmax Function	258
13.2.4	One-Hot Encoding and Decoding	259
13.2.5	Cross-entropy Revisited	260
13.2.6	Problem Formulation in Matrix Form (***)	262
13.3	Artificial Neural Networks	263
13.3.1	Artificial Neuron	263
13.3.2	Logistic Regression as a Neural Network	264
13.3.3	Example in R	265
13.4	Deep Neural Networks	267
13.4.1	Introduction	267
13.4.2	Activation Functions	267
13.4.3	Example in R - 2 Layers	268
13.4.4	Example in R - 6 Layers	270
13.5	Preprocessing of Data	271
13.5.1	Introduction	271
13.5.2	Image Deskewing	272
13.5.3	Summary of All the Models Considered	274
13.6	Outro	276
13.6.1	Remarks	276
13.6.2	Beyond MNIST	276
14	Recommender Systems	279
14.1	Introduction	279
14.1.1	The Netflix Prize	279
14.1.2	Main Approaches	280
14.1.3	Formalism	281
14.2	Collaborative Filtering	282
14.2.1	Example	282
14.2.2	Similarity Measures	283
14.2.3	User-Based Collaborative Filtering	284
14.2.4	Item-Based Collaborative Filtering	285
14.3	Exercise: The MovieLens Dataset (*)	286
14.3.1	Dataset	286
14.3.2	Data Cleansing	287
14.3.3	Item-Item Similarities	288

14.3.4	Example Recommendations	289
14.3.5	Clustering	290
14.4	Outro	292
14.4.1	Remarks	292
15	Natural Language Processing	295
15.1	TO DO	295
Appendix		295
A	Notation Convention	297
B	Setting Up the R Environment	301
B.1	Installing R	301
B.2	Installing an IDE	301
B.3	Installing Recommended Packages	302
B.4	First R Script in RStudio	302
B.5	Exercises	303
B.5.1	First Steps with Vectors	303
B.5.2	Basic Plotting	304
C	Vector Algebra in R	305
C.1	Motivation	305
C.2	Numeric Vectors	307
C.2.1	Creating Numeric Vectors	307
C.2.2	Vector-Scalar Operations	310
C.2.3	Vector-Vector Operations	311
C.2.4	Aggregation Functions	312
C.2.5	Special Functions	313
C.2.6	Norms and Distances	314
C.2.7	Dot Product (*)	315
C.2.8	Missing and Other Special Values	317
C.3	Logical Vectors	318
C.3.1	Creating Logical Vectors	318
C.3.2	Logical Operations	318
C.3.3	Comparison Operations	319
C.3.4	Aggregation Functions	320
C.4	Character Vectors	321
C.4.1	Creating Character Vectors	321
C.4.2	Concatenating Character Vectors	321
C.4.3	Collapsing Character Vectors	322
C.5	Vector Subsetting	322
C.5.1	Subsetting with Positive Indices	322
C.5.2	Subsetting with Negative Indices	323
C.5.3	Subsetting with Logical Vectors	323
C.5.4	Replacing Elements	324

C.5.5	Other Functions	324
C.6	Named Vectors	325
C.6.1	Creating Named Vectors	325
C.6.2	Subsetting Named Vectors with Character String Indices	326
C.7	Factors	327
C.7.1	Creating Factors	327
C.7.2	Levels	327
C.7.3	Internal Representation (*)	328
C.8	Lists	330
C.8.1	Creating Lists	330
C.8.2	Named Lists	331
C.8.3	Subsetting and Extracting From Lists	332
C.8.4	Common Operations	333
C.9	Exercises	335
C.9.1	AUD/EUR Exchange Rates	335
C.10	Further Reading	336
D	Matrix Algebra in R	337
D.1	Creating Matrices	338
D.1.1	<code>matrix()</code>	338
D.1.2	Stacking Vectors	338
D.1.3	Beyond Numeric Matrices	339
D.1.4	Naming Rows and Columns	339
D.1.5	Other Methods	340
D.1.6	Internal Representation (*)	342
D.2	Common Operations	343
D.2.1	Matrix Transpose	343
D.2.2	Matrix-Scalar Operations	343
D.2.3	Matrix-Matrix Operations	344
D.2.4	Matrix Multiplication (*)	345
D.2.5	Aggregation of Rows and Columns	346
D.2.6	Vectorised Special Functions	347
D.2.7	Matrix-Vector Operations	348
D.3	Matrix Subsetting	349
D.3.1	Selecting Individual Elements	349
D.3.2	Selecting Rows and Columns	349
D.3.3	Selecting Submatrices	351
D.3.4	Selecting Based on Logical Vectors and Matrices	351
D.3.5	Selecting Based on Two-Column Matrices	352
D.4	Exercises	352
D.4.1	Currency Exchange Rates	352
D.4.2	Currency Exchange Rates Relative to 1999	353
D.5	Further Reading	356
E	Data Frame Wrangling in R	357

E.1	Creating Data Frames	358
E.2	Importing Data Frames	359
E.3	Data Frame Subsetting	360
E.3.1	Each Data Frame is a List	360
E.3.2	Each Data Frame is Matrix-like	361
E.4	Common Operations	362
E.5	Metaprogramming and Formulas (*)	365
E.6	Exercises	368
E.6.1	Urban Forest	368
E.7	Air Quality	370
E.8	Further Reading	372
F	Datasets	375
F.1	Sustainable Society Indices	375
F.2	Air Quality	378
F.3	Currency Exchange Rates	380
F.4	Urban Forest	383
F.5	Wine Quality	384
F.6	The World Factbook (Countries of the World)	386
F.7	EdStats (Country-Level Education Statistics)	390
F.8	Food and Nutrient Database for Dietary Studies (FNDDS)	398
F.9	Clustering Benchmarks	400
F.10	Movie Lens (TODO)	401
F.11	Other (TODO)	401
References		405

List of Tables

2.1	The number of possible partitions of a dataset with n elements to K clusters, denoted $\{ \binom{n}{K} \}$, grows rapidly as n increases	19
3.1	Some examples of output labels in binary classification tasks	40
3.2	True vs. predicted labels in a binary classification task; ideally, the number of false positives and false negatives should be kept to a minimum	46

List of Figures

1.1	Unsupervised learning: “But what it is exactly that I have to do here?”	8
1.2	Principal component analysis (a dimensionality reduction technique) applied on 5 columns of the <code>wine_quality</code> dataset.	8
1.3	Outliers can be thought of as anomalies of some sort	9
1.4	Two clusters identified by the K-means algorithm ($K = 2$)	10
1.5	Quantitative (numeric) outputs lead to regression problems	14
1.6	Qualitative (here: binary) outputs lead to classification tasks	15
1.7	Ordinal variables constitute ordinal regression tasks	15
2.1	Example points and Euclidean distances between them	22
2.2	OECD countries grouped w.r.t. the SSI dimensions	25
2.3	Scatterplot matrix for the SSI dimensions with the 3-partition generated by complete linkage	26
2.4	Cluster dendrogram generated by complete linkage	27
2.5	In single linkage, we find the closest pair of points; in complete linkage, we seek the pair furthest away from each other; in average linkage, we determine the arithmetic mean of all inter-cluster pairwise distances .	30
2.6	Cluster dendograms generated by single and average linkages	31
3.1	A synthetic 2D dataset where each point is assigned one of two distinct labels	41
3.2	1-NN class bounds for the 2D synthetic dataset	42
3.3	3-NN class bounds the our 2D synthetic dataset	43
3.4	15-NN class bounds the our 2D synthetic dataset	43
3.5	Performance of K -NN classifiers on the validation set as a function of K	51
5.1	A synthetic 2D dataset with the true decision boundary at $X_1 = 0$	68
5.2	(#fig:plot_rpart) The simplest decision tree for the synthetic 2D dataset and the corresponding decision boundaries	71
5.3	(#fig:plot_rpart2) A more complicated decision tree for the synthetic 2D dataset and the corresponding decision boundaries	72
5.4	(#fig:plot_rpart3) An even more complicated decision tree for the synthetic 2D dataset and the corresponding decision boundaries	72
5.5	(#fig:plot_rpart1) A decision tree for the <code>white_wines</code> dataset	73
5.6	(#fig:plot_rpart22) A (simpler) decision tree for the <code>white_wines</code> data-set	74

5.7	A (more complex) decision tree for the <code>white_wines</code> dataset	75
5.8	A decision tree explaining the <code>girls_rule_maths</code> variable	79
5.9	Another decision tree explaining the <code>girls_rule_maths</code> variable	80
5.10	Yet another decision tree explaining the <code>girls_rule_maths</code> variable	83
6.1	K-NN regression example	88
6.2	(#fig:credit_scatter) A scatter plot of Rating vs. Balance	89
6.3	(#fig:credit_XY_plot) A scatter plot of Rating vs. Balance with clients of Balance=0 removed	90
6.4	(#fig:normal_distrib) Probability density functions of normal distributions with different standard deviations σ	91
6.5	(#fig:credit_different_models) Different polynomial models fitted to data	92
6.6	(#fig:credit_residuals) Residuals are defined as the differences between the predicted and observed outputs $\hat{y}_i - y_i$	92
6.7	(#fig:credit_different_lines_ss) Three simple linear models together with the corresponding SSRs	95
6.8	(#fig:credit_plot_lm) Fitted regression line	96
6.9	(#fig:anscombe_fit_apply) Fitted regression lines for the Anscombe quartet	103
6.10	(#fig:anscombe_resid_plot) Residuals vs. fitted values for the regression lines fitted to the Anscombe quartet	105
7.1	(#fig:simple_recap2) Fitted regression line for the Credit dataset	109
7.2	Fitted regression plane for the Credit dataset	110
7.3	(#fig:x12_y) Scatter plots of Y vs. X_1 and X_2	111
7.4	(#fig:x12_ycolmap) A heatmap for Rating as a function of Balance and Income; greens represent low credit ratings, whereas reds – high ones	112
7.5	(#fig:boxplot_explained) An example boxplot	116
7.6	(#fig:boxplot_residuals) Box plots of the residuals for the three models studied	116
7.7	(#fig:violinplot_residuals) Violin plots of the residuals for the three models studied	117
7.8	(#fig:absresiduals_boxplot) Box plots of the modules of the residuals for the three models studied	118
7.9	(#fig:resid_vs_fitted) Residuals vs. fitted outputs for the three regression models	120
7.10	Scatter plot matrix for the Credit dataset	122
7.11	(#fig:pearson_interpret) Different datasets and the corresponding Pearson's r coefficients	123
7.12	Polynomials of different degrees fitted to the Credit dataset	130
7.13	Synthetic data generated by means of the formula $Y = 3x^3 + 5 (+ \text{noise})$	131
7.14	Polynomials fitted to our synthetic dataset	132
7.15	MSE on the dataset used to construct the model vs. MSE on a whole range of points as function of the polynomial degree	133

7.16	Fitted regression line for ans1	135
7.17	Fitted quadratic model for ans2	135
7.18	Scatter plot for ans3	136
7.19	Scatter plot for ans3 with the outlier removed and the fitted linear model	137
7.20	A scatter plot matrix and regression lines for the 4 variables most correlated with the per capita GDP	140
7.21	Most correlated pair of variables and the fitted regression line	144
7.22	Histograms of the empirical distribution of the GDP per capita with linear (left) and log (right) scale on the X axis	145
7.23	Histogram of the empirical distribution of the GDP per capita now with human-readable X axis labels (not the logarithmic scale)	146
7.24	Linear model fitted for life expectancy vs. GDP/person	147
7.25	Scatter plot of life expectancy vs. GDP/person with log scale on the X axis	147
7.26	Linear model fitted for life expectancy vs. the logarithm of GDP/person	149
7.27	Logarithmic model fitted for life expectancy vs. GDP/person	149
7.28	Scatter plot matrix for GDP, imports and exports	151
8.1	A synthetic 2D dataset with the true decision boundary at $X_1 = 0$	162
8.2	Quality metrics for a binary classifier “Classify X as 1 iff $(X) > T$ and as 0 iff $(X) \leq T$ ”	165
8.3	The logistic sigmoid function, φ	166
8.4	The probability that a given wine is a high-alcohol one given its density; black and red points denote the actual observed data points from the class 0 and 1, respectively	167
9.1	(#fig:f_global_minimum) A function with the global minimum at $x^* = 1$	184
9.2	(#fig:f_global_minimum_not_unique) A function that has multiple minima	185
9.3	(#fig:f_global_local_minima) A function with two local minima	185
9.4	(#fig:smooth_vs_nonsmooth) Smooth vs. non-smooth vs. noisy objective functions	186
9.5	(#fig:contour_g) A contour plot and a heat map of $g(x_1, x_2)$	188
9.6	(#fig:perspective_g) Perspective plots of $g(x_1, x_2)$	188
9.7	Each plotting symbol marks a point where the objective function was evaluated by the BFGS method	191
9.8	(#fig:bfgs_multi_hist) A histogram of the objective function’s value at the local minimum found when using a random initial guess	192
9.9	(#fig:bfgs_multi_where) Each line segment connects a starting point to the point of where the BFGS algorithm converged	193
9.10	Scaled gradients (pink arrows) and minus gradients (blue arrows) of $g(x_1, x_2)$ at different points	197
9.11	Path taken by the gradient descent algorithm with $\eta = 0.01$	199
9.12	Path taken by the gradient descent algorithm with $\eta = 0.05$	199
9.13	Path taken by the gradient descent algorithm with $\eta = 0.1$	200

9.14 (#fig:mnist_sgd) Cross-entropy and accuracy on the train and test set in each iteration of SGD; batch size of 32	207
9.15 (#fig:mnist_sgd2b) Cross-entropy and accuracy on the train and test set in each iteration of SGD; batch size of 128	207
9.16 An illustration of the definition of a convex function	208
9.17 A convex and a non-convex set	210
10.1 3-means clustering on a projection of the Iris dataset	215
10.2 3-means clustering (colours) vs true Iris species (shapes)	215
10.3 (#fig:kmeans_problem1) Cluster centres (blue dots) identified by the 3-means algorithm	217
10.4 (#fig:kmeans_problem2) The division of the whole space into three sets based on the proximity to cluster centres (a so-called Voronoi diagram)	218
10.5 (#fig:kmeanimpl_plot) The arrows denote the cluster centres in each iteration of the Lloyd algorithm	222
10.6 (#fig:gendata_example) plot of chunk gendata_example	224
10.7 plot of chunk gendatas	225
10.8 (#fig:clustering_factbook7) Cluster dendrogram for the World Factbook dataset – Complete linkage	230
10.9 (#fig:clustering_factbook10) Cluster dendrogram for the World Factbook dataset – Genie algorithm	231
10.10 (#fig:clustering_factbook12) 3 clusters discovered by the Genie algorithm	232
10.11 (#fig:sipu_unbalance2) <code>sipu_unbalance</code> dataset	234
10.12 (#fig:sipu_unbalance3a) Results of K-means on the <code>sipu_unbalance</code> dataset	234
10.13 (#fig:sipu_unbalance3b) Results of K-means on the <code>sipu_unbalance</code> dataset – many more restarts	235
10.14 (#fig:sipu_unbalance6) Results of K-means on the <code>sipu_unbalance</code> dataset – an educated guess on the cluster centres' locations	236
10.15 (#fig:clustering_benchmarks_plot1) Clustering of the <code>wut_isolation</code> dataset	238
10.16 (#fig:clustering_benchmarks_plot2) Clustering of the <code>wut_mk2</code> dataset . .	239
10.17 (#fig:clustering_benchmarks_plot3) Clustering of the <code>wut_z3</code> dataset . .	239
10.18 (#fig:clustering_benchmarks_plot4) Clustering of the <code>sipu_aggregation</code> dataset	240
10.19 (#fig:clustering_benchmarks_plot5) Clustering of the <code>sipu_pathbased</code> dataset	240
10.20 (#fig:clustering_benchmarks_plot6) Clustering of the <code>sipu_unbalance</code> dataset	241
10.21 (#fig:kmeans_different_K) 3-means (colours) vs. 4-means (symbols) on example data; the “circle” cluster cannot decide if it likes the green or the black one more	243
13.1 (#fig:mnist_demo) Example images in the MNIST database	255
13.2 (#fig:mnist_info2b) Example image from the MNIST dataset	256

13.3 (#fig:cross_entropy_revisited_example3) The less the classifier is confident about the prediction of the actually true label, the greater the penalty	261
13.4 Neuron as a mathematical (black box) function; image based on: https://en.wikipedia.org/wiki/File:Neuron3.png by Egm4313.s12 at English Wikipedia, licensed under the Creative Commons Attribution-Share Alike 3.0 Unported license	263
13.5 A simple model of an artificial neuron	264
13.6 Binary logistic regression	264
13.7 Multinomial logistic regression	265
13.8 Performance metrics for multinomial logistic regression on MNIST	267
13.9 A multi-layer neural network	268
13.10 Performance metrics for a 2-layer net 784-800-10 [relu] on MNIST	269
13.11 Performance metrics for a 6-layer net 784-2500-2000-1500-1000-500-10 [relu] on MNIST	271
13.12 Deskewing of the MNIST digits	272
13.13 Performance of Multinomial Logistic Regression on the deskewed MNIST	274
13.14 Summary of F-measures for each classified digit and every method	275
13.15 A heat map of F-measures for each classified digit and each method	275
14.1 Cluster dendrogram for the movies	291
C.1 Uniformly vs. normally distributed random variables	310
C.2 An example plot of the sine and cosine functions	314
C.3 Example vectors in 2D	316
C.4 <code>as.numeric()</code> on factors can be used to create different plotting styles	330
D.1 Example plot with <code>matplotlib()</code>	347
D.2 Currency exchange rates relative to 1999-01-04 (with EUR as the base currency)	355
E.1 Metaprogramming in action: Take a look at the Y axis label	365
E.2 Example box plot created via the formula interface	367
E.3 The number of trees planted in the City of Melbourne each year	370
E.4 Monthly averages of the air quality parameters in Geelong, VIC, Australia	373

Preface

This is a draft version (distributed in the hope that it will be useful) of the book *Machine Learning Classics with R* by Marek Gagolewski.

Please submit any feature requests, remarks, and bug fixes via the project site at [github¹](https://github.com/gagolews/lmlcr). Thanks!

Copyright (C) 2020, Marek Gagolewski². This material is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International³ License (CC BY-NC-ND 4.0).

You can access this book at:

- <https://lmlcr.gagolewski.com/> (a browser-friendly version)
- <https://lmlcr.gagolewski.com/lmlcr.pdf> (PDF)
- <https://github.com/gagolews/lmlcr> (source code)

Aims and Scope

Machine learning has numerous exciting real-world applications, including stock market prediction, speech recognition, computer-aided medical diagnosis, content, and product recommendation, anomaly detection in security camera footage, game playing, autonomous vehicle operation, and many others.

In this book we will take an unpretentious glance at the most fundamental algorithms that have stood the test of time and which form the basis for state-of-the-art solutions

¹<https://github.com/gagolews/lmlcr/issues>

²<https://www.gagolewski.com>

³<https://creativecommons.org/licenses/by-nc-nd/4.0/>

of modern AI, which is principally (big) data-driven. We will learn how to use the R language (R Development Core Team 2020) for implementing various stages of data processing and modelling activities. For a more in-depth treatment of R, refer to this book's Appendices and, for instance, (Wickham & Grolemund 2017, Peng 2019, Venables et al. 2020).

These pages contain solid underpinnings for further studies related to statistical learning, machine learning, data science, data analytics, data mining, and artificial intelligence, including (Bishop 2006, Efron & Hastie 2016, Hastie et al. 2017). We will also appreciate the vital role of mathematics as a universal language for formalising data-intense problems and communicating their solutions. The book is aimed at readers who are yet to be fluent with university-level linear algebra, calculus as well as probability theory, such as 1st year undergrads or those who have forgotten all the maths they have learned and need a gentle, non-invasive, yet rigorous introduction to the topic. For a nice, machine learning-focused introduction to mathematics alone, see, e.g., (Deisenroth et al. 2020).

About the Author

I, Marek Gagolewski⁴, am currently a Senior Lecturer in Applied AI at Deakin University in Melbourne, VIC, Australia and an Associate Professor in Data Science at Warsaw University of Technology, Poland. My main passion is in research – my primary interests currently include machine learning and optimisation algorithms, data aggregation and clustering, statistical modelling, and scientific computing. *Explaining* of things matters to me more than merely tuning the knobs so as to increase a chosen performance metric (with uncontrollable consequences to other ones); the latter rather belongs to technology and wizardry, not science.

I'm an author of more than 70 publications. I've developed several open source R and Python packages, including `stringi`⁵ – one of the most often downloaded R extensions that aims at text/natural language processing, and `genieclust`⁶ – featuring a fast&robust implementation of the hierarchical clustering algorithm named *Genie*.

On top of that, I teach various courses related to R and Python programming, algorithms, data science, and machine learning – and I'm good at it. This book was also influenced by my teaching experience at Data Science Retreat⁷ in Berlin, Germany.

Acknowledgements

This book has been prepared with R version 4.0.2 (2020-06-22) as well as pandoc, Markdown, and GitBook. R code chunks have been processed with knitr. A little help of bookdown, good ol' Makefiles, and shell scripts did the trick.

The following R packages are used or referred to in the text: bookdown, Cairo, DEoptim,

⁴<http://www.gagolewski.com/>

⁵<https://stringi.gagolewski.com/>

⁶<https://genieclust.gagolewski.com/>

⁷<https://datascienceretreat.com>

fastcluster, FNN, genie, genieclust, glmnet, gsl, hydroPSO, ISLR, keras, knitr, MASS, Matrix, microbenchmark, pdist, randomForest, RColorBrewer, recommenderlab, rpart, rpart.plot, rworldmap, scatterplot3d, stringi, tensorflow, tidyverse, titanic, vioplot, xgboost.

During the writing of this book, I've been listening to the music featuring John Coltrane, Krzysztof Komeda, Henry Threadgill, Albert Ayler, Paco de Lucia, and Tomatito.

1

Introduction

An *algorithm* is a well-defined sequence of instructions that, for a given sequence of input arguments, yields some desired output. In other words, it is a specific recipe for what we call a *function* in mathematics. Unfortunately, algorithm development is a tedious task. We need to be super-precise about covering all the possible scenarios and modelling them accurately. When we are to build a self-driving vehicle, we need to be ready for an infinite number of situations that can arise on the road. When we are to build a system for diagnosing tumour on medical images, there are so many decisions to make of whether should we consider a particular area suspicious, how does it relate to the neighbouring cells, etc.

1.1 What is Machine Learning?

In *machine learning* (ML), we build and study computer algorithms that make *predictions* or *decisions* but whose *all tiny bits* are not manually programmed. Moreover, some algorithms might also be able to *discover* new interesting facts about a problem instance at hand.

Learning, however, needs some material based upon which new knowledge is to be developed. In other words, we need *data*.

1.1.1 Data Sources

Information can come from various sources, e.g., physical sensors, files, databases, or (pseudo)random number generators. It can take different forms, e.g., vectors, matrices and other tensors, graphs, audio/video streams, text etc. With the advent of the internet era, data have become ubiquitous.

Exercise 1.1 *Think of how much information you consume and generate when you interact with your social media or news feeds every day.*

Here are some application domains where machine learning has already proven itself very useful:

- Financial services (banking, insurance, investment funds)

- Oil, gas, and other energy (solar, mining)
- Real estate
- Pharmaceuticals
- Advertising
- Transportation
- Retail
- Healthcare
- Food production

Exercise 1.2 Think of different ways in which these sectors could benefit from ML solutions.

To be frank, the above list was generated by duckduckgoing the “biggest industries” query. That was a very easy task; ML is already everywhere. Basically, wherever we have data and there is a need to improve some processes or discover new aspects about a problem domain, there is a place for ML solutions.

Of course, it’s not all about business revenue (luckily). We can do a lot of great work for greater good; with the increased availability of open data, everyone can be a reporter, an engaged citizen that seeks for truth. There are NGOs. Finally, there are researchers (remember that the main role of most universities is still to spread the advancement of knowledge and not to make money!) that need these methods to make new discoveries, e.g., in psychology, sociology, agriculture, engineering, biotechnology, pharmacy, medicine, genetics, you name it.

1.1.2 Main Types of Machine Learning Problems

Machine Learning problems include, but are not limited to:

- *Supervised learning* – for every input point (e.g., a photo) there is an associated desired output (e.g., whether it depicts a crosswalk or how many cars can be seen on it); in other words, there is a “teacher” ready to instruct us about what is the expected behaviour in a given context.
- *Semi-supervised learning* – some inputs are labelled, other ones are not (definitely a less laborious/cheaper scenario).
- *Unsupervised learning* – inputs are unlabelled; the aim is to discover the underlying structure in the data (e.g., automatically group customers w.r.t. common behavioural patterns); think about how much you learnt by just interacting with your environment by wandering here and there.
- *Reinforcement learning* – learn to act based on a feedback given after the actual decision was made such as “that was nice! well done! look at you go!” (e.g., learn to play some video game by testing different hypotheses what to do in order to prosper/survive as long as possible).

In the course of this book, we're going to take a deep dive into the machine learning algorithms that stood the test of time and that are still used – with modifications – by researchers and practitioners. We will learn about how we can use them for improving different processes. Importantly, we will also get to know their limitations.

Arranging the material so as to maximise its usefulness (we'd like to empower ourselves with a set of tools that can get some work done), lay solid groundwork for further studies and develop the adjoining skills (the practice of data analysis certainly requires us to learn programming and be able not only to read but also understand mathematical notation) is a multi-objective optimisation problem. A problem involving many constraints though; as this is an introductory course, we obviously cannot cover everything, not yet. In particular, we will be learning maths and programming “from scratch”, our space-time is limited etc. We also want to enjoy our journey – to have some fun knowing that big projects never come into being in a day or seven.

Therefore, this time we will limit ourselves “just” to supervised and unsupervised learning, as these are the most common instances.

1.2 Input Data, \mathbf{X}

1.2.1 Abstract Formalism

Let $\mathcal{X} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}\}$ be an input sample (“a database”) that contains information on n entities or phenomena (in their entirety or at different points in time or space). Most often we assume that each object in \mathcal{X} , say $\mathbf{x}^{(i)}$, is represented by means of a sequence of p real numbers (for some p). We denote this fact as $\mathbf{x}^{(i)} \in \mathbb{R}^p$.

Remark 1.1 Here, the “ \in ” symbol stand for “is in” or “is a member of”, \mathbb{R} denotes the set of real numbers (the one that features, 0, -358745.2394 , 42 and π , amongst uncountably many others) and \mathbb{R}^p is the set of real-valued sequences of length p (i.e., p such numbers considered at a time). For instance, \mathbb{R}^2 includes pairs such as $(1, 2)$, $(-6.4, 0)$, and $(1/3, 10^3)$. More generally, we write $(x_1, \dots, x_p) \in \mathbb{R}^p$, where $x_j \in \mathbb{R}$ is the j -th component of the sequence, for any $j = 1, \dots, p$.

Remark 1.2 To put emphasis on the fact that (x_1, \dots, x_p) is not an “atomic” entity, but a one that a collection of many scalar components, we will use bold fonts for denoting whole sequences, e.g., write $\mathbf{x} = (x_1, \dots, x_p)$.

Remark 1.3 If $\mathbf{x} \in \mathbb{R}^p$, then we often say that \mathbf{x} is a sequence of p numbers, a (numeric) p -tuple, a p -dimensional real vector, a point in a p -dimensional real space, or an element of a real p -space etc. We explain the basics of notation convention used throughout this book (and in most math works in general) in Appendix A.

At a first glance, such a representation might seem overly simplistic, especially if we'd wish to store information on “very complex” objects. However, it turns out that in most cases expressing them as *feature vectors* (i.e., establishing a set of numeric attributes that

best describe them in a task at hand) is not only natural but also perfectly sufficient for achieving whatever we aim at.

Exercise 1.3 Consider the following problems:

- How would you represent a patient in a clinic (for the purpose of conducting research in cardiology)?
- How would you represent a car in an insurance company's database (for the purpose of determining how much a driver should pay annually for the mandatory policy)?
- How would you represent a student in an university (for the purpose of granting them scholarships)?

In each case, list a few numeric features that best describe the reality of concern. Note that descriptive labels can always be encoded as numbers, e.g., female = 1, male = 2.

The setting we've established is *abstract* in the sense that there might be different realities *hidden* behind the symbols we use. This is exactly what maths is for – creating *abstractions* or *models* of complex entities/phenomena so that they can be much more easily manipulated or understood.

This is very powerful – spend a moment contemplating how many real-world situations fit into our framework. In particular the last chapters are devoted to the less-obvious data domains, for instance text or image/video data, e.g., a 1920×1080 pixel image can be “unwound” to a “flat” sequence of length 2,073,600.

The whole dataset consisting of n observations (samples, cases, records) and p features (variables, attributes, characteristics) can be written as an $(n \times p)$ -matrix:

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{bmatrix}.$$

Mathematically, we denote this as $\mathbf{X} \in \mathbb{R}^{n \times p}$ (read “ \mathbf{X} is a real-valued matrix with n rows and p columns”).

Remark 1.4 A matrix is an array of rectangular shape, with items arranged in rows and columns. In cases such as this we say that we deal with structured (tabular) data (think: Excel/Calc spreadsheets, SQL tables etc.).

1.2.2 Concrete Example

Let's consider the Wine Quality dataset (see Section F.5 for more details), that stores information on 11 physicochemical features (amongst others – see below) of ca. 6500 Portuguese wines from different wineries.

```
wine_quality <- read.csv("datasets/wine_quality_all.csv",
  comment.char="#")
X <- wine_quality[, 1:11] # extract the first 11 columns
X <- as.matrix(X) # convert to numeric matrix
head(X) # display the first 6 rows, i.e., X[1:6,]

##      fixed.acidity volatile.acidity citric.acid ... alcohol
## [1,]      7.4          0.70     0.00 ...    9.4
## [2,]      7.8          0.88     0.00 ...    9.8
## [3,]      7.8          0.76     0.04 ...    9.8
## [4,]     11.2          0.28     0.56 ...    9.8
## [5,]      7.4          0.70     0.00 ...    9.4
## [6,]      7.4          0.66     0.00 ...    9.4
```

We can get to now the dimensionality (shape) of \mathbf{X} by calling:

```
dim(X) # gives n and p, respectively
## [1] 6497   11
nrow(X) # dim(X)[1] - 1st dimension - n
## [1] 6497
ncol(X) # dim(X)[2] - 2nd dimension - p
## [1] 11
```

Therefore, $\mathbf{X} \in \mathbb{R}^{6497 \times 11}$, i.e., $n = 6497$ and $p = 11$.

Remark 1.5 Obviously, displaying \mathbf{X} in its entirety would take up a lot of pages. Not only there are so many rows, also the number of columns makes the matrix too “wide” – mainly because of the presence of the descriptive feature labels (column names). Therefore, when printed out on the console, R will “wrap” the columns to fit the page width.

```
head(X, 3) # the first 3 rows
```

```
##      fixed.acidity volatile.acidity citric.acid residual.sugar
## [1,]      7.4          0.70     0.00           1.9
## [2,]      7.8          0.88     0.00           2.6
## [3,]      7.8          0.76     0.04           2.3
##      chlorides free.sulfur.dioxide total.sulfur.dioxide density pH
## [1,]     0.076              11            34  0.9978 3.51
## [2,]     0.098              25            67  0.9968 3.20
## [3,]     0.092              15            54  0.9970 3.26
##      sulphates alcohol
## [1,]     0.56      9.4
## [2,]     0.68      9.8
## [3,]     0.65      9.8
```

Let's forget about the column labels for the time being, as they might be distracting us from fully appreciating the abstract nature of the setting we're trying to develop.

```
dimnames(X) <- NULL # drop row and column names
head(X) # show the first 6 rows
```

```
## [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]
## [1,] 7.4 0.70 0.00 1.9 0.076 11 34 0.9978 3.51 0.56 9.4
## [2,] 7.8 0.88 0.00 2.6 0.098 25 67 0.9968 3.20 0.68 9.8
## [3,] 7.8 0.76 0.04 2.3 0.092 15 54 0.9970 3.26 0.65 9.8
## [4,] 11.2 0.28 0.56 1.9 0.075 17 60 0.9980 3.16 0.58 9.8
## [5,] 7.4 0.70 0.00 1.9 0.076 11 34 0.9978 3.51 0.56 9.4
## [6,] 7.4 0.66 0.00 1.8 0.075 13 40 0.9978 3.51 0.56 9.4
```

An element of \mathbf{X} located in the i -th row and the j -th column, denoted $x_{i,j} \in \mathbb{R}$, represents the j -th feature of the i -th observation, $i = 1, \dots, n, j = 1, \dots, p$.

For instance, the 3rd wine's volatile acidity (expressed in grams of acetic acid per dm³), $x_{3,2}$, is:

```
X[3, 2] # 3rd row, 2nd column
```

```
## [1] 0.76
```

The third observation (data point, row in \mathbf{X} , i.e., the 3rd wine in the dataset) consists of items $(x_{3,1}, \dots, x_{3,p})$ that can be extracted by calling:

```
X[3, ]
```

```
## [1] 7.800 0.760 0.040 2.300 0.092 15.000 54.000 0.997 3.260
## [10] 0.650 9.800
```

Moreover, the second feature/variable/column (in our case: volatile acidity) is comprised of $(x_{1,2}, x_{2,2}, \dots, x_{n,2})$:

```
X[, 2]
```

```
## [1] 0.700 0.880 0.760 0.280 0.700 0.660 0.600 0.650 0.580 0.500 0.580
## [12] 0.500 0.615 0.610 0.620 0.620 0.280 0.560 0.590 0.320 0.220 0.390
## [23] 0.430 0.490 0.400 0.390 0.410 0.430 0.710 0.645 0.675 0.685 0.655
## [34] 0.605 0.320 0.645 0.600 0.380 1.130 0.450 0.450 0.610 0.490 0.660
## [45] 0.670 0.520 0.935 0.290 0.400 0.310 0.660 0.520 0.500 0.380 0.510
## [56] 0.620 0.420 0.630 0.590 0.390 0.400 0.690 0.520 0.735 0.725 0.725
## [67] 0.520 0.705 0.320 0.705 0.630 0.670 0.690 0.675 0.320 0.410 0.410
## [78] 0.785 0.750 0.625 0.450 0.430 0.500 0.670 0.300 0.550 0.490 0.490
## [89] 0.390 0.620 0.520 0.490 0.490 0.490 1.020 0.600 0.775 0.500 0.900
## [ reached getOption("max.print") -- omitted 6398 entries ]
```

Exercise 1.4 Identify the precise location of the value 0.76 in all the above examples.

We will sometimes use the following notation to emphasise that the \mathbf{X} matrix consists of n rows or p columns:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_{1,\cdot} \\ \mathbf{x}_{2,\cdot} \\ \vdots \\ \mathbf{x}_{n,\cdot} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_{\cdot,1} & \mathbf{x}_{\cdot,2} & \cdots & \mathbf{x}_{\cdot,p} \end{bmatrix}.$$

Here, $\mathbf{x}_{i,\cdot}$ is a *row vector* of length p , i.e., a $(1 \times p)$ -matrix:

$$\mathbf{x}_{i,\cdot} = \begin{bmatrix} x_{i,1} & x_{i,2} & \cdots & x_{i,p} \end{bmatrix}.$$

Moreover, $\mathbf{x}_{\cdot,j}$ is a *column vector* of length n , i.e., an $(n \times 1)$ -matrix:

$$\mathbf{x}_{\cdot,j} = \begin{bmatrix} x_{1,j} & x_{2,j} & \cdots & x_{n,j} \end{bmatrix}^T = \begin{bmatrix} x_{1,j} \\ x_{2,j} \\ \vdots \\ x_{n,j} \end{bmatrix},$$

where $.^T$ denotes the *transpose* of a given matrix – think of this as a kind of rotation; it allows us to introduce a set of “vertically stacked” objects using a single inline formula.

Remark 1.6 Details on how to get started with the R environment are provided in Appendix B. A comprehensive introduction to vector and matrix algebra in R as well as basic data frame wrangling is given in Appendices C, D, and E. From now on, we assume that we are all familiar with this material.

1.3 Unsupervised Learning

In *unsupervised learning* (learning without a teacher), the input data vectors $\mathbf{x}_{1,\cdot}, \dots, \mathbf{x}_{n,\cdot}$ (rows in \mathbf{X}) are all we have. Our aim now is to discover the *underlying structure in the data*, whatever that means. For instance, Figure 1.1 depicts a scatter plot of two selected wine features. What interesting information can we extract from this figure? Perhaps not much, this is exactly why we need data mining methods!

1.3.1 Dimensionality Reduction

Certain classes of unsupervised learning problems are not only intellectually stimulating, but practically useful at the same time.

In particular, in *dimensionality reduction* we seek a meaningful projection of a high dimensional space (think: a matrix with many columns) to a space of lower dimension. For

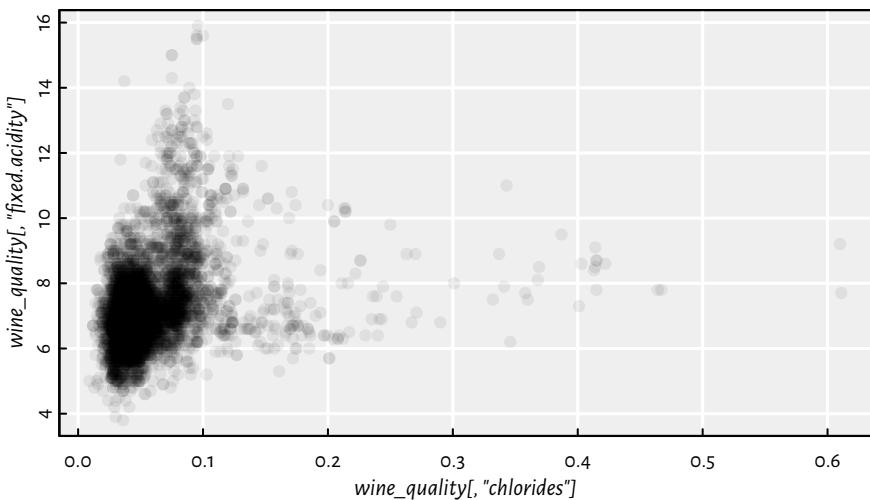


Figure 1.1: Unsupervised learning: “But what it is exactly that I have to do here?”

instance, if the output number of features is 2 (or maybe 3 as well), we will be able to visualise a complex dataset on a standard scatter plot.

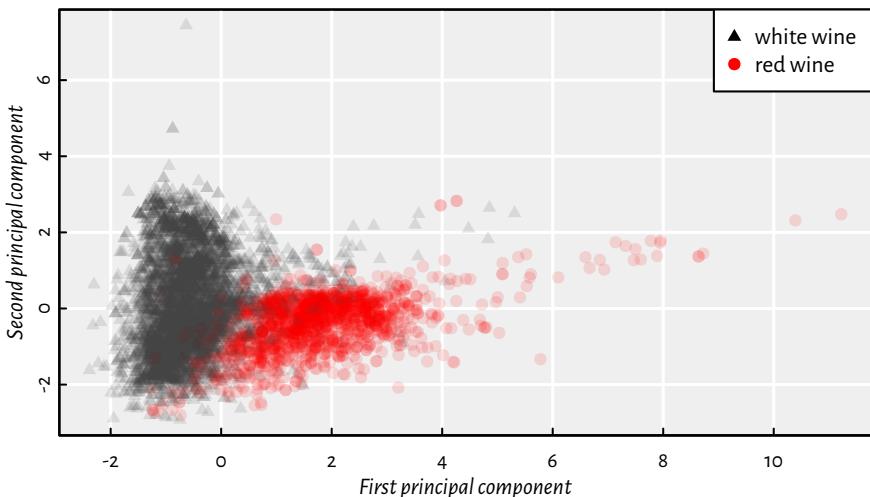


Figure 1.2: Principal component analysis (a dimensionality reduction technique) applied on 5 columns of the `wine_quality` dataset.

For instance, Figure 1.2 depicts the two dimensions “revealed” by the PCA method (Prin-

incipal Component Analysis) – a technique proposed by the mathematician Karl Pearson (the Father of mathematical statistics) at the beginning of the 20th century.

It's actually nice to see that the wines of different colours (see `wine_quality[, "color"]`) form two quite distinct “blobs”. By analysing the results returned by PCA in more detail, we could better understand the differences in physicochemical features of wines of different colours (amongst others).

Once we go through this course, we will be perfectly equipped for studying such dimensionality reduction methods as: multidimensional scaling (MDS), kernel PCA, t-SNE, autoencoders (deep learning). See, for example, (Hastie et al. 2017) for more details.

TODO: discuss PCA – iterative algorithm???

1.3.2 Anomaly Detection

Another family of unsupervised learning methods deals with *anomaly detection*, where our task is to identify rare, suspicious, ab-normal, novel or out-standing items. For example, these can be cars driving on walkways in a park’s security camera footage or fraudulent transactions identified in some financial service.

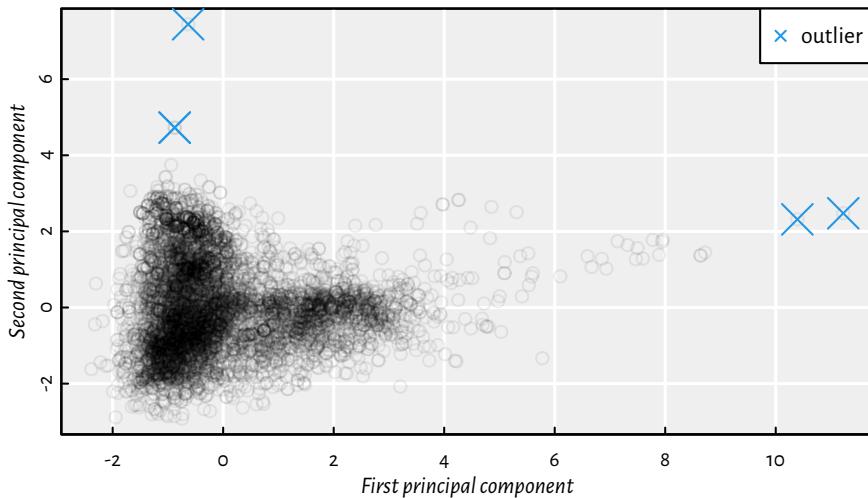


Figure 1.3: Outliers can be thought of as anomalies of some sort

Figure 1.3 highlights a few *outliers* – observations that differ (in the sense they are much farther away) significantly from the rest of the points. They might be due to some measurement or data input error or just be an inherent part of the dataset. Hence, it might be interesting to inspect them in more detail. Section 7.3.1.2 includes one of the simplest (and most popular) definition of an outlier for unidimensional data. More advanced

methods, see, e.g., (Chandola et al. 2009), are based upon the algorithms that we describe also in this book.

1.3.3 Clustering

Moreover, the aim of *clustering* (grouping or segmentation) is to automatically discover some *naturally occurring* subgroups in the data set. For example, these may be customers having different shopping patterns (such as “young parents”, “students”, and “elders”) or protein sequences that are to be grouped by their function.

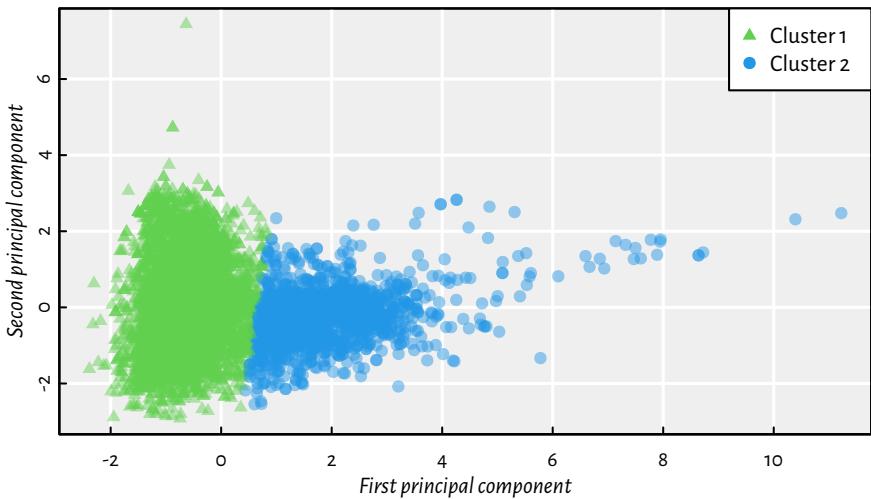


Figure 1.4: Two clusters identified by the K-means algorithm ($K = 2$)

Figure 1.4 gives the results of running the K-means algorithm (see Chapter 10) on the discussed dataset. The two identified clusters are quite similar to the wine colours, despite the fact that this information has not been used during the method’s operation.

We discuss the most seminal clustering techniques in Chapters 2 (hierarchical agglomerative algorithms) and 10 (K-means method), see also, e.g., (Jain 2010), (Wierzchoń & Kłopotek 2018), (Hastie et al. 2017).

1.4 Supervised Learning

1.4.1 Desired Outputs, \mathbf{y}

In supervised learning, there is a teacher; apart from the inputs (in the form of a matrix \mathbf{X}), we are also given the corresponding reference/desired outputs (a vector \mathbf{y}), which are usually encoded as individual numbers (scalars) or textual labels.

For instance, in the `wine_quality` dataset, this can be the `response` column, which gives the wine rating on the scale of 0 (bad) to 10 (excellent).

```
y <- wine_quality[, "response"] # extract the `response` column
Xy <- cbind(X, y) # join X and y (column-bind) - for printing
head(Xy)
```

```
##      fixed.acidity volatile.acidity citric.acid ... alcohol response
## [1,]      7.4          0.70     0.00 ...     9.4      5
## [2,]      7.8          0.88     0.00 ...     9.8      5
## [3,]      7.8          0.76     0.04 ...     9.8      5
## [4,]     11.2          0.28     0.56 ...     9.8      6
## [5,]      7.4          0.70     0.00 ...     9.4      5
## [6,]      7.4          0.66     0.00 ...     9.4      5
```

More formally, with each input $\mathbf{x}_{i \cdot}$, we associate the desired output y_i , $i = 1, \dots, n$. Hence, our dataset is $[\mathbf{X} \mathbf{y}]$, with the i -th row being $[\mathbf{x}_{i \cdot}, y_i]$:

$$[\mathbf{X} \mathbf{y}] = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} & y_1 \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} & y_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} & y_n \end{bmatrix},$$

where:

$$\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix}^T = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

Where do we get the reference outputs from? Sometimes they are the part of the dataset gathering process themselves. In other cases, they require a lot of work and money – in particular, in order to judge the wine quality, someone had to engage a couple of experts and ask them to drink 6500 wine bottles! This decent piece of research certainly required a lot of sacrifices.

Amazon's Mechanical Turk or Google's image reCAPTCHA (identify images with pedestrian crossings) are examples of services for crowdsourcing of such \mathbf{y} s.

1.4.2 Types of Supervised Learning Problems

The aim of supervised learning is to try to create an “algorithm” that, given an input point, generates an output that is as *close* (with respect to some useful metrics) as possible to the desired one. The given data sample will be used to “train” this “model”.

Depending on the type of the elements in \mathbf{y} (the domain of \mathbf{y}), supervised learning problems are usually pigeonholed as:

- *regression* – each y_i is a real number, e.g., y_i = future market stock price with $\mathbf{x}_{i,\cdot}$ = prices from p previous days;
- *classification* – each y_i is a discrete label, e.g., y_i = healthy (0) or ill (1) with $\mathbf{x}_{i,\cdot}$ = a patient’s health record;
- *ordinal regression* (a.k.a. ordinal classification) – each y_i is a rank, e.g., y_i = rating of a TV series on the scale 1–5 with $\mathbf{x}_{i,\cdot}$ = ratings of p most similar productions.

Exercise 1.5 Which of the following are instances of classification problems? Which of them are regression tasks?

- Detect email spam
- Predict a market stock price
- Predict the likeability of a new ad
- Assess credit risk
- Detect tumour tissues in medical images
- Predict time-to-recovery of cancer patients
- Recognise smiling faces on photographs
- Detect unattended luggage in airport security camera footage
- Turn on emergency braking to avoid a collision with pedestrians

What kind of data should you gather in order to tackle them?

Once we build a supervised learning algorithm, its outputs can be interpreted by a human being and used (critically) for whatever purpose (e.g., a GP can utilise it for diagnosing illnesses). Their outputs can be also provided to other algorithms to take immediate actions (e.g., buy those market shares asap!).

We are going to cover quite a decent number of classical approaches to data classification and regression, respectively:

- K-nearest neighbour classifiers (Chapter 3), decision trees (Chapter 5), logistic regression (Chapter 8), and its generalisation in form of feed-forward deep neural networks (Chapter 13);
- K-nearest neighbour regression (Chapter **TODO**), regression trees (Chapter **TODO**) and, linear models for regression (Chapter **TODO**).

1.4.3 One Dataset – Many Problems

The title of this subsection might sound worrying, but we need to be clear about one thing: in mathematics (and all fields that use maths as a universal language), we yearn for problems. If we have a problem, an interesting problem in particular, it means we can occupy ourselves for a long time doing something exciting and challenging. We just love what we do. It's not merely our work, it's our hobby and passion as well.

Therefore, by saying that a single dataset can become an instance of many different ML problems, we should be pretty excited.

For example, in the `wine_quality` dataset we have the following columns:

```
names(wine_quality) # names of the columns
```

```
## [1] "fixed.acidity"           "volatile.acidity"
## [3] "citric.acid"             "residual.sugar"
## [5] "chlorides"                "free.sulfur.dioxide"
## [7] "total.sulfur.dioxide"    "density"
## [9] "pH"                      "sulphates"
## [11] "alcohol"                  "response"
## [13] "color"
```

Any of the 11 physicochemical features can become the source of a regression task (with \mathbf{X} being all the remaining 10 – or less), because they are numeric (quantitative) variables. For instance, Figure 1.5 depicts the empirical distribution (density) of the `alcohol` variable in form of a histogram. We could be curious if, for example, any other chemical component predicts high alcohol content.

Here are some basic summary statistics of this variable:

```
hist(wine_quality[, "alcohol"], main="", col="white"); box()
summary(wine_quality[, "alcohol"])
```

```
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
##      8.0    9.5   10.3    10.5   11.3   14.9
```

The outputs generated by `summary()` include:

- **Min.** – sample minimum (the smallest value in the sample);
- **1st Qu.** – 1st quartile (25th percentile, quantile of order 0.25), being the observation q such that 25% values are not greater than q and 75% values are not smaller than q ;
- **Median** – median (2nd quartile, 50th percentile, quantile of order 0.5), being the observation “in the middle of the sample”;
- **Mean** – arithmetic mean;
- **3rd Qu.** – 3rd quartile (75th percentile, quantile of order 0.75);
- **Max.** – sample maximum.

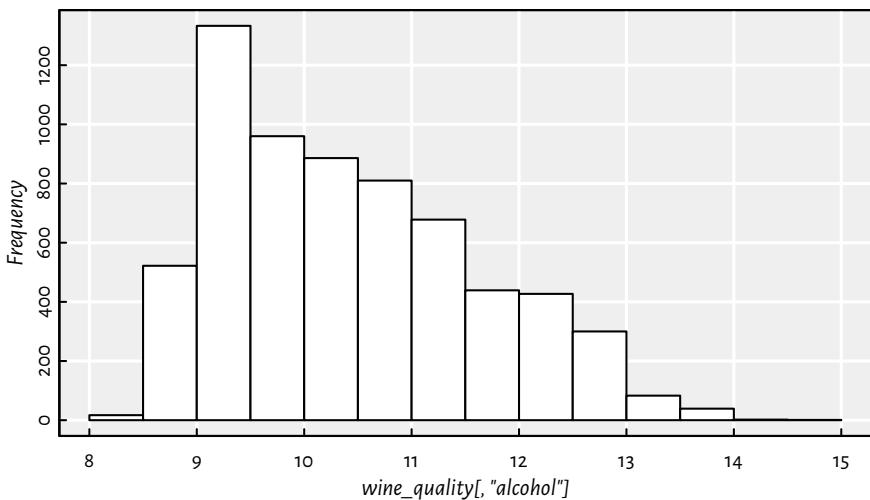


Figure 1.5: Quantitative (numeric) outputs lead to regression problems

Remark 1.7 The 1st, 2nd, and 3rd quartile of $(3, 2, 5, 4, 1)$ is 2, 3, 4, respectively (it's easiest to compute the quantiles by ordering the input sequence). Computing quantiles in the case where they are not defined in an unambiguous manner requires some (weighted) averaging, see `?quantile`. For instance, most often we'll say that the median of $(3, 2, 5, 6, 1, 4)$ is 3.5.

On the other hand, `color` is a qualitative variable with two possible outcomes (i.e., binary; see Figure 1.6 for a bar plot). Therefore, it can constitute a classification task.

Quantitative data can be summarised with a contingency table:

```
barplot(table(wine_quality[, "color"]), col="white", ylim=c(0, 6000))
table(wine_quality[, "color"])
```

```
##  
##   red white  
## 1599 4898
```

Moreover, `response` is an ordinal variable, representing a wine's rating as assigned by a wine expert (see Figure 1.7 for a barplot). Note that although the ranks are represented with numbers, we they are not continuous variables per se. However, these ranks are “something more” than just labels – they are linearly ordered, we know what’s the smallest rank and what’s the greatest one, ranks 3 and 7 are not just different: they are very different, etc.

```
barplot(table(wine_quality[, "response"]), col="white", ylim=c(0, 3000))
table(wine_quality[, "response"])
```

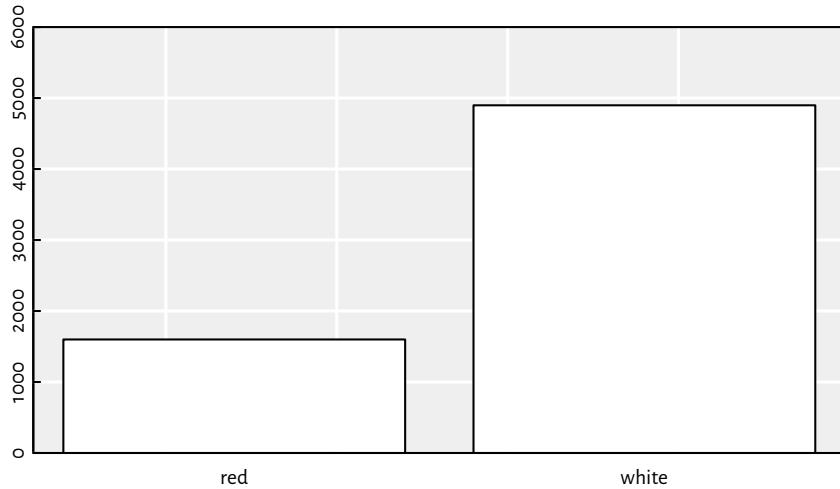


Figure 1.6: Qualitative (here: binary) outputs lead to classification tasks

```
##  
##      3      4      5      6      7      8      9  
##    30   216  2138  2836  1079   193     5
```

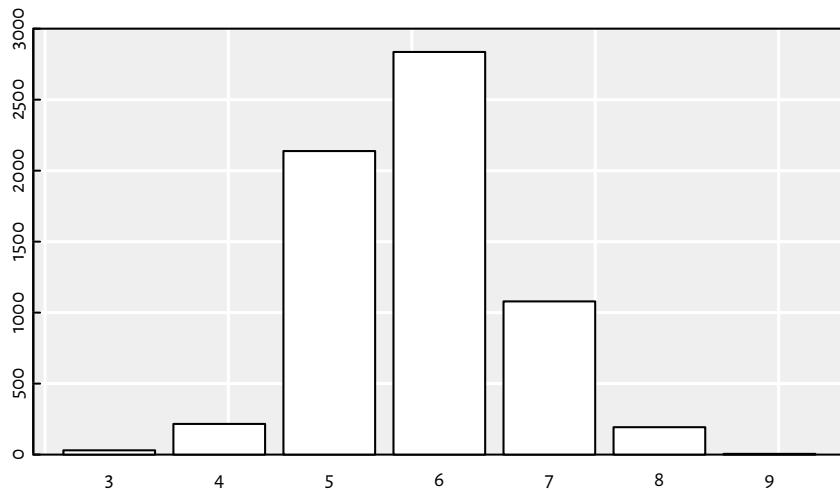


Figure 1.7: Ordinal variables constitute ordinal regression tasks

TODO

Machine learning is not a cure for all our pains and itches! Don't get the false impression that everything is in the data and "will discover itself". We still need expert knowledge at various stages of working with data: collection, preprocessing, inference, prediction, communication etc.; e.g., which models/variables we choose, which constraint we add etc.

Many "big" tasks are, and will be better off when modelled by human (e.g., using partial differential equations), such as prediction of weather. Machine learning could be a good addition to tune-up some of the underlying parameters or as a way to smartly aggregate multiple models into one forecast.

In the next chapter we will discuss our first family of clustering algorithms.

2

Agglomerative Hierarchical Clustering

TODO In this chapter, we will:

- concepts: clustering, data partition,
- Euclidean distance as a means to measure the (dis)similarity between two objects (rows in a matrix)
- algorithms: agglomerative hierarchical clustering with single, complete and average linkage

this is not our last encounter with clustering, more in chapter TODO

The aim of clustering (a.k.a. data segmentation or quantisation) is to split the input dataset into *interesting* subgroups in an unsupervised manner. Example applications of clustering include:

- taxonomisation, e.g., partition consumers to more “uniform” groups to better understand who they are and what do they need,
- aggregation, e.g., to reduce the number of observations by substituting them by “group representatives” or “prototypes”,
- object detection in images, e.g., tumour tissues on PET/CT scans,
- complex networks analysis, e.g., detecting communities in friendship, retweets, and other networks,
- spatial data analysis, e.g., identifying densely populated areas or traffic jams on maps,
- text analysis, e.g., for grouping documents into automatically detected topics,
- fine-tuning of supervised learning algorithms, e.g., recommender systems indicating content that was rated highly by users from the same group or learning multiple manifolds in a dimension reduction task.

2.1 Dataset Partitions

2.1.1 Label Vectors

Let's assume that the input data set $\mathbf{X} \in \mathbb{R}^{n \times p}$ consists of n observations and that we wish to identify K clusters, for some $K \geq 2$ which gives the number of subgroups. Such a clustering of \mathbf{X} will be represented by a vector $\mathbf{y} \in \{1, 2, \dots, K\}^n$ such that y_i gives the *cluster ID* (cluster identifier or number, an integer between 1 and K) of the i -th observation, $i = 1, 2, \dots, n$. Thus, we can think of clustering as of the process of labelling or assigning colours to each data point (for example, 1 = black, 2 = red, 3 = green, 4 = blue etc.). The only restriction we impose on each such *label* vector is that *each of the K labels must occur at least once*. This will guarantee that we have indeed a division of \mathbf{X} into K subgroups, and not less.

Remark 2.1 Given an integer K and a vector \mathbf{y} with elements $\{1, 2, \dots, K\}$, for instance:

```
y <- c(3, 1, 1, 2, 2, 1, 2, 3, 1, 1, 1)
K <- 3 # by the way, it's max(y)
```

we can check whether it represents a proper label vector by computing the number of occurrences of each label:

```
tabulate(y, K) # how many 1s, ..., Ks are there in y, respectively
```

```
## [1] 6 3 2
all(tabulate(y, K) > 0) # do we have at least 1 occurrence of each label?
```

```
## [1] TRUE
```

or, for example, determining the set of all unique values in \mathbf{y} :

```
unique(y)
```

```
## [1] 3 1 2
length(unique(y)) == K # do we have K unique values in y?
```

```
## [1] TRUE
```

Each such label vector yields a K -partition of the input data set, that is, its grouping into K disjoint subsets – every element in \mathbf{X} belongs to one (and only one) cluster.

2.1.2 K-Partitions

As it is beneficial for us to be *gently* exposed to as many easy mathematical formalisms as possible, let us use some set-theoretic notation to come up with a rigorous definition of a dataset grouping. Formally, *clustering* aims to find a *special kind* of a K -partition of the in-

put dataset, which – without loss of generality – we can identify with the set $\{1, 2, \dots, n\}$ of indexes of observations (rows) in \mathbf{X} .

Definition 2.1 We say that $\mathcal{C} = \{C_1, C_2, \dots, C_K\}$ forms a K -partition of the set $\{1, 2, \dots, n\}$, whenever:

- $\bigcup_{k=1}^K C_k = C_1 \cup C_2 \cup \dots \cup C_K = \{1, 2, \dots, n\}$ (the union of all clusters is the whole set; every point is assigned to some cluster; no point is neglected),
- $C_k \cap C_l = \emptyset$ for all $k \neq l$ (clusters are pairwise disjoint; their intersection is empty; they share no common elements),
- $C_k \neq \emptyset$ for all k (each cluster is nonempty).

In the label vector representation, we assume that $y_i = k$ iff $i \in C_k$, i.e., the i -th point is assigned the k -th label if and only if it belongs to the k -th cluster. It is easily seen that each label vector yields a grouping of $\{1, 2, \dots, n\}$ that obviously fulfil the first two conditions in the above definition, and that the third property is fulfilled by assuming that $(\forall k \in \{1, 2, \dots, K\}) (\exists i \in \{1, 2, \dots, n\}) y_i = k$ (read: for every k there exists i such that...), i.e., each label occurs at least once in \mathbf{y} .

Remark 2.2 The benefit of using mathematical notation is that we are super-precise, efficient, and universal. Natural tongue is vague and leaves much room for interpretation. We used so many words to explain the above concepts. As soon as we further develop our skills of speaking the language of mathematics (note that we've started with basic phrases such as Good Afternoon, I am Sorry, and Thank You), it'll be more natural for us to just write:

"Let $\mathcal{C} = \{C_1, \dots, C_K\}$ s.t. $\bigcup_{k=1}^K C_k = \{1, \dots, n\}$, $(\forall k \neq l) C_k \cap C_l = \emptyset$, $C_k \neq \emptyset$ "

without all the talking. This is what more advanced books related to machine learning and statistics do. We'll get to that. By the way, thanks to this it's relatively easy to read maths papers (especially those from the 1950s) written in other languages (such as Russian, German or French).

In theory, the number of possible K -partitions of a set with n elements is given by the Stirling number of the second kind:

$$\left\{ \begin{matrix} n \\ K \end{matrix} \right\} = \sum_{j=0}^K \frac{(-1)^j (K-j)^n}{j!(K-j)!} = \frac{K^n}{K!} - \frac{(K-1)^n}{(K-1)!} + \frac{(K-2)^n}{2(K-2)!} - \frac{(K-3)^n}{6(K-3)!} + \dots,$$

where $n! = 1 \cdot 2 \cdot 3 \cdots \cdot n$ with $0! = 1$ (n -factorial). In particular, already $\left\{ \begin{matrix} n \\ 2 \end{matrix} \right\} = 2^{n-1} - 1$ and $\left\{ \begin{matrix} n \\ K \end{matrix} \right\} \simeq K^n / K!$ for large n – that is a lot.

Table 2.1: The number of possible partitions of a dataset with n elements to K clusters, denoted $\left\{ \begin{matrix} n \\ K \end{matrix} \right\}$, grows rapidly as n increases

n	$\left\{ \begin{matrix} n \\ 2 \end{matrix} \right\}$	$\left\{ \begin{matrix} n \\ 3 \end{matrix} \right\}$	$\left\{ \begin{matrix} n \\ 4 \end{matrix} \right\}$	$\left\{ \begin{matrix} n \\ 5 \end{matrix} \right\}$	$\left\{ \begin{matrix} n \\ 6 \end{matrix} \right\}$	$\left\{ \begin{matrix} n \\ 7 \end{matrix} \right\}$	$\left\{ \begin{matrix} n \\ 8 \end{matrix} \right\}$	$\left\{ \begin{matrix} n \\ 9 \end{matrix} \right\}$	$\left\{ \begin{matrix} n \\ 10 \end{matrix} \right\}$
2	1								
3		3							
4			1		6		1		

n	$\{ \frac{n}{2} \}$	$\{ \frac{n}{3} \}$	$\{ \frac{n}{4} \}$	$\{ \frac{n}{5} \}$	$\{ \frac{n}{6} \}$	$\{ \frac{n}{7} \}$	$\{ \frac{n}{8} \}$	$\{ \frac{n}{9} \}$	$\{ \frac{n}{10} \}$
5	15	25	10	1					
6	31	90	65	15	1				
7	63	301	350	140	21	1			
8	127	966	1701	1050	266	28	1		
9	255	3025	7770	6951	2646	462	36	1	
10	511	9330	34105	42525	22827	5880	750	45	1

2.1.3 “Interesting” Partitions

We are not just interested in “any” partition, because there are simply way too plentiful (compare Table 2.1; some of them will certainly be more meaningful or valuable than others. However, even one of the most famous ML textbooks provides us with only a vague hint of what we might be looking for:

Definition 2.2 (or, rather, a “definition”). *Clustering concerns “segmenting a collection of objects into subsets so that those within each cluster are more **closely related** to one another than objects assigned to different clusters”* (Hastie et al. 2017).

TODO - change to: “that points within a cluster are close together, while the clusters themselves are far apart”* (Rao 1964)

It is not uncommon (see, e.g., (Estivill-Castro 2002)) to equate the general definition of data clustering problems with... the particular outputs generated by specific clustering algorithms. In some sense, that sounds fair – clustering is both art and science (compare (Luxburg et al. 2012)). From this perspective, we might be interested in identifying the two main types of clustering algorithms:

- *parametric* (model-based) – find clusters of specific shapes or following specific multidimensional probability distributions, e.g., K-means, expectation-maximisation for Gaussian mixtures (EM), average linkage agglomerative clustering;
- *nonparametric* (model-free) – identify high-density or well-separable regions, perhaps in the presence of noise points, e.g., single linkage agglomerative clustering, Genie, (H)DBSCAN, BIRCH.

In this chapter we’ll take a look at (*agglomerative*) *hierarchical clustering* algorithms which build a cluster structure *from the bottom*, i.e., by merging small points groups to form larger ones. In Chapter 10, we’ll discuss the K-means algorithm, which seeks “good” cluster prototypes.

2.2 Euclidean Distance

In order to build our first clustering algorithm, we need to refer to the notion of a *distance*, which quantifies the extent to which two observations in a dataset (two rows in \mathbf{X}) are different from or similar to each other.

In this chapter we will be dealing with the most natural, “school” straight-line distance, called the Euclidean metric. It works as if we were measuring how two points are far away from each other with a ruler.

Given two p -tuples $\mathbf{u} = (u_1, \dots, u_p)$ and $\mathbf{v} = (v_1, \dots, v_p)$, the *Euclidean metric* is a function d such that:

$$d(\mathbf{u}, \mathbf{v}) = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2 + \cdots + (u_p - v_p)^2} = \sqrt{\sum_{i=1}^p (u_i - v_i)^2}.$$

The Euclidean metric will often also be denoted with $\|\mathbf{u} - \mathbf{v}\|$.

Remark 2.3 \sum (Greek capital letter *Sigma*) denotes summation. Read “ $\sum_{i=1}^p z_i$ ” as “the sum of z_i for i from 1 to p ”; this is just a convenient shorthand for $z_1 + z_2 + \cdots + z_p$.

Exercise 2.1 Consider the following $\mathbf{X} \in \mathbb{R}^{3 \times 2}$:

$$\mathbf{X} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ \frac{1}{4} & 1 \end{bmatrix}$$

Calculate (by hand) $d(\mathbf{x}_{1,.}, \mathbf{x}_{2,.})$, $d(\mathbf{x}_{1,.}, \mathbf{x}_{3,.})$, $d(\mathbf{x}_{2,.}, \mathbf{x}_{3,.})$, $d(\mathbf{x}_{1,.}, \mathbf{x}_{1,.})$, and $d(\mathbf{x}_{2,.}, \mathbf{x}_{1,.})$.

In order to compute all pairwise distances between the rows in a matrix with R, we can call the `dist()` function (see also Figure 2.1 for a graphical illustration):

```
X <- rbind(
  c(0,      0),
  c(1,      0),
  c(0.25,   1)
)
print(X)

##      [,1] [,2]
## [1,] 0.00   0
## [2,] 1.00   0
## [3,] 0.25   1

dist(X)

##          1         2
## 2 1.0000
## 3 1.0308 1.2500
```

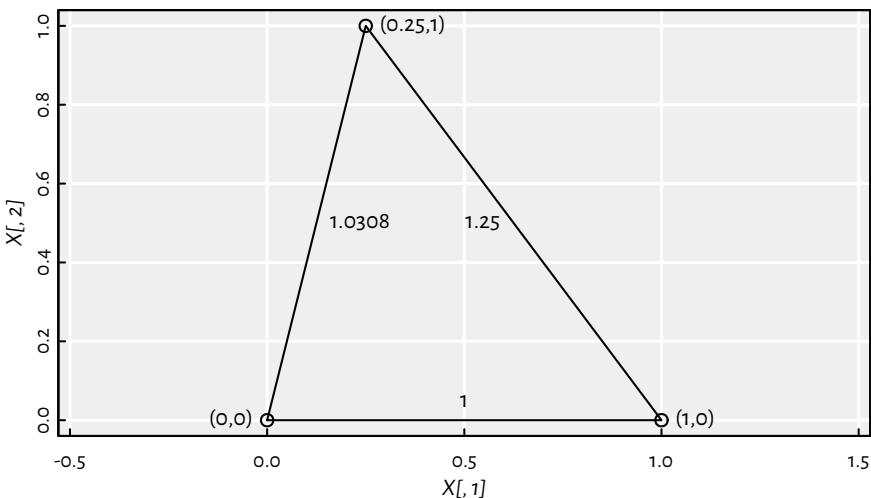


Figure 2.1: Example points and Euclidean distances between them

Note that only 3 values were reported. This is due to the fact that:

- the distance between a point and itself is always 0, i.e., $d(\mathbf{u}, \mathbf{u}) = 0$;
- the order of the points (from-to or to-from) doesn't matter, we have $d(\mathbf{u}, \mathbf{v}) = d(\mathbf{v}, \mathbf{u})$ (this is called *symmetry*).

If we want the full distance matrix, we can call:

```
as.matrix(dist(X))

##      1     2     3
## 1 0.0000 1.00 1.0308
## 2 1.0000 0.00 1.2500
## 3 1.0308 1.25 0.0000
```

Note the zeros on the main diagonal and that the matrix is symmetric around it.

Remark 2.4 Overall, there are $n(n-1)/2$ “interesting” pairwise distances for a dataset of n points. In R, these are stored using *numeric* type, each being an 8-byte floating point number. Therefore, by calling *dist()* on matrices with too many rows we can easily run out of memory; already for $n = 100,000$ we need ca. 40 GB of RAM.

Remark 2.5 Later we will mention that there are many possible distances, allowing to measure the similarity of points not only in \mathbb{R}^p , but also character strings, protein sequences, film ratings, etc.; there is even an encyclopedia of distances (see (Deza & Deza 2014))!

2.3 Hierarchical Clustering at a Glance

Hierarchical methods generate a whole hierarchy of *nested* partitions. A K -partition for any K can be extracted later at any time. In the case of *agglomerative* hierarchical algorithms:

- at the lowest level of the hierarchy, each point belongs to its own cluster (there are n singletons);
- at the highest level of the hierarchy, there is one cluster that embraces all the points;
- moving from the i -th to the $(i + 1)$ -th level, we select (somehow; see below) a pair of clusters to be merged.

Let's consider the Sustainable Society Indices Dataset (see Section F.1) that measures the Human, Environmental, and Economic Wellbeing in each country on the scale [0, 10].

```
ssi <- read.csv("datasets/ssi_2016_dimensions.csv",
                 comment.char="#")
head(ssi)

##      Country HumanWellbeing EnvironmentalWellbeing EconomicWellbeing
## 1    Albania        8.1162            5.3961        2.6317
## 2    Algeria        6.5283            3.5698        4.6622
## 3    Angola         4.1749            6.4411        2.8579
## 4 Argentina        6.8385            4.0929        5.7379
## 5 Armenia          7.6110            4.0443        3.1948
## 6 Australia         8.0647            2.4262        7.5716
```

In order to better understand how the algorithms of interest work, we'll restrict ourselves a smaller sample of countries, say, to 37 members of the OECD:

```
oecd <- c("Australia", "Austria", "Belgium", "Canada", "Chile", "Colombia",
         "Czech Republic", "Denmark", "Estonia", "Finland", "France", "Germany",
         "Greece", "Hungary", "Iceland", "Ireland", "Israel", "Italy", "Japan",
         "Korea, South", "Latvia", "Lithuania", "Luxembourg", "Mexico",
         "Netherlands", "New Zealand", "Norway", "Poland", "Portugal",
         "Slovak Republic", "Slovenia", "Spain", "Sweden", "Switzerland",
         "Turkey", "United Kingdom", "United States")
ssi <- ssi[ssi[, "Country"] %in% oecd, ]
X <- as.matrix(ssi[, -1]) # everything except the Country column
dimnames(X)[[1]] <- ssi[, 1] # set row names
head(X)

##             HumanWellbeing EnvironmentalWellbeing EconomicWellbeing
## Australia           8.0647            2.4262        7.5716
## Austria            8.4803            4.5422        5.6352
```

```

## Belgium      8.6593      2.6583      4.7965
## Canada      8.2536      2.4219      4.2430
## Chile       6.8764      4.3150      5.1530
## Colombia    5.9280      5.9480      4.1480

dim(X) # n and p

## [1] 37  3

```

Hence, we have $\mathbf{X} \in \mathbb{R}^{37 \times 3}$. Note that the matrix in R has row and column names set for greater readability.

The most basic implementation of a few agglomerative hierarchical clustering algorithms is provided by the `stats::hclust()` function, which works on a pairwise distance matrix. By default, the method computes the so-called *complete* linkage. We'll explain what that means later, let's just simply use it in a "black-box" fashion now, i.e., without going into details:

```
# Euclidean distances between all pairs of points:  
D <- dist(X)  
h <- hclust(D) # method="complete"  
print(h)
```

```

## Call:
## hclust(d = D)
##
## Cluster method : complete
## Distance       : euclidean
## Number of objects: 37

```

Remark 2.6 `stats::hclust()` refers to the `hclust()` function in the `stats` package. This is a package that comes with R, therefore there should be no need to load it explicitly. Otherwise, we would have to use the complete descriptor (with the `stats::` prefix) or call `library("stats")` to attach the package's namespace.

The obtained hierarchy can be *cut* at an arbitrary level by applying the `cutree()` function. Let's use $K = 3$:

```
K <- 3  
y <- cutree(h, K) # extract the 3-partition
```

A picture is worth a thousand words, therefore let's generate a map with the `rworldmap` package (see Figure 2.2):

```
library("rworldmap") # see the package's manual for details  
mapdata <- data.frame(Country=dimnames(X)[[1]], Cluster=y)  
mapdata <- joinCountryData2Map(mapdata, joinCode="NAME",  
    nameJoinColumn="Country")
```

```
mapCountryData(mapdata, nameColumnToPlot="Cluster",
  catMethod="categorical", missingCountryCol="white",
  colourPalette=palette.colors(K, "Okabe-Ito"),
  mapTitle="", nameColumnToHatch="Cluster")
```

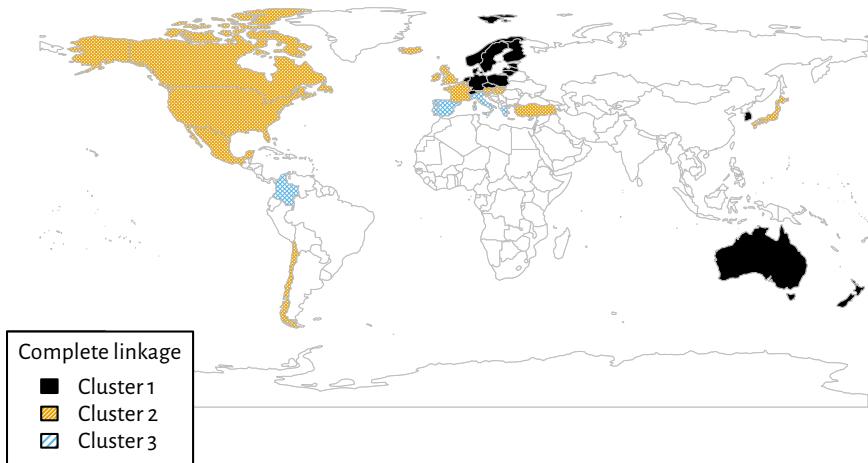


Figure 2.2: OECD countries grouped w.r.t. the SSI dimensions

We have so many questions now! First of all, why the countries were grouped in such a way? Well, this is because the algorithm is programmed to perform certain computational steps that lead to this particular solution. We'll discuss them in very detail soon.

A better question is: what characterises the countries in each cluster? To explore possible answers to this question, we can compute, e.g., the average indicators in each group:

```
aggregate(as.data.frame(X), list(Cluster=y), mean)
```

##	Cluster	HumanWellbeing	EnvironmentalWellbeing	EconomicWellbeing
## 1	1	8.4195	3.5158	7.2490
## 2	2	8.0008	3.9414	5.0226
## 3	3	7.4919	5.2490	3.7474

Therefore, on average, the countries in the 1st cluster have higher Economic Wellbeing scores than the countries in the 2nd cluster, and these outperform those from the 3rd cluster. On the other hand, the 3rd cluster countries score much more highly on the Environmental Wellbeing dimension. This summary, however, is valid only if the countries in each cluster are score similarly on each scale (are homogeneous).

Luckily, we have only 3 dimensions, therefore we can visualise different two-dimensional projections of this dataset. Figure 2.3 depicts a slightly “tuned up” (we've added the ISO

3-letter country codes) version of a scatter plot matrix that we would normally obtain by calling `pairs(X, col=y, pch=y)`.

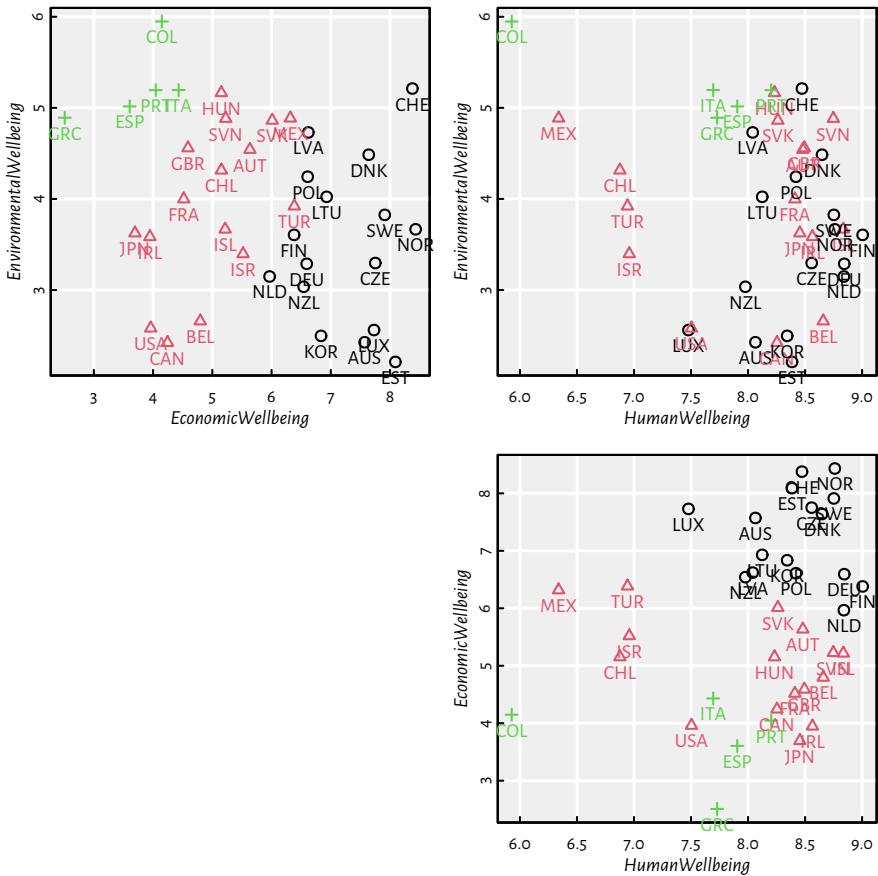


Figure 2.3: Scatterplot matrix for the SSI dimensions with the 3-partition generated by complete linkage

It seems that Economic Wellbeing is the deciding factor for distinguishing between these countries. We see on the Economic vs. Environmental Wellbeing plot that the clusters are nicely separated. Human Wellbeing, on the other hand, is pretty “mixed up” across the clusters.

2.4 Cluster Dendrograms

A *dendrogram* can provide us with some insight into the underlying data structure as well as with some hints about how many clusters (K) should extract.

```
# Plotting of dendrogram: essentially plot(hclust(dist(X)))
plot(h, xlab=NA, main=NA); box()
```

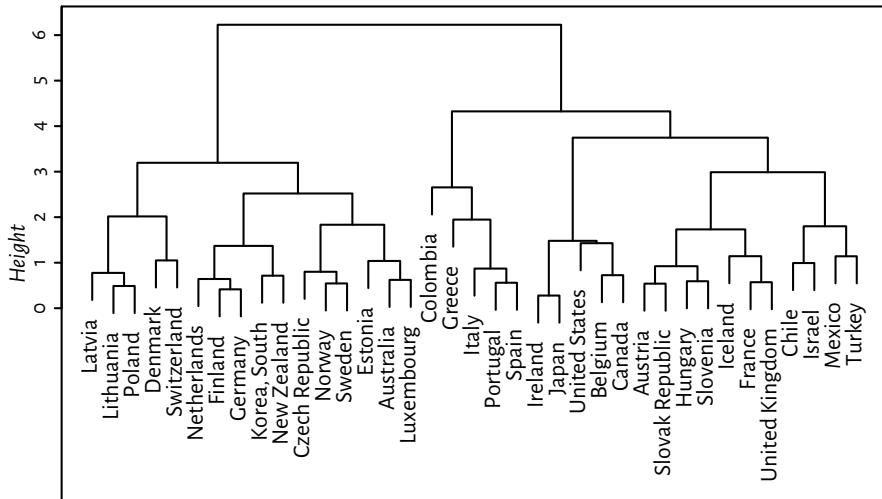


Figure 2.4: Cluster dendrogram generated by complete linkage

Figure 2.4 depicts the dendrogram corresponding to the performed agglomerative hierarchical clustering.

A dendrogram is a tree-like diagram whose nodes represent clusters at different stages of the algorithm's run:

- The *leaves* of the tree, i.e., the nodes at the very bottom, stand for the individual data points (recall that in agglomerative clustering we start with n singletons). In our case, these are the 37 OECD countries.
- Each two-teeth “fork” represents two clusters being merged at some iteration. For example, at some stages we have clusters like: {Italy} and {Portugal, Spain} which are merged so that we obtain {Italy, Portugal, Spain}. Moreover, the union of {Chile, Israel} and {Mexico, Turkey} gives {Chile, Israel, Mexico, Turkey}.
- At the tree top (“root” – the tree is plotted upside down) we have a state where all the clusters have been merged with each other (there is one cluster consisting of all the points).

- The tree can be “cut” at any level so as to obtain a clustering to any number of sub-groups. In particular, if we cut the tree at height ≈ 4 , we will get a 3-partition from Figure 2.3.
-

2.5 Linkage Functions

Let's formalise the agglomerative hierarchical clustering process. Initially (“iteration 0”), each of the n points is a member of its own cluster. Let $C^{(0)}$ denote this n -partition:

$$C^{(0)} = \{C_1^{(0)}, C_2^{(0)}, \dots, C_n^{(0)}\},$$

where $C_i^{(0)}$ is the cluster that consists of the i -th point only, i.e., $C_i^{(0)} = \{i\}$.

While an agglomerative hierarchical clustering algorithm is being computed, there are $n - k$ clusters at the k -th step of the procedure,

$$C^{(k)} = \{C_1^{(k)}, C_2^{(k)}, \dots, C_{u-1}^{(k)}, C_u^{(k)}, C_{u+1}^{(k)}, \dots, C_{v-1}^{(k)}, C_v^{(k)}, C_{v+1}^{(k)}, \dots, C_{n-k}^{(k)}\},$$

where $k = 0, 1, \dots, n - 1$.

When proceeding from step k to $k + 1$, we determine two clusters $C_u^{(k)}$ and $C_v^{(k)}$, $u < v$, to be *merged* so that the partition at the higher level is of the form:

$$C^{(k+1)} = \{C_1^{(k)}, C_2^{(k)}, \dots, C_{u-1}^{(k)}, C_u^{(k)} \cup C_v^{(k)}, C_{u+1}^{(k)}, \dots, C_{v-1}^{(k)}, C_{v+1}^{(k)}, \dots, C_{n-k}^{(k)}\}.$$

It is evident that this way we will arrive at a 1-partition (a single boring cluster consisting of all the points), $C^{(n-1)} = \{C_1^{(n-1)}\}$ with $C_1^{(n-1)} = \{1, 2, \dots, n\}$.

Example 2.1 Restricting ourselves to the first 5 countries on the left in Figure 2.4, we have:

$$\begin{aligned} C^{(0)} &= \{\{\text{Latvia}\}, \{\text{Lithuania}\}, \{\text{Poland}\}, \{\text{Denmark}\}, \{\text{Switzerland}\}\}, & (C_2^{(0)} \cup C_3^{(0)}) \\ C^{(1)} &= \{\{\text{Latvia}\}, \{\text{Lithuania, Poland}\}, \{\text{Denmark}\}, \{\text{Switzerland}\}\}, & (C_1^{(1)} \cup C_2^{(1)}) \\ C^{(2)} &= \{\{\text{Latvia, Lithuania, Poland}\}, \{\text{Denmark}\}, \{\text{Switzerland}\}\}, & (C_2^{(2)} \cup C_3^{(2)}) \\ C^{(3)} &= \{\{\text{Latvia, Lithuania, Poland}\}, \{\text{Denmark, Switzerland}\}\}, & (C_1^{(3)} \cup C_2^{(3)}) \\ C^{(4)} &= \{\{\text{Latvia, Lithuania, Poland, Denmark, Switzerland}\}\}. \end{aligned}$$

There is one component missing – how to determine the pair of clusters $C_u^{(k)}$ and $C_v^{(k)}$ to be merged at the k -th iteration? For classic agglomerative clustering algorithms, we choose two clusters that are *closest to each other* – they minimise the inter-cluster distance.

Formally, we write the condition upon our decision is based as:

$$\min_{u,v:u < v} d^*(C_u^{(k)}, C_v^{(k)}),$$

(read: “find u and v such that $u < v$ and for which $d^*(C_u^{(k)}, C_v^{(k)})$ is the smallest”), where $d^*(C_u^{(k)}, C_v^{(k)})$ is the *distance* between two clusters $C_u^{(k)}$ and $C_v^{(k)}$.

Note that usually we consider only the distances between *individual points*, not sets of points, at least this is what we’ve done in Section 2.2. Therefore, what we need to do is to introduce d^* as a suitable extension of the Euclidean metric d .

First, we will assume that $d^*(\{i\}, \{j\}) = d(\mathbf{x}_{i,\cdot}, \mathbf{x}_{j,\cdot})$, i.e., the distance between singleton clusters is the same as the distance between the corresponding points (rows in \mathbf{X}) themselves.

As far as more populous point groups are concerned, there are many popular choices of d^* (which in the context of hierarchical clustering we call *linkage functions*):

- single linkage – the distance between two clusters is equal to the distance between the closest pair of points:

$$d_S^*(C_u^{(k)}, C_v^{(k)}) = \min_{i \in C_u^{(k)}, j \in C_v^{(k)}} d(\mathbf{x}_{i,\cdot}, \mathbf{x}_{j,\cdot}),$$

- complete linkage – we choose the pair of points farthest away:

$$d_C^*(C_u^{(k)}, C_v^{(k)}) = \max_{i \in C_u^{(k)}, j \in C_v^{(k)}} d(\mathbf{x}_{i,\cdot}, \mathbf{x}_{j,\cdot}),$$

- average linkage – we consider the average distance:

$$d_A^*(C_u^{(k)}, C_v^{(k)}) = \frac{1}{|C_u^{(k)}||C_v^{(k)}|} \sum_{i \in C_u^{(k)}} \sum_{j \in C_v^{(k)}} d(\mathbf{x}_{i,\cdot}, \mathbf{x}_{j,\cdot})$$

where $|C_u^{(k)}|$ and $|C_v^{(k)}|$ denote the total number of points in these two clusters.

An illustration of the way different linkages are computed is given in Figure 2.5.

Exercise 2.2 Here are the pairwise distances between the first 5 countries (w.r.t. the SSI dimensions) on the left in Figure 2.4:

```
as.matrix(
  dist(X[c("Latvia", "Lithuania", "Poland", "Denmark", "Switzerland"),])
)
```

	Latvia	Lithuania	Poland	Denmark	Switzerland
## Latvia	0.00000	0.77607	0.61706	1.21210	1.8724
## Lithuania	0.77607	0.00000	0.48823	0.99683	1.9058
## Poland	0.61706	0.48823	0.00000	1.08317	2.0167
## Denmark	1.21210	0.99683	1.08317	0.00000	1.0502
## Switzerland	1.87241	1.90577	2.01673	1.05017	0.0000

Calculate (by hand):

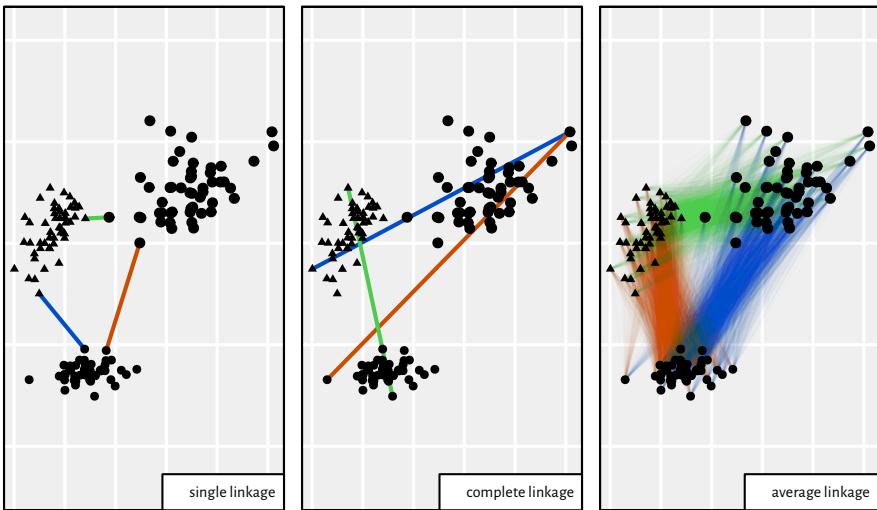


Figure 2.5: In single linkage, we find the closest pair of points; in complete linkage, we seek the pair furthest away from each other; in average linkage, we determine the arithmetic mean of all inter-cluster pairwise distances

- $d_S (\{\text{Latvia}\}, \{\text{Lithuania, Poland}\})$,
- $d_C (\{\text{Latvia}\}, \{\text{Lithuania, Poland}\})$,
- $d_A (\{\text{Latvia}\}, \{\text{Lithuania, Poland}\})$,
- $d_S (\{\text{Latvia, Lithuania, Poland}\}, \{\text{Denmark, Switzerland}\})$,
- $d_C (\{\text{Latvia, Lithuania, Poland}\}, \{\text{Denmark, Switzerland}\})$,
- $d_A (\{\text{Latvia, Lithuania, Poland}\}, \{\text{Denmark, Switzerland}\})$.

We can perform agglomerative clustering of our dataset with respect to different linkages by setting the `method` argument in the `hclust()` function:

```
D <- dist(X)
hs <- hclust(D, method="single")
hc <- hclust(D, method="complete") # the default
ha <- hclust(D, method="average")
```

Figure 2.6 depicts the dendograms corresponding to single and average linkages, which we can plot by calling `plot(hs)` and `plot(ha)`.

Remark 2.7 *The height on the Y axis in each dendrogram's plot represents the distance between the merged clusters (as measured by the assumed linkage function).*

Both complete and average linkages give us nicely “balanced” cluster hierarchy (average

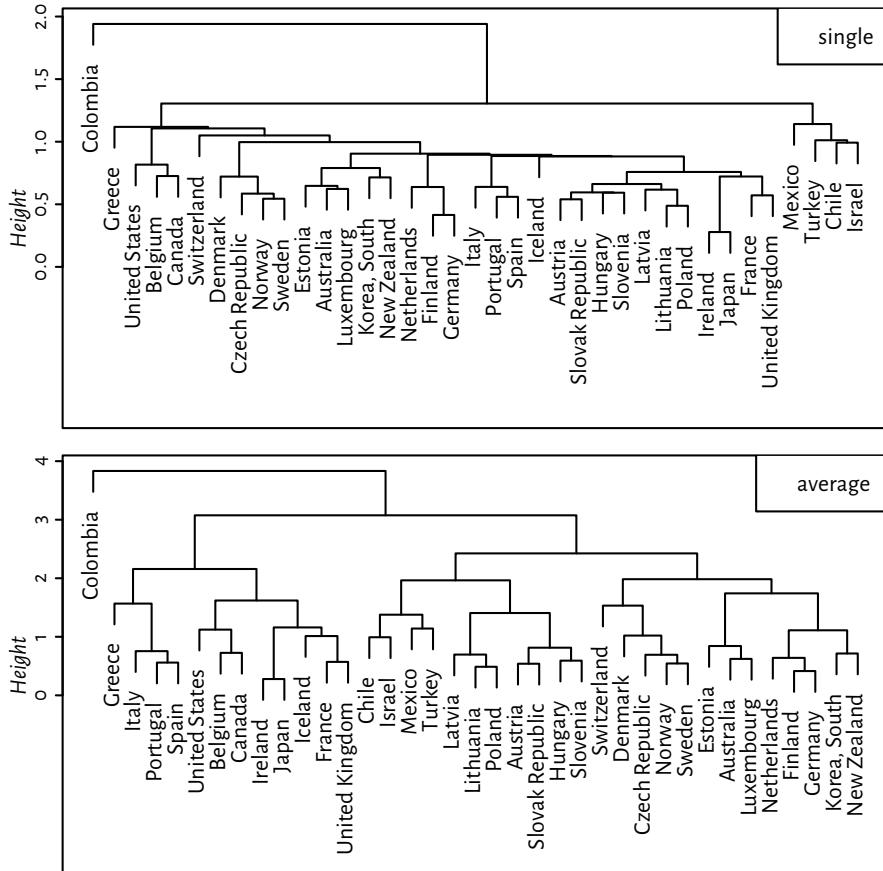


Figure 2.6: Cluster dendrograms generated by single and average linkages

linkage identified Colombia as an “outlier” though). The obtained groupings are similar to some extent, yet, they are not identical – both algorithms are based on different heuristics. However, single linkage tends to produce somehow “ugly” trees – it’s often the case that it’ll produce few large clusters and a lot of small ones.

The reader is encouraged to perform a similar “debugging” of the obtained clusterings (say, for the 3- and 4-partitions) as we performed above (compute the average features in each cluster, draw pairwise scatterplots).

2.6 Exercises

Exercise 2.3 Calculate the complete linkage agglomerative clustering algorithm by hand on the space of the 3 SSI dimensions for France, Ireland, Spain, Colombia, Austria, and Germany. For each iteration of the algorithm, print the members of the current partition and the distances between each pair of clusters.

Solution.

The pairwise distance matrix is:

```
X2 <- X[c("France", "Ireland", "Spain", "Colombia", "Austria", "Germany"),]
D2 <- dist(X2)
as.matrix(D2)
```

	France	Ireland	Spain	Colombia	Austria	Germany
## France	0.00000	0.72203	1.4567	3.1783	1.2450	2.2364
## Ireland	0.72203	0.00000	1.6126	3.5469	1.9434	2.6779
## Spain	1.45668	1.61261	0.0000	2.2527	2.1620	3.5767
## Colombia	3.17828	3.54690	2.2527	0.0000	3.2715	4.6430
## Austria	1.24497	1.94341	2.1620	3.2715	0.0000	1.6198
## Germany	2.23643	2.67788	3.5767	4.6430	1.6198	0.0000

$$\mathcal{C}^{(0)} = \{\{FRA\}, \{IRL\}, \{ESP\}, \{COL\}, \{AUT\}, \{DEU\}\}$$

Inter-cluster distances (complete linkage):

	{FRA}	{IRL}	{ESP}	{COL}	{AUT}	{DEU}
{FRA}	0.00	0.72	1.46	3.18	1.24	2.24
{IRL}	0.72	0.00	1.61	3.55	1.94	2.68
{ESP}	1.46	1.61	0.00	2.25	2.16	3.58
{COL}	3.18	3.55	2.25	0.00	3.27	4.64
{AUT}	1.24	1.94	2.16	3.27	0.00	1.62
{DEU}	2.24	2.68	3.58	4.64	1.62	0.00

Merging {IRL} and {FRA} (distance=0.72)

$$\mathcal{C}^{(1)} = \{\{FRA, IRL\}, \{ESP\}, \{COL\}, \{AUT\}, \{DEU\}\}$$

Inter-cluster distances (complete linkage):

	{FRA,IRL}	{ESP}	{COL}	{AUT}	{DEU}
{FRA,IRL}	0.00	1.61	3.55	1.94	2.68
{ESP}	1.61	0.00	2.25	2.16	3.58

	{FRA,IRL}	{ESP}	{COL}	{AUT}	{DEU}
{COL}	3.55	2.25	0.00	3.27	4.64
{AUT}	1.94	2.16	3.27	0.00	1.62
{DEU}	2.68	3.58	4.64	1.62	0.00

Merging {ESP} and {FRA,IRL} (distance=1.61)

$$\mathcal{C}^{(2)} = \{\{FRA, IRL, ESP\}, \{COL\}, \{AUT\}, \{DEU\}\}$$

Inter-cluster distances (complete linkage):

	{FRA,IRL,ESP}	{COL}	{AUT}	{DEU}
{FRA,IRL,ESP}	0.00	3.55	2.16	3.58
{COL}	3.55	0.00	3.27	4.64
{AUT}	2.16	3.27	0.00	1.62
{DEU}	3.58	4.64	1.62	0.00

Merging {DEU} and {AUT} (distance=1.62)

$$\mathcal{C}^{(3)} = \{\{FRA, IRL, ESP\}, \{COL\}, \{AUT, DEU\}\}$$

Inter-cluster distances (complete linkage):

	{FRA,IRL,ESP}	{COL}	{AUT,DEU}
{FRA,IRL,ESP}	0.00	3.55	3.58
{COL}	3.55	0.00	4.64
{AUT,DEU}	3.58	4.64	0.00

Merging {COL} and {FRA,IRL,ESP} (distance=3.55)

$$\mathcal{C}^{(4)} = \{\{FRA, IRL, ESP, COL\}, \{AUT, DEU\}\}$$

Inter-cluster distances (complete linkage):

	{FRA,IRL,ESP,COL}	{AUT,DEU}
{FRA,IRL,ESP,COL}	0.00	4.64
{AUT,DEU}	4.64	0.00

Merging {AUT,DEU} and {FRA,IRL,ESP,COL} (distance=4.64)

$$\mathcal{C}^{(5)} = \{\{FRA, IRL, ESP, COL, AUT, DEU\}\}$$

■

Exercise 2.4 As in the previous exercise, but apply the average linkage.

Solution.

$C^{(0)}$ is identical to the one in the previous exercise and the first merge decision is the same.

$$C^{(1)} = \{\{FRA, IRL\}, \{ESP\}, \{COL\}, \{AUT\}, \{DEU\}\}$$

Inter-cluster distances (average linkage):

	{FRA,IRL}	{ESP}	{COL}	{AUT}	{DEU}
{FRA,IRL}	0.00	1.53	3.36	1.59	2.46
{ESP}	1.53	0.00	2.25	2.16	3.58
{COL}	3.36	2.25	0.00	3.27	4.64
{AUT}	1.59	2.16	3.27	0.00	1.62
{DEU}	2.46	3.58	4.64	1.62	0.00

Merging {ESP} and {FRA,IRL} (distance=1.53)

$$C^{(2)} = \{\{FRA, IRL, ESP\}, \{COL\}, \{AUT\}, \{DEU\}\}$$

Inter-cluster distances (average linkage):

	{FRA,IRL,ESP}	{COL}	{AUT}	{DEU}
{FRA,IRL,ESP}	0.00	2.99	1.78	2.83
{COL}	2.99	0.00	3.27	4.64
{AUT}	1.78	3.27	0.00	1.62
{DEU}	2.83	4.64	1.62	0.00

Merging {DEU} and {AUT} (distance=1.62)

$$C^{(3)} = \{\{FRA, IRL, ESP\}, \{COL\}, \{AUT, DEU\}\}$$

Inter-cluster distances (average linkage):

	{FRA,IRL,ESP}	{COL}	{AUT,DEU}
{FRA,IRL,ESP}	0.00	2.99	2.31
{COL}	2.99	0.00	3.96
{AUT,DEU}	2.31	3.96	0.00

Merging {AUT,DEU} and {FRA,IRL,ESP} (distance=2.31)

$$C^{(4)} = \{\{FRA, IRL, ESP, AUT, DEU\}, \{COL\}\}$$

Inter-cluster distances (average linkage):

	{FRA,IRL,ESP,AUT,DEU}	{COL}
{FRA,IRL,ESP,AUT,DEU}	0.00	3.38
{COL}	3.38	0.00

Merging {COL} and {FRA,IRL,ESP,AUT,DEU} (distance=3.38)

$$\mathcal{C}^{(5)} = \{\{FRA, IRL, ESP, COL, AUT, DEU\}\}$$

■

Exercise 2.5 As in the previous exercise, but apply the single linkage.

Solution.

The first is identical to the above one.

$$\mathcal{C}^{(1)} = \{\{FRA, IRL\}, \{ESP\}, \{COL\}, \{AUT\}, \{DEU\}\}$$

Inter-cluster distances (single linkage):

	{FRA,IRL}	{ESP}	{COL}	{AUT}	{DEU}
{FRA,IRL}	0.00	1.46	3.18	1.24	2.24
{ESP}	1.46	0.00	2.25	2.16	3.58
{COL}	3.18	2.25	0.00	3.27	4.64
{AUT}	1.24	2.16	3.27	0.00	1.62
{DEU}	2.24	3.58	4.64	1.62	0.00

Merging {AUT} and {FRA,IRL} (distance=1.24)

$$\mathcal{C}^{(2)} = \{\{FRA, IRL, AUT\}, \{ESP\}, \{COL\}, \{DEU\}\}$$

Inter-cluster distances (single linkage):

	{FRA,IRL,AUT}	{ESP}	{COL}	{DEU}
{FRA,IRL,AUT}	0.00	1.46	3.18	1.62
{ESP}	1.46	0.00	2.25	3.58
{COL}	3.18	2.25	0.00	4.64
{DEU}	1.62	3.58	4.64	0.00

Merging {ESP} and {FRA,IRL,AUT} (distance=1.46)

$$\mathcal{C}^{(3)} = \{\{FRA, IRL, ESP, AUT\}, \{COL\}, \{DEU\}\}$$

Inter-cluster distances (single linkage):

	{FRA,IRL,ESP,AUT}	{COL}	{DEU}
{FRA,IRL,ESP,AUT}	0.00	2.25	1.62
{COL}	2.25	0.00	4.64
{DEU}	1.62	4.64	0.00

Merging {DEU} and {FRA,IRL,ESP,AUT} (distance=1.62)

$$\mathcal{C}^{(4)} = \{\{FRA, IRL, ESP, AUT, DEU\}, \{COL\}\}$$

Inter-cluster distances (single linkage):

	{FRA,IRL,ESP,AUT,DEU}	{COL}
{FRA,IRL,ESP,AUT,DEU}	0.00	2.25
{COL}	2.25	0.00

Merging {COL} and {FRA,IRL,ESP,AUT,DEU} (distance=2.25)

$$\mathcal{C}^{(5)} = \{\{FRA, IRL, ESP, COL, AUT, DEU\}\}$$

■

Exercise 2.6 Perform cluster analysis of the OECD countries based on the 7 SSI categories `ssi_2016_categories.csv`. Apply single, complete, and average linkage. Draw the corresponding dendograms. Compute the averages of all indicators in each point group for clusterings of different sizes. Draw scatter plots for different pairs of variables (e.g., by calling `pairs()`), where clusters are represented by points of different shapes and colours.

Exercise 2.7 Perform cluster analysis of the OECD countries based on the 21 SSI indicators `ssi_2016_indicators.csv`

Exercise 2.8 Perform cluster analysis for all the countries covered by the Sustainable Society Indices.

2.7 Remarks

Complete linkage dates back to Denmark in the year 1948; a similar idea (for different dissimilarity measures) was used in (Sørensen 1948) to “establish groups of equal amplitude in plant sociology based on similarity of species content”. Single linkage has been proposed by Polish mathematicians as “Wrocław taxonomy” (or the dendrite method) in 1951 (Florek et al. 1951). Average linkage, also known as UPGMA (unweighted pair group method with arithmetic mean), has been proposed in (Sokal & Michener 1958).

There are many other linkages, e.g., the Ward linkage minimises the variance of within-cluster distances (TODO: within cluster sum of squares, the same as k-means). Together with the linkages presented in this chapter it is generalised by the Lance–Williams formula (Lance & Williams 1967), see also (Müllner 2011).

Hierarchical clustering algorithms are fantastic, because they output a whole hierarchy of *nested* partitions – it's not that when we cut it at the k -th level we'll obtain something totally unrelated to what can we find at the l -th one ($k \neq l$). However, this comes at a price – these algorithms are quite slow for large datasets. In particular, the implementations included in the `stats::hclust()` function (which we have used above) have $O(n^3)$ time complexity. Therefore, in practice it's better to use function `fastcluster::hclust()` (Müllner 2013), which provides a drop-in replacement for the original routine.

Single linkage is the fastest – even though it has $O(n^2)$ worst-case time complexity in generic metric spaces, however, it can run much faster in real spaces of small dimensionality (e.g., when analysing spatial data), as it can be computed based on a minimum spanning tree of the pairwise distance graph.

Moreover, as we've said, there are $n(n - 1)/2$ unique pairwise distances between n points. Don't try calling `dist(hclust(X))` on large data matrices. `fastcluster::hclust.vector()` implements single and Ward (but not complete and average) linkages without the need of precomputing the whole distance matrix.

For a comprehensive overview of the efficient algorithms to compute the aforementioned linkages, see (Müllner 2011).

TODO

other distances/metrics

Note that all the indicators were on the same scale. We'll deal with different datasets in the next chapter

otherwise, we need to perform some feature weighting or calibration; e.g., standardisation

TODO: chapter 3 ????

Agglomerative clustering algorithms utilise the “bottom-up” approach

there are also divisive algorithms (top-down, start with the 1-partition, split clusters into smaller chunks), but they tend to be much more computationally intensive, see, Mueller's ITM, however

what we did above, we call a greedy strategy btw

There are modification to this scheme, e.g., the Genie algorithm (see package `genieclust`) (Gagolewski et al. 2016) merges clusters based on the closest pairs of points (just as single linkage), however under the constraint that the inequality of the cluster

size distribution cannot go beyond some threshold – this prevents leaving out very small clusters at higher levels of the hierarchy.

is the obtained clustering a good one? well, if it leads to interesting “discoveries” or something useful, it is; more discussion in Chapter 10

we mention genieclust and centroid-based linkages later

will we play with HDBSCAN*?

(Jain 2010) (Wierzchoń & Kłopotek 2018) (Blum et al. 2020)

Classification with Nearest Neighbours

TODO In this chapter, we will:

- solve a prediction task with k-nearest neighbour classifier
 - discuss performance metrics for binary classification: accuracy, precision, recall, and F-measure
 - introduce best practices of optimal classifier selection – in particular, explain the difference between training, validation, and test sets
-

3.1 Introduction

In the previous chapter, we were only given an input matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ representing n objects described by means of p numerical features.

In this chapter we are going to be interested in *supervised learning* tasks; we assume that with each $\mathbf{x}_{i,\cdot}$, ($i = 1, \dots, n$) we associate the desired output y_i :

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

More precisely, we'd now like to focus on *classification* – we assume that each y_i is of qualitative type (e.g., a label, a category, a colour).

The classes are usually *encoded* with consecutive integers, say, $1, 2, \dots, L$, where L is the total number of unique cases. Mathematically, we'll write that $y_i \in \{1, \dots, L\}$.

Remark 3.1 *factor* datatype in R (see Section C.7 for more details) gives a very convenient means to encode categorical data (such as \mathbf{y}):

```
y <- c("black", "green", "black", "red", "green", "green", "blue", "red")
(y <- factor(y, levels=c("black", "red", "green", "blue")))

## [1] black green black red  green green blue  red
## Levels: black red green blue
```

Internally, objects of type `factor` are represented as integer vectors with elements in $\{1, \dots, L\}$, where L is the number of possible levels. Labels, used to “decipher” the numeric codes, are stored separately and can always be replaced with other ones.

```
as.numeric(y) # 1st label, 3rd label, 1st label, and so forth
```

```
## [1] 1 3 1 2 3 3 4 2
```

```
length(levels(y)) # L = number of levels
```

```
## [1] 4
```

```
levels(y) # 1=black, 2=red, 3=green, 4=red
```

```
## [1] "black" "red"    "green"   "blue"
```

```
levels(y) <- c("k", "r", "g", "b") # re-encode
```

```
y
```

```
## [1] k g k r g g b r
```

```
## Levels: k r g b
```

In this chapter we'll assume $L = 2$, i.e., that each y_i can only take one of two possible values. Such a case is very common in practice; it has been even given a special name: *binary classification*. The two classes are traditionally denoted as 0 and 1, see Table 3.1 for some examples.

Table 3.1: Some examples of output labels in binary classification tasks

0	1
no	yes
false	true
failure	success
reject	accept
healthy	ill

Remark 3.2 0s and 1s are mathematically convenient, because instead of stating “if $y_i = 1$, then let $z = a$ and otherwise take $z = b$ ”, we can simply write “ $z = y_i a + (1 - y_i) b$ ” (see, e.g., the definition of cross-entropy in 8). Ah, those neat math tricks!

Figure 3.1 gives a scatter plot of an example synthetic two-dimensional dataset (i.e., $p = 2$) with the reference binary ys depicted using different plotting symbols. We see that the two classes overlap slightly – they cannot be separated by means of any simple boundary (e.g., a line).

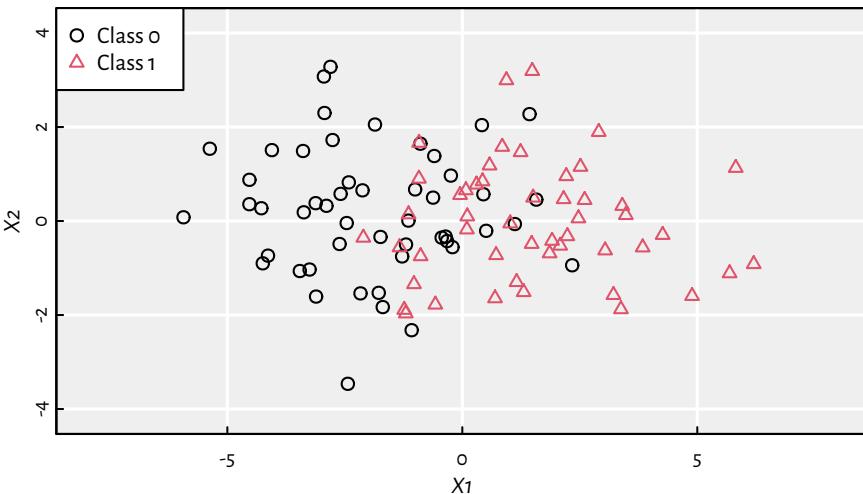


Figure 3.1: A synthetic 2D dataset where each point is assigned one of two distinct labels

3.2 K-Nearest Neighbours Classifier

Our main aim will be to “build” an algorithm that takes a previously unobserved input sequence \mathbf{x}' and which – based what we know already, i.e., on \mathbf{X} and \mathbf{y} generates a corresponding prediction, \hat{y}' (hopefully being equal to the true y' , whatever it is).

For instance, we may have a set of medical records, where we store patient data regarding their measurable health-related parameters like blood pressure, severity of different symptoms etc. Moreover, each patient is labelled – we know if they have been diagnosed a disease or not. But wait, there comes a new patient! We measure their blood pressure, as some questions, input the data into our algorithm and – based on the result – advise them to stay in bed.

In this chapter we will consider a very simple classifier, based on an idea which dates back to at least 1950s (see (Fix & Hodges 1951), (Fix & Hodges 1952)).

Rule. “If you don’t know what to do in a situation, just act like the people around you”

For some integer $K \geq 1$, the *K-Nearest Neighbour (K-NN) Classifier* proceeds as follows. To classify a new point \mathbf{x}' :

1. Find the K nearest neighbours of a given point \mathbf{x}' amongst the points in \mathbf{X} , i.e., points $\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_K}$, that are the closest to \mathbf{x}' w.r.t. the Euclidean distance:

- a. compute the Euclidean distances between each \mathbf{x}_{i_r} from the input set and \mathbf{x}' :

$$d_i = d(\mathbf{x}_{i_r}, \mathbf{x}'),$$

- b. order d_i 's in increasing order, and fetch the indices i_1, \dots, i_K which yield:

$$d_{i_1} \leq d_{i_2} \leq \dots \leq d_{i_K}.$$

2. Fetch the reference labels y_{i_1}, \dots, y_{i_K} corresponding to the K nearest neighbours.
3. Return their *mode* as a result, i.e., the most frequently occurring label (also called *majority vote*).

Let's illustrate how a K -NN classifier works on the above 2D synthetic dataset. First we consider $K = 1$, see Figure 3.2. Dark and light regions depict how new points would be classified (class 0 and 1, respectively). For instance, the point $(0, 2)$ lays in the “dark zone”, therefore we'd assign it label 0.

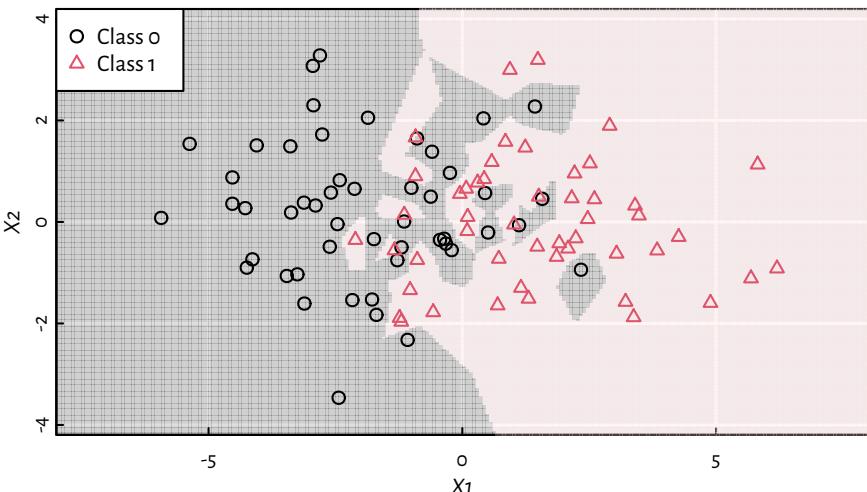


Figure 3.2: 1-NN class bounds for the 2D synthetic dataset

We see that 1-NN is “greedy” or “lazy” in the sense that we just locate the nearest point. Increasing K somehow smoothens the decision boundary (this makes it less “local” and more “global”). Figure 3.3 depicts the $K = 3$ case.

The 15-NN classifier, see Figure 3.4, does a quite good job with identifying a boundary between the two classes – the whole \mathbb{R}^2 space seems to be (more or less) split into two disconnected subregions; there are no “lakes” or “islands” or significant sizes.

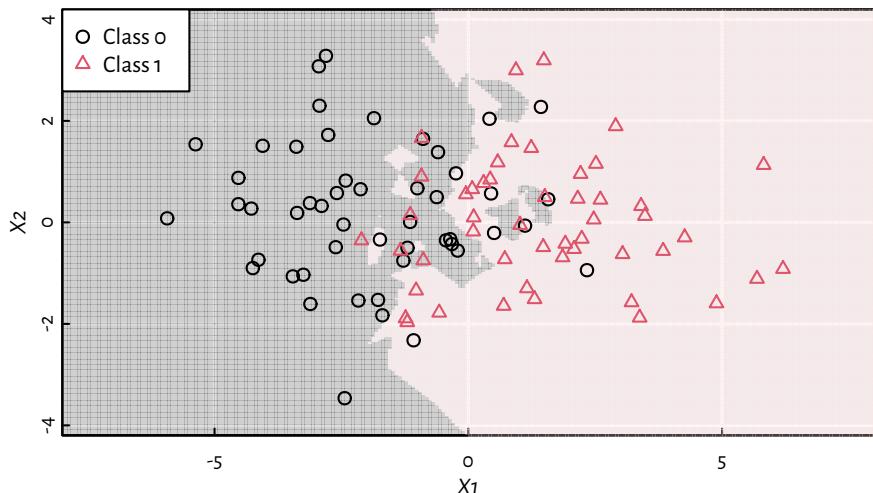


Figure 3.3: 3-NN class bounds the our 2D synthetic dataset

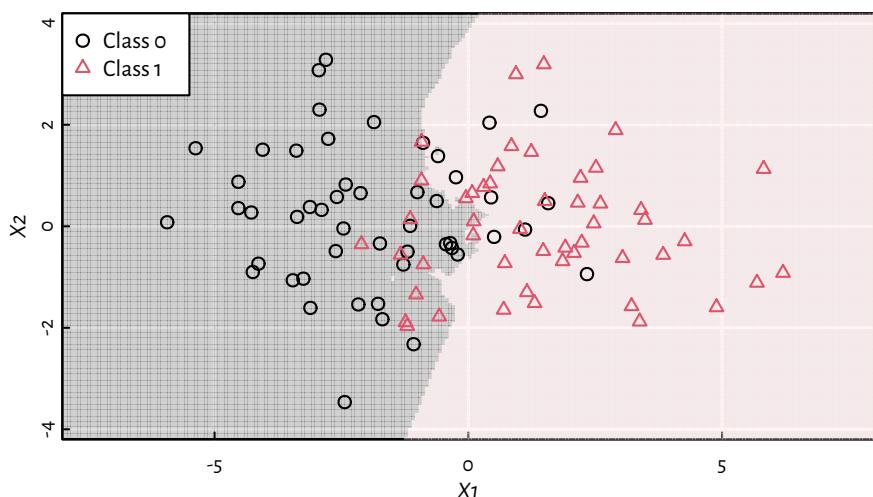


Figure 3.4: 15-NN class bounds the our 2D synthetic dataset

3.3 Example in R

As a real-world illustration, let's consider a subset of the Wine Quality dataset (see Appendix F.5 for more details) that features a sample of white wines and includes only 3 of their physicochemical characteristics:

```
wine_train <- read.csv("datasets/wine_train.csv")
head(wine_train)

##   chlorides density volatile.acidity bad
## 1      0.056  0.99680          0.40   1
## 2      0.040  0.99062          0.18   0
## 3      0.050  0.99830          0.24   0
## 4      0.022  0.98915          0.42   0
## 5      0.021  0.99021          0.25   0
## 6      0.025  0.99071          0.30   0

dim(wine_train) # number of rows, number of columns

## [1] 300    4
```

The last column determines whether experts claim that a given wine is of low quality (class 1, we don't want that if we make a party) or not (class 0). We will use it as a response variable in a wine classification task.

Remark 3.3 Note that above \mathbf{X} and \mathbf{y} are stored together as a single object, $[\mathbf{X} \mathbf{y}]$:

$$[\mathbf{X} \mathbf{y}] = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} & y_1 \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} & y_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} & y_n \end{bmatrix},$$

In R, matrices (see Appendix D) are used to store data of the same type. As \mathbf{X} consists of real numbers, we might sometimes run into a “type mismatch” error if we want to extend it with \mathbf{y} that is not numeric. This is why such “combined” objects are usually stored as R `data.frames` (see Appendix E); this class allows columns of mixed types.

Let's extract the input matrix $\mathbf{X} \in \mathbb{R}^{300 \times 3}$ by taking the first 3 columns from the `wine_train` data frame and the reference outputs $\mathbf{y} \in \{0, 1\}^{300}$ given by the last column.

```
X_train <- as.matrix(wine_train[, -4]) # all except the last column
y_train <- factor(wine_train[, 4])
```

Now $[\mathbf{X} \mathbf{y}]$ is a basis for an interesting (and challenging) binary classification task. It constitutes our *training sample* – one from which we actually *learn* what levels of chemical features make a gathering with some vino a disappointing experience. Yet, we don't get a certificate in wine tasting only to hang some fancy diploma in a golden frame on the

wall. We want to *apply* the knowledge we've gained for the bottles whose corks (or caps) are yet to be popped off.

Here are some bottles for which we can show off our skills:

```
X_test <- as.matrix(read.csv("datasets/wine_test_X.csv"))
head(X_test)

##      chlorides density volatile.acidity
## [1,]    0.056  0.99950          0.19
## [2,]    0.027  0.99240          0.33
## [3,]    0.054  0.99836          0.31
## [4,]    0.057  0.99654          0.28
## [5,]    0.041  0.99280          0.22
## [6,]    0.028  0.99176          0.14

dim(X_test)

## [1] 150   3
```

We will denote is as $\mathbf{X}' \in \mathbb{R}^{150 \times 3}$ and call it a *test sample*. Let's invoke the `knn()` function from package FNN to classify the points \mathbf{X}' with the 5-nearest neighbour rule so as to obtain the corresponding predicted \hat{y}' .

```
library("FNN") # load the package

y_pred <- knn(X_train, X_test, y_train, k=5)
head(y_pred, 32) # predicted outputs

##  [1] 0 0 1 1 0 0 0 1 1 0 0 0 0 0 1 0 1 0 1 0 0 1 1 1 0 0 0 0 0 0 0 0 0
## Levels: 0 1
```

Great, we can now enjoy the bottles no. 1, 2, 5, 6, 7, 10, 11, 12, 13, 14, ... because the classifier claims they're not bad.

3.4 Classifier Assessment

But wait, are the consumers *really* going to enjoy the aforementioned wines? The 5-NN classifier might as well be deceiving us or be generating random predictions. It wouldn't be responsible to put it in production without making sure it's doing its job reasonably well.

Luckily, we have gathered information on the *true* labels for the observations in the test set, y' :

```
y_test <- factor(read.csv("datasets/wine_test_y.csv")[,1])
head(y_test, 32) # true outputs
```

```
## [1] 0 0 1 1 0 0 0 1 1 1 0 0 1 0 1 0 1 0 0 1 0 1 1 1 0 0 0 0 0 0 0 0 0
## Levels: 0 1
```

Comparing the displayed fragments of `y_test` and `y_pred` vectors (32 items), we see that most of the predictions are valid, but there are 4 mismatches. As there are 150 observations in the test set, it'll be better if we come up with some synthetic metrics that could summarise the *overall* performance of our classifier.

Accuracy is perhaps the most straightforward descriptor, being the ratio of the correctly classified instances to all the instances.

```
mean(y_test == y_pred) # accuracy
```

```
## [1] 0.71333
```

Could be worse. In 71% cases we make a correct prediction – we classify a wine as “not bad” when it’s actually decent and label it as “bad” when it’s indeed horrible.

Remark 3.4 Recall from Appendix C that the `==` operator in R works in an elementwise manner. It outputs a logical vector whose i -th element is TRUE whenever `test[i]` is equal to `pred[i]`. Calling `mean()` on a logical vector converts it to a numeric one: each TRUE is replaced with 1 and FALSE becomes 0. Arithmetic mean is nothing else than the sum of elements divided by their count. The sum of 1s and 0s is actually the number of 1s, i.e., for how many i s it holds that `test[i]` is equal to `pred[i]`.

Recall that y'_i is the true label associated with the i -th observation in the test set. Let \hat{y}'_i denote the classifier’s output for a given $x'_{i, \cdot}$. In our case, the outputs are binary, i.e., $\hat{y}'_i, y'_i \in \{0, 1\}$. Table 3.2 lists the 4 possible scenarios – all the distinct pairs of (\hat{y}'_i, y'_i) .

Table 3.2: True vs. predicted labels in a binary classification task; ideally, the number of false positives and false negatives should be kept to a minimum

.	$y'_i = 0$	$y'_i = 1$
$\hat{y}'_i = 0$	True Negative	False Negative (Type II error)
$\hat{y}'_i = 1$	False Positive (Type I error)	True Positive

We’d be happy with as many true positives and true negatives as possible. Note that the terms **positive** and **negative** refer to the classifier’s output, i.e., they indicate whether \hat{y}_i is equal to 1 and 0, respectively.

Let’s take a deeper look at why the classifier’s accuracy is “only” 71%. To summarise the correctness of predictions for the whole sample, we can compute the *confusion matrix*:

```
(C <- table(y_pred, y_test))
```

```
##      y_test
## y_pred  0   1
```

```
##      0 76 29
##      1 14 31
```

Actually, the classifier labels 29 bad wines as not-bad and 14 not-bad wines as bad.

In many applications we deal with *unbalanced problems*, where the case $y_i = 1$ is relatively rare(r), yet predicting it correctly is much more important than being accurate with respect to class 0 – think of medical applications, e.g., HIV testing or tumour diagnosis.

```
table(y_train)
```

```
## y_train
##   0   1
## 201  99
```

```
table(y_test)
```

```
## y_test
##   0   1
## 90  60
```

In our case, the “not-bad” class is much more populous than the “bad” one. Moreover, most will agree that it’s better to be surprised with a vino labelled as bad, than disappointed with a highly recommended product. Therefore, *accuracy* as a metric may fail to quantify what we are aiming for.

Remark 3.5 If only 1% of the cases have true $y'_i = 1$, then a dummy classifier that always outputs $\hat{y}'_i = 0$ has 99% accuracy.

Remark 3.6 Accuracy can be computed from a confusion matrix based on the formula:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}.$$

```
(C[1,1]+C[2,2])/sum(C) # equivalent to the above
```

```
## [1] 0.71333
```

Metrics such as precision and recall (and their aggregated version, F-measure) aim to address the above problem problem:

- *Precision* answers the question: If the classifier outputs 1, what is the probability that this is indeed true?

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

```
C[2,2]/(C[2,2]+C[2,1]) # Precision
```

```
## [1] 0.68889
```

- *Recall* (a.k.a. sensitivity, hit rate or true positive rate) – If the true class is 1, what is the probability that the classifier will detect it?

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

```
C[2,2]/(C[2,2]+C[1,2]) # Recall
```

```
## [1] 0.51667
```

Exercise 3.1 Precision or recall? It depends on an application. In each of the following settings, which measure is more important?

- medical diagnosis,
- medical screening,
- suggestions of potential matches in a dating app,
- plagiarism detection,
- wine recommendation.

As too many metrics may be daunting for some, as a compromise, we can use the *F*-measure (a.k.a. F_1 -measure), which is the harmonic mean of precision and recall:

$$F = \frac{1}{\frac{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}{2}} = \left(\frac{1}{2} \left(\text{Precision}^{-1} + \text{Recall}^{-1} \right) \right)^{-1} = \frac{\text{TP}}{\text{TP} + \frac{\text{FP}+\text{FN}}{2}}.$$

```
C[2,2]/(C[2,2]+0.5*C[1,2]+0.5*C[2,1]) # F
```

```
## [1] 0.59048
```

Exercise 3.2 Show that the above equality indeed holds.

Remark 3.7 The following function will come in handy in the future:

```
get_metrics <- function(y_pred, y_test)
{
  # first, let's make sure that both inputs are encoded
  # as factors with the same levels:
  all_levels <- unique(c(as.character(y_pred), as.character(y_test)))
  y_pred <- factor(y_pred, levels=all_levels)
  y_test <- factor(y_test, levels=all_levels)
  # compute the confusion matrix:
  C <- table(y_pred, y_test)
  stopifnot(dim(C) == c(2, 2))
  # fetch the metrics:
  c(Acc=(C[1,1]+C[2,2])/sum(C), # accuracy
    Prec=C[2,2]/(C[2,2]+C[2,1]), # precision
    Rec=C[2,2]/(C[2,2]+C[1,2]), # recall
    F=C[2,2]/(C[2,2]+0.5*C[1,2]+0.5*C[2,1]), # F-measure
```

```
# Confusion matrix items:
TN=C[1,1], FN=C[1,2],
FP=C[2,1], TP=C[2,2]
) # return a named vector
}
```

For example:

```
get_metrics(y_pred, y_test)
```

	Acc	Prec	Rec	F	TN	FN	FP	TP
##	0.71333	0.68889	0.51667	0.59048	76.00000	29.00000	14.00000	31.00000

3.5 Classifier Selection

Note that the nearest neighbour scheme implies in fact a whole family of classifiers. For each K , the corresponding K-NN method can be thought of as a different algorithm. Therefore, it is wise to pose the question: how to choose the best K for K-NN classification?

Here, by the *best* we mean one that has the highest *predictive power*, which we quantify by means of some chosen metric (such as accuracy, recall, precision, F-measure, etc.).

Let's study how the performance metrics change when we vary the number of nearest neighbours, K . Then, we'll choose the parameter that corresponds to, say, the greatest F-measure. However, we should not compute this metric on the test set! There is always a risk that we can *overfit* to current data – construct a classifier that performs extremely well on the samples we have but does not *generalise* well to the ones to come. The bottles we inspected is just a drop in the wine ocean that our machine learning solution will be filtering.

We are lucky though as we have one more wine sample available!

```
X_validate <- as.matrix(read.csv("datasets/wine_validate_X.csv"))
y_validate <- factor(read.csv("datasets/wine_validate_y.csv")[,1])
dim(X_validate)

## [1] 150    3
```

We will call it a *validation* (or development) set and use it for determining the optimal K . Then the *test* set will be recalled so as to perform the final evaluation (it's going to mimic the wines to come).

The following function computes the performance metrics for the K-NN classifier as a function of K :

```
knn_metrics <- function(K, X_train, X_validate, y_train, y_validate)
{
  y_pred <- knn(X_train, X_validate, y_train, k=K) # classify
  get_metrics(y_pred, y_validate)
}
```

For example:

```
knn_metrics(5, X_train, X_validate, y_train, y_validate)
```

```
##      Acc     Prec     Rec      F      TN      FN      FP      TP
## 0.65333 0.40625 0.28261 0.33333 85.00000 33.00000 19.00000 13.00000
```

Let's evaluate the performance metrics as a function of different odd (compare Section @[\(sec:mode\)](#)) K s:

```
Ks <- seq(1, 19, by=2) # 1, 3, 5, ...
Ps <- sapply(Ks, # on each element in this vector
             knn_metrics, # apply this function
             X_train, X_validate, y_train, y_validate # aux args
           )
# convert to "vertical" form - each K in separate row:
Ps <- as.data.frame(t(Ps))
```

Remark 3.8 Note that $sapply(X, f, arg1, arg2, \dots)$ outputs a matrix Z such that it's i -th column is $f(X[i], arg1, arg2, \dots)$. We transpose it by calling $t()$ to get a “vertical” (long) representation.

Example results:

```
round(cbind(K=Ks, Ps), 2)
```

```
##      K   Acc  Prec  Rec      F      TN      FN      FP      TP
## 1    1 0.57 0.30 0.30 0.30 72 32 32 32 14
## 2    3 0.63 0.34 0.24 0.28 83 35 21 11
## 3    5 0.65 0.41 0.28 0.33 85 33 19 13
## 4    7 0.64 0.37 0.24 0.29 85 35 19 11
## 5    9 0.65 0.38 0.20 0.26 89 37 15  9
## 6   11 0.69 0.48 0.26 0.34 91 34 13 12
## 7   13 0.68 0.46 0.24 0.31 91 35 13 11
## 8   15 0.69 0.48 0.22 0.30 93 36 11 10
## 9   17 0.70 0.53 0.22 0.31 95 36  9 10
## 10  19 0.71 0.60 0.20 0.30 98 37  6  9
```

Figure 3.5 is worth a thousand tables though (see `?matplot` in R). It seems that precision tends to slightly increase, whereas recall – decrease as K increases (on this dataset, it's not a general rule).

```
matplotlib(Ks, Ps[,1:4], xlab="K", ylab="Metric",
          col=1:4, lty=1:4, pch=1:4, type="b", ylim=c(0,1))
legend("top", legend=names(Ps[,1:4]),
       col=1:4, lty=1:4, pch=1:4, ncol=4, bg="white")
```

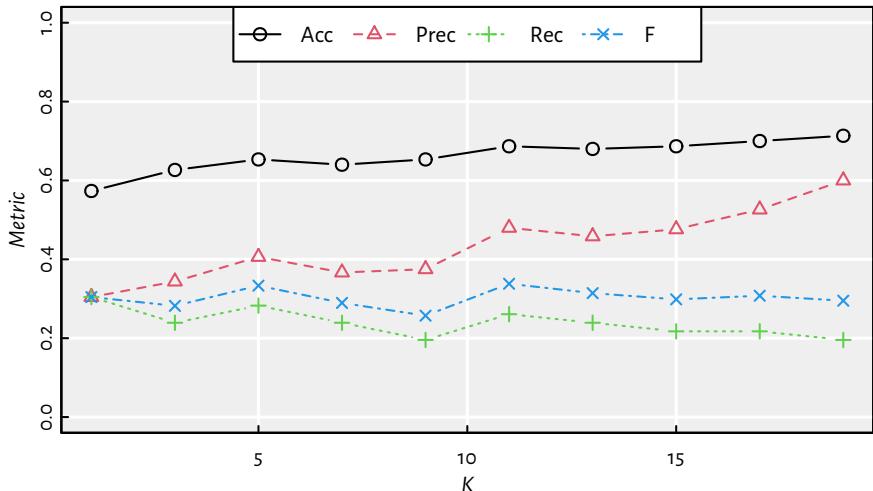


Figure 3.5: Performance of K -NN classifiers on the validation set as a function of K

It's very interesting that precision and recall (and hence F-measure) are much lower on the validation set than on the test set. This might be due to the small dataset sizes. As both of them have been generated independently and at random, and they constitute a representative sample of the population of all wines, the truth perhaps lies somewhere "in-between" the reported metrics.

Anyway, we get the best F-measure for $K = 11$. Let's see how well it is likely to perform "in production", i.e., on the test set:

```
knn_metrics(11, X_train, X_test, y_train, y_test)
```

```
##      Acc     Prec     Rec      F      TN      FN      FP      TP
## 0.66000 0.65517 0.31667 0.42697 80.00000 41.00000 10.00000 19.00000
```

This classifier is actually worse than the 5-NN one on the test set, but there is nothing we can do about it. It is an example of best practice to stick to the parameter we have identified as the optimal one on the validation sample – we have promised to treat the test set as an independent judge in our case (unless new data and more evidence come and we will be in a position to reevaluate/reconstruct our algorithm). We should treat this model with caution (actually, this is true for every model). No worries, we will encounter dilemma and disappointments of this kind in our everyday data science activities (see the next Chapter for more).

3.6 Implementing a K-NN Classifier (*)

3.6.1 Main Routine

To show that machine learning methods are not magical creatures who we should be terrified of, but rather the fruit of work of programmers just like us (have you ever thought of volunteering in an open source programming project for the good of the whole community?), let's implement a K-NN classifier ourselves, from scratch.

We'll use a top-bottom approach, starting with a general description of the admissible inputs and the expected output. Then we'll arrange the workflow for processing of data into conveniently manageable chunks.

The function's “declaration” will look like:

```
our_knn <- function(X_train, X_test, y_train, k=1) {
  # k=1 denotes a parameter with a default value
  # ... (see below) ...
}
```

It's advisable to specify the type and form of the arguments we're expecting on input. The `stopifnot()` function verifies if a given logical condition is true. If not, an error is raised.

```
# this is the body of our_knn() - part 1
stopifnot(is.numeric(X_train), is.matrix(X_train))
stopifnot(is.numeric(X_test), is.matrix(X_test))
stopifnot(is.factor(y_train))
stopifnot(ncol(X_train) == ncol(X_test))
stopifnot(nrow(X_train) == length(y_train))
stopifnot(k >= 1, k <= length(y_train))
n_train <- nrow(X_train)
n_test <- nrow(X_test)
p <- ncol(X_train)
L <- length(levels(y_train))
```

Therefore, we may assume from now on that $X_{\text{train}} \in \mathbb{R}^{n_{\text{train}} \times p}$, $X_{\text{test}} \in \mathbb{R}^{n_{\text{test}} \times p}$ and $y_{\text{train}} \in \{1, \dots, L\}^{n_{\text{train}}}$. Recall that R factor objects are internally encoded as integer vectors.

Next, we'll call the (to-be-done) function `our_get_knnx()`, which seeks nearest neighbours of all the points:

```
# our_get_knnx returns a matrix nn_indices of size n_test*k,
# where nn_indices[i, j] denotes the index of
# X_test[i, j]'s j-th nearest neighbour in X_train.
```

```
# (It is the point X_train[nn_indices[i,j],]).  
nn_indices <- our_get_knnx(X_train, X_test, k)
```

Then, for each point in `X_test`, we fetch the labels corresponding to its nearest neighbours and compute their mode (`our_mode()` function, wait for it):

```
y_pred <- numeric(n_test) # vector of length n_test  
# For now we will operate on the integer labels in {1,...,M}  
y_train_int <- as.numeric(y_train)  
for (i in 1:n_test) {  
  # Get the labels of the NNs of the i-th point:  
  nn_labels_i <- y_train_int[nn_indices[i,]]  
  # Compute the mode (majority vote):  
  y_pred[i] <- our_mode(nn_labels_i) # in {1,...,M}  
}
```

Finally, we should convert the resulting integer vector to an object of type `factor`:

```
# Convert y_pred to factor:  
return(factor(y_pred, labels=levels(y_train)))
```

3.6.2 Mode

To implement the mode, we can use the `tabulate()` function.

Exercise 3.3 *Read the function's documentation, see `?tabulate`.*

For example:

```
tabulate(c(1, 2, 1, 1, 1, 5, 2))
```

```
## [1] 4 2 0 0 1
```

We should keep in mind that there might be multiple modes – in such a case, it's best to pick one at random (to avoid any bias). For that, we can use the `sample()` function.

Exercise 3.4 *Read the function's man page, see `?sample`. Note that, dangerously, its behaviour is different when its first argument is a vector of length 1.*

An example implementation:

```
our_mode <- function(Y) {  
  # tabulate() will take care of  
  # checking the correctness of Y  
  t <- tabulate(Y) # factors == integer vectors here  
  # get indices corresponding to the maximal counts  
  mode_candidates <- which(t == max(t))  
  if (length(mode_candidates) == 1) return(mode_candidates)  
  else return(sample(mode_candidates, 1))  
}
```

Some tests:

```
our_mode(c(1, 1, 1, 1))
## [1] 1
our_mode(c(2, 2, 2, 2))
## [1] 2
our_mode(c(3, 1, 3, 3))
## [1] 3
our_mode(c(1, 1, 3, 3, 2))
## [1] 3
our_mode(c(1, 1, 3, 3, 2))
## [1] 1
```

3.6.3 NN Search Methods

Last but not least, we get to implement the `our_get_knnx()` function. It's the function responsible for seeking the indices of nearest neighbours.

```
# our_get_knnx returns a matrix nn_indices of size n_test*k,
# where nn_indices[i,j] denotes the index of
# X_test[i,]'s j-th nearest neighbour in X_train.
# (It is the point X_train[nn_indices[i,j],]).
our_get_knnx <- function(X_train, X_test, k) {
  # ...
}
```

A naive approach to `our_get_knnx()` relies on computing all pairwise distances, and sorting them.

```
our_get_knnx <- function(X_train, X_test, k) {
  n_test <- nrow(X_test)
  nn_indices <- matrix(NA_real_, nrow=n_test, ncol=k)
  for (i in 1:n_test) {
    d <- apply(X_train, 1, function(x)
      sqrt(sum((x-X_test[i,])^2)))
    # now d[j] is the distance
    # between X_train[j,] and X_test[i,]
    nn_indices[i,] <- order(d)[1:k]
  }
  return(nn_indices)
}
```

A comparison with FNN::knn() with regards to time needed to run the algorithms:

```
library("microbenchmark")
summary(microbenchmark(
  our={A<-our_knn(X_train, X_test, y_train, k=5)},
  fnn={B<-FNN::knn(X_train, X_test, y_train, k=5)},
  unit="ms" # # milliseconds
))

##   expr      min       lq     mean   median      uq     max neval
## 1  our 101.70 107.4 112.859 110.641 114.03 200.462    100
## 2  fnn 10.09 10.6 11.638 11.124 12.33 17.205    100
mean(A == B) # 1.0 on perfect match

## [1] 1
```

Both functions return identical results but our implementation is 10x slower. It turns out that our_get_knnx() is the part that actually constitutes the K-NN classifier's performance bottleneck in case on big data samples. FNN::knn() is efficiently written in C++, which is a compiled programming language.

R, on the other hand (just like Python and Matlab) is interpreted, therefore – as a rule of thumb – we should consider it an order of magnitude slower (see, however, the Julia language).

Let's substitute our naive implementation of our_get_knnx() with the equivalent one, but written in C++ (available in the FNN package).

Remark 3.9 Note that we can write a C++ implementation ourselves, see the Rcpp package (Eddelbuettel 2013) for convenient R and C++ integration.

```
our_get_knnx <- function(X_train, X_test, k) {
  # this is used by our_knn()
  FNN::get.knnx(X_train, X_test, k, algorithm="brute")$nn.index
}

summary(microbenchmark(
  our = {A<-our_knn(X_train, X_test, y_train, k=5)},
  fnn1={B<-FNN::knn(X_train, X_test, y_train, k=5)}, # kd_tree, see below
  fnn2={C<-FNN::knn(X_train, X_test, y_train, k=5, algorithm="brute")},
  unit="ms" # milliseconds
))

##   expr      min       lq     mean   median      uq     max neval
## 1  our 0.84474 0.90961 0.95866 0.93309 0.96688 2.1751    100
## 2 fnn1 9.76447 10.33771 10.98574 10.71992 11.23307 17.1421    100
## 3 fnn2 9.81180 10.40005 10.96215 10.66986 11.20543 13.7569    100
```

```
mean(A == B) # 1.0 on perfect match
```

```
## [1] 1
```

The timings are really interesting, taking into account that `FNN::knn()` uses `FNN::get.knnx()` as well (the R package ecosystem is comprised of open source software, we have the freedom to read the function's source code by calling `print(FNN::knn())`). Of course, before drawing conclusions about the quality of our implementation, we should test the aforementioned procedures on samples of different sizes and dimensionalities. This is left to the reader as an:

Exercise 3.5 *Test the run-times of the aforementioned procedures on samples of different sizes and dimensionalities, for example, on datasets generated randomly.*

Before we conclude, let us note that there are special *spatial search data structures* – such as metric trees – that aim to speed up searching for nearest neighbours in *low-dimensional spaces* (for small p). For example, `FNN::get.knnx()` also implements the so-called kd-trees.

Here is a function that generates n random points in a p dimensional hypercube. The function fill report time taken to look up k nearest neighbours based on a brute-force algorithm (all pairs of points considered) vs. on the kd-tree.

```
test_speed <- function(n, p, k) {
  A <- matrix(runif(n*p), nrow=n, ncol=p)
  s <- summary(microbenchmark::microbenchmark(
    brute_force=FNN::get.knnx(A, A, k, algorithm="brute"),
    kd_tree=FNN::get.knnx(A, A, k, algorithm="kd_tree"),
    times=3
  ), unit="s")
  # report minimum of each 3 time measurements:
  return(structure(s$min, names=as.character(s$expr)))
}
```

Example timings:

```
test_speed(10000, 2, 5)
```

```
## brute_force      kd_tree
##     0.288927     0.012373
```

```
test_speed(10000, 5, 5)
```

```
## brute_force      kd_tree
##     0.406603     0.060394
```

```
test_speed(10000, 10, 5)
```

```
## brute_force      kd_tree
```

```
##      0.63985      0.63335
test_speed(10000, 20, 5)

## brute_force      kd_tree
##      1.2351      5.2283
```

In spaces of higher dimensionality, the brute force algorithm is actually faster. It turns out that searching in high-dimensional spaces is hard due to the various phenomena collectively referred to as the *curse of dimensionality* (see, e.g., (Blum et al. 2020)). Yet, is low-dimensional data boring? Well, our physical world is perceived as 3-dimensional, spatial data (on maps) is 2-dimensional, therefore, there are interesting use cases of kd-trees anyway.

3.7 Remarks

TODO

Note that the K-nearest neighbour method is suitable for any multiclass classification, i.e., for any number of levels L of y . However, precision and recall can only be computed if $L = 2$.

In practice searching for nearest neighbours is time-consuming for larger datasets – to classify a single point we have to query the whole training set (unless it's a space of low dimensionality).

Note that our implementation requires $c \cdot n_{\text{test}} \cdot n_{\text{train}} \cdot p$ arithmetic operations for some $c > 1$. The overall cost of sorting is at least $d \cdot n_{\text{test}} \cdot n_{\text{train}} \cdot \log n_{\text{train}}$ for some $d > 1$. This does not scale well with both n_{test} and n_{train} (think – big data).

Moreover, the training set should be available at all times. Other algorithms discussed in this book will try to come up with a synthetic/compressed representation (a model) of the training set.

However, the advantage of K-NN is that it naturally adapts to new training points – we don't have to recompute any models to take them into account (hence, we can say, that K-NN can also work in an *online mode).

K-NN is an influential concept – see Chapter 14 for how it is naturally employed in recommender systems.

Instead of nearest neighbours, we could be also considering so-called ϵ -neighbourhoods – all the points from \mathbf{X} whose distance to a given x' is not greater than ϵ (for example, DBSCAN (Ling 1973) (Ester et al. 1996) is a clustering method based on this concept). However, in classification task we should be aware of the fact that some test points might have empty ϵ -neighbourhoods and that this issue should be resolved somehow.

There are also approximate nearest neighbours, e.g., nmslib, faiss, etc., see also <https://github.com/erikbern/ann-benchmarks>

(*) 1-NN classification is essentially based on a dataset's so-called Voronoi diagram. Interestingly, in single linkage clustering, we also seek 1-nearest neighbours (between clusters).

Recommended further reading: (Hastie et al. 2017: Section 13.3)

Next Chapter....

Further we will discuss some other well-known classifiers:

- *Decision trees*
- *Logistic regression*

4

Feature Engineering

TODO In this chapter, we will:

- ...

train-test-validate split manually?

K-NN and back to hclust?

doesn't answer why?

Actually it's not that bad At least we can take a look at the most *similar* wines to the tested ones (neighbours) and inspect their characteristics

models don't have to work well out of the box – feature engineering: feature selection, standardisation

in `wine_train` all variables used the same units

normalisation – how SSI is created?

log-normal distribution, power law?

normal distribution

you can compare apples with oranges if you have apple and orange *counts*

even if on the same scale, 1 kg of gold is not the same as 1 kg of citric acid

A random **train-test split** of the original dataset:

- *training sample* (usually 60-80% of the observations) – used to construct a model,
- *test sample* (remaining 40-20%) – used to assess the goodness of fit.

In order to overcome this problem, we can perform a random **train-validation-test split** of the original dataset:

- *training sample* (e.g., 60%) – used to construct the models
- *validation sample* (e.g., 20%) – used to tune the hyperparameters of the classifier
- *test sample* (e.g., 20%) – used to assess the goodness of fit

In the K-NN classification task, there are many hyperparameters to tune up:

- Which K should we choose?
- Should we standardise the dataset?

- Which variables should be taken into account when computing the Euclidean distance?

An example way to perform a 60/20/20% train-validation-test split:

```
# set.seed(123) # reproducibility matters
# random_indices <- sample(n)
# n1 <- floor(n*0.6)
# n2 <- floor(n*0.8)
# X2_train <- X[random_indices[1 :n1], ]
# Y2_train <- Y[random_indices[1 :n1] ]
# X2_valid <- X[random_indices[(n1+1):n2], ]
# Y2_valid <- Y[random_indices[(n1+1):n2] ]
# X2_test <- X[random_indices[(n2+1):n ], ]
# Y2_test <- Y[random_indices[(n2+1):n ] ]
# stopifnot(nrow(X2_train)+nrow(X2_valid)+nrow(X2_test)
#           == nrow(X))
```

Remark 4.1 *Test sample must not be used in the training phase! (No cheating!)*

60/40% train-test split in R:

```
# set.seed(123) # reproducibility matters
# random_indices <- sample(n)
# head(random_indices) # preview
# # first 60% of the indices (they are arranged randomly)
# # will constitute the train sample:
# train_indices <- random_indices[1:floor(n*0.6)]
# X_train <- X[train_indices,]
# Y_train <- Y[train_indices]
# # the remaining indices (40%) go to the test sample:
# X_test <- X[-train_indices,]
# Y_test <- Y[-train_indices]
```

Remark 4.2 (*) If our dataset is too small, we can use various cross-validation techniques instead of a train-validate-test split.

4.0.1 Feature Engineering

Note that the Euclidean distance that we used above implicitly assumes that every feature (independent variable) is on the same scale.

However, when dealing with, e.g., physical quantities, we often perform conversions of units of measurement (kg → g, feet → m etc.).

Transforming a single feature may drastically change the metric structure of the dataset and therefore highly affect the obtained predictions.

To “bring data to the same scale”, we often apply a trick called **standardisation**.

Computing the so-called **Z-scores** of the j -th feature, $x_{\cdot j}$, is done by subtracting from

each observation the sample mean and dividing the result by the sample standard deviation:

$$z_{i,j} = \frac{x_{i,j} - \bar{x}_{\cdot,j}}{s_{x_{\cdot,j}}}$$

This is a new feature $\mathbf{z}_{\cdot,j}$ that always has mean 0 and standard deviation of 1.

Moreover, it is *unit-less* (e.g., we divide a value in kgs by a value in kgs, the units are cancelled out). This, amongst others, prevents one of the features from dominating the other ones.

Z-scores are easy to interpret, e.g., 0.5 denotes an observation that is 0.5 standard deviations above the mean and -3 informs us that a value is 3 standard deviations below the mean.

Remark 4.3 (*) If data are normally distributed (bell-shaped histogram), with very high probability, most (expected value is 99.74%) observations should have Z-scores between -3 and 3. Those that don't, are "suspicious", maybe they are outliers? We should inspect them manually.

Let's compute $\mathbf{Z}_{\text{train}}$ and \mathbf{Z}_{test} , being the standardised versions of $\mathbf{X}_{\text{train}}$ and \mathbf{X}_{test} , respectively.

```
# means <- apply(X_train, 2, mean) # column means
# sds   <- apply(X_train, 2, sd)   # column standard deviations
# Z_train <- X_train # copy
# Z_test  <- X_test # copy
# for (j in 1:ncol(X)) {
#   Z_train[,j] <- (Z_train[,j]-means[j])/sds[j]
#   Z_test[,j]  <- (Z_test[,j] -means[j])/sds[j]
# }
```

Note that we have transformed the training and test sample in the very same way. Computing means and standard deviations separately for these two datasets is a common error – it is the training set that we use in the course of the learning process. The above can be re-written as:

```
# Z_train <- t(apply(X_train, 1, function(r) (r-means)/sds))
# Z_test  <- t(apply(X_test, 1, function(r) (r-means)/sds))
```

See Figure ?? for an illustration. Note that the righthand figures (histograms of standardised variables) are on the same scale now.

Remark 4.4 Of course, standardisation is only about shifting and scaling, it preserves the shape of the distribution. If the original variable is right skewed or bimodal, its standardised version will remain as such.

Let's compute the accuracy of K-NN classifiers acting on standardised data.

```
# Y_knn5s <- knn(Z_train, Z_test, Y_train, k=5)
# mean(Y_test == Y_knn5s) # accuracy
# Y_knn9s <- knn(Z_train, Z_test, Y_train, k=9)
# mean(Y_test == Y_knn9s) # accuracy
```

The accuracy is much better.

Standardisation is an example of *feature engineering*.

Good models rarely work well “straight out of the box” – if that was the case, we wouldn’t need data scientists and machine learning engineers!

To increase models’ accuracy, we often spend a lot of time:

- cleansing data (e.g., removing outliers)
- extracting new features
- transforming existing features
- trying to find a set of features that are relevant

This is the “more art than science” part of data science (sic!), and hence most textbooks are not really eager for discussing such topics (including this one).

Sorry, this is sad but true. The solutions that work well in the case of dataset A may fail in the B case and vice versa. However, the more exercises you solve, the greater the arsenal of ideas/possible approaches you will have at hand when dealing with real-world problems.

Feature selection – example (manually selected columns):

```
# features <- c("chlorides", "volatile.acidity")
# Y_knn5s <- knn(Z_train[,features], Z_test[,features],
#   Y_train, k=5)
# mean(Y_test == Y_knn5s) # accuracy
# Y_knn9s <- knn(Z_train[,features], Z_test[,features],
#   Y_train, k=9)
# mean(Y_test == Y_knn9s) # accuracy
```

Exercise 4.1 Try to find a combination of 2-4 features (by guessing or applying magic tricks) that increases the accuracy of a K-NN classifier on this dataset.

4.0.2 Different Metrics (*)

(this is like metric space engineering)

The Euclidean distance is just one particular example of many possible **metrics** (metric == a mathematical term, above we have used this term in a more relaxed fashion, when referring to accuracy etc.).

Mathematically, we say that d is a metric on a set X (e.g., \mathbb{R}^p), whenever it is a function $d : X \times X \rightarrow [0, \infty]$ such that for all $x, x', x'' \in X$:

- $d(x, x') = 0$ if and only if $x = x'$,
- $d(x, x') = d(x', x)$ (it is symmetric)
- $d(x, x'') \leq d(x, x') + d(x', x'')$ (it fulfils the triangle inequality)

Remark 4.5 (*) Not all the properties are required in all the applications; sometimes we might need a few additional ones.

We can easily generalise the way we introduced the K-NN method to have a classifier that is based on a point's neighbourhood with respect to any metric.

Example metrics on \mathbb{R}^p :

- **Euclidean**

$$d_2(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\| = \|\mathbf{x} - \mathbf{x}'\|_2 = \sqrt{\sum_{i=1}^p (x_i - x'_i)^2}$$

- **Manhattan** (taxicab)

$$d_1(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|_1 = \sum_{i=1}^p |x_i - x'_i|$$

- **Chebyshev** (maximum)

$$d_\infty(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|_\infty = \max_{i=1, \dots, p} |x_i - x'_i|$$

We can define metrics on different spaces too.

For example, the **Levenshtein distance** is a popular choice for comparing character strings (also DNA sequences etc.)

It is an *edit distance* – it measures the minimal number of single-character insertions, deletions or substitutions to change one string into another.

For instance:

```
adist("happy", "nap")
```

```
##      [,1]
## [1,]    3
```

This is because we need 1 substitution and 2 deletions,

happy → nappy → napp → nap.

See also:

- the Hamming distance for categorical vectors (or strings of equal lengths),
- the Jaccard distance for sets,
- the Kendall tau rank distance for rankings.

Moreover, R package `stringdist` includes implementations of numerous string metrics.

4.1 Exercises

4.1.1 Wine Quality – Best K-NN Parameters via Cross-Validation (*)

Consider the Wine Quality dataset (see Appendix F for details):

```
wine_quality <- read.csv("datasets/wine_quality_all.csv",
  comment.char="#")
head(wine_quality, 3)

##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1          7.4            0.70      0.00           1.9     0.076
## 2          7.8            0.88      0.00           2.6     0.098
## 3          7.8            0.76      0.04           2.3     0.092
##   free.sulfur.dioxide total.sulfur.dioxide density      pH sulphates
## 1                  11            0.9978  3.51      0.56
## 2                  25            0.9968  3.20      0.68
## 3                  15            0.9970  3.26      0.65
##   alcohol response color
## 1      9.4      5  red
## 2      9.8      5  red
## 3      9.8      5  red
```

Exercise 4.2 Add a new column named *quality*. A wine should get a *quality* of 1 if its rating is greater than or equal to 7 (a good wine) and a quality of 0 otherwise.

Exercise 4.3 Perform a random train-test split of size 60-40%: create the matrices *X_train* and *X_test* containing the 11 physicochemical wine features and the corresponding label vectors *Y_train* and *Y_test* that inform on the wines' quality.

Exercise 4.4 Determine the best parameter setting for the K-nearest neighbour classification of the *quality* variable based on the 11 physicochemical features. Perform the so-called grid (exhaustive) search over all the possible combinations of the following parameters:

1. $K: 1, 3, 5, 7 \text{ or } 9$,
2. preprocessing: none (raw input data), standardised variables or robustly standardised variables,
3. metric: L_2 (Euclidean) or L_1 (Manhattan).

In other words, there are $5 \cdot 3 \cdot 2 = 30$ combinations of parameters in total, and hence – 30 different scenarios to consider. By the best classifier we mean the one that maximises the F-measure obtained by the so-called 5-fold cross-validation (see below).

Robust standardisation. To perform a robust standardisation, for each column individually, subtract its median and then divide it by its median absolute deviation (MAD, i.e.,

`median(abs(x-median(x)))`). This data preprocessing scheme is less sensitive to outliers than the classic standardisation.

Note that the L_1 metric-based K-nearest neighbour method is not available in the FNN package. You need to implement it yourself.

Cross-validation. We have discussed that it would not be fair to use the test set for choosing of the optimal parameters (we would be overfitting to the test set). We know that one possible way to assure the transparent evaluation of a classifier is to perform a train-validate-test split and use the validation set for parameter tuning.

Here we will use a different technique – one that estimates the methods’ “true” predictive performance more accurately, yet at the cost of significantly increased run-time. Namely, in *5-fold cross-validation*, we split the original train set randomly into 5 disjoint parts: A, B, C, D, E (more or less of the same number of observations). We use each combination of 4 chunks as training sets and the remaining part as the validation set, on which we compute the F-measure:

train set	validation set	F-measure
B, C, D, E	A	F_A
A, C, D, E	B	F_B
A, B, D, E	C	F_C
A, B, C, E	D	F_D
A, B, C, D	E	F_E

At the end we report the average F-measure, $(F_A + F_B + F_C + F_D + F_E)/5$.

4.2 Remarks

TODO

mention: kernels

Recommended further reading: (Hastie et al. 2017: Section 13.3)

Next Chapter....

5

Classification with Decision Trees

TODO In this chapter, we will:

- ...
- ...

data generators – why 100% accuracy might not be possible, why we need statistics or information theory (amongst others) (Devroye et al. 1996), (Blum et al. 2020) – research in machine learning vs. research with machine learning (citations! - asymptotics etc.) – 2 normal distributions, non-separable illustration

in (Devroye et al. 1996) presented are several interesting results of the asymptotic behaviour of the K-NN classifier when $K \rightarrow \infty$ and $K/n \rightarrow \infty$ as $n \rightarrow \infty$ (e.g., both training sample size n and K grow but n grows at least an order of magnitude faster than K)

maths can be used to answer questions about the general behaviour of the method: “it is always true that”, “with probability XX it holds”, “if XXX, then the expected is YYY” etc.

prediction vs. description

decision trees – towards models (explanation)

5.1 Introduction

5.1.1 Classification Task

Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be an input matrix that consists of n points in a p -dimensional space (each of the n objects is described by means of p numerical features)

Recall that in supervised learning, with each $\mathbf{x}_{i \cdot}$ we associate the desired output y_i .

Hence, our dataset is $[\mathbf{X} \mathbf{y}]$ – where each object is represented as a row vector $[\mathbf{x}_{i \cdot}, y_i]$, $i = 1, \dots, n$:

$$[\mathbf{X} \mathbf{y}] = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} & y_1 \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} & y_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} & y_n \end{bmatrix}.$$

In this chapter we are still interested in **classification** tasks; we assume that each y_i is a descriptive label.

Let's assume that we are faced with **binary classification** tasks.

Hence, there are only two possible labels that we traditionally denote with 0s and 1s.

For example:

○	1
no	yes
false	true
failure	success
healthy	ill

Let's recall the synthetic 2D dataset from the previous chapter (true decision boundary is at $X_1 = 0$), see Figure 5.1.

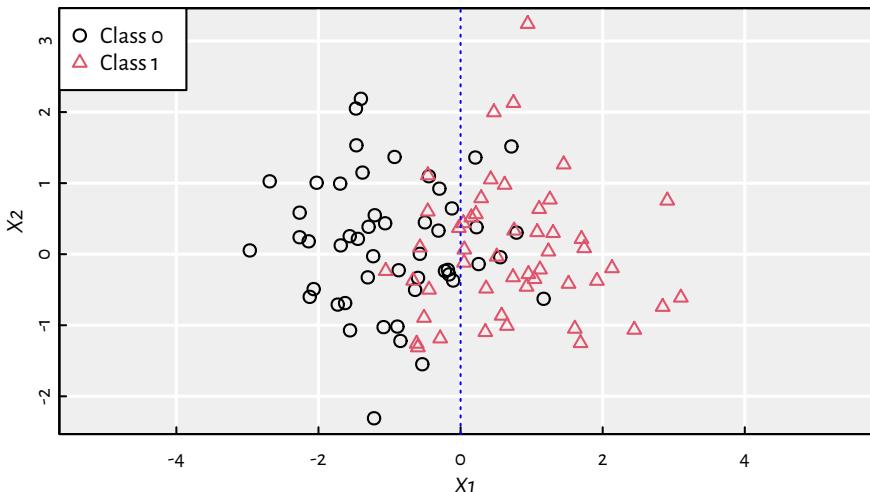


Figure 5.1: A synthetic 2D dataset with the true decision boundary at $X_1 = 0$

5.1.2 Data

For illustration, we'll be considering the Wine Quality dataset (white wines only):

```
wine_quality <- read.csv("datasets/wine_quality_all.csv",
  comment.char="#")
white_wines <- wine_quality[wine_quality$color == "white",]
(n <- nrow(white_wines)) # number of samples

## [1] 4898
```

The input matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ consists of the first 10 numeric variables:

```
X <- as.matrix(white_wines[,1:10])
dim(X)
```

```
## [1] 4898   10
head(X, 2) # first two rows

##      fixed.acidity volatile.acidity citric.acid residual.sugar
## 1600          7.0            0.27       0.36        20.7
## 1601          6.3            0.30       0.34        1.6
##      chlorides free.sulfur.dioxide total.sulfur.dioxide density pH
## 1600     0.045             45         170    1.001 3.0
## 1601     0.049             14         132    0.994 3.3
##      sulphates
## 1600     0.45
## 1601     0.49
```

The 11th variable measures the amount of alcohol (in %).

We will convert this dependent variable to a binary one:

- 0 == (alcohol < 12) == lower-alcohol wines,
- 1 == (alcohol >= 12) == higher-alcohol wines

```
# recall that TRUE == 1
Y <- factor(as.character(as.numeric(white_wines$alcohol >= 12)))
table(Y)
```

```
## Y
##   0   1
## 4085 813
```

60/40% train-test split:

```
set.seed(123) # reproducibility matters
random_indices <- sample(n)
head(random_indices) # preview

## [1] 2463 2511 2227 526 4291 2986
```

```
# first 60% of the indices (they are arranged randomly)
# will constitute the train sample:
train_indices <- random_indices[1:floor(n*0.6)]
X_train <- X[train_indices,]
Y_train <- Y[train_indices]
# the remaining indices (40%) go to the test sample:
X_test <- X[-train_indices,]
Y_test <- Y[-train_indices]
```

Let's also compute `Z_train` and `Z_test`, being the standardised versions of `X_train` and `X_test`, respectively.

```
means <- apply(X_train, 2, mean) # column means
sds   <- apply(X_train, 2, sd)   # column standard deviations
Z_train <- t(apply(X_train, 1, function(r) (r-means)/sds))
Z_test  <- t(apply(X_test, 1, function(r) (r-means)/sds))

get_metrics <- function(Y_pred, Y_test)
{
  C <- table(Y_pred, Y_test) # confusion matrix
  stopifnot(dim(C) == c(2, 2))
  c(Acc=(C[1,1]+C[2,2])/sum(C), # accuracy
    Prec=C[2,2]/(C[2,2]+C[2,1]), # precision
    Rec=C[2,2]/((C[2,2]+C[1,2])), # recall
    F=C[2,2]/((C[2,2]+0.5*C[1,2]+0.5*C[2,1])), # F-measure
    # Confusion matrix items:
    TN=C[1,1], FN=C[1,2],
    FP=C[2,1], TP=C[2,2]
  ) # return a named vector
}
```

5.2 Decision Trees

5.2.1 Introduction

Note that a K-NN classifier discussed in the previous chapter is **model-free**. The whole training set must be stored and referred to at all times.

Therefore, it doesn't *explain* the data we have – we may use it solely for the purpose of *prediction*.

Perhaps one of the most interpretable (and hence human-friendly) models consist of decision rules of the form:

IF $x_{i,j_1} \leq v_1$ **AND ... AND** $x_{i,j_r} \leq v_r$ **THEN** $\hat{y}_i = 1$.

These can be organised into a **hierarchy** for greater readability.

This idea inspired the notion of **decision trees** (Breiman et al. 1984).

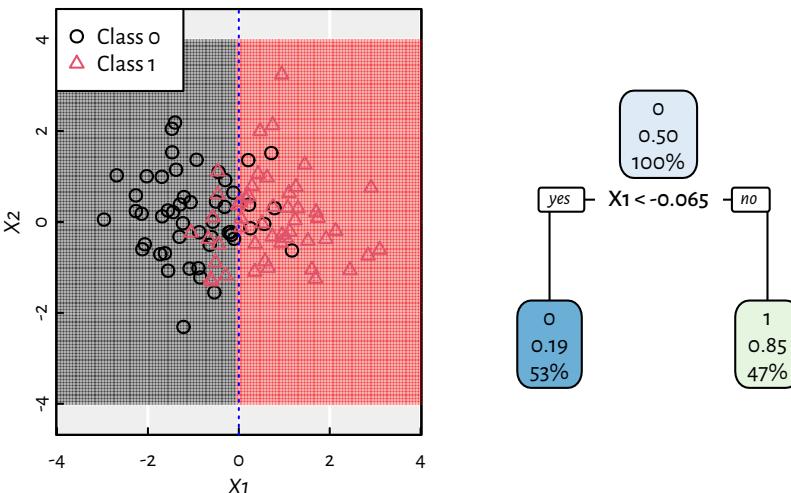


Figure 5.2: (#fig:plot_rpart) The simplest decision tree for the synthetic 2D dataset and the corresponding decision boundaries

Figure @ref(fig:plot_rpart2) depicts a very simple decision tree for the aforementioned synthetic dataset. There is only one decision boundary (based on X_1) that splits data into the “left” and “right” sides. Each tree node reports 3 pieces of information:

- dominating class (0 or 1)
- (relative) proportion of 0s represented in a node
- (absolute) proportion of all observations in a node

Figures @ref(fig:plot_rpart2) and @ref(fig:plot_rpart3) depict trees with more decision rules. Take a moment to contemplate how the corresponding decision boundaries changed with the introduction of new decision rules.

5.2.2 Example in R

We will use the `rpart()` function from the `rpart` package to build a classification tree.

```
library("rpart")
library("rpart.plot")
set.seed(123)
```

`rpart()` uses a formula (~) interface, hence it will be easier to feed it with data in a `data.frame` form.

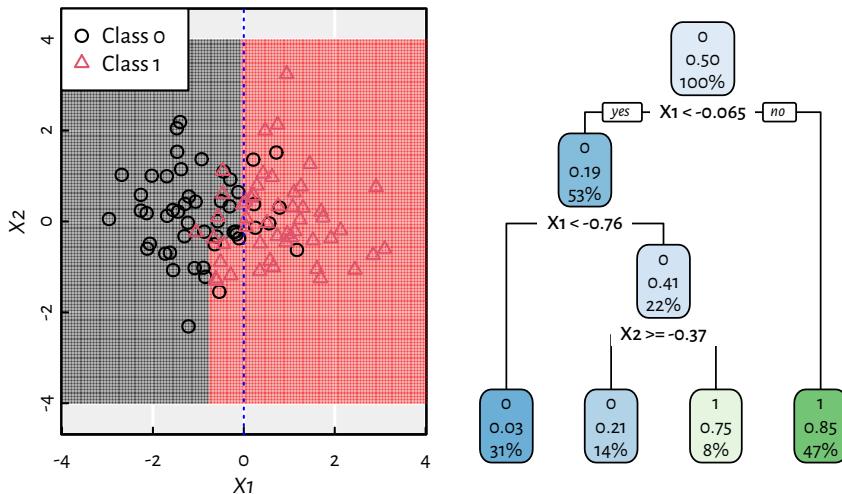


Figure 5.3: (#fig:plot_rpart2) A more complicated decision tree for the synthetic 2D dataset and the corresponding decision boundaries

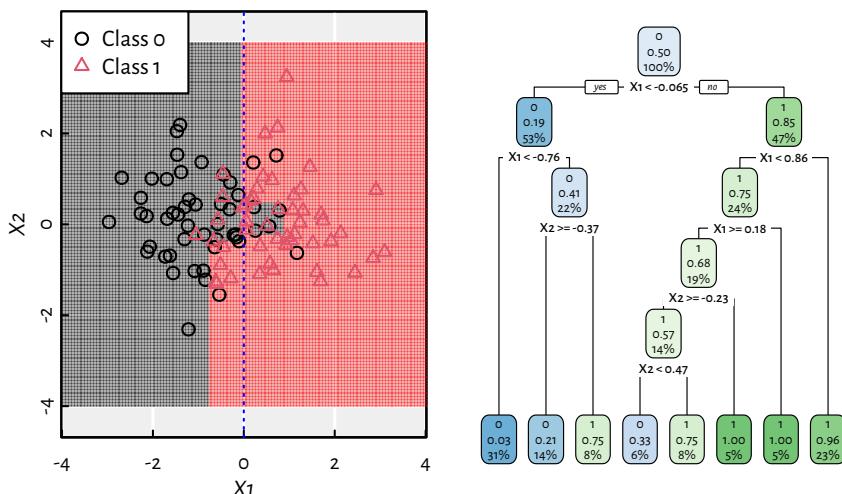


Figure 5.4: (#fig:plot_rpart3) An even more complicated decision tree for the synthetic 2D dataset and the corresponding decision boundaries

```
XY_train <- cbind(as.data.frame(X_train), Y=Y_train)
XY_test <- cbind(as.data.frame(X_test), Y=Y_test)
```

Fit and plot a decision tree, see Figure @ref(fig:plot_rpart1).

```
t1 <- rpart(Y~., data=XY_train, method="class")
rpart.plot(t1, tweak=1.1, fallen.leaves=FALSE, digits=3)
```

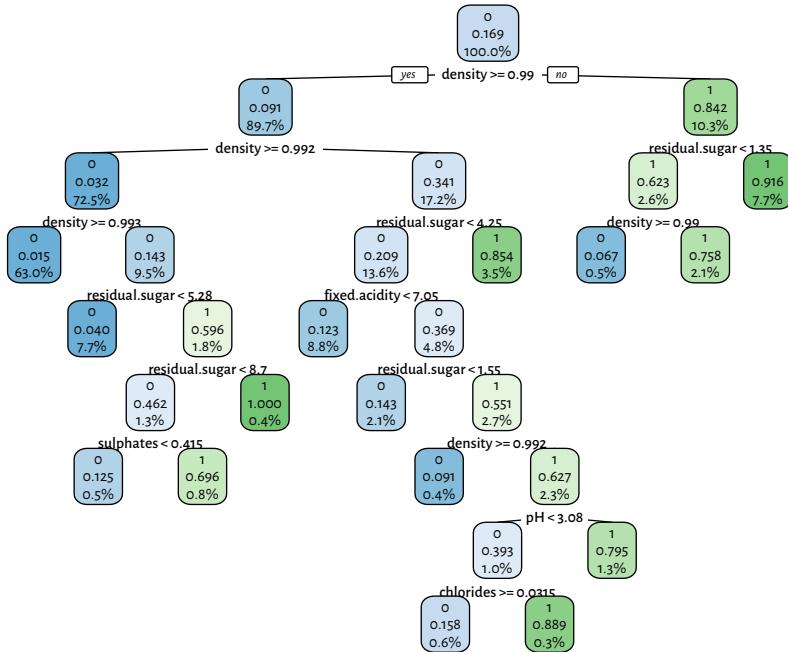


Figure 5.5: (#fig:plot_rpart1) A decision tree for the `white_wines` dataset

We can build less or more complex trees by playing with the `cp` parameter, see Figures @ref(fig:plot_rpart222) and 5.7.

```
# cp = complexity parameter, smaller □ more complex tree
t2 <- rpart(Y~., data=XY_train, method="class", cp=0.1)
rpart.plot(t2, tweak=1.1, fallen.leaves=FALSE, digits=3)
```

```
# cp = complexity parameter, smaller □ more complex tree
t3 <- rpart(Y~., data=XY_train, method="class", cp=0.00001)
rpart.plot(t3, tweak=1.1, fallen.leaves=FALSE, digits=3)
```

Trees with few decision rules actually are very nicely interpretable. On the other hand,

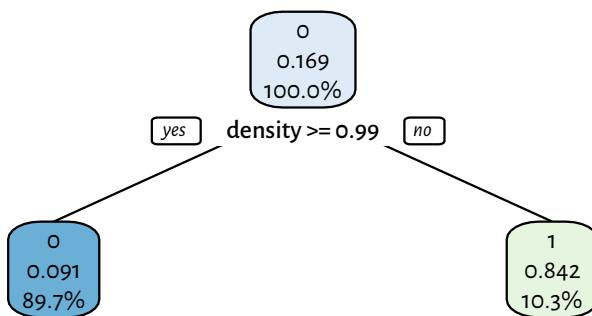


Figure 5.6: (#fig:plot_rpart222) A (simpler) decision tree for the `white_wines` dataset

plotting of the complex ones is just hopeless; we should treat them as “black boxes” instead.

Let's make some predictions:

```

Y_pred <- predict(t1, XY_test, type="class")
get_metrics(Y_pred, Y_test)

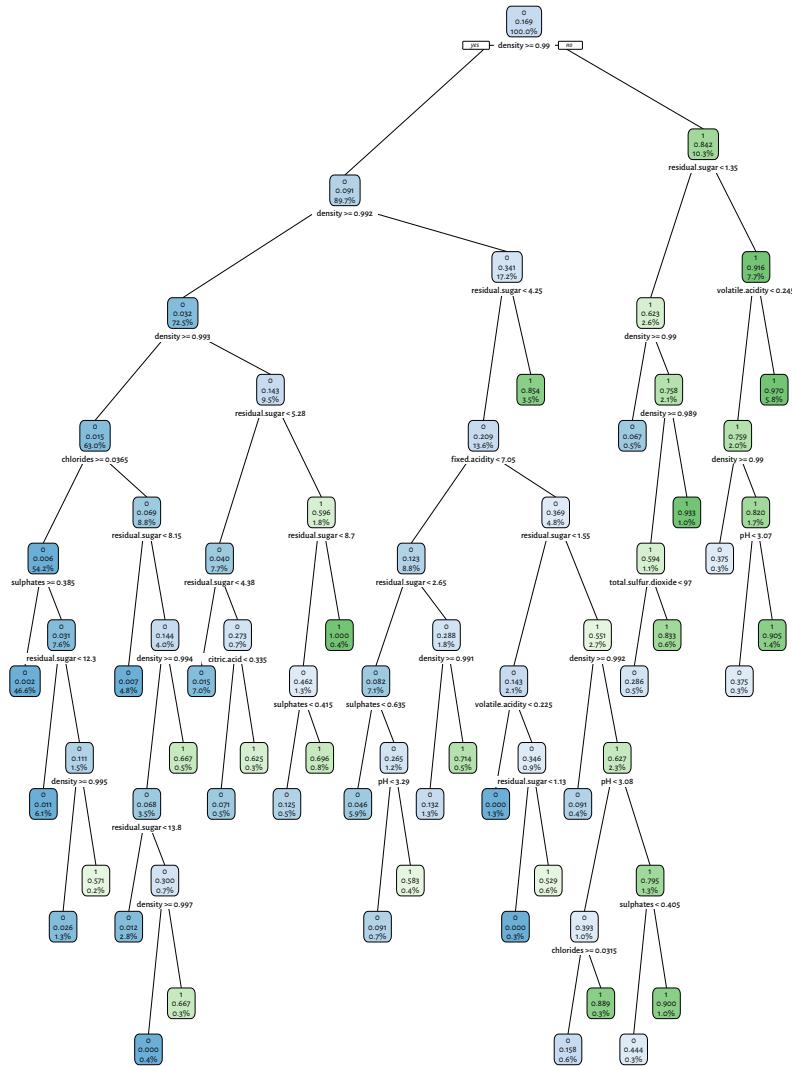
##      Acc      Prec      Rec      F      TN      FN
##  0.92857  0.80623  0.73502  0.76898 1587.00000  84.00000
##      FP      TP
##  56.00000 233.00000

Y_pred <- predict(t2, XY_test, type="class")
get_metrics(Y_pred, Y_test)

##      Acc      Prec      Rec      F      TN      FN
##  0.90255  0.83871  0.49211  0.62028 1613.00000 161.00000
##      FP      TP
##  30.00000 156.00000

Y_pred <- predict(t3, XY_test, type="class")
get_metrics(Y_pred, Y_test)

##      Acc      Prec      Rec      F      TN      FN
##  0.91837  0.73433  0.77603  0.75460 1554.00000  71.00000
##      FP      TP
##  89.00000 246.00000
  
```

Figure 5.7: A (more complex) decision tree for the `white_wines` dataset

Remark. (*) Interestingly, `rpart()` also provides us with information about the importance degrees of each independent variable.

```
t1$variable.importance/sum(t1$variable.importance)
```

```
##           density      residual.sugar      fixed.acidity
## 0.6562490          0.1984221          0.0305167
## chlorides          pH      volatile.acidity
## 0.0215008          0.0209678          0.0192880
## sulphates total.sulfur.dioxide citric.acid
## 0.0184293          0.0140482          0.0119201
## free.sulfur.dioxide
## 0.0086579
```

5.2.3 A Note on Decision Tree Learning

Learning an optimal decision tree is a computationally hard problem – we need some heuristics.

Examples:

- ID3 (Iterative Dichotomiser 3) (Quinlan 1986)
- C4.5 algorithm (Quinlan 1993)
- CART by Leo Breiman et al., (Breiman et al. 1984)

(***) Decision trees are most often constructed by a *greedy, top-down recursive partitioning*, see., e.g., (Therneau & Atkinson 2019).

5.3 Exercises

5.3.1 EdStats – Where Girls Are Better at Maths Than Boys?

In this task we will consider the “wide” version of the EdStats dataset:

```
edstats <- read.csv("datasets/edstats_2019_wide.csv",
                     comment.char="#")
edstats[1, 1:6]

##   CountryName HD.HCI.AMRT HD.HCI.AMRT.FE HD.HCI.AMRT.MA HD.HCI.EYRS
## 1 Afghanistan     0.7797        0.8018      0.7597       8.58
##   HD.HCI.EYRS.FE
## 1                  6.73

meta <- read.csv("datasets/edstats_meta.csv",
                  comment.char="#")
```

This dataset is small, moreover, we'll be more interested in the description (understanding) of data, not prediction of the response variable to unobserved samples. Note that we have the *population* of the World countries at hand here (new countries do not arise on a daily basis). Therefore, a train-test split won't be performed.

Exercise 5.1 Add a 0/1 factor-type variable *girls_rule_maths* that is equal to 1 if and only if a country's average score of 15-year-old female students on the PISA mathematics scale is greater than the corresponding indicator for the male ones.

Solution.

Recall that a conversion of a logical value to a number yields 1 for TRUE and 0 for FALSE. Hence:

```
edstats$girls_rule_maths <-
  factor(as.numeric(
    edstats$L0.PISA.MAT.FE>edstats$L0.PISA.MAT.MA
  ))
head(edstats$girls_rule_maths, 10)

## [1] <NA> 1     1     <NA> <NA> <NA> <NA> 0     <NA>
## Levels: 0 1
```

Unfortunately, there are many missing values in the dataset. More precisely:

```
sum(is.na(edstats$girls_rule_maths)) # count

## [1] 187

mean(is.na(edstats$girls_rule_maths)) # proportion

## [1] 0.69776
```

Countries such as Egypt, India, Iran or Venezuela are not amongst the 79 members of the Programme for International Student Assessment. Thus, we'll have to deal with the data we have.

The percentage of countries where "girls rule" is equal to:

```
mean(edstats$girls_rule_maths==1, na.rm=TRUE)

## [1] 0.33333
```

Here is the list of those counties:

```
as.character(na.omit(
  edstats[edstats$girls_rule_maths==1, "CountryName"]
))

## [1] "Albania"           "Algeria"
## [3] "Brunei Darussalam" "Bulgaria"
## [5] "Cyprus"             "Dominican Republic"
## [7] "Finland"            "Georgia"
## [9] "Hong Kong SAR, China" "Iceland"
```

```

## [11] "Indonesia"           "Israel"
## [13] "Jordan"              "Lithuania"
## [15] "Malaysia"             "Malta"
## [17] "Moldova"              "North Macedonia"
## [19] "Norway"               "Philippines"
## [21] "Qatar"                "Saudi Arabia"
## [23] "Sweden"               "Thailand"
## [25] "Trinidad and Tobago" "United Arab Emirates"
## [27] "Vietnam"

```



Exercise 5.2 Learn a decision tree that distinguishes between the countries where girls are better at maths than boys and assess the quality of this classifier.

Solution.

Let's first create a subset of `edstats` that doesn't include the country names as well as the boys' and girls' math scores.

```

edstats_subset <- edstats[!(names(edstats) %in%
  c("CountryName", "LO.PISA.MAT.FE", "LO.PISA.MAT.MA"))]

```

Fitting and plotting (see Figure 5.8) of the tree can be performed as follows:

```

library("rpart")
library("rpart.plot")
tree <- rpart(girls_rule_maths~, data=edstats_subset,
  method="class", model=TRUE)
rpart.plot(tree)

```

The variables included in the model are:

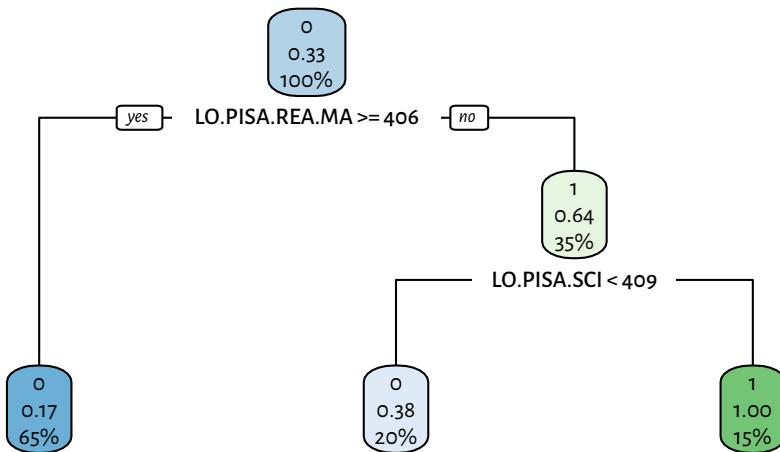
Note that the decision rules are well-interpretable, we can make a whole story around it. Whether or not it is actually true – is a different... story.

To compute the basic classifier performance scores, let's recall the `get_metrics()` function:

```

get_metrics <- function(Y_pred, Y_test)
{
  C <- table(Y_pred, Y_test) # confusion matrix
  stopifnot(dim(C) == c(2, 2))
  c(Acc=(C[1,1]+C[2,2])/sum(C), # accuracy
    Prec=C[2,2]/(C[2,2]+C[2,1]), # precision
    Rec=C[2,2]/(C[2,2]+C[1,2]), # recall
    F=C[2,2]/(C[2,2]+0.5*C[1,2]+0.5*C[2,1]), # F-measure
    # Confusion matrix items:
    TN=C[1,1], FN=C[1,2],
    FP=C[2,1], TP=C[2,2])
}

```

Figure 5.8: A decision tree explaining the `girls_rule_maths` variable

```

    ) # return a named vector
}
  
```

Now we can judge the tree's character:

```

Y_pred <- predict(tree, edstats_subset, type="class")
get_metrics(Y_pred, edstats_subset$girls_rule_maths)
  
```

	Acc	Prec	Rec	F	TN	FN	FP	TP
##	0.81481	1.00000	0.44444	0.61538	54.00000	15.00000	0.00000	12.00000

■

Exercise 5.3 Learn a decision tree that this time doesn't rely on any of the PISA indicators.

Solution.

Let's remove the unwanted variables:

```

edstats_subset <- edstats[!(names(edstats) %in%
  c("LO.PISA.MAT", "LO.PISA.MAT.FE", "LO.PISA.MAT.MA",
  "LO.PISA.REA", "LO.PISA.REA.FE", "LO.PISA.REA.MA",
  "LO.PISA.SCI", "LO.PISA.SCI.FE", "LO.PISA.SCI.MA",
  "CountryName"))]
  
```

On a side note, this could be done more easily by calling, e.g., `stri_startswith_fixed(names(edstats), "LO.PISA")` from the `stringi` package.

Fitting and plotting (see Figure 5.9) of the tree:

```
tree <- rpart(girls_rule_maths~, data=edstats_subset,
               method="class", model=TRUE)
rpart.plot(tree)
```

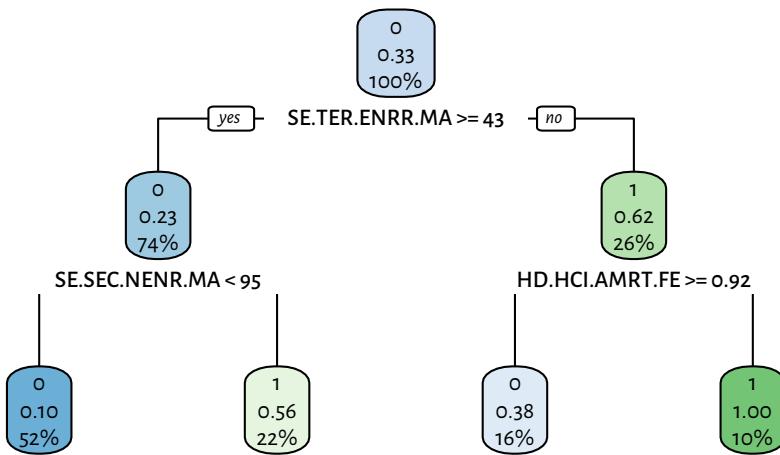


Figure 5.9: Another decision tree explaining the `girls_rule_maths` variable

Performance metrics:

```
Y_pred <- predict(tree, edstats, type="class")
get_metrics(Y_pred, edstats_subset$girls_rule_maths)
```

##	Acc	Prec	Rec	F	TN	FN	FP	TP
##	0.79012	0.69231	0.66667	0.67925	46.00000	9.00000	8.00000	18.00000

It's interesting to note that some of the goodness-of-fit measures are actually higher now.

The variables included in the model are:

5.3.2 EdStats and World Factbook – Joining Forces

In the course of our data science journey, we have considered two datasets dealing with country-level indicators: the World Factbook and World Bank's EdStats.

```
factbook <- read.csv("datasets/world_factbook_2020.csv",
                      comment.char="#")
edstats <- read.csv("datasets/edstats_2019_wide.csv",
                     comment.char="#")
```

Let's combine the information they provide and see if we come up with a better model of where girls' math scores are higher.

Exercise 5.4 Some country names in one dataset don't match those in the other one, for instance: Czech Republic vs. Czechia, Myanmar vs. Burma, etc. Resolve these conflicts as best you can.

Solution.

To get a list of the mismatched country names, we can call either:

```
factbook$country[!(factbook$country %in% edstats$CountryName)]
```

or:

```
edstats$CountryName[!(edstats$CountryName %in% factbook$country)]
```

Unfortunately, the data need to be cleaned manually – it's a tedious task. The following consists of what we hope are the best matches between the two datasets (yet, the list is not perfect; in particular, the Republic of North Macedonia is completely missing in one of the datasets):

```
from_to <- matrix(ncol=2, byrow=TRUE, c(
  # FROM (edstats)                      # TO (factbook)
  "Brunei Darussalam"                  , "Brunei"                ,
  "Congo, Dem. Rep."                  , "Congo, Democratic Republic of the" ,
  "Congo, Rep."                      , "Congo, Republic of the"        ,
  "Czech Republic"                   , "Czechia"                 ,
  "Egypt, Arab Rep."                 , "Egypt"                  ,
  "Hong Kong SAR, China"            , "Hong Kong"               ,
  "Iran, Islamic Rep."              , "Iran"                   ,
  "Korea, Dem. People's Rep."       , "Korea, North"             ,
  "Korea, Rep."                      , "Korea, South"             ,
  "Kyrgyz Republic"                 , "Kyrgyzstan"              ,
  "Lao PDR"                         , "Laos"                   ,
  "Macao SAR, China"                , "Macau"                  ,
  "Micronesia, Fed. Sts."           , "Micronesia, Federated States of" ,
  "Myanmar"                          , "Burma"                  ,
  "Russian Federation"              , "Russia"                 ,
  "Slovak Republic"                 , "Slovakia"                ,
  "St. Kitts and Nevis"             , "Saint Kitts and Nevis"   ,
  "St. Lucia"                        , "Saint Lucia"              ,
  "St. Martin (French part)"        , "Saint Martin"             ,
  "St. Vincent and the Grenadines", "Saint Vincent and the Grenadines" ,
  "Syrian Arab Republic"            , "Syria"                  ,
  "Venezuela, RB"                  , "Venezuela"               ,
  "Virgin Islands (U.S.)"          , "Virgin Islands"          ,
  "Yemen, Rep."                     , "Yemen"                  )
))
```

Conversion of the names:

```
for (i in 1:nrow(from_to)) {
  edstats$CountryName[edstats$CountryName==from_to[i,1]] <- from_to[i,2]
}
```

On a side note (*), this could be done with a single call to a function in the *stringi* package:

```
library("stringi")
edstats$CountryName <- stri_replace_all_fixed(edstats$CountryName,
from_to[,1], from_to[,2], vectorize_all=FALSE)
```



Exercise 5.5 Merge (join) the two datasets based on the country names.

Solution.

This can be done by means of the *merge()* function.

```
edbook <- merge(edstats, factbook, by.x="CountryName", by.y="country")
ncol(edbook) # how many columns we have now
## [1] 157
```



Exercise 5.6 Learn a decision tree that distinguishes between the countries where girls are better at maths than boys and assess the quality of this classifier.

Solution.

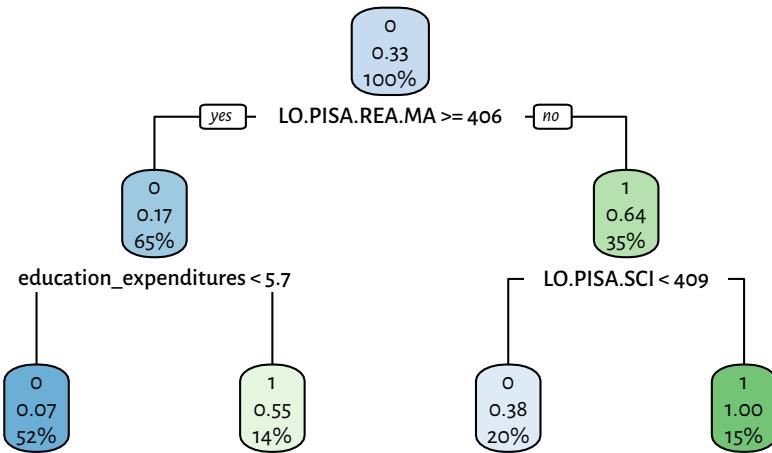
We proceed as in one of the previous exercises:

```
edbook$girls_rule_maths <-
  factor(as.numeric(
    edbook$L0.PISA.MAT.FE>edbook$L0.PISA.MAT.MA
  ))
edbook_subset <- edbook[!(names(edbook) %in%
  c("CountryName", "L0.PISA.MAT.FE", "L0.PISA.MAT.MA"))]
```

Fitting and plotting (see Figure 5.10):

```
library("rpart")
library("rpart.plot")
tree <- rpart(girls_rule_maths~, data=edbook_subset,
  method="class", model=TRUE)
rpart.plot(tree)
```

Performance metrics:

Figure 5.10: Yet another decision tree explaining the `girls_rule_maths` variable

```

Y_pred <- predict(tree, edbook_subset, type="class")
get_metrics(Y_pred, edbook_subset$girls_rule_maths)
  
```

```

##      Acc      Prec      Rec      F      TN      FN      FP      TP
## 0.82716 0.78261 0.66667 0.72000 49.00000 9.00000 5.00000 18.00000
  
```

The variables included in the model are:

This is... not at all enlightening. Rest assured that experts in education or econometrics for whom we work in this (imaginary) project would raise many questions at this very point. Merely applying some computational procedure on a dataset doesn't cut it; it's too early to ask for a paycheque. Classifiers are just blind tools in our gentle yet firm hands; new questions are risen, new answers must be sought. Further explorations are of course left as an exercise to the kind reader.

■

5.3.3 Wine Quality – Random Forest and XGBoost (*)

Let's consider the Wine Quality dataset:

```

wine_quality <- read.csv("datasets/wine_quality_all.csv",
  comment.char="#")
head(wine_quality, 3)

##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.4            0.70       0.00          1.9     0.076
## 2           7.8            0.88       0.00          2.6     0.098
  
```

```

## 3           7.8          0.76        0.04         2.3      0.092
## free.sulfur.dioxide total.sulfur.dioxide density     pH sulphates
## 1            11           34  0.9978 3.51      0.56
## 2            25           67  0.9968 3.20      0.68
## 3            15           54  0.9970 3.26      0.65
## alcohol response color
## 1    9.4      5   red
## 2    9.8      5   red
## 3    9.8      5   red

```

Recall that there are 11 physicochemical features of wines reported (columns 1-11). Moreover, there is a wine rating (variable `response`) on the scale 0 (bad) to 10 (excellent) given by wine experts.

Exercise 5.7 Add a new column named `quality`. A wine should get a `quality` of 1 if its rating is greater than or equal to 7 (a good wine) and a `quality` of 0 otherwise.

Exercise 5.8 Perform a random train-test split of size 60-40%: create the matrices `X_train` and `X_test` containing the 11 physicochemical wine features and the corresponding label vectors `Y_train` and `Y_test` that inform on the wines' quality.

Exercise 5.9 Construct (of course, on the train set) a decision tree that models wine quality as a function the 11 physicochemical features. Assess the quality of the obtained model (of course, based on the test set) by computing the basic performance metrics for this classifier (accuracy, precision, recall, F-measure). Also, print the confusion matrix (see `table()`). Discuss the obtained results. Is this a good model?

Ensemble learning, bagging and random forest. Ensemble learning is a rather simple yet appealing idea that emphasises on the fact that no single model will likely ever be omniscient. Instead, we can construct numerous diverse models explaining a given dependent variable and then properly aggregate the obtained classifier committee to obtain a single answer.

Bagging (bootstrap aggregating) is one of the most popular ensemble types: here, we fit numerous models, each time to a different randomly selected subset of observations in the input dataset. In order to classify a new point, we ask each classifier what it “thinks” the corresponding label should be and report the mode (majority vote) of the candidate outputs.

The famous *random forest* algorithm is an example application of the bagging technique that is based on decision trees (with an extension that not only random subsets of observations, but also random columns in the dataset are used to learn the underlying models).

Exercise 5.10 Ronstruct a random forest that models `quality` as a function of the 11 physicochemical features – see R package `randomForest`. Assess and discuss the classifier's performance.

Boosting. Another powerful ensemble technique is called *boosting*. Here, we construct a supervised learner in an iterative manner; in each step a simple (“weak”) classifier is ad-

ded to the ensemble by learning “something new” about the dataset, e.g., by uplifting its performance in the cases where frequent misclassifications occur. When determining the final outcome, each underlying model might be assigned a different weight.

Exercise 5.11 *The XGBoost algorithm is a modern implementation of the gradient boosting approach based on decision trees. Fit a model with XGBoost (see package `xgboost`) that describes quality as a function of the 11 physicochemical features. Assess and discuss the classifier’s performance.*

5.4 Outro

The state-of-the art classifiers called *Random Forests* and *XGBoost* (see also: *AdaBoost*) are based on decision trees. They tend to be more accurate but – at the same time – they fail to exhibit the decision trees’ important feature: interpretability.

Trees can also be used for regression tasks, see R package `rpart`.

TODO

Recommended further reading: (James et al. 2017: Chapters 4 and 8)

Other: (Hastie et al. 2017: Chapters 4 and 7 as well as (*) Chapters 9, 10, 13, 15)

Next Chapter

6

Simple Linear Regression

TODO In this chapter, we will:

- ...
- ...

overfitting, generalisation

6.1 Simple Regression

6.1.1 Introduction

6.1.2 Side Note: K-NN Regression

TODO: later

The K-Nearest Neighbour scheme is intuitively pleasing.

No wonder it has inspired a similar approach for solving a regression task.

In order to make a prediction for a new point \mathbf{x}' :

1. find the K-nearest neighbours of \mathbf{x}' amongst the points in the train set, denoted $\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_K}$,
2. fetch the corresponding reference outputs y_{i_1}, \dots, y_{i_K} ,
3. return their arithmetic mean as a result,

$$\hat{y} = \frac{1}{K} \sum_{j=1}^K y_{i_j}.$$

Recall our modelling of the Credit Rating (Y) as a function of the average Credit Card Balance (X) based on the `ISLR::Credit` dataset.

```
library("ISLR") # Credit dataset
Xc <- as.matrix(as.numeric(Credit$Balance[Credit$Balance>0]))
Yc <- as.matrix(as.numeric(Credit$Rating[Credit$Balance>0]))
```

```
library("FNN") # knn.reg function
x <- as.matrix(seq(min(Xc), max(Xc), length.out=101))
y1 <- knn.reg(Xc, x, Yc, k=1)$pred
y5 <- knn.reg(Xc, x, Yc, k=5)$pred
y25 <- knn.reg(Xc, x, Yc, k=25)$pred
```

The three models are depicted in Figure 6.1. Again, the higher the K , the smoother the curve. On the other hand, for small K we adapt better to what's in a point's neighbourhood.

```
plot(Xc, Yc, col="#666666c0",
      xlab="Balance", ylab="Rating")
lines(x, y1, col=2, lwd=3)
lines(x, y5, col=3, lwd=3)
lines(x, y25, col=4, lwd=3)
legend("topleft", legend=c("K=1", "K=5", "K=25"),
       col=c(2, 3, 4), lwd=3, bg="white")
```

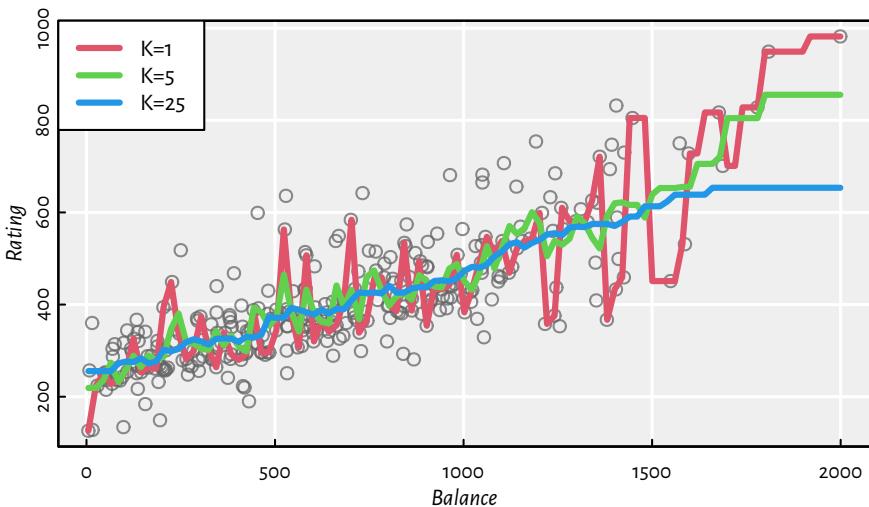


Figure 6.1: K-NN regression example

Simple regression is the easiest setting to start with – let's assume $p = 1$, i.e., all inputs are 1-dimensional. Denote $x_i = x_{i,1}$.

We will use it to build many intuitions, for example, it'll be easy to illustrate all the concepts graphically.

```
library("ISLR") # Credit dataset
plot(Credit$Balance, Credit$Rating) # scatter plot
```

In what follows we will be modelling the Credit Rating (Y) as a function of the average

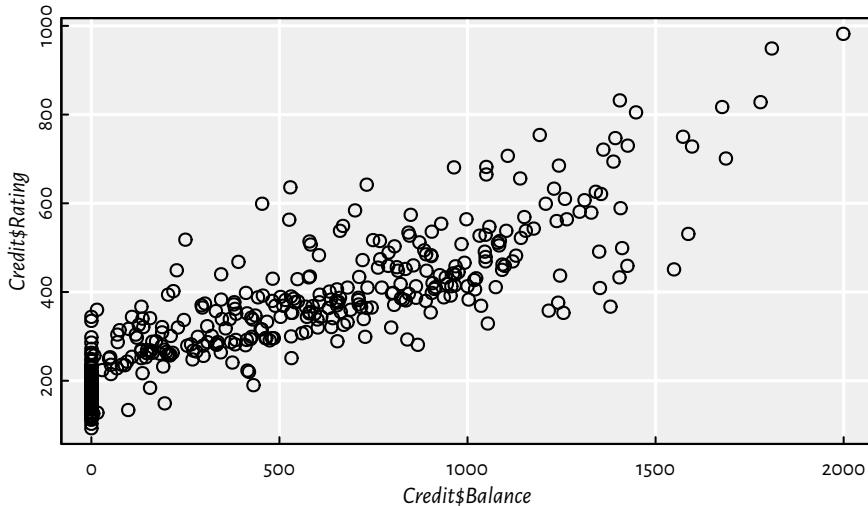


Figure 6.2: (#fig:credit_scatter) A scatter plot of Rating vs. Balance

Credit Card Balance (X) in USD for customers with *positive Balance only*. It is because it is evident from Figure @ref(fig:credit_scatter) that some customers with zero balance obtained a credit rating based on some external data source that we don't have access to in our very setting.

```
X <- as.matrix(as.numeric(Credit$Balance[Credit$Balance>0]))
Y <- as.matrix(as.numeric(Credit$Rating[Credit$Balance>0]))
```

Figure @ref(fig:credit_XY_plot) gives the updated scatter plot with the zero-balance clients “taken care of”.

```
plot(X, Y, xlab="X (Balance)", ylab="Y (Rating)")
```

Our aim is to construct a function f that **models** Rating as a function of Balance, $f(X) = Y$.

We are equipped with $n = 310$ reference (observed) Ratings $\mathbf{y} = [y_1 \dots y_n]^T$ for particular Balances $\mathbf{x} = [x_1 \dots x_n]^T$.

Note the following naming conventions:

- Variable types:
 - X – independent/explanatory/predictor variable
 - Y – dependent/response/predicted variable
- Also note that:
 - \hat{Y} – idealisation (any possible Rating)

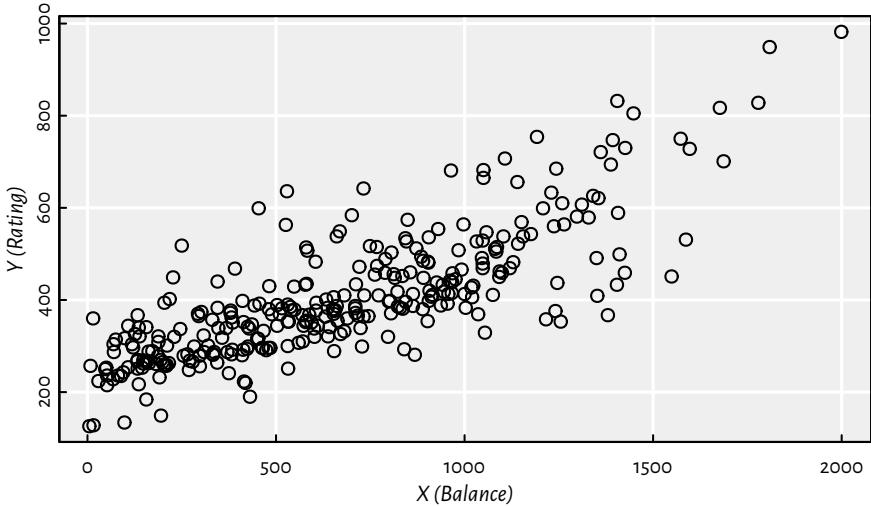


Figure 6.3: (#fig:credit_XY_plot) A scatter plot of Rating vs. Balance with clients of Balance=0 removed

– $\mathbf{y} = [y_1 \cdots y_n]^T$ – values actually observed

The model will not be ideal, but it might be usable:

- We will be able to **predict** the rating of any new client.

What should be the rating of a client with Balance of \$1500?

What should be the rating of a client with Balance of \$2500?

- We will be able to **describe** (understand) this reality using a single mathematical formula so as to infer that there is an association between X and Y

Think of “data compression” and laws of physics, e.g., $E = mc^2$.

(*) Mathematically, we will assume that there is some “true” function that models the data (true relationship between Y and X), but the observed outputs are subject to **additive error**:

$$Y = f(X) + \varepsilon.$$

ε is a random term, classically we assume that errors are independent of each other, have expected value of 0 (there is no systematic error = unbiased) and that they follow a normal distribution.

(*) We denote this as $\varepsilon \sim \mathcal{N}(0, \sigma)$ (read: random variable ε follows a normal distribution with expected value of 0 and standard deviation of σ for some $\sigma \geq 0$).

σ controls the amount of noise (and hence, uncertainty). Figure

@ref{fig:normal_distrib} gives the plot of the probability distribution function (PDFs, densities) of $\mathcal{N}(0, \sigma)$ for different σ s:

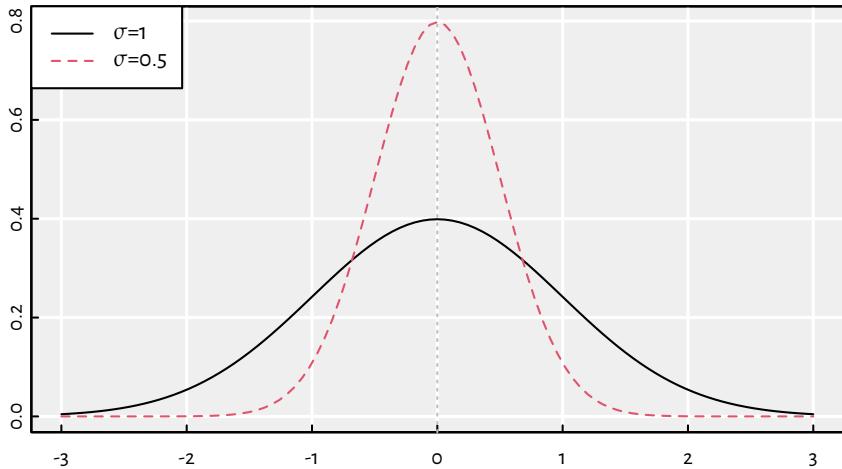


Figure 6.4: (#fig:normal_distrib) Probability density functions of normal distributions with different standard deviations σ .

6.1.3 Search Space and Objective

There are many different functions that can be **fitted** into the observed (\mathbf{x}, \mathbf{y}) , compare Figure @ref{fig:credit_different_models}. Some of them are better than the other (with respect to different aspects, such as fit quality, simplicity etc.).

Thus, we need a formal **model selection criterion** that could enable us to tackle the model fitting task on a computer.

Usually, we will be interested in a model that minimises appropriately aggregated **residuals** $f(x_i) - y_i$, i.e., **predicted outputs minus observed outputs**, often denoted with $\hat{y}_i - y_i$, for $i = 1, \dots, n$.

In Figure @ref{fig:credit_residuals}, the residuals correspond to the lengths of the dashed line segments – they measure the discrepancy between the outputs generated by the model (what we get) and the true outputs (what we want).

Top choice: sum of squared residuals:

$$\begin{aligned} \text{SSR}(f|\mathbf{x}, \mathbf{y}) &= (f(x_1) - y_1)^2 + \dots + (f(x_n) - y_n)^2 \\ &= \sum_{i=1}^n (f(x_i) - y_i)^2 \end{aligned}$$

Remark. The notation $\text{SSR}(f|\mathbf{x}, \mathbf{y})$ means that it is the error measure corresponding to

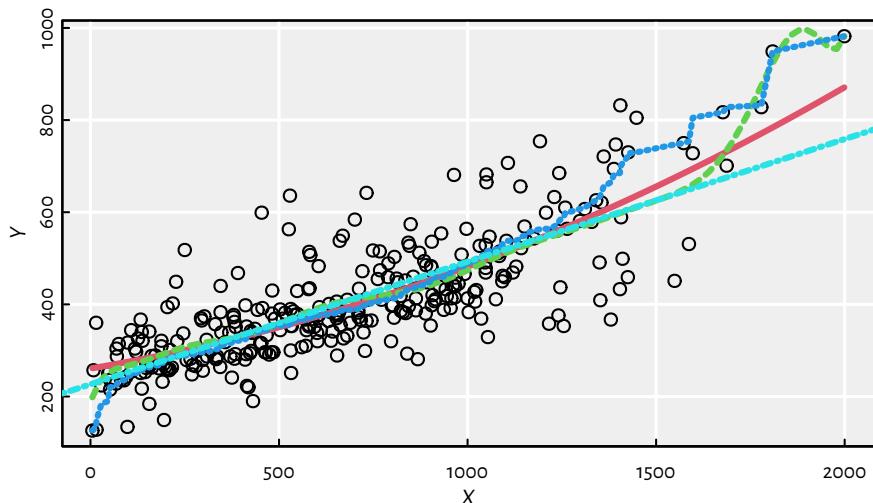


Figure 6.5: (#fig:credit_different_models) Different polynomial models fitted to data

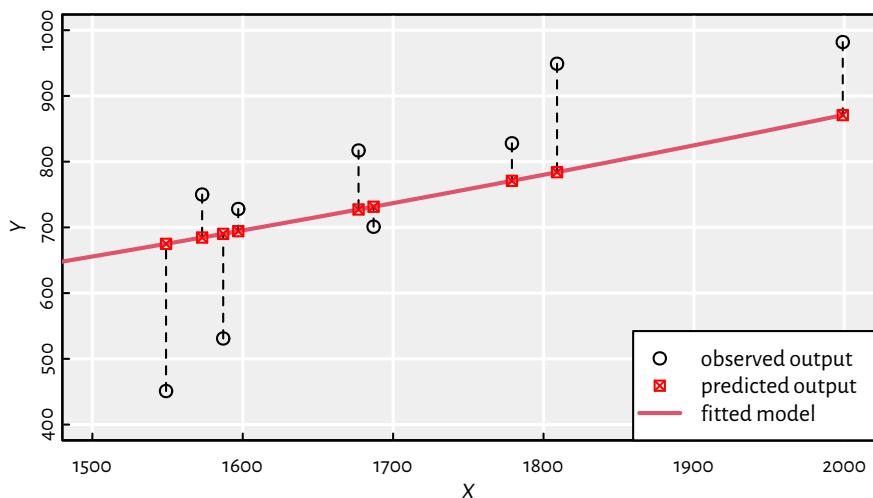


Figure 6.6: (#fig:credit_residuals) Residuals are defined as the differences between the predicted and observed outputs $\hat{y}_i - y_i$

the model (f) given our data.

We could've denoted it with $\text{SSR}_{\mathbf{x}, \mathbf{y}}(f)$ or even $\text{SSR}(f)$ to emphasise that \mathbf{x}, \mathbf{y} are just fixed values and we are not interested in changing them at all (they are “global variables”).

We enjoy SSR because (amongst others):

- larger errors are penalised much more than smaller ones
-

(this can be considered a drawback as well)

- (***) statistically speaking, this has a clear underlying interpretation
-

(assuming errors are normally distributed, finding a model minimising the SSR is equivalent to maximum likelihood estimation)

- the models minimising the SSR can often be found easily
-

(corresponding optimisation tasks have an analytic solution – studied already by Gauss in the late 18th century)

(***) Other choices:

- regularised SSR, e.g., lasso or ridge regression (in the case of multiple input variables)
- sum or median of absolute values (robust regression)

Fitting a model to data can be written as an optimisation problem:

$$\min_{f \in \mathcal{F}} \text{SSR}(f | \mathbf{x}, \mathbf{y}),$$

i.e., find f minimising the SSR (**seek “best” f**) amongst the set of admissible models \mathcal{F} .

Example \mathcal{F} s:

- $\mathcal{F} = \{\text{All possible functions of one variable}\}$ – if there are no repeated x_i 's, this corresponds to data *interpolation*; note that there are many functions that give SSR of 0.
- $\mathcal{F} = \{x \mapsto x^2, x \mapsto \cos(x), x \mapsto \exp(2x + 7) - 9\}$ – obviously an ad-hoc choice but you can easily choose the best amongst the 3 by computing 3 sums of squares.
- $\mathcal{F} = \{x \mapsto a + bx\}$ – the space of linear functions of one variable
- etc.

(e.g., $x \mapsto x^2$ is read “ x maps to x^2 ” and is an elegant way to define an inline function f such that $f(x) = x^2$)

6.2 Simple Linear Regression

6.2.1 Introduction

If the family of admissible models \mathcal{F} consists only of all linear functions of one variable, we deal with a **simple linear regression**.

Our problem becomes:

$$\min_{a,b \in \mathbb{R}} \sum_{i=1}^n (a + bx_i - y_i)^2$$

In other words, we seek best fitting line in terms of the squared residuals.

This is the **method of least squares**.

This is particularly nice, because our search space is just \mathbb{R}^2 – easy to handle both analytically and numerically.

Exercise 6.1 Which of lines in Figure @ref(fig:credit_different_lines_ss) is the least squares solution?

6.2.2 Solution in R

Let's fit the linear model minimising the SSR in R. The `lm()` function (linear models) has a convenient formula-based interface.

```
f <- lm(Y~X)
```

In R, the expression “`Y~X`” denotes a formula, which we read as: variable `Y` is a function of `X`. Note that the dependent variable is on the left side of the formula. Here, `X` and `Y` are two R numeric vectors of identical lengths.

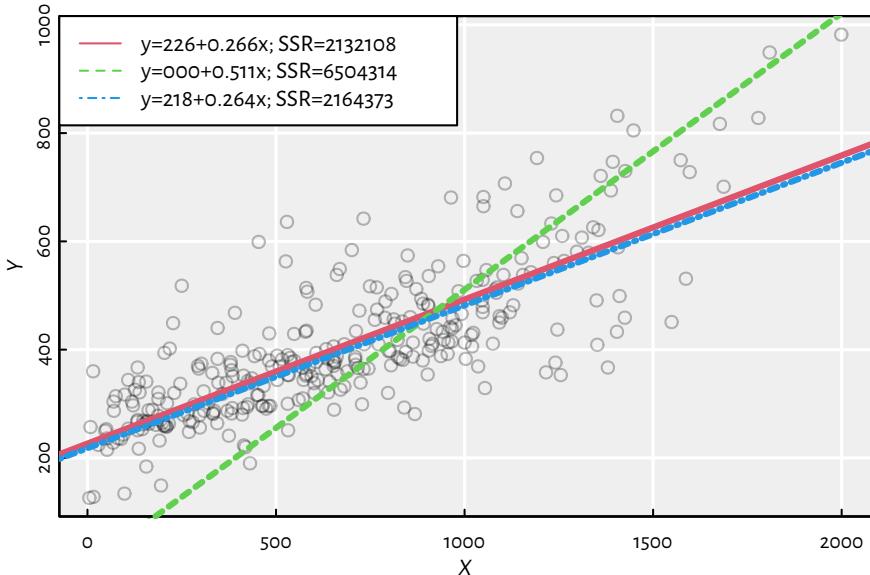


Figure 6.7: (#fig:credit_different_lines_ssrs) Three simple linear models together with the corresponding SSRs

Let's print the fitted model:

```
print(f)

##
## Call:
## lm(formula = Y ~ X)
##
## Coefficients:
## (Intercept)          X
##     226.471           0.266
```

Hence, the fitted model is:

$$Y = f(X) = 226.47114 + 0.26615X \quad (+\varepsilon)$$

Coefficient a (intercept):

```
f$coefficient[1]
```

```
## (Intercept)
##     226.47
```

Coefficient b (slope):

```
f$coefficient[2]
```

```
##      X
## 0.26615
```

Plotting, see Figure @ref(fig:credit_plot_lm):

```
plot(X, Y, col="#000000aa")
abline(f, col=2, lwd=3)
```

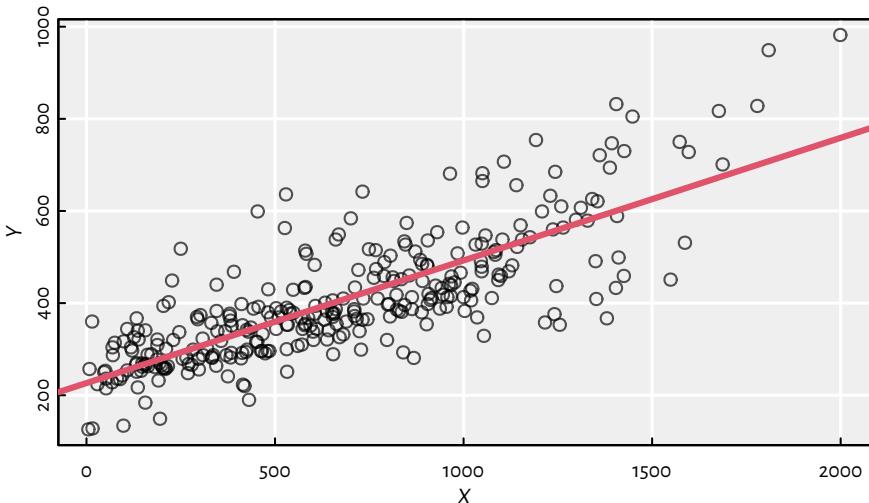


Figure 6.8: (#fig:credit_plot_lm) Fitted regression line

SSR:

```
sum(f$residuals^2)

## [1] 2132108
sum((f$coefficient[1]+f$coefficient[2]*X-Y)^2) # equivalent

## [1] 2132108
```

We can predict the model's output for yet-unobserved inputs by writing:

```
X_new <- c(1500, 2000, 2500) # example inputs
f$coefficient[1] + f$coefficient[2]*X_new

## [1] 625.69 758.76 891.84
```

Note that linear models can also be fitted based on formulas that refer to a data frame's columns. For example, let us wrap both `x` and `y` inside a data frame:

```
XY <- data.frame(Balance=X, Rating=Y)
head(XY, 3)
```

```
##   Balance Rating
## 1      333    283
## 2      903    483
## 3      580    514
```

By writing:

```
f <- lm(Rating~Balance, data=XY)
```

now Balance and Rating refer to column names in the XY data frame, and not the objects in R's "workspace".

Based on the above, we can make a prediction using the `predict()` function"

```
X_new <- data.frame(Balance=c(1500, 2000, 2500))
predict(f, X_new)
```

```
##      1      2      3
## 625.69 758.76 891.84
```

Interestingly:

```
predict(f, data.frame(Balance=c(5000)))
```

```
##      1
## 1557.2
```

This is more than the highest possible rating – we have extrapolated way beyond the observable data range.

Note that our $Y = a + bX$ model is **interpretable** and **well-behaving** (not all machine learning models will have this feature, think: deep neural networks, which we rather conceive as *black boxes*):

- we know that by increasing X by a small amount, Y will also increase (positive correlation),
- the model is continuous – small change in X doesn't yield any drastic change in Y ,
- we know what will happen if we increase or decrease X by, say, 100,
- the function is invertible – if we want Rating of 500, we can compute the associated preferred Balance that should yield it (provided that the model is valid).

6.2.3 Analytic Solution

It may be shown (which we actually do below) that the solution is:

$$\begin{cases} b = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n y_i \sum_{i=1}^n x_i}{n \sum_{i=1}^n x_i x_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i} \\ a = \frac{1}{n} \sum_{i=1}^n y_i - b \frac{1}{n} \sum_{i=1}^n x_i \end{cases}$$

Which can be implemented in R as follows:

```
n <- length(X)
b <- (n*sum(X*Y)-sum(X)*sum(Y))/(n*sum(X*X)-sum(X)^2)
a <- mean(Y)-b*mean(X)
c(a, b) # the same as f$coefficients

## [1] 226.47114 0.26615
```

6.2.4 Derivation of the Solution (**)

Remark. You can safely skip this part if you are yet to know how to search for a minimum of a function of many variables and what are partial derivatives.

Denote with:

$$E(a, b) = \text{SSR}(x \mapsto a + bx | \mathbf{x}, \mathbf{y}) = \sum_{i=1}^n (a + bx_i - y_i)^2.$$

We seek the minimum of E w.r.t. both a, b .

Theorem. If E has a (local) minimum at (a^*, b^*) , then its partial derivatives vanish therein, i.e., $\partial E / \partial a(a^*, b^*) = 0$ and $\partial E / \partial b(a^*, b^*) = 0$.

We have:

$$E(a, b) = \sum_{i=1}^n (a + bx_i - y_i)^2.$$

We need to compute the partial derivatives $\partial E / \partial a$ (derivative of E w.r.t. variable a – all other terms treated as constants) and $\partial E / \partial b$ (w.r.t. b).

Useful rules – derivatives w.r.t. a (denote $f'(a) = (f(a))'$):

- $(f(a) + g(a))' = f'(a) + g'(a)$ (derivative of sum is sum of derivatives)
- $(f(a)g(a))' = f'(a)g(a) + f(a)g'(a)$ (derivative of product)
- $(f(g(a)))' = f'(g(a))g'(a)$ (chain rule)
- $(c)' = 0$ for any constant c (expression not involving a)
- $(a^p)' = pa^{p-1}$ for any p

- in particular: $(ca^2 + d)' = 2ca$, $(ca)' = c$, $((ca + d)^2)' = 2(ca + d)c$ (application of the above rules)

We seek a, b such that $\frac{\partial E}{\partial a}(a, b) = 0$ and $\frac{\partial E}{\partial b}(a, b) = 0$.

$$\begin{cases} \frac{\partial E}{\partial a}(a, b) = 2 \sum_{i=1}^n (a + bx_i - y_i) = 0 \\ \frac{\partial E}{\partial b}(a, b) = 2 \sum_{i=1}^n (a + bx_i - y_i) x_i = 0 \end{cases}$$

This is a system of 2 linear equations. Easy.

Rearranging like back in the school days:

$$\begin{cases} b \sum_{i=1}^n x_i + an = \sum_{i=1}^n y_i \\ b \sum_{i=1}^n x_i x_i + a \sum_{i=1}^n x_i = \sum_{i=1}^n x_i y_i \end{cases}$$

It is left as an exercise to show that the solution is:

$$\begin{cases} b^* = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n y_i \sum_{i=1}^n x_i}{n \sum_{i=1}^n x_i x_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i} \\ a^* = \frac{1}{n} \sum_{i=1}^n y_i - b^* \frac{1}{n} \sum_{i=1}^n x_i \end{cases}$$

(we should additionally perform the second derivative test to assure that this is the minimum of E – which is exactly the case though)

(**) In the next chapter, we will introduce the notion of Pearson's linear coefficient, r (see `cor()` in R). It might be shown that a and b can also be rewritten as:

```
(b <- cor(X,Y)*sd(Y)/sd(X))
```

```
## [1] 0.26615
(a <- mean(Y)-b*mean(X))

## [1] 226.47
```

6.3 Exercises

6.3.1 The Anscombe Quartet

Here is a famous illustrative example proposed by the statistician Francis Anscombe in the early 1970s.

```
print(anscombe) # `anscombe` is a built-in object
```

```
##   x1 x2 x3 x4     y1    y2    y3    y4
## 1 10 10 10  8 8.04 9.14 7.46 6.58
## 2  8  8  8  8 6.95 8.14 6.77 5.76
## 3 13 13 13  8 7.58 8.74 12.74 7.71
## 4  9  9  9  8 8.81 8.77 7.11 8.84
## 5 11 11 11  8 8.33 9.26 7.81 8.47
## 6 14 14 14  8 9.96 8.10 8.84 7.04
## 7  6  6  6  8 7.24 6.13 6.08 5.25
## 8  4  4  4 19 4.26 3.10 5.39 12.50
## 9 12 12 12  8 10.84 9.13 8.15 5.56
## 10 7  7  7  8 4.82 7.26 6.42 7.91
## 11 5  5  5  8 5.68 4.74 5.73 6.89
```

What we see above is a single data frame that encodes four separate datasets: `anscombe$x1` and `anscombe$y1` define the first pair of variables, `anscombe$x2` and `anscombe$y2` define the second pair and so forth.

Exercise 6.2 Split the above data (manually) into four data frames `ans1`, ..., `ans4` with columns `x` and `y`.

For example, `ans1` should look like:

```
print(ans1)
```

```
##   x     y
## 1 10 8.04
## 2  8 6.95
## 3 13 7.58
## 4  9 8.81
## 5 11 8.33
## 6 14 9.96
## 7  6 7.24
## 8  4 4.26
## 9 12 10.84
## 10 7 4.82
## 11 5 5.68
```

Solution.

```
ans1 <- data.frame(x=anscombe$x1, y=anscombe$y1)
ans2 <- data.frame(x=anscombe$x2, y=anscombe$y2)
ans3 <- data.frame(x=anscombe$x3, y=anscombe$y3)
ans4 <- data.frame(x=anscombe$x4, y=anscombe$y4)
print(ans1)
```



Exercise 6.3 Compute the mean of each x and y variable.

Solution.

```
mean(ans1$x) # individual column
```

```
## [1] 9
```

```
mean(ans1$y) # individual column
```

```
## [1] 7.5009
```

```
sapply(ans2, mean) # all columns in ans2
```

```
##      x      y
```

```
## 9.0000 7.5009
```

```
sapply(anscombe, mean) # all columns in the full anscombe dataset
```

```
##      x1      x2      x3      x4      y1      y2      y3      y4
```

```
## 9.0000 9.0000 9.0000 9.0000 7.5009 7.5009 7.5000 7.5009
```

Comment: This is really interesting, all the means of x columns as well as the means of y s are almost identical.



Exercise 6.4 Compute the standard deviation of each x and y variable.

Solution.

The solution is similar to the previous one, just replace `mean` with `sd`. Here, to learn something new, we will use the `knitr::kable()` function that pretty-prints a given matrix or data frame:

```
results <- sapply(anscombe, sd)
knitr::kable(results, col.names="standard deviation")
```

standard deviation	
x_1	3.3166
x_2	3.3166
x_3	3.3166

	standard deviation
x4	3.3166
y1	2.0316
y2	2.0317
y3	2.0304
y4	2.0306

Comment: This is even more interesting, because the numbers agree up to 2 decimal digits.



Exercise 6.5 Fit a simple linear regression model for each data set. Draw the scatter plots again (`plot()`) and add the regression lines (`lines()` or `abline()`).

Solution.

To recall, this can be done with the `lm()` function explained in Lecture 2.

At this point we should already have become lazy – the tasks are very repetitious. Let's automate them by writing a single function that does all the above for any data set:

```
fit_models <- function(ans) {
  # ans is a data frame with columns x and y
  f <- lm(y~x, data=ans) # fit linear model
  print(f$coefficients) # estimated coefficients
  plot(ans$x, ans$y) # scatter plot
  abline(f, col="red") # regression line
  return(f)
}
```

Now we can apply it on the four particular examples.

```
par(mfrow=c(2, 2)) # four plots on 1 figure (2x2 grid)
f1 <- fit_models(ans1)
```

```
## (Intercept)          x
##     3.00009      0.50009
f2 <- fit_models(ans2)
```

```
## (Intercept)          x
##     3.0009      0.5000
f3 <- fit_models(ans3)
```

```
## (Intercept)          x
##     3.00245      0.49973
```

```
f4 <- fit_models(ans4)
```

```
## (Intercept)          x
##      3.00173     0.49991
```

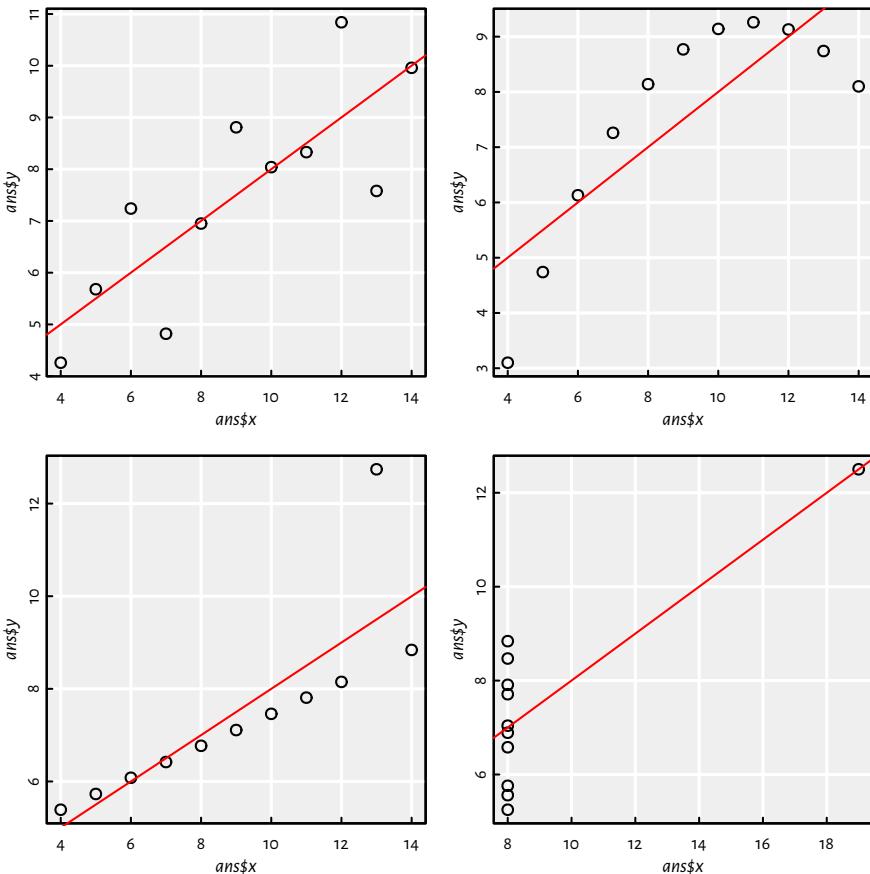


Figure 6.9: (#fig:anscombe_fit_apply) Fitted regression lines for the Anscombe quartet

Comment: All the estimated models are virtually the same, the regression lines are $y = 0.5x + 3$, compare Figure @ref(fig:anscombe_fit_apply).

Exercise 6.6 Create scatter plots of the residuals (\hat{y}_i minus true y_i) as a function of the predicted $\hat{y}_i = f(x_i)$ for every $i = 1, \dots, 11$.

Solution.

To recall, the model predictions can be generated by (amongst others) calling the `predict()` function.

```
y_pred1 <- f1$fitted.values # predict(f1, ans1)
y_pred2 <- f2$fitted.values # predict(f2, ans2)
y_pred3 <- f3$fitted.values # predict(f3, ans3)
y_pred4 <- f4$fitted.values # predict(f4, ans4)
```

Plots of residuals as a function of the predicted (fitted) values are given in Figure @reff{fig:anscombe_resid_plot}.

```
par(mfrow=c(2, 2)) # four plots on 1 figure (2x2 grid)
plot(y_pred1, y_pred1-ans1$y)
plot(y_pred2, y_pred2-ans2$y)
plot(y_pred3, y_pred3-ans3$y)
plot(y_pred4, y_pred4-ans4$y)
```

Comment: Ideally, the residuals shouldn't be correlated with the predicted values – they should “oscillate” randomly around 0. This is only the case of the first dataset. All the other cases are “alarming” in the sense that they suggest that the obtained models are “suspicious” (perhaps data cleansing is needed or a linear model is not at all appropriate).

Exercise 6.7 Draw conclusions (in your own words). ■

Solution.

We're being taught a lesson here: don't perform data analysis tasks automatically, don't look at bare numbers only, visualise your data first! ■

Exercise 6.8 Read more about Anscombe's quartet at https://en.wikipedia.org/wiki/Anscombe%27s_quartet

6.3.2 Median House Value in Boston

The famous Boston dataset from the MASS package records the historical (in the 1970s) median house value (`medv` column, in 1000s USD) for 506 suburbs around Boston, MA, USA.

You can access the dataset by calling:

```
# call install.packages("MASS") first (only once)
library("MASS")
head(Boston, 3)

##      crim zn indus chas   nox     rm    age     dis   rad tax ptratio   black
## 1 0.00632 18  2.31     0 0.538 6.575 65.2 4.0900    1 296    15.3 396.90
```

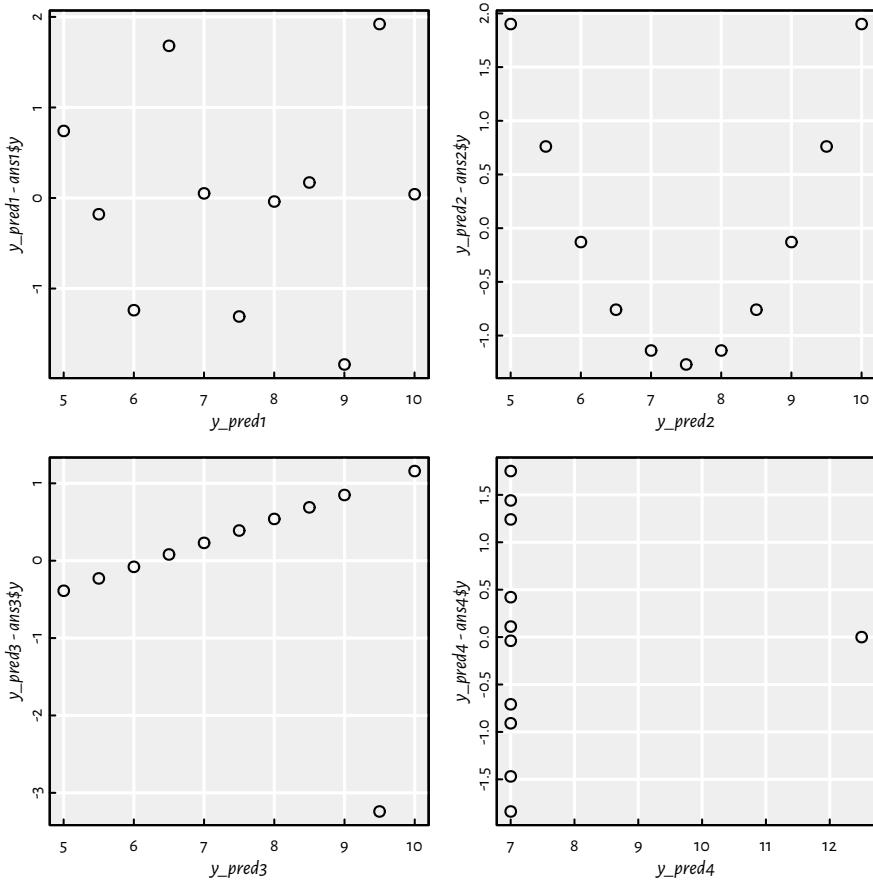


Figure 6.10: (#fig:anscombe_resid_plot) Residuals vs. fitted values for the regression lines fitted to the Anscombe quartet

```
## 2 0.02731 0 7.07 0 0.469 6.421 78.9 4.9671 2 242 17.8 396.90
## 3 0.02729 0 7.07 0 0.469 7.185 61.1 4.9671 2 242 17.8 392.83
##   lstat medv
## 1 4.98 24.0
## 2 9.14 21.6
## 3 4.03 34.7
```

Read the description of each of the 14 columns in the dataset's manual, see `?Boston`.

Exercise 6.9 Fit a simple linear model of $Y = \text{medv}$ as a function of $X = \text{lstat}$.

Exercise 6.10 Fit a quadratic polynomial model ($aX^2 + bX + c$) for the same pair of variables.

Exercise 6.11 Fit a 5-degree polynomial model ($aX^5 + bX^4 + cX^3 + dX^2 + eX + f$) for the same pair of variables.

Exercise 6.12 Draw the scatter plot of the two variables and add the fitted regression curves (all three on a single plot, use different colours).

Exercise 6.13 Compute RMSE, MAE and R^2 of each model. Provide the interpretations of the obtained values.

Exercise 6.14 For each model, draw the plot of the residuals ($\hat{y}_i - y_i$) as a function of the predicted outputs (\hat{y}_i). Describe these plots in your own words.

Exercise 6.15 Predict the medv values for lstat of 0, 25, 50 and 75 using all the models. Compare and discuss the results. Which of the predictions seem trustworthy?

6.4 Outro

In supervised learning, with each input point, there's an associated reference output value.

Learning a model = constructing a function that approximates (minimising some error measure) the given data.

Regression = the output variable Y is continuous.

We studied linear models with a single independent variable based on the least squares (SSR) fit.

In the next part we will extend this setting to the case of many variables, i.e., $p > 1$, called multiple regression.

TODO

Recommended further reading: (James et al. 2017: Chapters 1, 2 and 3)

Other: (Hastie et al. 2017: Chapter 1, Sections 3.2 and 3.3)

Next Chapter....

Multiple Regression

TODO In this chapter, we will:

- ...
 - ...
-

7.1 Introduction

7.1.1 Formalism

Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be an input matrix that consists of n points in a p -dimensional space.

In other words, we have a database on n objects, each of which being described by means of p numerical features.

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{bmatrix}$$

Recall that in supervised learning, apart from \mathbf{X} , we are also given the corresponding \mathbf{y} ; with each input point \mathbf{x}_i , we associate the desired output y_i .

In this chapter we are still interested in **regression** tasks; hence, we assume that each y_i it is a real number, i.e., $y_i \in \mathbb{R}$.

Hence, our dataset is $[\mathbf{X} \mathbf{y}]$ – where each object is represented as a row vector $[\mathbf{x}_i, y_i]$, $i = 1, \dots, n$:

$$[\mathbf{X} \mathbf{y}] = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} & y_1 \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} & y_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} & y_n \end{bmatrix}.$$

7.1.2 Simple Linear Regression - Recap

In a simple regression task, we have assumed that $p = 1$ – there is only one independent variable, denoted $x_i = x_{i,1}$.

We restricted ourselves to linear models of the form $Y = f(X) = a + bX$ that minimised the sum of squared residuals (SSR), i.e.,

$$\min_{a,b \in \mathbb{R}} \sum_{i=1}^n (a + bx_i - y_i)^2.$$

The solution is:

$$\begin{cases} b = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n y_i \sum_{i=1}^n x_i}{n \sum_{i=1}^n x_i x_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i} \\ a = \frac{1}{n} \sum_{i=1}^n y_i - b \frac{1}{n} \sum_{i=1}^n x_i \end{cases}$$

Fitting in R can be performed by calling the `lm()` function:

```
library("ISLR") # Credit dataset
X <- as.numeric(Credit$Balance[Credit$Balance>0])
Y <- as.numeric(Credit$Rating[Credit$Balance>0])
f <- lm(Y~X) # Y~X is a formula, read: Y is a function of X
print(f)

##
## Call:
## lm(formula = Y ~ X)
##
## Coefficients:
## (Intercept)          X
##       226.471        0.266
```

Figure @ref(fig:simple_recap2) gives the scatter plot of Y vs. X together with the fitted simple linear model.

```
plot(X, Y, xlab="X (Balance)", ylab="Y (Credit)")
abline(f, col=2, lwd=3)
```

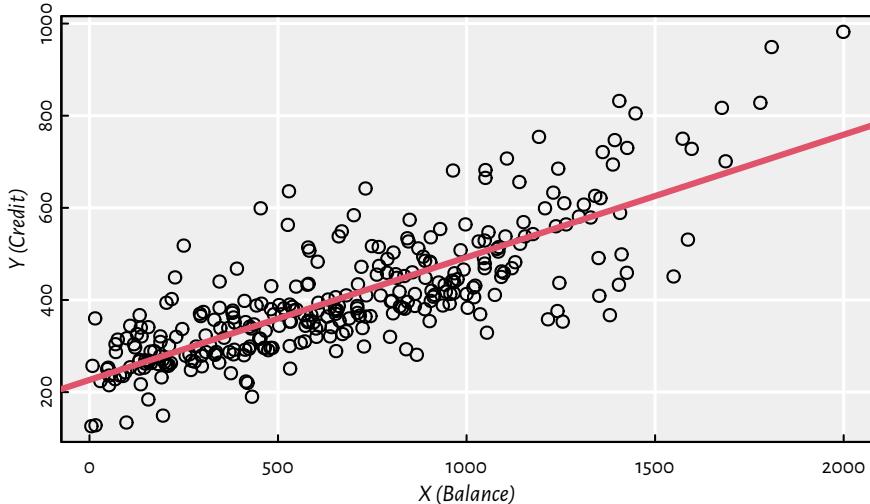


Figure 7.1: (#fig:simple_recap2) Fitted regression line for the Credit dataset

7.2 Multiple Linear Regression

7.2.1 Problem Formulation

Let's now generalise the above to the case of many variables X_1, \dots, X_p .

We wish to model the dependent variable as a function of p independent variables.

$$Y = f(X_1, \dots, X_p) \quad (+\varepsilon)$$

Restricting ourselves to the class of **linear models**, we have

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p.$$

Above we studied the case where $p = 1$, i.e., $Y = a + bX_1$ with $\beta_0 = a$ and $\beta_1 = b$.

The above equation defines:

- $p = 1$ — a line (see Figure @ref(fig:simple_recap2)),
- $p = 2$ — a plane (see Figure 7.2),
- $p \geq 3$ — a hyperplane (well, most people find it difficult to imagine objects in high dimensions, but we are lucky to have this thing called maths).

7.2.2 Fitting a Linear Model in R

`lm()` accepts a formula of the form $Y \sim X_1 + X_2 + \dots + X_p$.

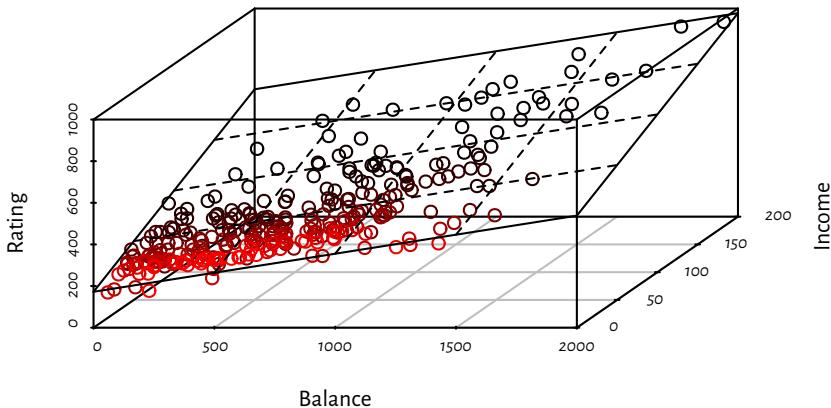


Figure 7.2: Fitted regression plane for the Credit dataset

It finds the least squares fit, i.e., solves

$$\min_{\beta_0, \beta_1, \dots, \beta_p \in \mathbb{R}} \sum_{i=1}^n (\beta_0 + \beta_1 x_{i,1} + \dots + \beta_p x_{i,p} - y_i)^2$$

```
X1 <- as.numeric(Credit$Balance[Credit$Balance>0])
X2 <- as.numeric(Credit$Income[Credit$Balance>0])
Y <- as.numeric(Credit$Rating[Credit$Balance>0])
f <- lm(Y~X1+X2)
f$coefficients # β₀, β₁, β₂
```

## (Intercept)	X1	X2
## 172.5587	0.1828	2.1976

By the way, the 3D scatter plot in Figure 7.2 was generated by calling:

```
library("scatterplot3d")
s3d <- scatterplot3d(X1, X2, Y,
  angle=60, # change angle to reveal more
  highlight.3d=TRUE, xlab="Balance", ylab="Income",
  zlab="Rating")
s3d$plane3d(f, lty.box="solid")
```

(s3d is an R list, one of its elements named plane3d is a function object – this is legal)

7.3 Finding the Best Model

7.3.1 Model Diagnostics

Here is Rating (Y) as function of Balance (X_1 , lefthand side of Figure @ref(fig:x12_y)) and Income (X_2 , righthand side of Figure @ref(fig:x12_y)).

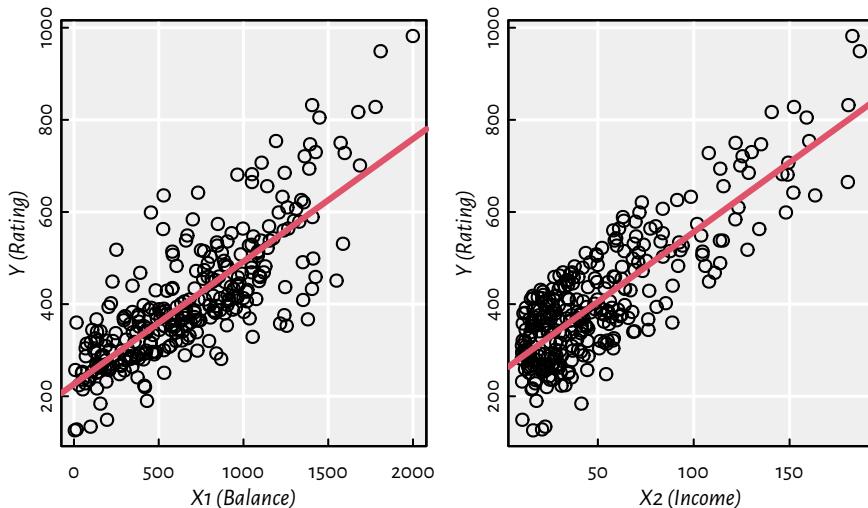


Figure 7.3: (#fig:x12_y) Scatter plots of Y vs. X_1 and X_2

Moreover, Figure @ref(fig:x12_ycolmap) depicts (in a hopefully readable manner) both X_1 and X_2 with Rating Y encoded with a colour (low ratings are green, high ratings are red; some rating values are explicitly printed out within the plot).

Consider the three following models.

Formula	Equation
Rating ~ Balance + Income	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
Rating ~ Balance	$Y = a + bX_1 (\beta_0 = a, \beta_1 = b, \beta_2 = 0)$
Rating ~ Income	$Y = a + bX_2 (\beta_0 = a, \beta_1 = 0, \beta_2 = b)$

```
f12 <- lm(Y~X1+X2) # Rating ~ Balance + Income
f12$coefficients
```

```
## (Intercept)          X1          X2
##     172.5587     0.1828     2.1976
```

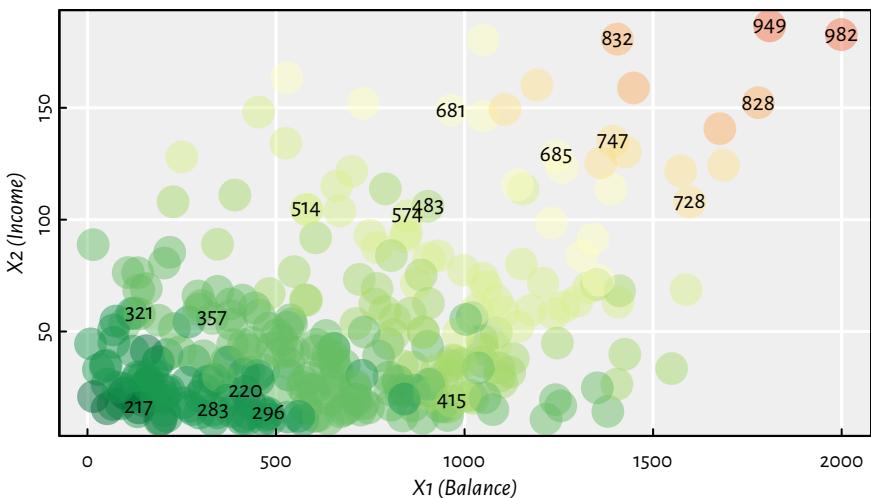


Figure 7.4: (#fig:x12_ycolmap) A heatmap for Rating as a function of Balance and Income; greens represent low credit ratings, whereas reds – high ones

```
f1 <- lm(Y~X1)      # Rating ~ Balance
f1$coefficients

## (Intercept)      X1
## 226.47114     0.26615

f2 <- lm(Y~X2)      # Rating ~ Income
f2$coefficients

## (Intercept)      X2
## 253.8514      3.0253
```

Which of the three models is the best? Of course, by using the word “best”, we need to answer the question “best?... but with respect to what kind of measure?”

So far we were fitting w.r.t. SSR, as the multiple regression model generalises the two simple ones, the former must yield a not-worse SSR. This is because in the case of $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$, setting β_1 to 0 (just one of uncountably many possible β_1 s, if it happens to be the *best* one, good for us) gives $Y = a + bX_2$ whereas by setting β_2 to 0 we obtain $Y = a + bX_1$.

```
sum(f12$residuals^2)

## [1] 358261
sum(f1$residuals^2)

## [1] 2132108
```

```
sum(f2$residuals^2)
```

```
## [1] 1823473
```

We get that, in terms of SSRs, f_{12} is better than f_2 , which in turn is better than f_1 . However, these error values per se (sheer numbers) are meaningless (not meaningful).

Remark. Interpretability in ML has always been an important issue, think the EU General Data Protection Regulation (GDPR), amongst others.

7.3.1.1 SSR, MSE, RMSE and MAE

The quality of fit can be assessed by performing some *descriptive statistical analysis of the residuals*, $\hat{y}_i - y_i$, for $i = 1, \dots, n$.

I know how to summarise data on the residuals! Of course I should compute their arithmetic mean and I'm done with that shtask! Interestingly, the mean of residuals (this can be shown analytically) in the least squared fit is always equal to 0:

$$\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i) = 0.$$

Therefore, we need a different metric.

Exercise 7.1 (*) A proof of this fact is left as an exercise to the curious; assume $p = 1$ just as in the previous chapter and note that $\hat{y}_i = ax_i + b$.

```
mean(f12$residuals) # almost zero numerically
```

```
## [1] -2.0867e-16
```

```
all.equal(mean(f12$residuals), 0)
```

```
## [1] TRUE
```

We noted that sum of squared residuals (SSR) is not interpretable, but the mean squared residuals (MSR) – also called mean squared error (MSE) regression loss – is a little better. Recall that mean is defined as the sum divided by number of samples.

$$\text{MSE}(f) = \frac{1}{n} \sum_{i=1}^n (f(\mathbf{x}_{i,\cdot}) - y_i)^2.$$

```
mean(f12$residuals^2)
```

```
## [1] 1155.7
```

```
mean(f1$residuals^2)
```

```
## [1] 6877.8
```

```
mean(f2$residuals^2)
```

```
## [1] 5882.2
```

This gives an information of how much do we err *per sample*, so at least this measure does not depend on n anymore. However, if the original Y s are, say, in metres [m], MSE is expressed in metres squared [m^2].

To account for that, we may consider the root mean squared error (RMSE):

$$\text{RMSE}(f) = \sqrt{\frac{1}{n} \sum_{i=1}^n (f(\mathbf{x}_{i,\cdot}) - y_i)^2}.$$

This is just like with the sample variance vs. standard deviation – recall the latter is defined as the square root of the former.

```
sqrt(mean(f12$residuals^2))
```

```
## [1] 33.995
```

```
sqrt(mean(f1$residuals^2))
```

```
## [1] 82.932
```

```
sqrt(mean(f2$residuals^2))
```

```
## [1] 76.695
```

The interpretation of the RMSE is rather quirky; it is some-sort-of-averaged *deviance* from the true rating (which is on the scale 0–1000, hence we see that the first model is not that bad). Recall that the square function is sensitive to large observations, hence, it penalises notable deviations more heavily.

As still we have a problem with finding something easily interpretable (your non-technical boss or client may ask you: but what do these numbers mean??), we suggest here that the mean absolute error (MAE) might be a better idea than the above:

$$\text{MAE}(f) = \frac{1}{n} \sum_{i=1}^n |f(\mathbf{x}_{i,\cdot}) - y_i|.$$

```
mean(abs(f12$residuals))
```

```
## [1] 22.863
```

```
mean(abs(f1$residuals))
```

```
## [1] 61.489
```

```
mean(abs(f2$residuals))
```

```
## [1] 64.151
```

With the above we may say “On average, the predicted rating differs from the observed one by...”. That is good enough.

Remark. (*) You may ask why don't we fit models so as to minimise the MAE and we minimise the RMSE instead (note that minimising RMSE is the same as minimising the SSR, one is a strictly monotone transformation of the other and do not affect the solution). Well, it is possible. It turns out that, however, minimising MAE is more computationally expensive and the solution may be numerically unstable. So it's rarely an analyst's first choice (assuming they are well-educated enough to know about the mean absolute error regression task). However, it may be worth trying it out sometimes.

Sometimes we might prefer MAE regression to the classic one if our data is heavily contaminated by outliers. But in such cases it is worth checking if proper data cleansing does the trick.

7.3.1.2 Graphical Summaries of Residuals

If we are not happy with single numerical aggregated of the residuals or their absolute values, we can (and should) always compute a whole bunch of descriptive statistics:

```
summary(f12$residuals)
```

```
##    Min. 1st Qu. Median     Mean 3rd Qu.    Max.
## -108.10   -1.94    7.81     0.00   20.25   50.62
```

```
summary(f1$residuals)
```

```
##    Min. 1st Qu. Median     Mean 3rd Qu.    Max.
## -226.8   -48.3   -10.1     0.0    42.6   268.7
```

```
summary(f2$residuals)
```

```
##    Min. 1st Qu. Median     Mean 3rd Qu.    Max.
## -195.16  -57.34   -1.28     0.00   64.01  175.34
```

Graphically, it is nice to summarise the empirical distribution of the residuals on a **box and whisker plot**. Here is the key to decipher Figure @ref(fig:boxplot_explained):

- IQR == Interquartile range == $Q_3 - Q_1$ (box width)
- The box contains 50% of the “most typical” observations
- Box and whiskers altogether have width $\leq 4 \text{ IQR}$
- Outliers == observations potentially worth inspecting (is it a bug or a feature?) – values $< Q_1 - 1.5 \text{ IQR}$ or $> Q_3 + 1.5 \text{ IQR}$

Figure @ref(fig:boxplot_residuals) is worth a thousand words:

```
boxplot(horizontal=TRUE, xlab="residuals", col="white",
  list(f12=f12$residuals, f1=f1$residuals, f2=f2$residuals))
abline(v=0, lty=3)
```

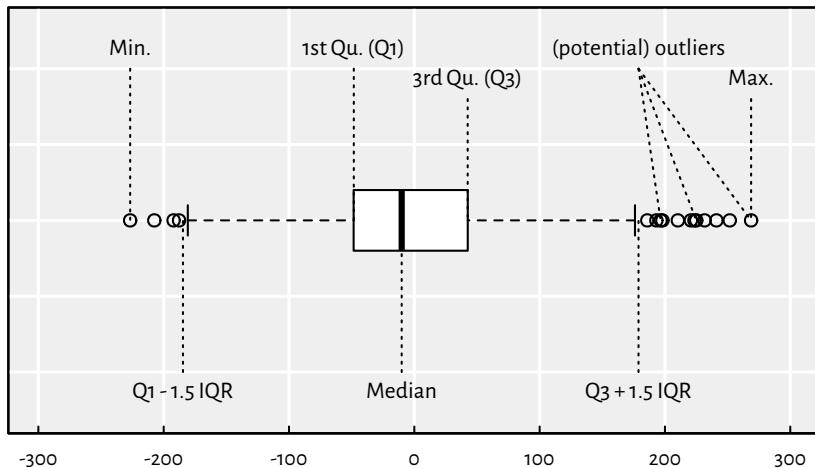


Figure 7.5: (#fig:boxplot_explained) An example boxplot

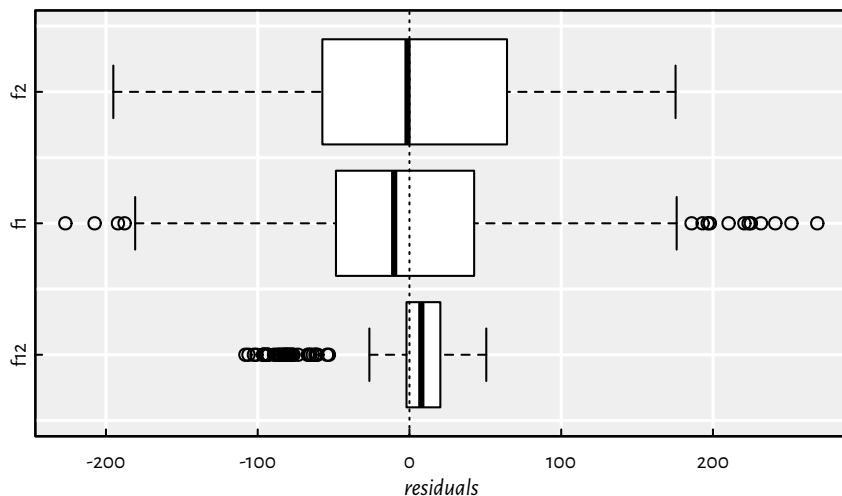


Figure 7.6: (#fig:boxplot_residuals) Box plots of the residuals for the three models studied

Figure @ref(fig:violinplot_residuals) gives a *violin plot* – a blend of a box plot and a (kernel) density estimator (histogram-like):

```
library("vioplot")
vioplot(horizontal=TRUE, xlab="residuals", col="white",
        list(f12=f12$residuals, f1=f1$residuals, f2=f2$residuals))
abline(v=0, lty=3)
```

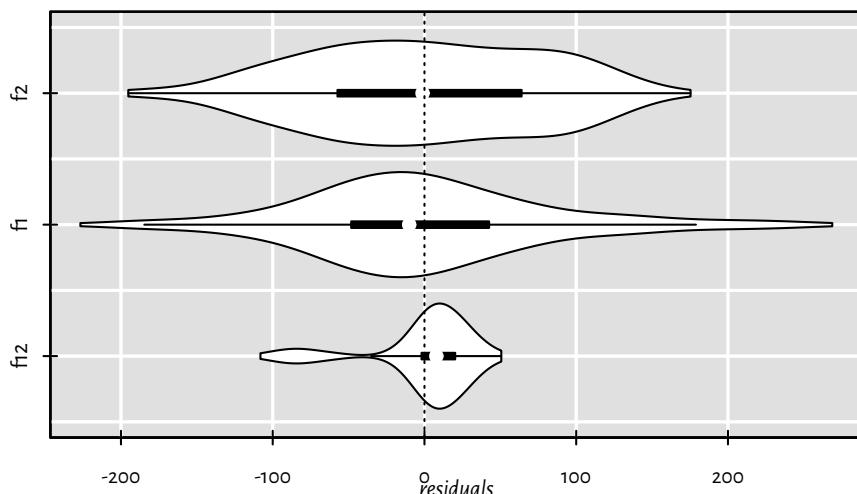


Figure 7.7: (#fig:violinplot_residuals) Violin plots of the residuals for the three models studied

We can also take a look at the absolute values of the residuals. Here are some descriptive statistics:

```
summary(abs(f12$residuals))

##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
##  0.065   6.464  14.071  22.863  26.418 108.100

summary(abs(f1$residuals))

##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
##  0.506  19.664  45.072  61.489  80.124 268.738

summary(abs(f2$residuals))

##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
##  0.655  29.854  59.676  64.151  95.738 195.156
```

Figure @ref(fig:absresiduals_boxplot) is worth \$1000:

```
boxplot(horizontal=TRUE, col="white", xlab="abs(residuals)",
  list(f12=abs(f12$residuals), f1=abs(f1$residuals),
       f2=abs(f2$residuals)))
abline(v=0, lty=3)
```

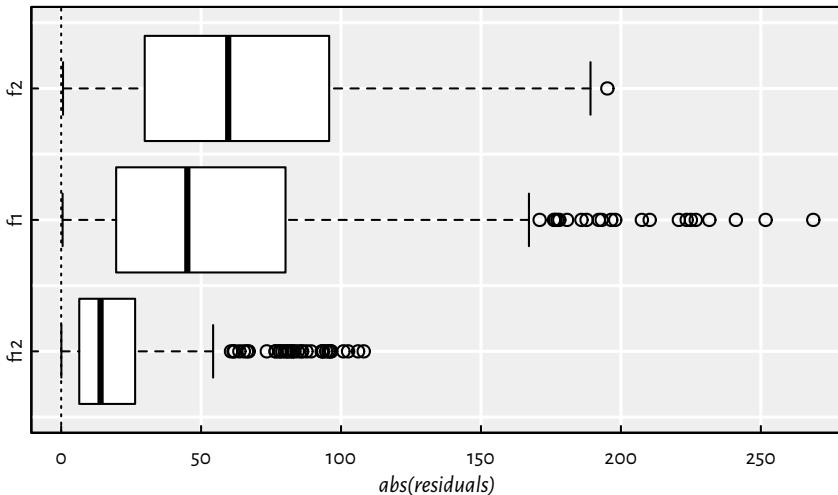


Figure 7.8: (#fig:absresiduals_boxplot) Box plots of the modules of the residuals for the three models studied

7.3.1.3 Coefficient of Determination (R-squared)

If we didn't know the range of the dependent variable (in our case we do know that the credit rating is on the scale 0–1000), the RMSE or MAE would be hard to interpret.

It turns out that there is a popular *normalised* (unit-less) measure that is somehow easy to interpret with no domain-specific knowledge of the modelled problem. Namely, the (unadjusted) R^2 score (the coefficient of determination) is given by:

$$R^2(f) = 1 - \frac{\sum_{i=1}^n (y_i - f(\mathbf{x}_{i,.}))^2}{\sum_{i=1}^n (y_i - \bar{y})^2},$$

where \bar{y} is the arithmetic mean $\frac{1}{n} \sum_{i=1}^n y_i$.

```
(r12 <- summary(f12)$r.squared)
```

```
## [1] 0.93909
1 - sum(f12$residuals^2)/sum((Y-mean(Y))^2) # the same
## [1] 0.93909
```

```
(r1 <- summary(f1)$r.squared)
## [1] 0.63751
(r2 <- summary(f2)$r.squared)
## [1] 0.68998
```

The coefficient of determination gives the proportion of variance of the dependent variable explained by independent variables in the model; $R^2(f) \approx 1$ indicates a perfect fit. The first model is a very good one, the simple models are “more or less okay”.

Unfortunately, R^2 tends to automatically increase as the number of independent variables increase (recall that the more variables in the model, the better the SSR must be). To correct for this phenomenon, we sometimes consider the **adjusted R^2** :

$$\bar{R}^2(f) = 1 - (1 - R^2(f)) \frac{n - 1}{n - p - 1}$$

```
summary(f12)$adj.r.squared
## [1] 0.93869
n <- length(x); 1 - (1 - r12)*(n-1)/(n-3) # the same
## [1] 0.93869
summary(f1)$adj.r.squared
## [1] 0.63633
summary(f2)$adj.r.squared
## [1] 0.68897
```

In other words, the adjusted R^2 penalises for more complex models.

Remark. (*) Side note – results of some statistical tests (e.g., significance of coefficients) are reported by calling `summary(f12)` etc. — refer to a more advanced source to obtain more information. These, however, require the verification of some assumptions regarding the input data and the residuals.

```
summary(f12)
##
## Call:
## lm(formula = Y ~ X1 + X2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -108.10    -1.94     7.81    20.25    50.62
```

```

## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.73e+02  3.95e+00   43.7 <2e-16 ***
## X1          1.83e-01  5.16e-03   35.4 <2e-16 ***
## X2          2.20e+00  5.64e-02   39.0 <2e-16 ***  
## ---        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 34.2 on 307 degrees of freedom
## Multiple R-squared:  0.939, Adjusted R-squared:  0.939 
## F-statistic: 2.37e+03 on 2 and 307 DF,  p-value: <2e-16

```

7.3.1.4 Residuals vs. Fitted Plot

We can also create scatter plots of the residuals (predicted \hat{y}_i minus true y_i) as a function of the predicted $\hat{y}_i = f(\mathbf{x}_i, \cdot)$, see Figure @ref(fig:resid_vs_fitted).

```

Y_pred12 <- f12$fitted.values # predict(f12, data.frame(X1, X2))
Y_pred1  <- f1$fitted.values # predict(f1, data.frame(X1))
Y_pred2  <- f2$fitted.values # predict(f2, data.frame(X2))
par(mfrow=c(1, 3))
plot(Y_pred12, Y_pred12-Y)
plot(Y_pred1,  Y_pred1 -Y)
plot(Y_pred2,  Y_pred2 -Y)

```

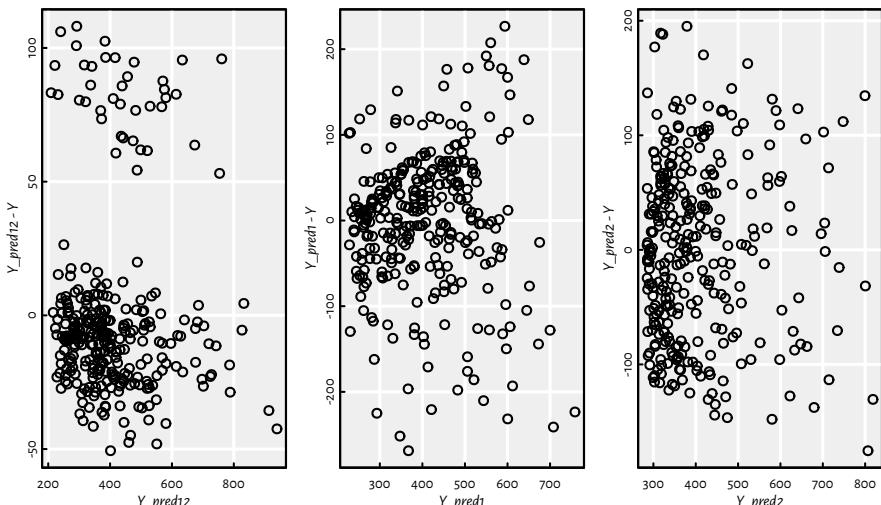


Figure 7.9:(#fig:resid_vs_fitted) Residuals vs. fitted outputs for the three regression models

Ideally (provided that the hypothesis that the dependent variable is indeed a linear function of the dependent variable(s) is true), we would expect to see a point cloud that spread around 0 in a very much unorderly fashion.

7.3.2 Variable Selection

Okay, up to now we've been considering the problem of modelling the `Rating` variable as a function of `Balance` and/or `Income`. However, in the `Credit` data set there are other variables possibly worth inspecting.

Consider all quantitative (numeric-continuous) variables in the `Credit` data set.

```
C <- Credit[Credit$Balance>0,
  c("Rating", "Limit", "Income", "Age",
    "Education", "Balance")]
head(C)
```

	Rating	Limit	Income	Age	Education	Balance
## 1	283	3606	14.891	34	11	333
## 2	483	6645	106.025	82	15	903
## 3	514	7075	104.593	71	11	580
## 4	681	9504	148.924	36	11	964
## 5	357	4897	55.882	68	16	331
## 6	569	8047	80.180	77	10	1151

Obviously there are many possible combinations of the variables upon which regression models can be constructed (precisely, for p variables there are 2^p such models). How do we choose the *best* set of inputs?

Remark. We should already be suspicious at this point: wait... *best* requires some sort of criterion, right?

First, however, let's draw a matrix of scatter plots for every pair of variables – so as to get an impression of how individual variables interact with each other, see Figure 7.10.

```
pairs(C)
```

It seems like `Rating` depends on `Limit` almost linearly... We have a tool to actually quantify the degree of linear dependence between a pair of variables – Pearson's r – the linear correlation coefficient:

$$r(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}.$$

It holds $r \in [-1, 1]$, where:

- $r = 1$ – positive linear dependence (y increases as x increases)
- $r = -1$ – negative linear dependence (y decreases as x increases)

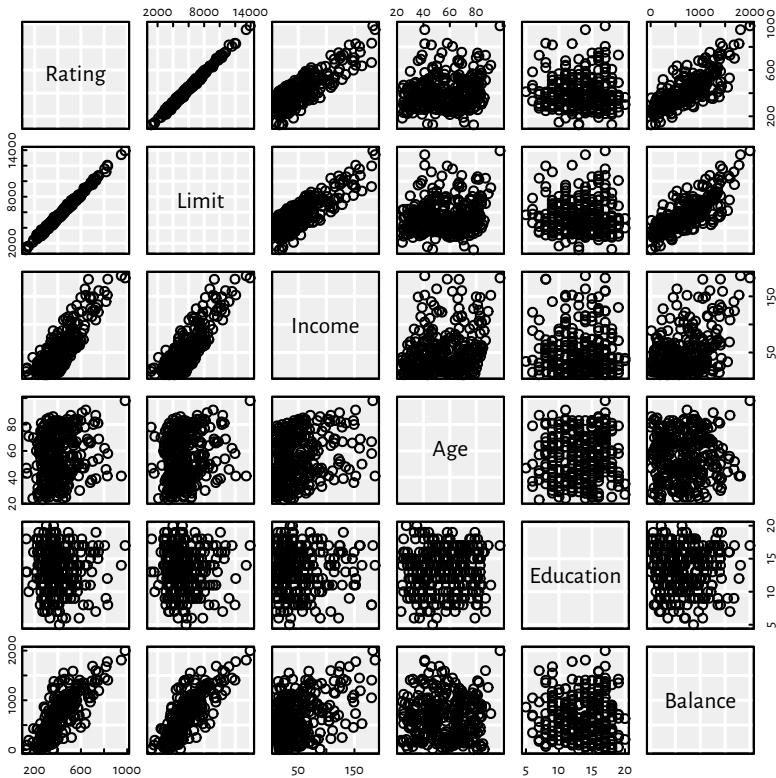


Figure 7.10: Scatter plot matrix for the Credit dataset

- $r \approx 0$ – uncorrelated or non-linearly dependent

Figure @ref{fig:pearson_interpret} gives an illustration of the above.

To compute Pearson's r between all pairs of variables, we call:

```
round(cor(C), 3)
```

	Rating	Limit	Income	Age	Education	Balance
## Rating	1.000	0.996	0.831	0.167	-0.040	0.798
## Limit	0.996	1.000	0.834	0.164	-0.032	0.796
## Income	0.831	0.834	1.000	0.227	-0.033	0.414
## Age	0.167	0.164	0.227	1.000	0.024	0.008
## Education	-0.040	-0.032	-0.033	0.024	1.000	0.001
## Balance	0.798	0.796	0.414	0.008	0.001	1.000

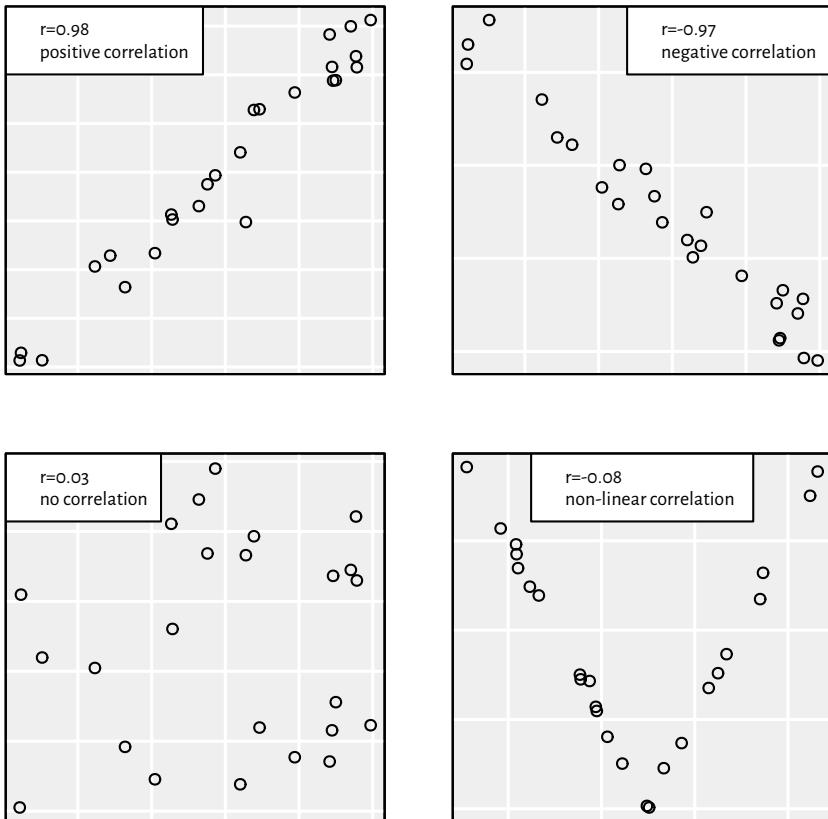


Figure 7.11: (#fig:pearson_interpret) Different datasets and the corresponding Pearson's r coefficients

Rating and Limit are almost perfectly linearly correlated, and both seem to describe the same thing.

For practical purposes, we'd rather model Rating as a function of the other variables. For simple linear regression models, we'd choose either Income or Balance. How about multiple regression though?

The best model:

- has high predictive power,
- is simple.

These two criteria are often mutually exclusive.

Which variables should be included in the optimal model?

Again, the definition of the “best” object needs a *fitness* function.

For fitting a single model to data, we use the SSR.

We need a metric that takes the number of dependent variables into account.

Remark. (*) Unfortunately, the adjusted R^2 , despite its interpretability, is not really suitable for this task. It does not penalise complex models heavily enough to be really useful.

Here we'll be using **the Akaike Information Criterion** (AIC).

For a model f with p' independent variables:

$$\text{AIC}(f) = 2(p' + 1) + n \log(\text{SSR}(f)) - n \log n$$

Our task is to find the combination of independent variables that minimises the AIC.

Remark. (***) Note that this is a bi-level optimisation problem – for every considered combination of variables (which we look for), we must solve another problem of finding the best model involving these variables – the one that minimises the SSR.

$$\min_{s_1, s_2, \dots, s_p \in \{0,1\}} \left(\begin{array}{l} 2 \left(\sum_{j=1}^p s_j + 1 \right) + \\ n \log \left(\min_{\beta_0, \beta_1, \dots, \beta_p \in \mathbb{R}} \sum_{i=1}^n (\beta_0 + s_1 \beta_1 x_{i,1} + \dots + s_p \beta_p x_{i,p} - y_i)^2 \right) \end{array} \right)$$

We dropped the $n \log n$ term, because it is always constant and hence doesn't affect the solution. If $s_j = 0$, then the $s_j \beta_j x_{i,j}$ term is equal to 0, and hence is not considered in the model. This plays the role of including $s_j = 1$ or omitting $s_j = 0$ the j -th variable in the model building exercise.

For p variables, the number of their possible combinations is equal to 2^p (grows exponentially with p). For large p (think big data), an extensive search is impractical (in our case we could get away with this though – left as an exercise to a slightly more advanced reader). Therefore, to find the variable combination minimising the AIC, we often rely on one of the two following greedy heuristics:

- forward selection:

1. start with an empty model
2. find an independent variable whose addition to the current model would yield the highest decrease in the AIC and add it to the model
3. go to step 2 until AIC decreases

- backward elimination:

1. start with the full model

2. find an independent variable whose removal from the current model would decrease the AIC the most and eliminate it from the model
3. go to step 2 until AIC decreases

Remark. (***) The above bi-level optimisation problem can be solved by implementing a genetic algorithm – see further chapter for more details.

Remark. (*) There are of course many other methods which also perform some form of variable selection, e.g., lasso regression. But these minimise a different objective.

First, a forward selection example. We need a data sample to work with:

```
C <- Credit[Credit$Balance>0,
  c("Rating", "Income", "Age",
    "Education", "Balance")]
```

Then, a formula that represents a model with no variables (model from which we'll start our search):

```
(model_empty <- Rating~1)
```

```
## Rating ~ 1
```

Last, we need a model that includes all the variables. We're too lazy to list all of them manually, therefore, we can use the `model.frame()` function to generate a corresponding formula:

```
(model_full <- formula(model.frame(Rating~, data=C))) # all variables
```

```
## Rating ~ Income + Age + Education + Balance
```

Now we are ready.

```
step(lm(model_empty, data=C), # starting model
  scope=model_full,          # gives variables to consider
  direction="forward")
```

```
## Start:  AIC=3055.8
## Rating ~ 1
##
##           Df Sum of Sq      RSS   AIC
## + Income     1   4058342  1823473 2695
## + Balance    1   3749707  2132108 2743
## + Age        1    164567  5717248 3049
## <none>                   5881815 3056
## + Education  1     9631  5872184 3057
##
## Step:  AIC=2694.7
## Rating ~ Income
##
##           Df Sum of Sq      RSS   AIC
```

```

## + Balance    1   1465212  358261 2192
## <none>                 1823473 2695
## + Age       1      2836 1820637 2696
## + Education 1     1063 1822410 2697
##
## Step: AIC=2192.3
## Rating ~ Income + Balance
##
##          Df Sum of Sq   RSS   AIC
## + Age     1      4119 354141 2191
## + Education 1     2692 355568 2192
## <none>             358261 2192
##
## Step: AIC=2190.7
## Rating ~ Income + Balance + Age
##
##          Df Sum of Sq   RSS   AIC
## + Education 1     2926 351216 2190
## <none>             354141 2191
##
## Step: AIC=2190.1
## Rating ~ Income + Balance + Age + Education
##
## Call:
## lm(formula = Rating ~ Income + Balance + Age + Education, data = C)
##
## Coefficients:
## (Intercept)      Income      Balance        Age      Education
##     173.830      2.167      0.184      0.223      -0.960
formula(lm(Rating~., data=C))

## Rating ~ Income + Age + Education + Balance

```

The full model has been selected.

And now for something completely different – a backward elimination example:

```

step(lm(model_full, data=C), # from
     scope=model_empty,      # to
     direction="backward")

## Start: AIC=2190.1
## Rating ~ Income + Age + Education + Balance
##
##          Df Sum of Sq   RSS   AIC

```

```

## <none>                      351216 2190
## - Education     1      2926  354141 2191
## - Age          1      4353  355568 2192
## - Balance       1    1468466 1819682 2698
## - Income        1    1617191 1968406 2722

##
## Call:
## lm(formula = Rating ~ Income + Age + Education + Balance, data = C)
##
## Coefficients:
## (Intercept)      Income         Age   Education      Balance
## 173.830        2.167        0.223      -0.960       0.184

```

The full model is considered the best again.

Forward selection example – full dataset:

```

C <- Credit[, # do not restrict to Credit$Balance>0
  c("Rating", "Income", "Age",
    "Education", "Balance")]
step(lm(model_empty, data=C),
  scope=model_full,
  direction="forward")

## Start: AIC=4034.3
## Rating ~ 1
##
##           Df Sum of Sq    RSS  AIC
## + Balance   1  7124258 2427627 3488
## + Income    1  5982140 3569744 3643
## + Age       1   101661 9450224 4032
## <none>                  9551885 4034
## + Education 1     8675 9543210 4036
##
## Step: AIC=3488.4
## Rating ~ Balance
##
##           Df Sum of Sq    RSS  AIC
## + Income    1  1859749  567878 2909
## + Age       1    98562 2329065 3474
## <none>                  2427627 3488
## + Education 1     5130 2422497 3490
##
## Step: AIC=2909.3
## Rating ~ Balance + Income
##

```

```

##           Df Sum of Sq    RSS   AIC
## <none>              567878 2909
## + Age      1     2142 565735 2910
## + Education 1     1209 566669 2910

##
## Call:
## lm(formula = Rating ~ Balance + Income, data = C)
##
## Coefficients:
## (Intercept)      Balance        Income
##       145.351       0.213        2.186

```

This procedure suggests including only the Balance and Income variables.

Backward elimination example – full dataset:

```

step(lm(model_full, data=C), # full model
      scope=model_empty, # empty model
      direction="backward")

## Start:  AIC=2910.9
## Rating ~ Income + Age + Education + Balance
##
##           Df Sum of Sq    RSS   AIC
## - Education 1     1238 565735 2910
## - Age       1     2172 566669 2910
## <none>                564497 2911
## - Income    1   1759273 2323770 3475
## - Balance   1   2992164 3556661 3645
##
## Step:  AIC=2909.8
## Rating ~ Income + Age + Balance
##
##           Df Sum of Sq    RSS   AIC
## - Age      1     2142 567878 2909
## <none>                565735 2910
## - Income   1   1763329 2329065 3474
## - Balance   1   2991523 3557259 3643
##
## Step:  AIC=2909.3
## Rating ~ Income + Balance
##
##           Df Sum of Sq    RSS   AIC
## <none>                567878 2909
## - Income   1   1859749 2427627 3488
## - Balance   1   3001866 3569744 3643

```

```

## 
## Call:
## lm(formula = Rating ~ Income + Balance, data = C)
## 
## Coefficients:
## (Intercept)      Income      Balance
##       145.351        2.186        0.213

```

This procedure gives the same results as forward selection (however, for other data sets this might not necessarily be the case).

7.3.3 Variable Transformation

So far we have been fitting linear models of the form:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

What about some non-linear models such as polynomials etc.? For example:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1^3 + \beta_4 X_2.$$

Solution: pre-process inputs by setting $X'_1 := X_1$, $X'_2 := X_1^2$, $X'_3 := X_1^3$, $X'_4 := X_2$ and fit a linear model:

$$Y = \beta_0 + \beta_1 X'_1 + \beta_2 X'_2 + \beta_3 X'_3 + \beta_4 X'_4.$$

This trick works for every model of the form $Y = \sum_{i=1}^k \sum_{j=1}^p \varphi_{i,j}(X_j)$ for any k and any univariate functions $\varphi_{i,j}$.

Also, with a little creativity (and maths), we might be able to transform a few other models to a linear one, e.g.,

$$Y = b e^{aX} \quad \rightarrow \quad \log Y = \log b + aX \quad \rightarrow \quad Y' = aX + b'$$

This is an example of a model's **linearisation**. However, not every model can be linearised. In particular, one that involves functions that are not invertible.

For example, here's a series of simple ($p = 1$) degree- d polynomial regression models of the form:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \cdots + \beta_d X^d.$$

Such models can be fitted with the `lm()` function based on the formula of the form `Y~poly(X, d, raw=TRUE)` or `Y~X+I(X^2)+I(X^3)+...`

```
f1_1 <- lm(Y~X1)
f1_3 <- lm(Y~X1+I(X1^2)+I(X1^3)) # also: Y~poly(X1, 3, raw=TRUE)
f1_10 <- lm(Y~poly(X1, 10, raw=TRUE))
```

Above we have fitted the polynomials of degrees 1, 3 and 10. Note that a polynomial of degree 1 is just a line.

Let us depict the three models:

```
plot(X1, Y, col="#00000066", ylim=c(0, 1100))
x <- seq(min(X1), max(X1), length.out=101)
lines(x, predict(f1_1, data.frame(X1=x)), col="red", lwd=3, lty=1)
lines(x, predict(f1_3, data.frame(X1=x)), col="blue", lwd=3, lty=2)
lines(x, predict(f1_10, data.frame(X1=x)), col="darkgreen", lwd=3, lty=4)
legend("topleft", c("degree 1", "degree 3", "degree 10"),
lwd=3, col=c("red", "blue", "darkgreen"), lty=c(1,2,4), bg="white")
```

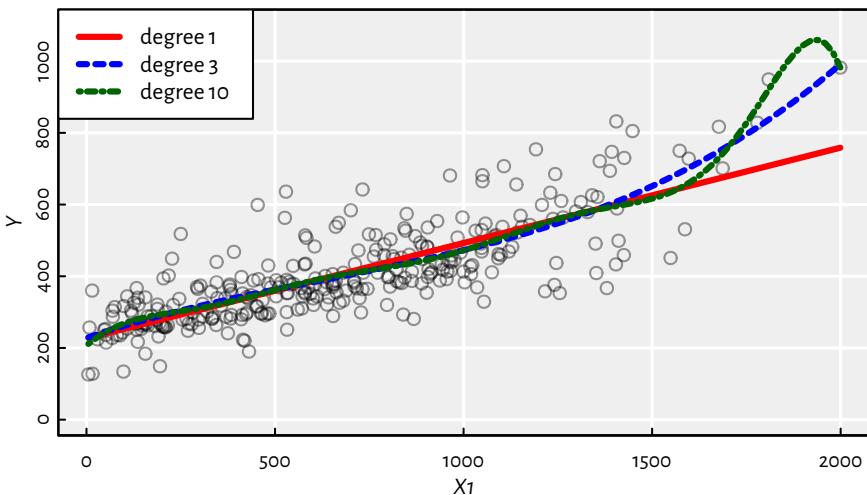


Figure 7.12: Polynomials of different degrees fitted to the Credit dataset

From Figure 7.12 we see that the models tend to agree on what the predictions should be for low and moderate balances (X_1). However, there's clearly a problem with the behaviour of the degree-10 polynomial on the righthand side of the figure.

7.3.4 Predictive vs. Descriptive Power

The above high-degree polynomial model (`f1_10`) is a typical instance of a phenomenon called an **overfit**.

Clearly (based on our expert knowledge), the Rating shouldn't decrease as Balance increases.

In other words, `f1_10` gives a better fit to data actually observed, but fails to produce good results for the points that are yet to come.

We say that it **generalises** poorly to unseen data.

Assume our true model is of the form:

```
true_model <- function(x) 3*x^3+5
```

Let's generate the following random sample from this model (with Y subject to error), see Figure 7.13:

```
set.seed(1234) # to assure reproducibility
n <- 25
X <- runif(n, min=0, max=1)
Y <- true_model(X)+rnorm(n, sd=0.2) # add normally-distributed noise

plot(X, Y)
x <- seq(0, 1, length.out=101)
lines(x, true_model(x), col=2, lwd=3, lty=2)
```

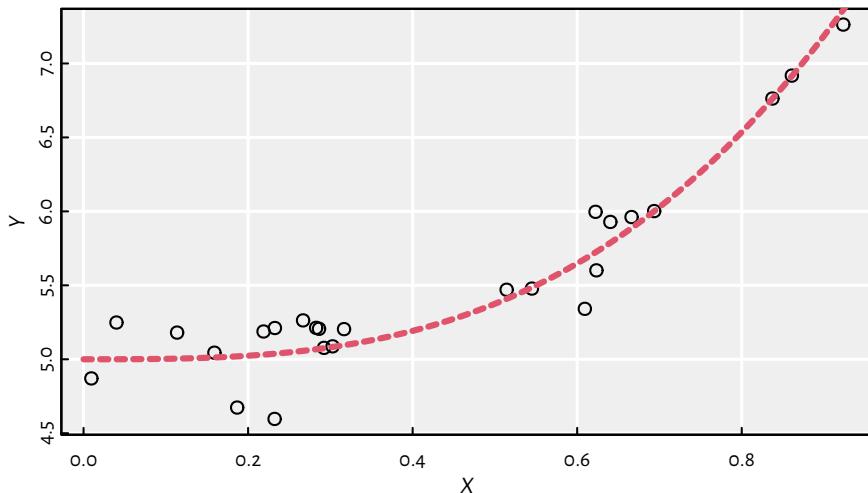


Figure 7.13: Synthetic data generated by means of the formula $Y = 3x^3 + 5$ (+ noise)

Let's fit polynomials of different degrees, see Figure 7.14.

```
plot(X, Y)
lines(x, true_model(x), col=2, lwd=3, lty=2)
```

```
dmax <- 11 # maximal polynomial degree
MSE_train <- numeric(dmax)
MSE_test <- numeric(dmax)
for (d in 1:dmax) { # for every polynomial degree
  f <- lm(Y~poly(X, d, raw=TRUE)) # fit a d-degree polynomial
  y <- predict(f, data.frame(X=x))
  lines(x, y, col=d)
  # MSE on given random X,Y:
  MSE_train[d] <- mean(f$residuals^2)
  # MSE on many more points:
  MSE_test[d] <- mean((y-true_model(x))^2)
}
}
```

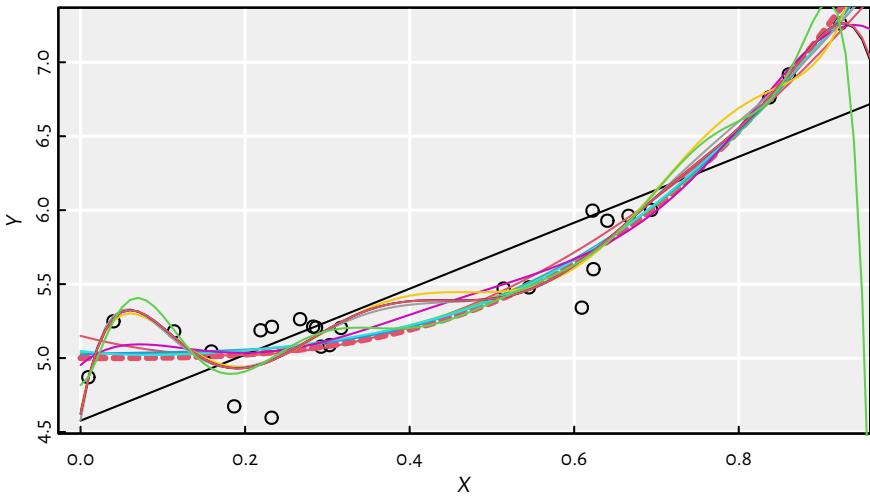


Figure 7.14: Polynomials fitted to our synthetic dataset

Some of the polynomials are fitted too well!

Remark (*) The oscillation of the high-degree polynomials at the domain boundaries is known as the Runge phenomenon.

Compare the mean squared error (MSE) for the observed vs. future data points, see Figure 7.15.

```
matplot(1:dmax, cbind(MSE_train, MSE_test), type="b",
       ylim=c(1e-3, 2e3), log="y", pch=1:2,
       xlab="Model complexity (polynomial degree)",
       ylab="MSE")
legend("topleft", legend=c("MSE on original data", "MSE on the whole range"),
       lty=1:2, col=1:2, pch=1:2, bg="white")
```

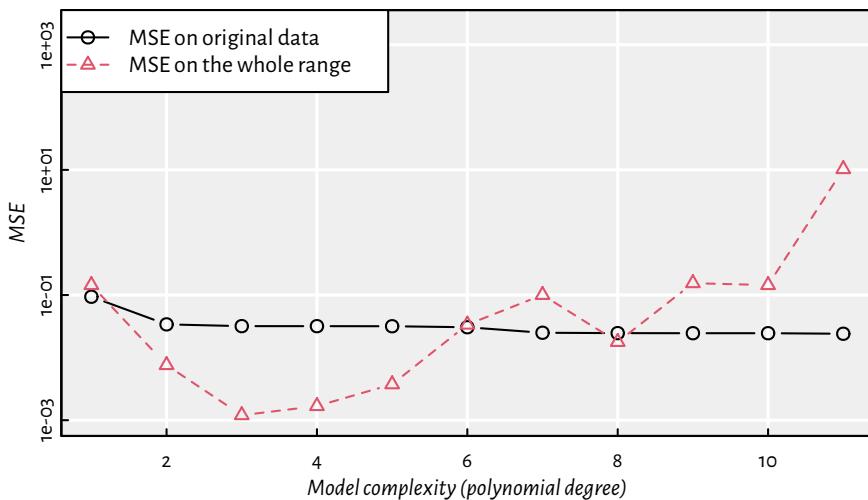


Figure 7.15: MSE on the dataset used to construct the model vs. MSE on a whole range of points as function of the polynomial degree

Note the logarithmic scale on the y axis.

This is a very typical behaviour!

- A model's fit to observed data improves as the model's complexity increases.
- A model's generalisation to unseen data initially improves, but then becomes worse.
- In the above example, the sweet spot is at a polynomial of degree 3, which is exactly our true underlying model.

Hence, most often we should be interested in the accuracy of the predictions made in the case of unobserved data.

If we have a data set of a considerable size, we can divide it (randomly) into two parts:

- *training sample* (say, 60% or 80%) – used to fit a model
- *test sample* (the remaining 40% or 20%) – used to assess its quality (e.g., using MSE)

More on this issue in the chapter on Classification.

Remark. (*) We shall see that sometimes a train-test-validate split will be necessary, e.g., 60-20-20%.

7.4 Exercises

7.4.1 Anscombe's Quartet Revisited

Consider the `anscombe` database once again:

```
print(anscombe) # `anscombe` is a built-in object
```

```
##   x1 x2 x3 x4     y1     y2     y3     y4
## 1 10 10 10  8 8.04 9.14 7.46 6.58
## 2  8  8  8  8 6.95 8.14 6.77 5.76
## 3 13 13 13  8 7.58 8.74 12.74 7.71
## 4  9  9  9  8 8.81 8.77 7.11 8.84
## 5 11 11 11  8 8.33 9.26 7.81 8.47
## 6 14 14 14  8 9.96 8.10 8.84 7.04
## 7  6  6  6  8 7.24 6.13 6.08 5.25
## 8  4  4  4 19 4.26 3.10 5.39 12.50
## 9 12 12 12  8 10.84 9.13 8.15 5.56
## 10 7  7  7  8 4.82 7.26 6.42 7.91
## 11 5  5  5  8 5.68 4.74 5.73 6.89
```

Recall that in the previous Chapter we have split the above data into four data frames `ans1`, ..., `ans4` with columns `x` and `y`.

Exercise 7.2 In `ans1`, fit a regression line to the data set as-is.

Solution.

We've done that already, see Figure 7.16. What a wonderful exercise, thank you – effective learning is often done by repeating stuff.

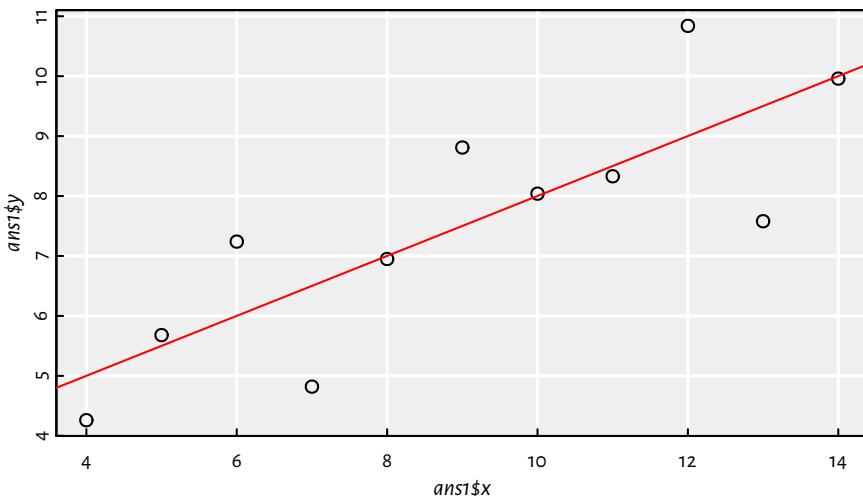
```
ans1 <- data.frame(x=anscombe$x1, y=anscombe$y1)
f1 <- lm(y~x, data=ans1)
plot(ans1$x, ans1$y)
abline(f1, col="red")
```

Exercise 7.3 In `ans2`, fit a quadratic model ($y = a + bx + cx^2$).

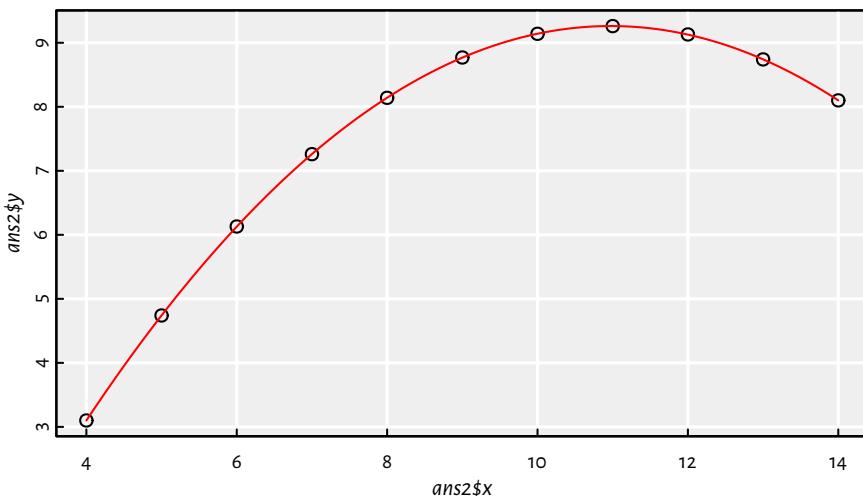
Solution.

How to fit a polynomial model is explained above.

```
ans2 <- data.frame(x=anscombe$x2, y=anscombe$y2)
f2 <- lm(y~x+I(x^2), data=ans2)
plot(ans2$x, ans2$y)
```

Figure 7.16: Fitted regression line for `ans1`

```
x_plot <- seq(4, 14, by=0.1)
y_plot <- predict(f2, data.frame(x=x_plot))
lines(x_plot, y_plot, col="red")
```

Figure 7.17: Fitted quadratic model for `ans2`

Comment: From Figure 7.17 we see that it's an almost-perfect fit! Clearly, the second Anscombe dataset isn't a case of linearly dependent variables.

Exercise 7.4 In `ans3`, remove the obvious outlier from data and fit a regression line.

Solution.

Let's plot the data set first, see Figure 7.18.

```
ans3 <- data.frame(x=anscombe$x3, y=anscombe$y3)
plot(ans3$x, ans3$y)
```

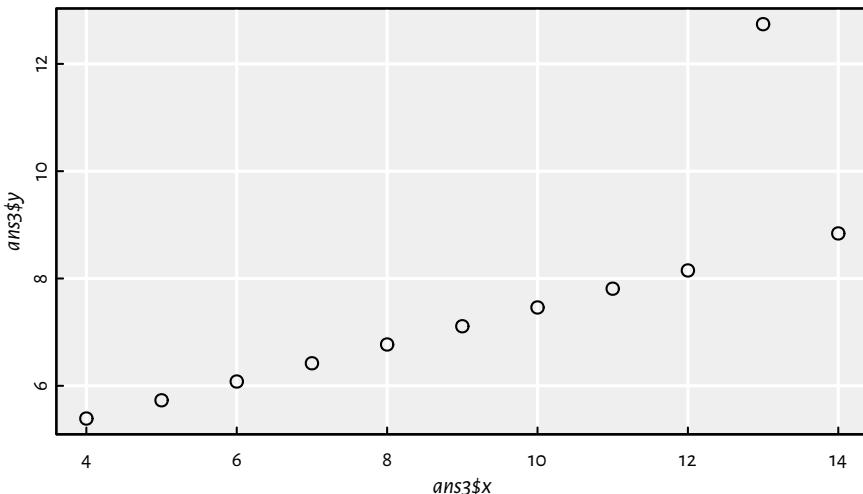


Figure 7.18: Scatter plot for `ans3`

Indeed, the observation at $x \approx 13$ is an obvious outlier. Perhaps the easiest way to remove it is to call:

```
ans3b <- ans3[ans3$y<=12,] # the outlier is definitely at y>12
```

We could also use the condition `y < max(y)`, amongst others.

Now let's fit the linear model:

```
f3b <- lm(y~x, data=ans3b)
plot(ans3b$x, ans3b$y)
abline(f3b, col="red")
```

Comment: Now Figure 7.19 is what we call linearly correlated data. By the way, Pearson's coefficient now equals 1.

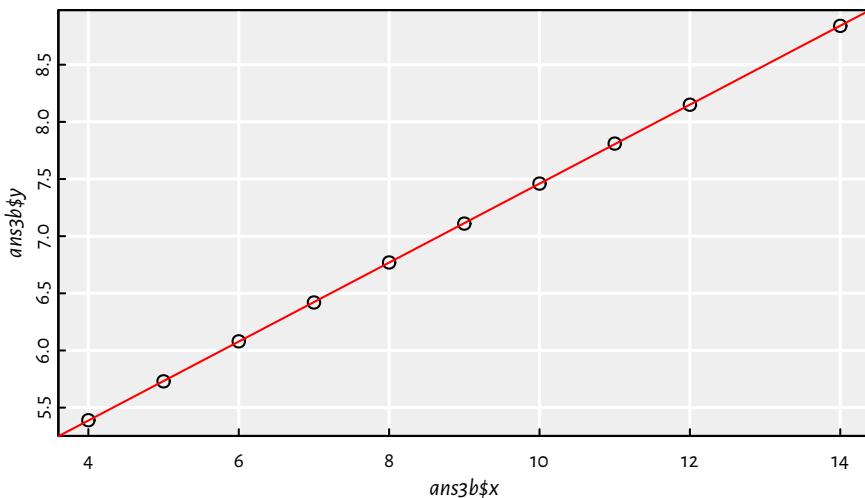


Figure 7.19: Scatter plot for `ans3` with the outlier removed and the fitted linear model

7.4.2 Countries of the World – Simple models involving the GDP per capita

Let's consider the World Factbook 2020 dataset (see Appendix F), which lists country names, their population, area, GDP, mortality rates and so forth:

```
factbook <- read.csv("datasets/world_factbook_2020.csv",
  comment.char="#")
```

Here is a preview of a few features for 3 selected countries (see `help("%in%")`):

```
factbook[factbook$country %in%
  c("Australia", "New Zealand", "United States"),
  c("country", "area", "population", "gdp_per_capita_ppp")]

##           country     area population gdp_per_capita_ppp
## 15      Australia 7741220    25466459          50400
## 169     New Zealand 268838     4925477          39000
## 247 United States 9833517   332639102         59800
```

Exercise 7.5 List the 10 countries with the highest GDP per capita.

Solution.

To recall, to generate a list of indexes that produce an ordered version of a numeric vector, we need to call the `order()` function.

```
which_top <- tail(order(factbook$gdp_per_capita_ppp, na.last=FALSE), 10)
factbook[which_top, c("country", "gdp_per_capita_ppp")]
```

```

##           country gdp_per_capita_ppp
## 113       Ireland            73200
## 35        Brunei            78900
## 114    Isle of Man          84600
## 211     Singapore          94100
## 26       Bermuda          99400
## 141 Luxembourg          105100
## 157      Monaco          115700
## 142      Macau          122000
## 192      Qatar          124100
## 139 Liechtenstein         139100

```

By the way, the reported values are in USD.

Question: Which of these countries are tax havens?



Exercise 7.6 Find the 5 most positively and the 5 most negatively correlated variables with the `gdp_per_capita_ppp` feature (of course, with respect to the Pearson coefficient).

Solution.

This can be solved via a call to `cor()`. Note that we need to make sure that missing values are omitted from computations. A quick glimpse at the manual page (`?cor`) reveals that computing the correlation between a column and all the other ones (of course, except `country`, which is non-numeric) can be performed as follows.

```

r <- cor(factbook$gdp_per_capita_ppp,
          factbook[, !(names(factbook) %in% c("country", "gdp_per_capita_ppp"))],
          use="complete.obs")[1,]
or <- order(r) # ordering permutation (indexes)
r[head(or, 5)] # first 5 ordered indexes

##   infant_mortality_rate maternal_mortality_rate           birth_rate
##                 -0.74658                  -0.67005             -0.60822
##   death_rate      total_fertility_rate
##                 -0.57216                  -0.56725

r[tail(or, 5)] # last 5 ordered indexes

##   natural_gas_production      gross_national_saving
##                   0.56898                  0.61133
##   median_age obesity_adult_prevalence_rate
##                   0.62090                  0.63681
##   life_expectancy_at_birth
##                   0.75461

```

Comment: “Live long and prosper” just gained a new meaning. Richer countries have

lower infant and maternal mortality rates, lower birth rates, but higher life expectancy and obesity prevalence. Note, however, that correlation is not causation: we are unlikely to increase the GDP by asking people to put on weight.



Exercise 7.7 Fit simple regression models where the per capita GDP explains its four most correlated variables (four individual models). Draw them on a scatter plot. Compute the root mean squared errors (RMSE), mean absolute errors (MAE) and the coefficients of determination (R^2).

Solution.

The four most correlated variables (we should look at the absolute value of the correlation coefficient now – recall that it is the correlation of 0 that means no linear dependence; 1 and -1 show a strong association between a pair of variables) are:

```
(most_correlated <- names(r)[tail(order(abs(r)), 4)])
```

```
## [1] "obesity_adult_prevalence_rate" "maternal_mortality_rate"
## [3] "infant_mortality_rate"           "life_expectancy_at_birth"
```

We could take the above column names and construct four formulas manually, e.g., by writing $gdp_per_capita_ppp \sim life_expectancy_at_birth$, but we are lazy. Being lazy when it comes to computer programming is often a virtue, not a flaw in one's character.

Instead, we will run a `for` loop that extracts the pairs of interesting columns and constructs a formula based on two vectors ($lm(Y \sim X)$), see Figure 7.20.

```
par(mfrow=c(2, 2)) # 4 plots on a 2x2 grid
for (i in 1:4) {
  print(most_correlated[i])
  X <- factbook[, "gdp_per_capita_ppp"]
  Y <- factbook[, most_correlated[i]]
  f <- lm(Y ~ X)
  print(cbind(RMSE=sqrt(mean(f$residuals^2)),
              MAE=mean(abs(f$residuals)),
              R2=summary(f)$r.squared))
  plot(X, Y, xlab="gdp_per_capita_ppp",
        ylab=most_correlated[i])
  abline(f, col="red")
}

## [1] "obesity_adult_prevalence_rate"
##      RMSE     MAE      R2
## [1,] 11.041 8.1589 0.062196

## [1] "maternal_mortality_rate"
##      RMSE     MAE      R2
## [1,] 204.93 146.53 0.21481
```

```
## [1] "infant_mortality_rate"
##      RMSE     MAE     R2
## [1,] 15.746 12.166 0.3005

## [1] "life_expectancy_at_birth"
##      RMSE     MAE     R2
## [1,] 5.4292 4.3727 0.43096
```

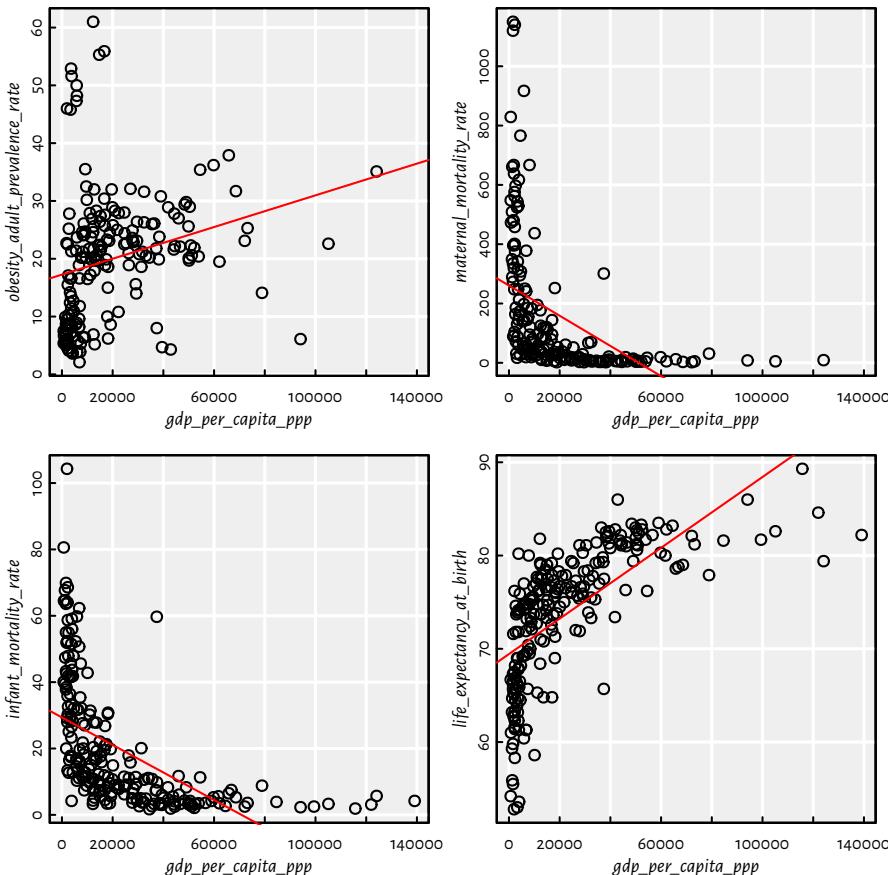


Figure 7.20: A scatter plot matrix and regression lines for the 4 variables most correlated with the per capita GDP

Recall that the root mean squared error is the square root of the arithmetic mean of the squared residuals. Mean absolute error is the average of the absolute values of the residuals. The coefficient of determination is given by: $R^2(f) = 1 - \frac{\sum_{i=1}^n (y_i - f(\mathbf{x}_{i,.}))^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$.

Comment: Unfortunately, we were misled by the high correlation coefficients between

the Xs and Ys: the low actual R^2 scores indicate that these models should not be deemed trustworthy. Note that 3 of the plots are evidently L-shaped.

Fun fact: (*) Interestingly, it can be shown that R^2 (in the case of the linear models fitted by minimising the SSR) is the square of the correlation between the true Ys and the predicted Ys:

```
X <- factbook[, "gdp_per_capita_ppp"]
Y <- factbook[, most_correlated[i]]
f <- lm(Y~X, y=TRUE)
print(summary(f)$r.squared)

## [1] 0.43096
print(cor(f$fitted.values, f$y)^2)

## [1] 0.43096
```

Side note: Do note that RMSE and MAE are interpretable: for instance, average error of life expectancy prediction based on the GDP is 4-5 years. Recall that you can find the information on the variables' units of measure at <https://www.cia.gov/library/publications/the-world-factbook/docs/rankorderguide.html>.



7.4.3 Countries of the World – Most correlated variables (*)

Let's get back to the World Factbook 2020 dataset.

```
factbook <- read.csv("datasets/world_factbook_2020.csv",
comment.char="#" )
```

Exercise 7.8 Create a data frame C with three columns named $col1$, $col2$ and r and $p(p - 1)/2$ rows, where p is the number of numeric features in $factbook$. Every row should represent a unique pair of column names in $factbook$ (we do not distinguish between a, b and b, a) of correlation coefficients between them.

Solution.

First we will solve this exercise considering only 4 numeric features in our dataset, so that we can keep track of how the R expressions we evaluate actually work.

Let us compute the Pearson coefficients between chosen pairs of variables.

```
R <- cor(factbook[,c("area", "median_age", "birth_rate", "exports")],
use="complete.obs") # 4 selected columns
print(R)

##                  area median_age birth_rate  exports
## area       1.000000  0.044524 -0.031995  0.49259
```

```
## median_age  0.044524  1.000000 -0.921592  0.29973
## birth_rate -0.031995 -0.921592  1.000000 -0.24296
## exports      0.492586  0.299727 -0.242955  1.00000
```

Note that the *R* matrix has 1.0 on the diagonal (where each entry represents a correlation between a variable and itself). Moreover, it is symmetric around the diagonal— $R[i, j] == R[j, i]$, because it is the correlation between the same pair of variables. Hence, from now on we may be interested in the elements below the diagonal. We can get access to them by using `lower.tri()` (“lower triangle”).

```
R[lower.tri(R)]
```

```
## [1]  0.044524 -0.031995  0.492586 -0.921592  0.299727 -0.242955
```

This is already the 3rd column of the data frame we are asked to generate, which should look like:

```
##          col1      col2      r
## 1 median_age    area  0.044524
## 2 birth_rate    area -0.031995
## 3 exports       area  0.492586
## 4 birth_rate median_age -0.921592
## 5 exports median_age  0.299727
## 6 exports birth_rate -0.242955
```

How the generate `col1` and `col2`? One idea is to take the “lower triangles” of the following matrices:

```
##      [,1]      [,2]      [,3]      [,4]
## [1,] "area"    "area"    "area"    "area"
## [2,] "median_age" "median_age" "median_age" "median_age"
## [3,] "birth_rate" "birth_rate" "birth_rate" "birth_rate"
## [4,] "exports"   "exports"   "exports"   "exports"
```

and:

```
##      [,1]      [,2]      [,3]      [,4]
## [1,] "area"    "median_age" "birth_rate" "exports"
## [2,] "area"    "median_age" "birth_rate" "exports"
## [3,] "area"    "median_age" "birth_rate" "exports"
## [4,] "area"    "median_age" "birth_rate" "exports"
```

Here is a complete solution for all the features is `factbook`:

```
R <- cor(factbook[,-1], use="complete.obs") # skip the `country` column
rrr <- matrix(dimnames(R)[[1]], nrow=nrow(R), ncol=ncol(R))
ccc <- matrix(dimnames(R)[[2]], byrow=TRUE, nrow=nrow(R), ncol=ncol(R))
C <- data.frame(col1=rrr[lower.tri(rrr)],
                 col2=ccc[lower.tri(ccc)],
                 r=R[lower.tri(R)])
```

Comment: In “classical” programming languages we would perhaps have used of a double (nested) for loop here (a less readable solution).

Exercise 7.9 Find the 5 most correlated pairs of variables.

Solution.

This can be done by ordering the rows of C in decreasing order of absolute values of $C\$r$, and then choosing the first 5 rows.

```
C_top <- head(C[order(abs(C$r), decreasing=TRUE),], 5)
knitr::kable(C_top)
```

	col1	col2	r
1687	electricity_installed_generating_capacity	electricity_production	0.99942
1684	electricity_consumption	electricity_production	0.99921
88	labor_force	population	0.99862
1718	electricity_installed_generating_capacity	electricity_consumption	0.99815
1300	telephones_mobile_cellular	labor_force	0.99793

Comment: The most correlated pairs of features are not really “mind-blowing” ...

Exercise 7.10 Fit simple regression models for the most correlated pair of variables.

Solution.

There is a degree of ambiguity here: should `col1` or rather `col2` be treated as the dependent variable in our model? Let's do it either way.

To learn something new, which is exactly why we are all here, we will create the formulas programmatically, by first concatenating (joining) appropriate strings (note that in order to input a double quotes character, we need to proceed in with a backslash), and then calling the `formula()` function.

```
form <- formula(paste(C_top[1,2], " ~ ", C_top[1,1]))
f <- lm(form, data=factbook)
print(f)
```

```
##
## Call:
## lm(formula = form, data = factbook)
##
## Coefficients:
##                               (Intercept)
##                               7.95e+08
## electricity_installed_generating_capacity
##                               3.63e+03
```

```
plot(factbook[,C_top[1,1]], factbook[,C_top[1,2]],
      xlab=C_top[1,1], ylab=C_top[1,2])
abline(f, col="red")
```

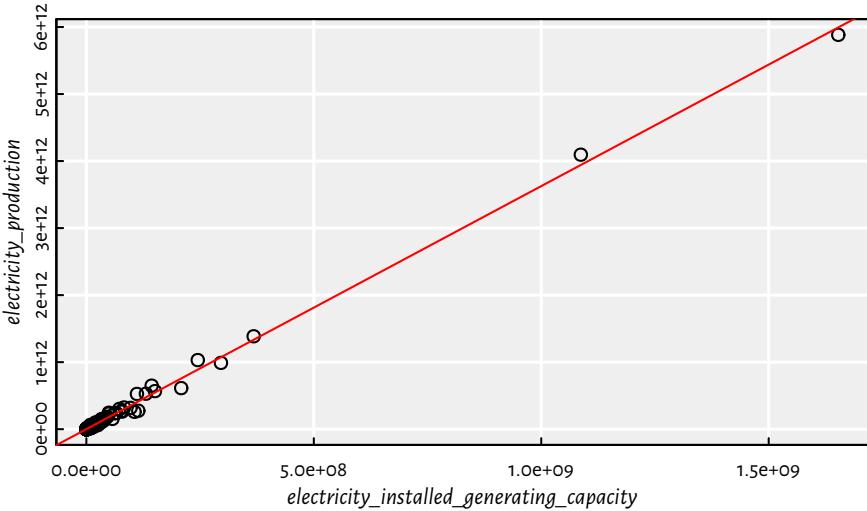


Figure 7.21: Most correlated pair of variables and the fitted regression line

See Figure 7.21 for the result.

7.4.4 Countries of the World – A non-linear model based on the GDP per capita

Let's revisit the World Factbook 2020 dataset.

```
factbook <- read.csv("datasets/world_factbook_2020.csv",
                      comment.char="#")
```

Exercise 7.11 Draw a histogram of the empirical distribution of the GDP per capita. Moreover, draw a histogram of the logarithm of the GDP/person.

Solution.

```
par(mfrow=c(1,2))
hist(factbook$gdp_per_capita_ppp, col="white", main=NA)
hist(log(factbook$gdp_per_capita_ppp), col="white", main=NA)
```

Comment: In Figure 7.22 we see that distribution of the GDP is right-skewed: most countries have small GDP. However, few of them (those in the “right tail” of the distribution) are very very rich (hey, how about taxing the richest countries?!). There is the famous

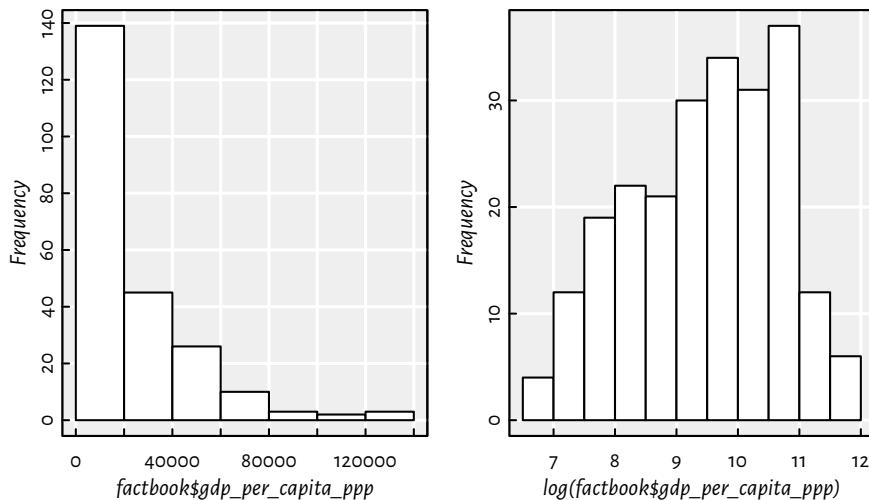


Figure 7.22: Histograms of the empirical distribution of the GDP per capita with linear (left) and log (right) scale on the X axis

observation made by V. Pareto stating that most assets are in the hands of the “wealthy minority” (compare: power law, rich-get-richer rule, preferential attachment in complex networks). Interestingly, many real-world-phenomena are distributed similarly (e.g., the popularity of web pages, the number of followers of Instagram profiles). It is frequently the case that the logarithm of the aforementioned variable looks more “normal” (is bell-shaped).

Side note: “The” logarithm most often refers to the logarithm base e , $\log x = \log_e x$, where $e \approx 2.72$ is the Euler constant, see `exp(1)` in R. Note that you can only compute logarithms of positive real numbers.

Non-technical audience might be confused when asked to contemplate the distribution of the logarithm of a variable. Let's make it more user-friendly (on the other hand, we could've asked them to harden up...) by nicely re-labelling the X axis, see Figure 7.23.

```
hist(log(factbook$gdp_per_capita_ppp), axes=FALSE,
     xlab="GDP per capita (thousands USD)", main=NA, col="white")
box()
axis(2) # Y axis
at <- c(1000, 2000, 5000, 10000, 20000, 50000, 100000, 200000)
axis(1, at=log(at), labels=at/1000)
```

Comment: This is still a plot of the logarithm of the distribution of the per capita GDP, but it's somehow “hidden” behind the human-readable axis labels. Nice. ■

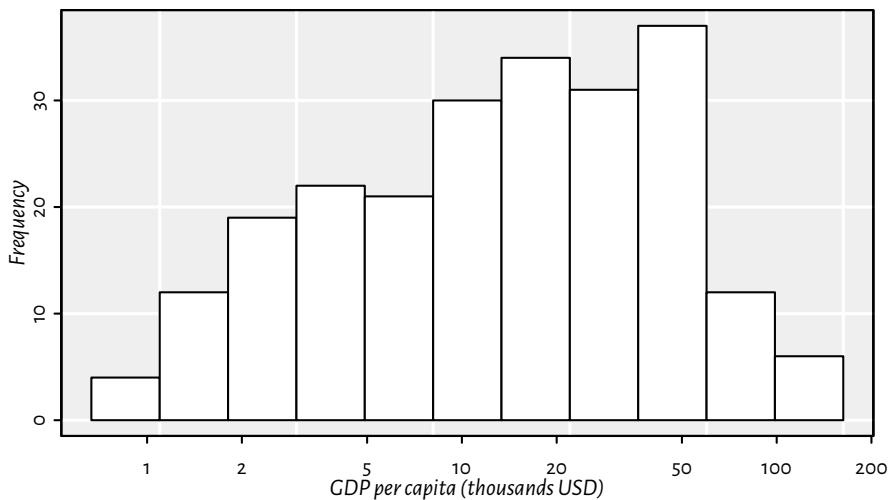


Figure 7.23: Histogram of the empirical distribution of the GDP per capita now with human-readable X axis labels (not the logarithmic scale)

Exercise 7.12 Fit a simple linear model of `life_expectancy_at_birth` as a function of `gdp_per_capita_ppp`.

Solution.

Easy. We have already done than in one of the previous exercises. Yet, to learn something new, let's note that the `plot()` function accepts formulas as well.

```
f <- lm(life_expectancy_at_birth ~ gdp_per_capita_ppp, data=factbook)
plot(life_expectancy_at_birth ~ gdp_per_capita_ppp, data=factbook)
abline(f, col="purple")
summary(f)$r.squared
```

```
## [1] 0.43096
```

Comment: From Figure 7.24 we see that this is not a good model. ■

Exercise 7.13 Draw a scatter plot of `life_expectancy_at_birth` as a function `gdp_per_capita_ppp`, with the X axis being logarithmic. Compute the correlation coefficient between `log(gdp_per_capita_ppp)` and `life_expectancy_at_birth`.

Solution.

We could apply the `log()`-transformation manually and generate fancy X axis labels ourselves.

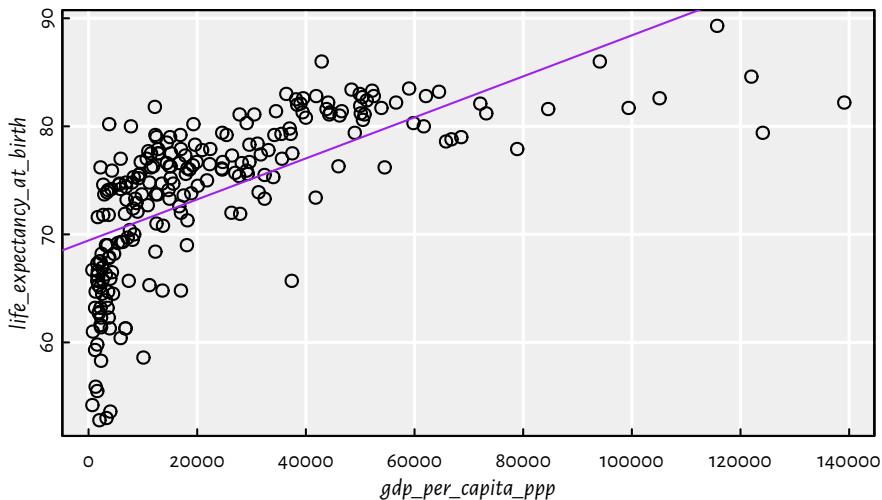


Figure 7.24: Linear model fitted for life expectancy vs. GDP/person

However, the `plot()` function has the `log` argument (see `?plot.default`) which provides us with all we need, see Figure 7.25.

```
plot(factbook$gdp_per_capita_ppp,
      factbook$life_expectancy_at_birth,
      log="x")
```

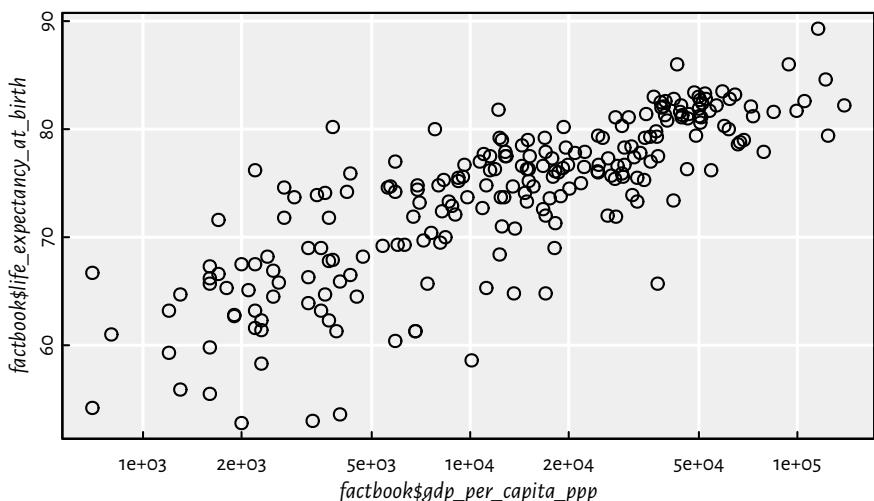


Figure 7.25: Scatter plot of life expectancy vs. GDP/person with log scale on the X axis

Here is the linear correlation coefficient between the logarithm of the GDP/person and the life expectancy.

```
cor(log(factbook$gdp_per_capita_ppp), factbook$life_expectancy_at_birth,
  use="complete.obs")
```

```
## [1] 0.80665
```

The correlation is quite high, hence the following task. ■

Exercise 7.14 Fit a model predicting `life_expectancy_at_birth` by means of `log(gdp_per_capita_ppp)`.

Solution.

We would like to fit a model of the form $Y = a \log X + b$. The formula `life_expectancy_at_birth ~ log(gdp_per_capita_ppp)` should do the trick here.

```
f <- lm(life_expectancy_at_birth ~ log(gdp_per_capita_ppp), data=factbook)
plot(life_expectancy_at_birth ~ log(gdp_per_capita_ppp), data=factbook)
abline(f, col="red", lty=3)
f$coefficients
```

```
##             (Intercept) log(gdp_per_capita_ppp)
##                  28.3064                 4.8178
```

```
summary(f)$r.squared
```

```
## [1] 0.65069
```

Comment: That is an okay model (in terms of the coefficient of determination), see Figure 7.26. ■

Exercise 7.15 Draw the fitted logarithmic model on a scatter plot with a standard, non-logarithmic X axis.

Solution.

The model fitted above is of the form $Y \approx 4.82 \log X + 28.31$. To depict it on a plot with linear (non-logarithmic) axes, we can compute this formula on multiple points by hand, see Figure 7.27.

```
plot(factbook$gdp_per_capita_ppp, factbook$life_expectancy_at_birth)
```

```
# many points on the X axis:
```

```
xxx <- seq(min(factbook$gdp_per_capita_ppp, na.rm=TRUE),
            max(factbook$gdp_per_capita_ppp, na.rm=TRUE),
            length.out=101)
```

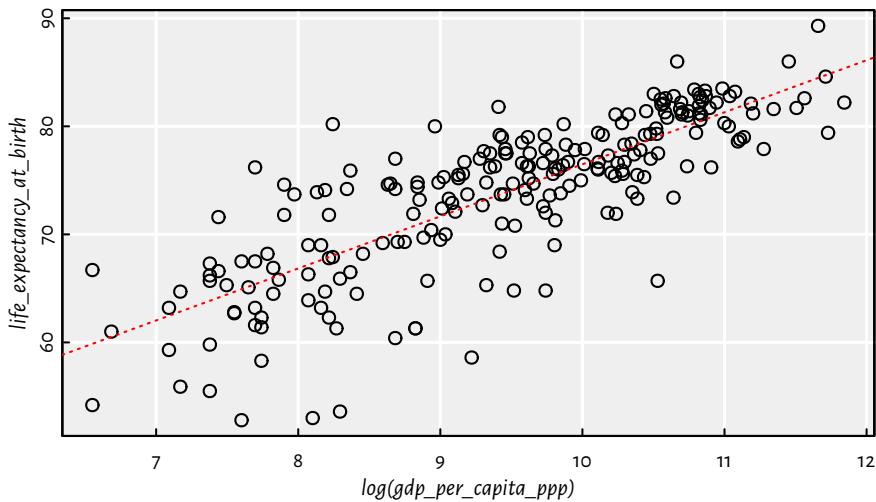


Figure 7.26: Linear model fitted for life expectancy vs. the logarithm of GDP/person

```
yyy <- f$coefficients[1] + f$coefficients[2]*log(xxx)
lines(xxx, yyy, col="red", lty=3)
```

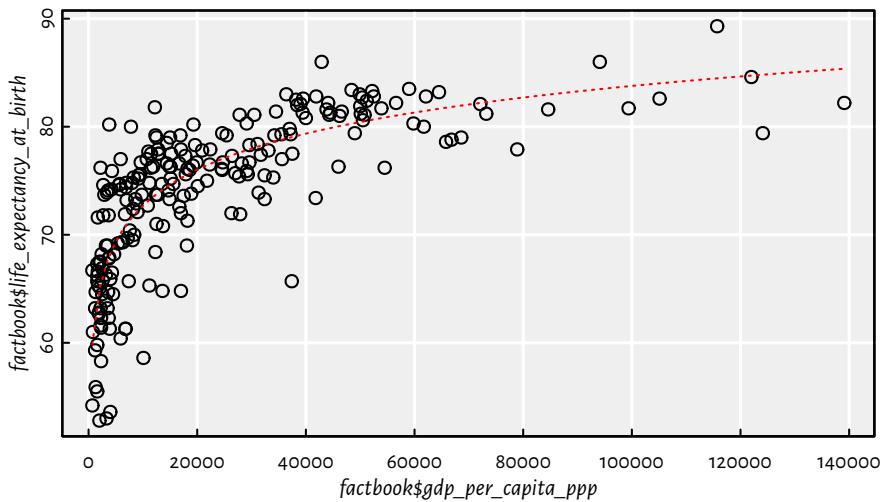


Figure 7.27: Logarithmic model fitted for life expectancy vs. GDP/person

Comment: Well, people are not immortal... The original (linear) model didn't really take that into account. Also, recall that correlation is not causation. Moreover, there is a lot of variability at an individual level. Being born in a less-wealthy country (e.g., not in a


```
## exports          0.81899 0.94241 1.00000
```

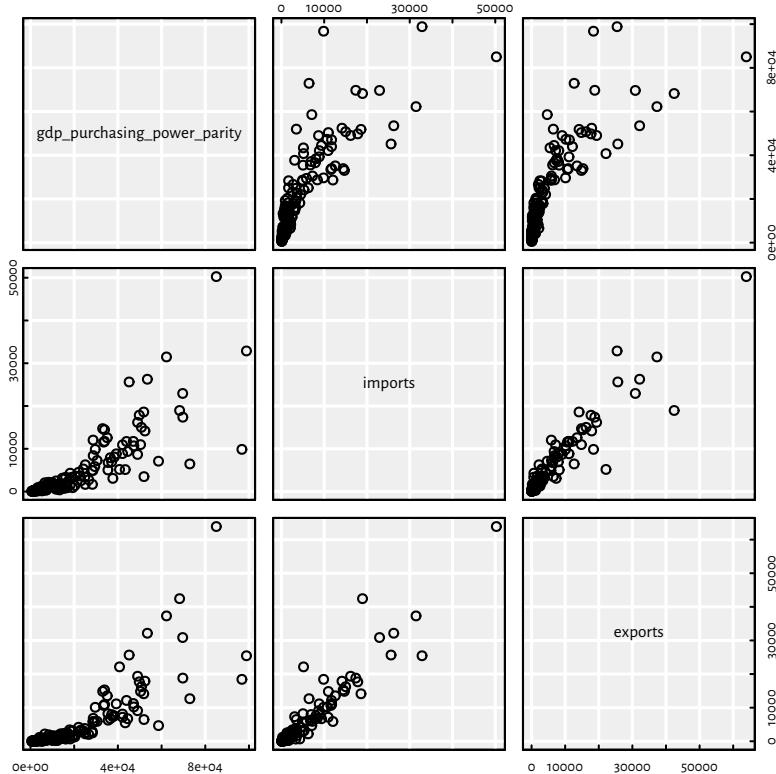


Figure 7.28: Scatter plot matrix for GDP, imports and exports

They are nicely correlated. Moreover, they are on a similar scale (“tens of thousands of USD per capita”).

Fitting the requested model yields:

```
options(scipen=10) # prefer "decimal" over "scientific" notation
f1 <- lm(gdp_purchasing_power_parity~imports+exports, data=factbookn)
f1$coefficients

## (Intercept)      imports      exports
##  9852.53813     1.44194     0.78067

summary(f1)$adj.r.squared

## [1] 0.69598
```

Exercise 7.17 Use forward selection to come up with a model for *gdp_purchasing_power_parity per capita*.

Solution.

```
(model_empty <- gdp_purchasing_power_parity~1)

## gdp_purchasing_power_parity ~ 1

(model_full <- formula(model.frame(gdp_purchasing_power_parity~, data=factbookn)))

## gdp_purchasing_power_parity ~ imports + exports + electricity_exports +
##     electricity_imports + military_expenditures + crude_oil_exports +
##     crude_oil_imports + natural_gas_exports + natural_gas_imports +
##     reserves_of_foreign_exchange_and_gold

f2 <- step(lm(model_empty, data=factbookn),
           scope=model_full,
           direction="forward", trace=0)
f2

## 
## Call:
## lm(formula = gdp_purchasing_power_parity ~ imports + crude_oil_exports +
##     crude_oil_imports + electricity_imports + natural_gas_imports,
##     data = factbookn)
## 
## Coefficients:
##             (Intercept)      imports    crude_oil_exports
##                 7603.24        1.77        128472.22
##     crude_oil_imports  electricity_imports  natural_gas_imports
##                 100781.64        1.62          3.13
summary(f2)$adj.r.squared

## [1] 0.7865
```

Comment: Interestingly, it's mostly the import-related variables that contribute to the GDP per capita. However, the model is not perfect, so we should refrain ourselves from building a brand new economic theory around this "discovery". On the other hand, you know what they say: all models are wrong, but some might be useful. Note that we used the adjusted R^2 coefficient to correct for the number of variables in the model so as to make it more comparable with the coefficient corresponding to the f1 model.

Exercise 7.18 Use backward elimination to construct a model for *gdp_purchasing_power_parity per capita*.

Solution.

```
f3 <- step(lm(model_full, data=factbookn),
            scope=model_empty,
            direction="backward", trace=0)
f3

## 
## Call:
## lm(formula = gdp_purchasing_power_parity ~ imports + electricity_imports +
##      crude_oil_exports + crude_oil_imports + natural_gas_imports,
##      data = factbookn)
##
## Coefficients:
##             (Intercept)           imports   electricity_imports
##                   7603.24              1.77                  1.62
##      crude_oil_exports  crude_oil_imports natural_gas_imports
##                   128472.22             100781.64                  3.13
summary(f3)$adj.r.squared
```

```
## [1] 0.7865
```

Comment: This is the same model as the one found by forward selection, i.e., f2.



7.4.6 Median House Value in Boston (Continued)

Let's get back to the Boston dataset from the MASS package that we considered in the chapter on simple linear regression:

```
library("MASS")
head(Boston, 3)

##      crim zn indus chas   nox     rm    age     dis rad tax ptratio black
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900    1 296  15.3 396.90
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671    2 242  17.8 396.90
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671    2 242  17.8 392.83
##      lstat medv
## 1  4.98 24.0
## 2  9.14 21.6
## 3  4.03 34.7
```

Read the description of each of the 14 columns in the dataset's manual, see `?Boston`.

Exercise 7.19 Construct a multiple regression model for `medv` as a function of `lstat`, `rm` and `tax`.

Exercise 7.20 Use forward selection (with respect to the AIC criterion) to come up with a multiple regression model for `medv` as a function of other variables.

Exercise 7.21 Use backward elimination (with respect to the AIC criterion) to construct a multiple regression model for *medv* as a function of other variables.

Exercise 7.22 Construct a multiple regression model for *medv* as a function of *lstat*, *rm* and *tax* transformed in various ways. Apply logarithms, squares, exponential (amongst others) functions on the variables and try to come up with the best model by trial and error (you can also use forward selection and/or backward elimination for this).

Exercise 7.23 For each model, draw the plot of the residuals ($\hat{y}_i - y_i$) as a function of the predicted outputs (\hat{y}_i). Describe these plots in your own words.

Exercise 7.24 Compare the four models in terms of AIC, RMSE, MAE and adjusted R^2 . Which model is the best with regards to each metric? Draw conclusions in your own words.

7.5 Outro

7.5.1 Remarks

Multiple regression is simple, fast to apply and interpretable.

Linear models go beyond fitting of straight lines and other hyperplanes!

A complex model may overfit and hence generalise poorly to unobserved inputs.

Note that the SSR criterion makes the models sensitive to outliers.

Remember:

good models

=

better understanding of the modelled reality + better predictions

=

more revenue, your boss' happiness, your startup's growth etc.

7.5.2 Other Methods for Regression

Other example approaches to regression:

- ridge regression,
- lasso regression,
- least absolute deviations (LAD) regression,
- multiadaptive regression splines (MARS),
- K-nearest neighbour (K-NN) regression, see `FNN::knn.reg()` in R,
- regression trees,
- support-vector regression (SVR),
- neural networks (also deep) for regression.

7.5.3 Derivation of the Solution (**)

We would like to find an analytical solution to the problem of minimising of the sum of squared residuals:

$$\min_{\beta_0, \beta_1, \dots, \beta_p \in \mathbb{R}} E(\beta_0, \beta_1, \dots, \beta_p) = \sum_{i=1}^n (\beta_0 + \beta_1 x_{i,1} + \dots + \beta_p x_{i,p} - y_i)^2$$

This requires computing the $p + 1$ partial derivatives $\partial E / \partial \beta_j$ for $j = 0, \dots, p$.

The partial derivatives are very similar to each other; $\frac{\partial E}{\partial \beta_0}$ is given by:

$$\frac{\partial E}{\partial \beta_0}(\beta_0, \beta_1, \dots, \beta_p) = 2 \sum_{i=1}^n (\beta_0 + \beta_1 x_{i,1} + \dots + \beta_p x_{i,p} - y_i)$$

and $\frac{\partial E}{\partial \beta_j}$ for $j > 0$ is equal to:

$$\frac{\partial E}{\partial \beta_j}(\beta_0, \beta_1, \dots, \beta_p) = 2 \sum_{i=1}^n x_{i,j} (\beta_0 + \beta_1 x_{i,1} + \dots + \beta_p x_{i,p} - y_i)$$

Then all we need to do is to solve the system of linear equations:

$$\begin{cases} \frac{\partial E}{\partial \beta_0}(\beta_0, \beta_1, \dots, \beta_p) = 0 \\ \frac{\partial E}{\partial \beta_1}(\beta_0, \beta_1, \dots, \beta_p) = 0 \\ \vdots \\ \frac{\partial E}{\partial \beta_p}(\beta_0, \beta_1, \dots, \beta_p) = 0 \end{cases}$$

The above system of $p + 1$ linear equations, which we are supposed to solve for $\beta_0, \beta_1, \dots, \beta_p$:

$$\begin{cases} 2 \sum_{i=1}^n (\beta_0 + \beta_1 x_{i,1} + \dots + \beta_p x_{i,p} - y_i) = 0 \\ 2 \sum_{i=1}^n x_{i,1} (\beta_0 + \beta_1 x_{i,1} + \dots + \beta_p x_{i,p} - y_i) = 0 \\ \vdots \\ 2 \sum_{i=1}^n x_{i,p} (\beta_0 + \beta_1 x_{i,1} + \dots + \beta_p x_{i,p} - y_i) = 0 \end{cases}$$

can be rewritten as:

$$\begin{cases} \sum_{i=1}^n (\beta_0 + \beta_1 x_{i,1} + \dots + \beta_p x_{i,p}) = \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i,1} (\beta_0 + \beta_1 x_{i,1} + \dots + \beta_p x_{i,p}) = \sum_{i=1}^n x_{i,1} y_i \\ \vdots \\ \sum_{i=1}^n x_{i,p} (\beta_0 + \beta_1 x_{i,1} + \dots + \beta_p x_{i,p}) = \sum_{i=1}^n x_{i,p} y_i \end{cases}$$

and further as:

$$\begin{cases} \beta_0 n + \beta_1 \sum_{i=1}^n x_{i,1} + \dots + \beta_p \sum_{i=1}^n x_{i,p} = \sum_{i=1}^n y_i \\ \beta_0 \sum_{i=1}^n x_{i,1} + \beta_1 \sum_{i=1}^n x_{i,1} x_{i,1} + \dots + \beta_p \sum_{i=1}^n x_{i,1} x_{i,p} = \sum_{i=1}^n x_{i,1} y_i \\ \vdots \\ \beta_0 \sum_{i=1}^n x_{i,p} + \beta_1 \sum_{i=1}^n x_{i,p} x_{i,1} + \dots + \beta_p \sum_{i=1}^n x_{i,p} x_{i,p} = \sum_{i=1}^n x_{i,p} y_i \end{cases}$$

Note that the terms involving $x_{i,j}$ and y_i (the sums) are all constant – these are some fixed real numbers. We have learned how to solve such problems in high school.

Exercise 7.25 Try deriving the analytical solution and implementing it for $p = 2$. Recall that in the previous chapter we solved the special case of $p = 1$.

7.5.4 Solution in Matrix Form (***)

Assume that $\mathbf{X} \in \mathbb{R}^{n \times p}$ (a matrix with inputs), $\mathbf{y} \in \mathbb{R}^{n \times 1}$ (a column vector of reference outputs) and $\boldsymbol{\beta} \in \mathbb{R}^{(p+1) \times 1}$ (a column vector of parameters).

Firstly, note that a linear model of the form:

$$f_{\boldsymbol{\beta}}(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

can be rewritten as:

$$f_{\boldsymbol{\beta}}(\mathbf{x}) = \beta_0 1 + \beta_1 x_1 + \cdots + \beta_p x_p = \dot{\mathbf{x}} \boldsymbol{\beta},$$

where $\dot{\mathbf{x}} = [1 \ x_1 \ x_2 \ \cdots \ x_p]$.

Similarly, if we assume that $\dot{\mathbf{X}} = [1 \ \mathbf{X}] \in \mathbb{R}^{n \times (p+1)}$ is the input matrix with a prepended column of 1s, i.e., $1 = [1 \ 1 \ \cdots \ 1]^T$ and $\dot{x}_{i,0} = 1$ (for brevity of notation the columns added will have index 0), $\dot{x}_{i,j} = x_{i,j}$ for all $j \geq 1$ and all i , then:

$$\hat{\mathbf{y}} = \dot{\mathbf{X}} \boldsymbol{\beta}$$

gives the vector of predicted outputs for every input point.

This way, the sum of squared residuals

$$E(\beta_0, \beta_1, \dots, \beta_p) = \sum_{i=1}^n (\beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} - y_i)^2$$

can be rewritten as:

$$E(\boldsymbol{\beta}) = \|\dot{\mathbf{X}} \boldsymbol{\beta} - \mathbf{y}\|^2,$$

where as usual $\|\cdot\|^2$ denotes the squared Euclidean norm.

Recall that this can be re-expressed as:

$$E(\boldsymbol{\beta}) = (\dot{\mathbf{X}} \boldsymbol{\beta} - \mathbf{y})^T (\dot{\mathbf{X}} \boldsymbol{\beta} - \mathbf{y}).$$

In order to find the minimum of E w.r.t. $\boldsymbol{\beta}$, we need to find the parameters that make the partial derivatives vanish, i.e.:

$$\left\{ \begin{array}{lcl} \frac{\partial E}{\partial \beta_0}(\boldsymbol{\beta}) & = & 0 \\ \frac{\partial E}{\partial \beta_1}(\boldsymbol{\beta}) & = & 0 \\ \vdots & & \\ \frac{\partial E}{\partial \beta_p}(\boldsymbol{\beta}) & = & 0 \end{array} \right.$$

Remark. (***) Interestingly, the above can also be expressed in matrix form, using the special notation:

$$\nabla E(\boldsymbol{\beta}) = 0$$

Here, ∇E (nabla symbol = differential operator) denotes the function gradient, i.e., the vector of all partial derivatives. This is nothing more than syntactic sugar for this quite commonly applied operator.

Anyway, the system of linear equations we have derived above:

$$\left\{ \begin{array}{lcl} \beta_0 n + \beta_1 \sum_{i=1}^n x_{i,1} + \cdots + \beta_p \sum_{i=1}^n x_{i,p} & = & \sum_{i=1}^n y_i \\ \beta_0 \sum_{i=1}^n x_{i,1} + \beta_1 \sum_{i=1}^n x_{i,1}x_{i,1} + \cdots + \beta_p \sum_{i=1}^n x_{i,1}x_{i,p} & = & \sum_{i=1}^n x_{i,1}y_i \\ \vdots \\ \beta_0 \sum_{i=1}^n x_{i,p} + \beta_1 \sum_{i=1}^n x_{i,p}x_{i,1} + \cdots + \beta_p \sum_{i=1}^n x_{i,p}x_{i,p} & = & \sum_{i=1}^n x_{i,p}y_i \end{array} \right.$$

can be rewritten in matrix terms as:

$$\left\{ \begin{array}{lcl} \beta_0 \dot{\mathbf{x}}_{:,0}^T \dot{\mathbf{x}}_{:,0} + \beta_1 \dot{\mathbf{x}}_{:,0}^T \dot{\mathbf{x}}_{:,1} + \cdots + \beta_p \dot{\mathbf{x}}_{:,0}^T \dot{\mathbf{x}}_{:,p} & = & \dot{\mathbf{x}}_{:,0}^T \mathbf{y} \\ \beta_0 \dot{\mathbf{x}}_{:,1}^T \dot{\mathbf{x}}_{:,0} + \beta_1 \dot{\mathbf{x}}_{:,1}^T \dot{\mathbf{x}}_{:,1} + \cdots + \beta_p \dot{\mathbf{x}}_{:,1}^T \dot{\mathbf{x}}_{:,p} & = & \dot{\mathbf{x}}_{:,1}^T \mathbf{y} \\ \vdots \\ \beta_0 \dot{\mathbf{x}}_{:,p}^T \dot{\mathbf{x}}_{:,0} + \beta_1 \dot{\mathbf{x}}_{:,p}^T \dot{\mathbf{x}}_{:,1} + \cdots + \beta_p \dot{\mathbf{x}}_{:,p}^T \dot{\mathbf{x}}_{:,p} & = & \dot{\mathbf{x}}_{:,p}^T \mathbf{y} \end{array} \right.$$

This can be restated as:

$$\left\{ \begin{array}{lcl} (\dot{\mathbf{x}}_{:,0}^T \dot{\mathbf{X}}) \boldsymbol{\beta} & = & \dot{\mathbf{x}}_{:,0}^T \mathbf{y} \\ (\dot{\mathbf{x}}_{:,1}^T \dot{\mathbf{X}}) \boldsymbol{\beta} & = & \dot{\mathbf{x}}_{:,1}^T \mathbf{y} \\ \vdots \\ (\dot{\mathbf{x}}_{:,p}^T \dot{\mathbf{X}}) \boldsymbol{\beta} & = & \dot{\mathbf{x}}_{:,p}^T \mathbf{y} \end{array} \right.$$

which in turn is equivalent to:

$$(\dot{\mathbf{X}}^T \dot{\mathbf{X}}) \boldsymbol{\beta} = \dot{\mathbf{X}}^T \mathbf{y}.$$

Such a system of linear equations in matrix form can be solved numerically using, amongst others, the `solve()` function.

Remark. (***) In practice, we'd rather rely on QR or SVD decompositions of matrices for efficiency and numerical accuracy reasons.

Numeric example – solution via `lm()`:

```
X1 <- as.numeric(Credit$Balance[Credit$Balance>0])
X2 <- as.numeric(Credit$Income[Credit$Balance>0])
Y <- as.numeric(Credit$Rating[Credit$Balance>0])
lm(Y~X1+X2)$coefficients
```

	X1	X2
## (Intercept)		
## 172.5587	0.1828	2.1976

Recalling that $\mathbf{A}^T \mathbf{B}$ can be computed by calling `t(A) %*% B` or – even faster – by calling `crossprod(A, B)`, we can also use `solve()` to obtain the same result:

```
X_dot <- cbind(1, X1, X2)
solve( crossprod(X_dot, X_dot), crossprod(X_dot, Y) )

##      [,1]
## 172.5587
## X1   0.1828
## X2   2.1976
```

7.5.5 Pearson's r in Matrix Form (**)

Recall the Pearson linear correlation coefficient:

$$r(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

Denote with \mathbf{x}° and \mathbf{y}° the centred versions of \mathbf{x} and \mathbf{y} , respectively, i.e., $x_i^\circ = x_i - \bar{x}$ and $y_i^\circ = y_i - \bar{y}$.

Rewriting the above yields:

$$r(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^n x_i^\circ y_i^\circ}{\sqrt{\sum_{i=1}^n (x_i^\circ)^2} \sqrt{\sum_{i=1}^n (y_i^\circ)^2}}$$

which is exactly:

$$r(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^\circ \cdot \mathbf{y}^\circ}{\|\mathbf{x}^\circ\| \|\mathbf{y}^\circ\|}$$

i.e., the normalised dot product of the centred versions of the two vectors.

This is the cosine of the angle between the two vectors (in n -dimensional spaces)!

(**) Recalling from the previous chapter that $\mathbf{A}^T \mathbf{A}$ gives the dot product between all the pairs of columns in a matrix \mathbf{A} , we can implement an equivalent version of `cor(C)` as follows:

```
C <- Credit[Credit$Balance>0,
  c("Rating", "Limit", "Income", "Age",
    "Education", "Balance")]
C_centred <- apply(C, 2, function(c) c-mean(c))
C_normalised <- apply(C_centred, 2, function(c)
  c/sqrt(sum(c^2)))
round(t(C_normalised) %*% C_normalised, 3)

##          Rating Limit Income Age Education Balance
## Rating     1.000  0.996  0.831  0.167    -0.040   0.798
## Limit      0.996  1.000  0.834  0.164    -0.032   0.796
```

```
## Income    0.831  0.834  1.000  0.227    -0.033  0.414
## Age       0.167  0.164  0.227  1.000     0.024  0.008
## Education -0.040 -0.032 -0.033  0.024     1.000  0.001
## Balance    0.798  0.796  0.414  0.008     0.001  1.000
```

TODO

Recommended further reading: (James et al. 2017: Chapters 1, 2 and 3)

Other: (Hastie et al. 2017: Chapter 1, Sections 3.2 and 3.3)

Next chapter....

8

Classification with Linear Models

TODO In this chapter, we will:

- ...
 - ...
-

8.1 Introduction

8.1.1 Classification Task

Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be an input matrix that consists of n points in a p -dimensional space (each of the n objects is described by means of p numerical features)

Recall that in supervised learning, with each \mathbf{x}_i , we associate the desired output y_i .

Hence, our dataset is $[\mathbf{X} \mathbf{y}]$ – where each object is represented as a row vector $[\mathbf{x}_i, y_i]$, $i = 1, \dots, n$:

$$[\mathbf{X} \mathbf{y}] = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} & y_1 \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} & y_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} & y_n \end{bmatrix}.$$

In this chapter we are still interested in **classification** tasks; we assume that each y_i is a descriptive label.

Let's assume that we are faced with **binary classification** tasks.

Hence, there are only two possible labels that we traditionally denote with 0s and 1s.

For example:

0	1
no	yes
false	true

o	1
failure	success
healthy	ill

Let's recall the synthetic 2D dataset from the previous chapter (true decision boundary is at $X_1 = 0$), see Figure 8.1.

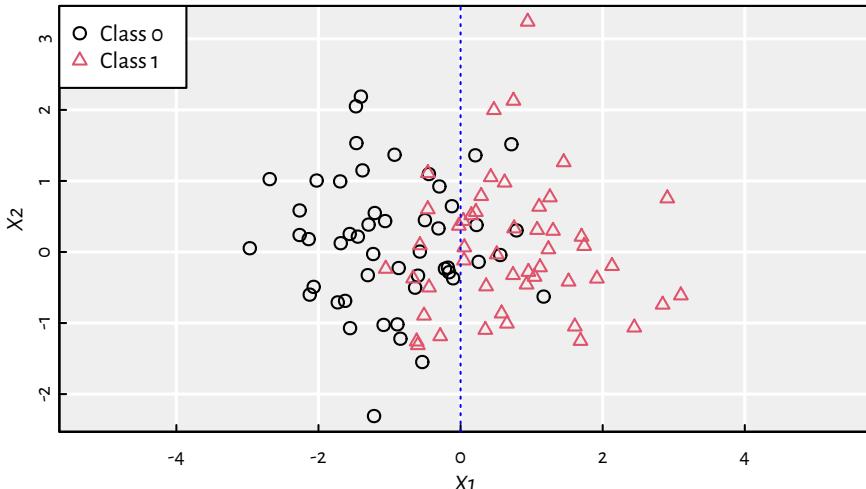


Figure 8.1: A synthetic 2D dataset with the true decision boundary at $X_1 = 0$

8.1.2 Data

For illustration, we'll be considering the Wine Quality dataset (white wines only):

```
wine_quality <- read.csv("datasets/wine_quality_all.csv",
  comment.char="#")
white_wines <- wine_quality[wine_quality$color == "white",]
(n <- nrow(white_wines)) # number of samples
```

[1] 4898

The input matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ consists of the first 10 numeric variables:

```
X <- as.matrix(white_wines[,1:10])
dim(X)
```

[1] 4898 10

```
head(X, 2) # first two rows

##      fixed.acidity volatile.acidity citric.acid residual.sugar
## 1600        7.0          0.27       0.36        20.7
## 1601        6.3          0.30       0.34        1.6
##      chlorides free.sulfur.dioxide total.sulfur.dioxide density pH
## 1600     0.045            45        170   1.001 3.0
## 1601     0.049            14        132   0.994 3.3
##      sulphates
## 1600     0.45
## 1601     0.49
```

The 11th variable measures the amount of alcohol (in %).

We will convert this dependent variable to a binary one:

- 0 == (alcohol < 12) == lower-alcohol wines,
- 1 == (alcohol >= 12) == higher-alcohol wines

```
# recall that TRUE == 1
Y <- factor(as.character(as.numeric(white_wines$alcohol >= 12)))
table(Y)
```

```
## Y
##   0   1
## 4085 813
```

60/40% train-test split:

```
set.seed(123) # reproducibility matters
random_indices <- sample(n)
head(random_indices) # preview

## [1] 2463 2511 2227  526 4291 2986

# first 60% of the indices (they are arranged randomly)
# will constitute the train sample:
train_indices <- random_indices[1:floor(n*0.6)]
X_train <- X[train_indices,]
Y_train <- Y[train_indices]
# the remaining indices (40%) go to the test sample:
X_test <- X[-train_indices,]
Y_test <- Y[-train_indices]

XY_train <- cbind(as.data.frame(X_train), Y=Y_train)
XY_test <- cbind(as.data.frame(X_test), Y=Y_test)
```

Let's also compute Z_train and Z_test, being the standardised versions of X_train and X_test, respectively.

```

means <- apply(X_train, 2, mean) # column means
sds   <- apply(X_train, 2, sd)   # column standard deviations
Z_train <- t(apply(X_train, 1, function(r) (r-means)/sds))
Z_test  <- t(apply(X_test, 1, function(r) (r-means)/sds))

get_metrics <- function(Y_pred, Y_test)
{
  C <- table(Y_pred, Y_test) # confusion matrix
  stopifnot(dim(C) == c(2, 2))
  c(Acc=(C[1,1]+C[2,2])/sum(C), # accuracy
    Prec=C[2,2]/(C[2,2]+C[2,1]), # precision
    Rec=C[2,2]/(C[2,2]+C[1,2]), # recall
    F=C[2,2]/(C[2,2]+0.5*C[1,2]+0.5*C[2,1]), # F-measure
    # Confusion matrix items:
    TN=C[1,1], FN=C[1,2],
    FP=C[2,1], TP=C[2,2]
  ) # return a named vector
}

```

8.2 Binary Logistic Regression

8.2.1 Motivation

Recall that for a regression task, we fitted a very simple family of models – the linear ones – by minimising the sum of squared residuals.

This approach was pretty effective.

(Very) theoretically, we could treat the class labels as numeric 0s and 1s and apply regression models in a binary classification task.

```

XY_train_r <- cbind(as.data.frame(X_train),
  Y=as.numeric(Y_train)-1 # 0.0 or 1.0
)
XY_test_r <- cbind(as.data.frame(X_test),
  Y=as.numeric(Y_test)-1 # 0.0 or 1.0
)
f_r <- lm(Y~density+residual.sugar+pH, data=XY_train_r)

Y_pred_r <- predict(f_r, XY_test_r)
summary(Y_pred_r)

##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
## -3.0468 -0.0211  0.1192  0.1645  0.3491  0.8892

```

The predicted outputs, \hat{Y} , are arbitrary real numbers, but we can convert them to binary ones by checking if, e.g., $\hat{Y} > 0.5$.

```
Y_pred <- as.numeric(Y_pred_r>0.5)
round(get_metrics(Y_pred, XY_test_r$Y), 3)
```

	Acc	Prec	Rec	F	TN	FN	FP	TP
##	0.927	0.865	0.647	0.740	1611.000	112.000	32.000	205.000

Remark. (*) The threshold $T = 0.5$ could even be treated as a free parameter we optimise for (w.r.t. different metrics over the validation sample), see Figure 8.2.

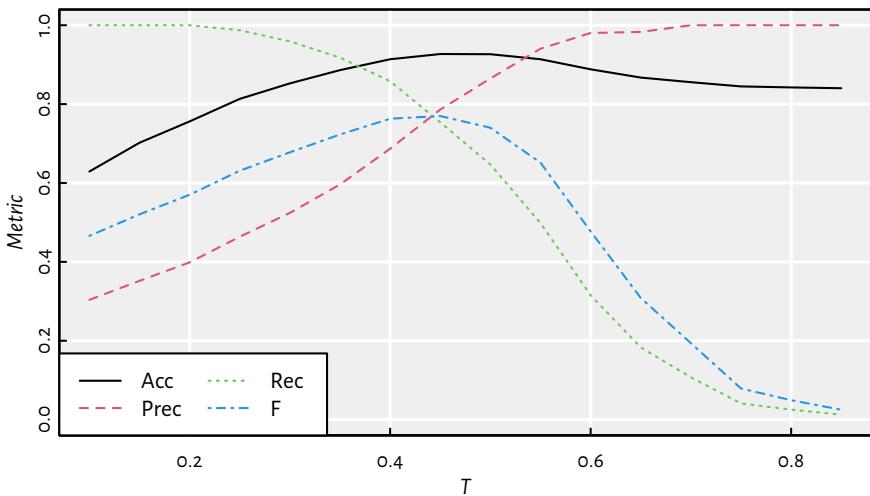


Figure 8.2: Quality metrics for a binary classifier “Classify X as 1 iff $f(X) > T$ and as 0 if $f(X) \leq T$ ”

Despite we can, we shouldn't use linear regression for classification. Treating class labels “0” and “1” as ordinary real numbers just doesn't cut it – we intuitively feel that we are doing something *ugly*. Luckily, there is a better, more meaningful approach that still relies on a linear model, but has the *right* semantics.

8.2.2 Logistic Model

Inspired by this idea, we could try modelling the **probability that a given point belongs to class 1**.

This could also provide us with the *confidence* in our prediction.

Probability is a number in $[0, 1]$, but the outputs of a linear model are arbitrary real numbers.

However, we could transform those real-valued outputs by means of some function $\phi : \mathbb{R} \rightarrow [0, 1]$ (preferably S-shaped == sigmoid), so as to get:

$$\Pr(Y = 1 | \mathbf{X}, \boldsymbol{\beta}) = \phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p).$$

Remark. The above reads as “Probability that Y is from class 1 given \mathbf{X} and $\boldsymbol{\beta}$ ”.

A popular choice is the **logistic sigmoid function**, see Figure 8.3:

$$\phi(t) = \frac{1}{1 + e^{-t}} = \frac{e^t}{1 + e^t}.$$

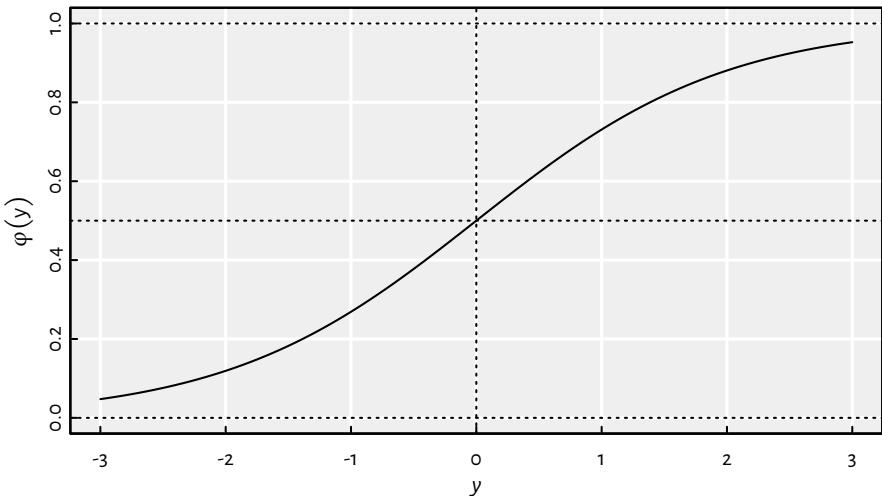


Figure 8.3: The logistic sigmoid function, φ

Hence our model becomes:

$$Y = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)}}$$

It is an instance of a **generalised linear model** (glm) (there are of course many other possible generalisations).

8.2.3 Example in R

Let us first fit a simple (i.e., $p = 1$) logistic regression model using the **density** variable. The goodness-of-fit measure used in this problem will be discussed a bit later.

```
(f <- glm(Y~density, data=XY_train, family=binomial("logit")))
##
## Call: glm(formula = Y ~ density, family = binomial("logit"), data = XY_train)
##
## Coefficients:
## (Intercept)      density
##       1173        -1184
##
## Degrees of Freedom: 2937 Total (i.e. Null);  2936 Residual
## Null Deviance:      2670
## Residual Deviance: 1420  AIC: 1420
```

"logit" above denotes the inverse of the logistic sigmoid function. The fitted coefficients are equal to:

```
f$coefficients
```

```
## (Intercept)      density
##       1173.2        -1184.2
```

Figure 8.4 depicts the obtained model, which can be written as:

$$\Pr(Y = 1|x) = \frac{1}{1 + e^{-(1173.21 - 1184.21x)}}$$

with $x = \text{density}$.

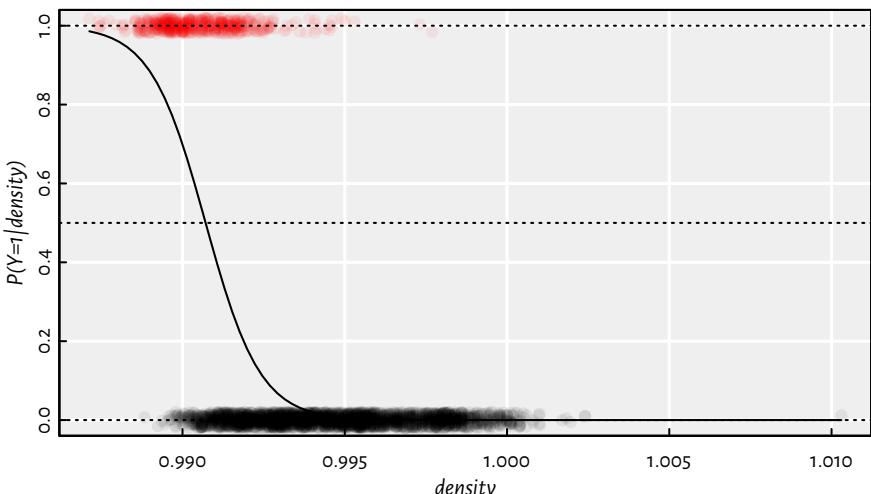


Figure 8.4: The probability that a given wine is a high-alcohol one given its density; black and red points denote the actual observed data points from the class 0 and 1, respectively

Some predicted probabilities:

```
round(head(predict(f, XY_test, type="response"), 12), 2)

## 1602 1605 1607 1608 1609 1613 1614 1615 1621 1622 1623 1627
## 0.01 0.01 0.00 0.02 0.03 0.36 0.00 0.31 0.36 0.06 0.03 0.00
```

We classify Y as 1 if the corresponding membership probability is greater than 0.5.

```
Y_pred <- as.numeric(predict(f, XY_test, type="response")>0.5)
get_metrics(Y_pred, Y_test)
```

	Acc	Prec	Rec	F	TN	FN
##	0.89796	0.72763	0.58991	0.65157	1573.00000	130.00000
##		FP	TP			
##	70.00000	187.00000				

And now a fit based on some other input variables:

```
(f <- glm(Y~density+residual.sugar+total.sulfur.dioxide,
  data=XY_train, family=binomial("logit"))

## 
## Call: glm(formula = Y ~ density + residual.sugar + total.sulfur.dioxide,
##   family = binomial("logit"), data = XY_train)
##
## Coefficients:
##             (Intercept)           density           residual.sugar
##                 2.50e+03            -2.53e+03            8.58e-01
##   total.sulfur.dioxide
##                 9.74e-03
##
## Degrees of Freedom: 2937 Total (i.e. Null); 2934 Residual
## Null Deviance: 2670
## Residual Deviance: 920 AIC: 928

Y_pred <- as.numeric(predict(f, XY_test, type="response")>0.5)
get_metrics(Y_pred, Y_test)

## 
## Acc Prec Rec F TN FN
## 0.93214 0.82394 0.73817 0.77870 1593.00000 83.00000
## FP TP
## 50.00000 234.00000
```

Exercise 8.1 Try fitting different models based on other sets of features.

8.2.4 Loss Function: Cross-entropy

The fitting of the model can be written as an optimisation task:

$$\min_{\beta_0, \beta_1, \dots, \beta_p \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n \epsilon(\hat{y}_i, y_i)$$

where $\epsilon(\hat{y}_i, y_i)$ denotes the penalty that measures the “difference” between the true y_i and its predicted version $\hat{y}_i = \Pr(Y = 1 | \mathbf{x}_{i,.}, \boldsymbol{\beta})$.

In the ordinary regression, we used the squared residual $\epsilon(\hat{y}_i, y_i) = (\hat{y}_i - y_i)^2$. In **logistic regression** (the kind of a classifier we are interested in right now), we use the **cross-entropy** (a.k.a. **log-loss**, binary cross-entropy),

$$\epsilon(\hat{y}_i, y_i) = -(y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i))$$

The corresponding loss function has not only many nice statistical properties (** related to maximum likelihood estimation etc.) but also an intuitive interpretation.

Note that the predicted \hat{y}_i is in $(0, 1)$ and the true y_i equals to either 0 or 1. Recall also that $\log t \in (-\infty, 0)$ for $t \in (0, 1)$. Therefore, the formula for $\epsilon(\hat{y}_i, y_i)$ has a very intuitive behaviour:

- if true $y_i = 1$, then the penalty becomes $\epsilon(\hat{y}_i, 1) = -\log(\hat{y}_i)$
 - \hat{y}_i is the probability that the classified input is indeed from class 1
 - we'd be happy if the classifier outputted $\hat{y}_i \approx 1$ in this case; this is not penalised as $-\log(t) \rightarrow 0$ as $t \rightarrow 1$
 - however, if the classifier is totally wrong, i.e., it thinks that $\hat{y}_i \approx 0$, then the penalty will be very high, as $-\log(t) \rightarrow +\infty$ as $t \rightarrow 0$
- if true $y_i = 0$, then the penalty becomes $\epsilon(\hat{y}_i, 0) = -\log(1 - \hat{y}_i)$
 - $1 - \hat{y}_i$ is the predicted probability that the input is from class 0
 - we penalise heavily the case where $1 - \hat{y}_i$ is small (we'd be happy if the classifier was sure that $1 - \hat{y}_i \approx 1$, because this is the ground-truth)

(*) Having said that, let's expand the above formulae. The task of minimising cross-entropy in the binary logistic regression can be written as $\min_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} E(\boldsymbol{\beta})$ with:

$$E(\boldsymbol{\beta}) = -\frac{1}{n} \sum_{i=1}^n y_i \log \Pr(Y = 1 | \mathbf{x}_{i,.}, \boldsymbol{\beta}) + (1 - y_i) \log(1 - \Pr(Y = 1 | \mathbf{x}_{i,.}, \boldsymbol{\beta}))$$

Taking into account that:

$$\Pr(Y = 1 | \mathbf{x}_{i,.}, \boldsymbol{\beta}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{i,1} + \dots + \beta_p x_{i,p})}},$$

we get:

$$E(\boldsymbol{\beta}) = -\frac{1}{n} \sum_{i=1}^n \left(\begin{array}{c} y_i \log \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{i,1} + \dots + \beta_p x_{i,p})}} \\ + (1 - y_i) \log \frac{e^{-(\beta_0 + \beta_1 x_{i,1} + \dots + \beta_p x_{i,p})}}{1 + e^{-(\beta_0 + \beta_1 x_{i,1} + \dots + \beta_p x_{i,p})}} \end{array} \right).$$

Logarithms are really practitioner-friendly functions, it holds:

- $\log 1 = 0$,
- $\log e = 1$ (where $e \approx 2.71828$ is the Euler constant; note that by writing \log we mean the natural a.k.a. base- e logarithm),
- $\log xy = \log x + \log y$,
- $\log x^p = p \log x$ (this is $\log(x^p)$, not $(\log x)^p$).

These facts imply, amongst others that:

- $\log e^x = x \log e = x$,
- $\log \frac{x}{y} = \log xy^{-1} = \log x + \log y^{-1} = \log x - \log y$ (of course for $y \neq 0$),
- $\log \frac{1}{y} = -\log y$

and so forth. Therefore, based on the fact that $1/(1 + e^{-x}) = e^x/(1 + e^x)$, the above optimisation problem can be rewritten as:

$$E(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \left(\begin{array}{c} y_i \log \left(1 + e^{-(\beta_0 + \beta_1 x_{i,1} + \dots + \beta_p x_{i,p})} \right) \\ + (1 - y_i) \log \left(1 + e^{+(\beta_0 + \beta_1 x_{i,1} + \dots + \beta_p x_{i,p})} \right) \end{array} \right)$$

or, if someone prefers:

$$E(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \left((1 - y_i) (\beta_0 + \beta_1 x_{i,1} + \dots + \beta_p x_{i,p}) + \log \left(1 + e^{-(\beta_0 + \beta_1 x_{i,1} + \dots + \beta_p x_{i,p})} \right) \right).$$

It turns out that there is no analytical formula for the optimal set of parameters ($\beta_0, \beta_1, \dots, \beta_p$ minimising the log-loss). In the chapter on optimisation, we shall see that the solution to the logistic regression can be solved numerically by means of quite simple iterative algorithms. The two expanded formulae have lost the appealing interpretation of the original one, however, it's more numerically well-behaving, see, e.g., the `log1p()` function in base R or, even better, `fermi_dirac_0()` in the `gsl` package.

8.3 Exercises

8.3.1 EdStats – Fitting of Binary Logistic Regression Models

In this task we're going to consider the “wide” version of the EdStats dataset again:

```
edstats <- read.csv("datasets/edstats_2019_wide.csv",
  comment.char="#")
```

Let's re-add the `girls_rule_maths` column just as in the previous exercise. Then, let's create a subset of `edstats` that doesn't include the country names as well as the boys' and girls' math scores.

```
edstats$girls_rule_maths <-
  factor(as.numeric(
    edstats$LO.PISA.MAT.FE>edstats$LO.PISA.MAT.MA
  ))
edstats_subset <- edstats[!(names(edstats) %in%
  c("CountryName", "LO.PISA.MAT.FE", "LO.PISA.MAT.MA"))]
```

Exercise 8.2 Fit and assess a logistic regression model for `girls_rule_maths` as a function of `LO.PISA.REA.MA+LO.PISA.SCI`.

Solution.

Fitting of the model:

```
(f1 <- glm(girls_rule_maths~LO.PISA.REA.MA+LO.PISA.SCI,
  data=edstats_subset, family=binomial("logit")))
```

```
##
## Call: glm(formula = girls_rule_maths ~ LO.PISA.REA.MA + LO.PISA.SCI,
##           family = binomial("logit"), data = edstats_subset)
##
## Coefficients:
## (Intercept) LO.PISA.REA.MA      LO.PISA.SCI
##          3.0927        -0.0882         0.0755
##
## Degrees of Freedom: 80 Total (i.e. Null); 78 Residual
##   (187 observations deleted due to missingness)
## Null Deviance: 103
## Residual Deviance: 77.9  AIC: 83.9
```

Performance metrics:

```
Y_pred <- as.numeric(predict(f1, edstats_subset, type="response")>0.5)
get_metrics(Y_pred, edstats_subset$girls_rule_maths)
```

	Acc	Prec	Rec	F	TN	FN	FP	TP
##	0.79012	0.75000	0.55556	0.63830	49.00000	12.00000	5.00000	15.00000

Relate the above numbers to those reported for the fitted decision trees.

Note that the fitted model is nicely interpretable: the lower the boys' average result on the Read-

ing Scale or the higher the country's result on the Science Scale, the higher the probability for *girls_rule_maths*:

```
example_X <- data.frame(
  LO.PISA.REA.MA=c(475, 450, 475, 500),
  LO.PISA.SCI= c(525, 525, 550, 500)
)
cbind(example_X,
  `Pr(Y=1)` =predict(f1, example_X, type="response"))

##   LO.PISA.REA.MA LO.PISA.SCI  Pr(Y=1)
## 1          475      525 0.703342
## 2          450      525 0.955526
## 3          475      550 0.939986
## 4          500      500 0.038094
```

■

Exercise 8.3 (*) Fit and assess a logistic regression model for *girls_rule_maths* featuring all *LO.PISA.REA** and *LO.PISA.SCI** as independent variables.

Solution.

Model fitting:

```
(f2 <- glm(girls_rule_maths ~ LO.PISA.REA + LO.PISA.REA.FE + LO.PISA.REA.MA +
            LO.PISA.SCI + LO.PISA.SCI.FE + LO.PISA.SCI.MA,
            data = edstats_subset, family = binomial("logit")))

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##
## Call: glm(formula = girls_rule_maths ~ LO.PISA.REA + LO.PISA.REA.FE +
##           LO.PISA.REA.MA + LO.PISA.SCI + LO.PISA.SCI.FE + LO.PISA.SCI.MA,
##           family = binomial("logit"), data = edstats_subset)
##
## Coefficients:
## (Intercept) LO.PISA.REA LO.PISA.REA.FE LO.PISA.REA.MA
## -2.265       1.268      -0.544      -0.734
## LO.PISA.SCI LO.PISA.SCI.FE LO.PISA.SCI.MA
## 1.269       -0.157      -1.112
##
## Degrees of Freedom: 80 Total (i.e. Null); 74 Residual
## (187 observations deleted due to missingness)
## Null Deviance: 103
## Residual Deviance: 33 AIC: 47
```

The mysterious fitted probabilities numerically 0 or 1 occurred warning denotes convergence problems of the underlying optimisation (fitting) procedure: at least one of the model

coefficients has had a fairly large order of magnitude and hence the fitted probabilities has come very close to 0 or 1. Recall that the probabilities are modelled by means of the logistic sigmoid function applied on the output of a linear combination of the dependent variables. Moreover, cross-entropy features a logarithm, and $\log 0 = -\infty$.

This can be due to the fact that all the variables in the model are very correlated with each other (multicollinearity; an ill-conditioned problem). The obtained solution might be unstable – there might be many local optima and hence, different parameter vectors might be equally good. Moreover, it is likely that a small change in one of the inputs might lead to large change in the estimated model (* normally, we would attack this problem by employing some regularisation techniques).

Of course, the model's performance metrics can still be computed, but then it's better if we treat it as a black box. Or, even better, reduce the number of independent variables and come up with a simpler model that serves its purpose better than this one.

```
Y_pred <- as.numeric(predict(f2, edstats_subset, type="response")>0.5)
get_metrics(Y_pred, edstats_subset$girls_rule_maths)
```

	Acc	Prec	Rec	F	TN	FN	FP	TP
##	0.86420	0.83333	0.74074	0.78431	50.00000	7.00000	4.00000	20.00000



8.3.2 EdStats – Variable Selection in Binary Logistic Regression (*)

Back to our `girls_rule_maths` example, we still have so much to learn!

```
edstats <- read.csv("datasets/edstats_2019_wide.csv",
  comment.char="#")
edstats$girls_rule_maths <-
  factor(as.numeric(
    edstats$LO.PISA.MAT.FE>edstats$LO.PISA.MAT.MA
  ))
edstats_subset <- edstats[!(names(edstats) %in%
  c("CountryName", "LO.PISA.MAT.FE", "LO.PISA.MAT.MA"))]
```

Exercise 8.4 Construct a binary logistic regression model via forward selection of variables.

Solution.

Just as in the linear regression case, we can rely on the `step()` function.

```
model_empty <- girls_rule_maths~1
(model_full <- formula(model.frame(girls_rule_maths~.,
  data=edstats_subset)))

## girls_rule_maths ~ HD.HCI.AMRT + HD.HCI.AMRT.FE + HD.HCI.AMRT.MA +
##      HD.HCI.EYRS + HD.HCI.EYRS.FE + HD.HCI.EYRS.MA + HD.HCI.HLOS +
##      HD.HCI.HLOS.FE + HD.HCI.HLOS.MA + HD.HCI.MORT + HD.HCI.MORT.FE +
```

```

##  HD.HCI.MORT.MA + HD.HCI.OVRL + HD.HCI.OVRL.FE + HD.HCI.OVRL.MA +
##  IT.CMP.PCMP.P2 + IT.NET.USER.P2 + LO.PISA.MAT + LO.PISA.REA +
##  LO.PISA.REA.FE + LO.PISA.REA.MA + LO.PISA.SCI + LO.PISA.SCI.FE +
##  LO.PISA.SCI.MA + NY.GDP.MKTP.CD + NY.GDP.PCAP.CD + NY.GDP.PCAP.PP.CD +
##  NY.GNP.PCAP.CD + NY.GNP.PCAP.PP.CD + SE.COM.DURS + SE.PRM.CMPT.FE.ZS +
##  SE.PRM.CMPT.MA.ZS + SE.PRM.CMPT.ZS + SE.PRM.ENRL.TC.ZS +
##  SE.PRM.ENRR + SE.PRM.ENRR.FE + SE.PRM.ENRR.MA + SE.PRM.NENR +
##  SE.PRM.NENR.FE + SE.PRM.NENR.MA + SE.PRM.PRIV.ZS + SE.SEC.ENRL.TC.ZS +
##  SE.SEC.ENRR + SE.SEC.ENRR.FE + SE.SEC.ENRR.MA + SE.SEC.NENR +
##  SE.SEC.NENR.MA + SE.SEC.PRIV.ZS + SE.TER.ENRR + SE.TER.ENRR.FE +
##  SE.TER.ENRR.MA + SE.TER.PRIV.ZS + SE.XPD.TOTL.GD.ZS + SL.TLF.ADVN.FE.ZS +
##  SL.TLF.ADVN.MA.ZS + SL.TLF.ADVN.ZS + SP.POP.TOTL + SP.POP.TOTL.FE.IN +
##  SP.POP.TOTL.MA.IN + SP.PRM.TOTL.FE.IN + SP.PRM.TOTL.IN +
##  SP.PRM.TOTL.MA.IN + SP.SEC.TOTL.FE.IN + SP.SEC.TOTL.IN +
##  SP.SEC.TOTL.MA.IN + UIS.PTRHC.56 + UIS.SAP.CE + UIS.SAP.CE.F +
##  UIS.SAP.CE.M + UIS.X.PPP.1.FSGOV + UIS.X.PPP.2T3.FSGOV +
##  UIS.X.PPP.5T8.FSGOV + UIS.X.US.1.FSGOV + UIS.X.US.2T3.FSGOV +
##  UIS.X.US.5T8.FSGOV + UIS.XGDP.1.FSGOV + UIS.XGDP.23.FSGOV +
##  UIS.XGDP.56.FSGOV + UIS.XUNIT.GDPCAP.1.FSGOV + UIS.XUNIT.GDPCAP.23.FSGOV +
##  UIS.XUNIT.GDPCAP.5T8.FSGOV + UIS.XUNIT.PPP.1.FSGOV.FFNTR +
##  UIS.XUNIT.PPP.2T3.FSGOV.FFNTR + UIS.XUNIT.PPP.5T8.FSGOV.FFNTR +
##  UIS.XUNIT.US.1.FSGOV.FFNTR + UIS.XUNIT.US.23.FSGOV.FFNTR +
##  UIS.XUNIT.US.5T8.FSGOV.FFNTR

f <- step(glm(model_empty, data=edstats_subset, family=binomial("logit")),
  scope=model_full, direction="forward")

## Start: AIC=105.12
## girls_rule_maths ~ 1

## Error in model.matrix.default(Terms, m, contrasts.arg = object$contrasts): variable 1

```

Melbourne, we have a problem! Our dataset has too many missing values, and those cannot be present in a logistic regression model (it's based on a linear combination of variables, and sums/products involving NAs yield NAs...).

Looking at the manual of ?step, we see that the default NA handling is via na.omit(), and that, when applied on a data frame, results in the removal of all the rows, where there is at least one NA. Sadly, it's too invasive.

We should get rid of the data blanks manually. First, definitely, we should remove all the rows where girls_rule_maths is unknown:

```

edstats_subset <-
  edstats_subset[!is.na(edstats_subset$girls_rule_maths),]

```

We are about to apply the forward selection process, whose purpose is to choose variables for a model. Therefore, instead of removing any more rows, we should remove the... columns with missing data:

```
edstats_subset <-
  edstats_subset[, colSums(sapply(edstats_subset, is.na))==0]
```

(*) Alternatively, we could apply some techniques of missing data imputation; this is beyond the scope of this book. For instance, NAs could be replaced by the averages of their respective columns.

We are ready now to make use of `step()`.

```
model_empty <- girls_rule_maths~1
(model_full <- formula(model.frame(girls_rule_maths~.,
  data=edstats_subset)))

## girls_rule_maths ~ IT.NET.USER.P2 + LO.PISA.MAT + LO.PISA.REA +
##      LO.PISA.REA.FE + LO.PISA.REA.MA + LO.PISA.SCI + LO.PISA.SCI.FE +
##      LO.PISA.SCI.MA + NY.GDP.MKTP.CD + NY.GDP.PCAP.CD + SP.POP.TOTL

f <- step(glm(model_empty, data=edstats_subset, family=binomial("logit")),
  scope=model_full, direction="forward")

## Start: AIC=105.12
## girls_rule_maths ~ 1
##
##                                Df Deviance   AIC
## + LO.PISA.REA.MA  1     90.9  94.9
## + LO.PISA.SCI.MA  1     93.3  97.3
## + NY.GDP.MKTP.CD  1     94.2  98.2
## + LO.PISA.REA     1     95.0  99.0
## + LO.PISA.SCI     1     96.9 100.9
## + LO.PISA.MAT     1     97.2 101.2
## + LO.PISA.REA.FE  1     97.9 101.9
## + LO.PISA.SCI.FE  1     99.4 103.4
## <none>                  103.1 105.1
## + SP.POP.TOTL     1    101.9 105.9
## + NY.GDP.PCAP.CD  1    102.3 106.3
## + IT.NET.USER.P2  1    103.1 107.1
##
## Step: AIC=94.93
## girls_rule_maths ~ LO.PISA.REA.MA
##
##                                Df Deviance   AIC
## + LO.PISA.REA     1     42.8  48.8
## + LO.PISA.REA.FE  1     50.5  56.5
## + LO.PISA.SCI.FE  1     65.4  71.4
## + LO.PISA.SCI     1     77.9  83.9
## + LO.PISA.MAT     1     83.5  89.5
## + NY.GDP.MKTP.CD  1     87.4  93.4
## + IT.NET.USER.P2  1     87.5  93.5
```

```

## <none>          90.9 94.9
## + NY.GDP.PCAP.CD 1   89.2 95.2
## + LO.PISA.SCI.MA 1   89.2 95.2
## + SP.POP.TOTL    1   90.2 96.2
##
## Step: AIC=48.83
## girls_rule_maths ~ LO.PISA.REA.MA + LO.PISA.REA
##
##           Df Deviance AIC
## <none>          42.8 48.8
## + LO.PISA.SCI.FE 1   40.9 48.9
## + SP.POP.TOTL    1   41.2 49.2
## + NY.GDP.PCAP.CD 1   41.3 49.3
## + LO.PISA.SCI     1   42.0 50.0
## + LO.PISA.MAT     1   42.4 50.4
## + IT.NET.USER.P2  1   42.7 50.7
## + LO.PISA.SCI.MA  1   42.7 50.7
## + NY.GDP.MKTP.CD  1   42.7 50.7
## + LO.PISA.REA.FE  1   42.8 50.8
print(f)

##
## Call: glm(formula = girls_rule_maths ~ LO.PISA.REA.MA + LO.PISA.REA,
##           family = binomial("logit"), data = edstats_subset)
##
## Coefficients:
## (Intercept) LO.PISA.REA.MA      LO.PISA.REA
##           -0.176        -0.600         0.577
##
## Degrees of Freedom: 80 Total (i.e. Null);  78 Residual
## Null Deviance:      103
## Residual Deviance: 42.8  AIC: 48.8
Y_pred <- as.numeric(predict(f, edstats_subset, type="response")>0.5)
get_metrics(Y_pred, edstats_subset$girls_rule_maths)

##      Acc      Prec      Rec      F      TN      FN      FP      TP
## 0.88889  0.84615  0.81481  0.83019 50.00000  5.00000  4.00000 22.00000

```

Exercise 8.5 Choose a model via backward elimination.

Solution.

Having a dataset with missing values removed, this is easy now:

```

f <- suppressWarnings( # yeah, yeah, yeah...
  # fitted probabilities numerically 0 or 1 occurred
  step(glm(model_full, data=edstats_subset, family=binomial("logit")),
    scope=model_empty, direction="backward")
)

## Start: AIC=50.83
## girls_rule_maths ~ IT.NET.USER.P2 + LO.PISA.MAT + LO.PISA.REA +
##   LO.PISA.REA.FE + LO.PISA.REA.MA + LO.PISA.SCI + LO.PISA.SCI.FE +
##   LO.PISA.SCI.MA + NY.GDP.MKTP.CD + NY.GDP.PCAP.CD + SP.POP.TOTL
##
##          Df Deviance AIC
## - LO.PISA.MAT  1  26.8 48.8
## - LO.PISA.SCI.MA 1  26.8 48.8
## - NY.GDP.PCAP.CD 1  26.9 48.9
## - LO.PISA.SCI  1  26.9 48.9
## - LO.PISA.SCI.FE 1  27.1 49.1
## - LO.PISA.REA.FE 1  27.4 49.4
## - LO.PISA.REA  1  27.5 49.5
## - LO.PISA.REA.MA 1  27.6 49.6
## <none>            26.8 50.8
## - IT.NET.USER.P2 1  29.3 51.3
## - NY.GDP.MKTP.CD 1  29.9 51.9
## - SP.POP.TOTL    1  31.7 53.7
##
## Step: AIC=48.84
## girls_rule_maths ~ IT.NET.USER.P2 + LO.PISA.REA + LO.PISA.REA.FE +
##   LO.PISA.REA.MA + LO.PISA.SCI + LO.PISA.SCI.FE + LO.PISA.SCI.MA +
##   NY.GDP.MKTP.CD + NY.GDP.PCAP.CD + SP.POP.TOTL
##
##          Df Deviance AIC
## - LO.PISA.SCI.MA 1  26.8 46.8
## - NY.GDP.PCAP.CD 1  26.9 46.9
## - LO.PISA.SCI  1  27.0 47.0
## - LO.PISA.SCI.FE 1  27.1 47.1
## - LO.PISA.REA.FE 1  27.4 47.4
## - LO.PISA.REA  1  27.5 47.5
## - LO.PISA.REA.MA 1  27.6 47.6
## <none>            26.8 48.8
## - IT.NET.USER.P2 1  29.3 49.3
## - NY.GDP.MKTP.CD 1  29.9 49.9
## - SP.POP.TOTL    1  31.7 51.7
##
## Step: AIC=46.84
## girls_rule_maths ~ IT.NET.USER.P2 + LO.PISA.REA + LO.PISA.REA.FE +

```

```

##      LO.PISA.REA.MA + LO.PISA.SCI + LO.PISA.SCI.FE + NY.GDP.MKTP.CD +
##      NY.GDP.PCAP.CD + SP.POP.TOTL
##
##          Df Deviance AIC
## - NY.GDP.PCAP.CD  1    26.9 44.9
## <none>            26.8 46.8
## - IT.NET.USER.P2  1    29.3 47.3
## - NY.GDP.MKTP.CD  1    29.9 47.9
## - LO.PISA.REA.FE  1    31.0 49.0
## - SP.POP.TOTL     1    31.8 49.8
## - LO.PISA.SCI     1    35.6 53.6
## - LO.PISA.SCI.FE  1    36.1 54.1
## - LO.PISA.REA     1    37.5 55.5
## - LO.PISA.REA.MA  1    50.9 68.9
##
## Step: AIC=44.87
## girls_rule_maths ~ IT.NET.USER.P2 + LO.PISA.REA + LO.PISA.REA.FE +
##      LO.PISA.REA.MA + LO.PISA.SCI + LO.PISA.SCI.FE + NY.GDP.MKTP.CD +
##      SP.POP.TOTL
##
##          Df Deviance AIC
## <none>            26.9 44.9
## - NY.GDP.MKTP.CD  1    30.5 46.5
## - IT.NET.USER.P2  1    31.0 47.0
## - LO.PISA.REA.FE  1    31.1 47.1
## - SP.POP.TOTL     1    33.0 49.0
## - LO.PISA.SCI     1    35.9 51.9
## - LO.PISA.SCI.FE  1    36.4 52.4
## - LO.PISA.REA     1    37.5 53.5
## - LO.PISA.REA.MA  1    50.9 66.9

```

The obtained model and its quality metrics:

```

print(f)

##
## Call: glm(formula = girls_rule_maths ~ IT.NET.USER.P2 + LO.PISA.REA +
##      LO.PISA.REA.FE + LO.PISA.REA.MA + LO.PISA.SCI + LO.PISA.SCI.FE +
##      NY.GDP.MKTP.CD + SP.POP.TOTL, family = binomial("logit"),
##      data = edstats_subset)
##
## Coefficients:
## (Intercept) IT.NET.USER.P2      LO.PISA.REA  LO.PISA.REA.FE
## -1.66e+01    1.61e-01     1.85e+00   -8.00e-01
## LO.PISA.REA.MA  LO.PISA.SCI  LO.PISA.SCI.FE  NY.GDP.MKTP.CD
## -1.03e+00    -1.35e+00    1.32e+00   -4.95e-12

```

```
##      SP.POP.TOTL
##      6.20e-08
##
## Degrees of Freedom: 80 Total (i.e. Null);  72 Residual
## Null Deviance:      103
## Residual Deviance: 26.9  AIC: 44.9
Y_pred <- as.numeric(predict(f, edstats_subset, type="response"))>0.5)
get_metrics(Y_pred, edstats_subset$girls_rule_maths)
```

```
##      Acc      Prec      Rec      F      TN      FN      FP      TP
## 0.91358 0.88462 0.85185 0.86792 51.00000 4.00000 3.00000 23.00000
```

Note that we got a better (lower) AIC than in the forward selection case, which means that backward elimination was better this time. On the other hand, we needed to suppress the fitted probabilities numerically 0 or 1 occurred warnings. The returned model is perhaps unstable as well and consists of too many variables.

■

8.3.3 Currency Exchange Rates Growth/Fall

Let's consider a modified version of the Currency Exchange Rates dataset released by the European Central Bank System:

```
currency_exchange <- read.csv("datasets/currency_exchange_diff.csv.gz",
  comment.char="#")
head(currency_exchange[,1:8]) # few first rows and columns

##      Day      AUD Change Dir      Lag1      Lag2      Lag3      Lag4
## 1 1999-01-04 1.9100    NA <NA>    NA      NA      NA      NA
## 2 1999-01-05 1.8944 -0.0156 dec     NA      NA      NA      NA
## 3 1999-01-06 1.8820 -0.0124 dec -0.0156    NA      NA      NA
## 4 1999-01-07 1.8474 -0.0346 dec -0.0124 -0.0156    NA      NA
## 5 1999-01-08 1.8406 -0.0068 dec -0.0346 -0.0124 -0.0156    NA
## 6 1999-01-11 1.8134 -0.0272 dec -0.0068 -0.0346 -0.0124 -0.0156

tail(currency_exchange[,1:8]) # a couple of last rows and a few columns

##      Day      AUD Change Dir      Lag1      Lag2      Lag3      Lag4
## 5528 2020-05-12 1.6625 -0.0084 dec  0.0096 -0.0091 -0.0342  0.0221
## 5529 2020-05-13 1.6687  0.0062 inc -0.0084  0.0096 -0.0091 -0.0342
## 5530 2020-05-14 1.6805  0.0118 inc  0.0062 -0.0084  0.0096 -0.0091
## 5531 2020-05-15 1.6805  0.0000 dec  0.0118  0.0062 -0.0084  0.0096
## 5532 2020-05-18 1.6736 -0.0069 dec  0.0000  0.0118  0.0062 -0.0084
## 5533 2020-05-19 1.6751  0.0015 inc -0.0069  0.0000  0.0118  0.0062
```

We have discussed this dataset in Appendix F. We will investigate how accurately we can predict if the EUR/AUD exchange rate increases (Dir equal to inc) or decreases (Dir equal

to dec) based on the price changes in preceding days (Lag1 – yesterday, Lag2 – two days ago etc.).

Exercise 8.6 Perform a train-test split. Create a data frame `currency_exchange_train` that consists of the first 80% rows in `currency_exchange` and a data frame `currency_exchange_test` that includes the remaining 20%.

Remark. This is a time series prediction task; we are not assigning the observations to either of these two subsets in a random manner. Instead, we use the “past” data for training (models’ construction) and the “future” values for testing (models’ performance evaluation). This is exactly how we often proceed in real life.

Exercise 8.7 Fit and plot a decision tree that models `Dir` as a function of `Lag1, ..., Lag5`. Compute the accuracy, precision, recall and F-measure of this classifier.

Exercise 8.8 Fit 5 binary logistic regression models for `Dir` as a function of the percentage return from: (i) the previous day, (ii) the last three days, (iii) five days, (iv) seven days, (v) all the previous 10 days. Print the estimated coefficients (parameters). Evaluate the performance of these 5 classifiers.

Exercise 8.9 Use the 5- and 15-nearest neighbours algorithm to predict `Dir` based on the percentage returns from the 5 previous days. Evaluate the performance of these 2 classifiers.

Exercise 8.10 Discuss the obtained results. Would you be keen to use any of the constructed classifiers for making real money?

8.4 Outro

8.4.1 Remarks

Other prominent classification algorithms:

- Naive Bayes and other probabilistic approaches,
- Support Vector Machines (SVMs) and other kernel methods,
- (Artificial) (Deep) Neural Networks.

Interestingly, in the next chapter we will note that the logistic regression model is a special case of a *feed-forward single layer neural network*.

We will also generalise the binary logistic regression to the case of a multiclass classification.

The state-of-the art classifiers called *Random Forests* and *XGBoost* (see also: *AdaBoost*) are based on decision trees. They tend to be more accurate but – at the same time – they fail to exhibit the decision trees’ important feature: interpretability.

Trees can also be used for regression tasks, see R package `rpart`.

TODO

Recommended further reading: (James et al. 2017: Chapters 4 and 8)

Other: (Hastie et al. 2017: Chapters 4 and 7 as well as (*) Chapters 9, 10, 13, 15)

Next Chapter...

9

Continuous Optimisation with Iterative Algorithms

TODO In this chapter, we will:

- ...
 - ...
-

9.1 Introduction

9.1.1 Optimisation Problems

Mathematical optimisation (a.k.a. mathematical programming) deals with the study of algorithms to solve problems related to selecting the *best* element amongst the set of available alternatives.

Most frequently “best” is expressed in terms of an *error* or *goodness of fit* measure:

$$f : \mathbb{D} \rightarrow \mathbb{R}$$

called an **objective function**, where \mathbb{D} is the **search space** (problem domain, feasible set).

An **optimisation task** deals with finding an element $\mathbf{x} \in \mathbb{D}$ amongst the set of possible candidate solutions, that minimises or maximises f :

$$\min_{\mathbf{x} \in \mathbb{D}} f(\mathbf{x}) \quad \text{or} \quad \max_{\mathbf{x} \in \mathbb{D}} f(\mathbf{x}),$$

In this chapter we will deal with **unconstrained continuous optimisation**, i.e., we will assume the search space is $\mathbb{D} = \mathbb{R}^p$ for some p – we’ll be optimising over p real-valued parameters.

9.1.2 Types of Minima and Maxima

Note that minimising f is the same as maximising $\bar{f} = -f$.

In other words, $\min_{\mathbf{x} \in \mathbb{D}} f(\mathbf{x})$ and $\max_{\mathbf{x} \in \mathbb{D}} -f(\mathbf{x})$ represent the same optimisation problems (and hence have identical solutions).

Definition. A **minimum** of f is a point \mathbf{x}^* such that $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{D}$. On the other hand, a **maximum** of f is a point \mathbf{x}^* such that $f(\mathbf{x}^*) \geq f(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{D}$.

Assuming that $\mathbb{D} = \mathbb{R}$, Figure @ref(fig:f_global_minimum) shows an example objective function, $f : \mathbb{D} \rightarrow \mathbb{R}$, that has a minimum at $x^* = 1$ with $f(x^*) = -2$.

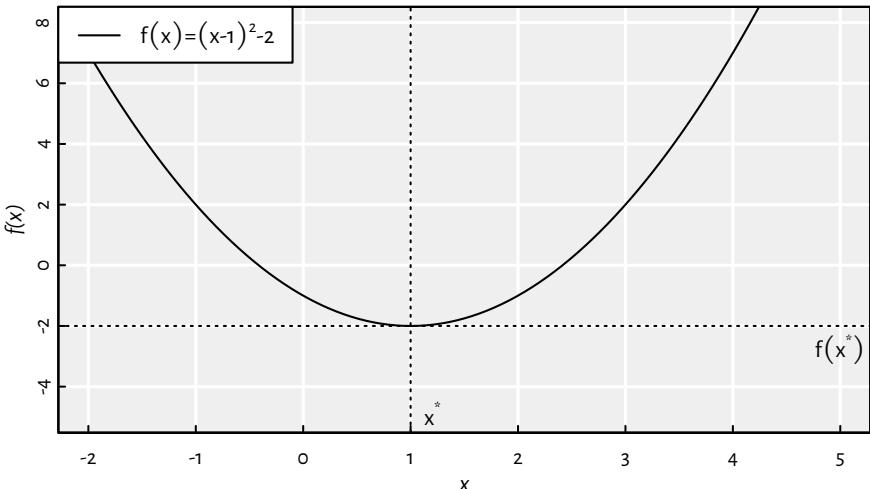


Figure 9.1: (#fig:f_global_minimum) A function with the global minimum at $x^* = 1$

Remark. We can denote these two facts as follows:

- $(\min_{x \in \mathbb{R}} f(x)) = -2$ (value of f at the minimum is -2),
- $(\arg \min_{x \in \mathbb{R}} f(x)) = 1$ (location of the minimum, i.e., *argument minimum*, is 1).

By definition, a minimum/maximum *might not necessarily be unique*. This depends on a problem.

Assuming that $\mathbb{D} = \mathbb{R}$, Figure @ref(fig:f_global_minimum_not_unique) gives an example objective function, $f : \mathbb{D} \rightarrow \mathbb{R}$, that has multiple minima; every $x^* \in [1 - \sqrt{2}, 1 + \sqrt{2}]$ yields $f(x^*) = 0$.

Remark. If this was the case of some machine learning problem, it'd mean that we could have many equally well-performing models, and hence many equivalent explanations of the same phenomenon.

Moreover, it may happen that a function has *multiple local minima*, compare Figure @ref(fig:f_global_local_minima).

Definition. We say that f has a **local minimum** at $\mathbf{x}^+ \in \mathbb{D}$, if for some neighbourhood $B(\mathbf{x}^+)$ of \mathbf{x}^+ it holds $f(\mathbf{x}^+) \leq f(\mathbf{x})$ for each $\mathbf{x} \in B(\mathbf{x}^+)$.

If $\mathbb{D} = \mathbb{R}$, by neighbourhood $B(x)$ of x we mean an open interval centred at x of width $2r$ for some small $r > 0$, i.e., $(x - r, x + r)$

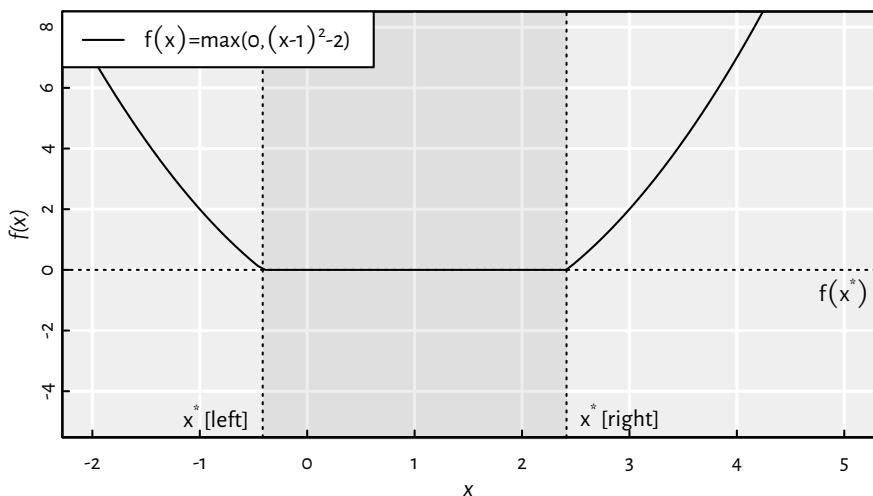


Figure 9.2: (#fig:f_global_minimum_not_unique) A function that has multiple minima

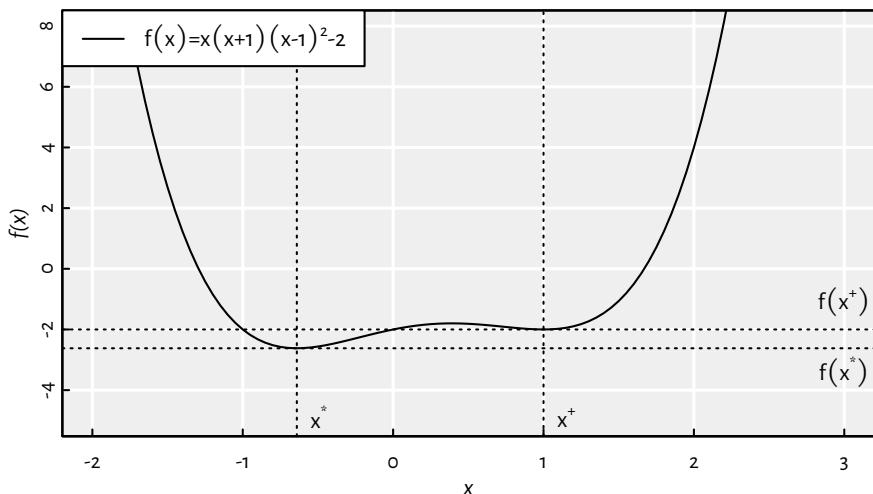


Figure 9.3: (#fig:f_global_local_minima) A function with two local minima

Definition. (*) If $\mathbb{D} = \mathbb{R}^p$ (for any $p \geq 1$), by neighbourhood $B(\mathbf{x})$ of \mathbf{x} we mean an *open ball* centred at \mathbf{x}^+ of some small radius $r > 0$, i.e., $\{\mathbf{y} : \|\mathbf{x} - \mathbf{y}\| < r\}$ (read: the set of all the points with Euclidean distances to \mathbf{x} less than r).

To avoid ambiguity, the “true” minimum (a point \mathbf{x}^* such that $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{D}$) is sometimes also referred to as a **global** minimum.

Remark. Of course, the global minimum is also a function’s local minimum.

The existence of local minima is problematic as most of the optimisation methods might get stuck there and fail to return the global one.

Moreover, we cannot often be sure if the result returned by an algorithm is indeed a global minimum. Maybe there exists a better solution that hasn’t been considered yet? Or maybe the function is very noisy (see Figure @ref(fig:smooth_vs_nonsmooth))?

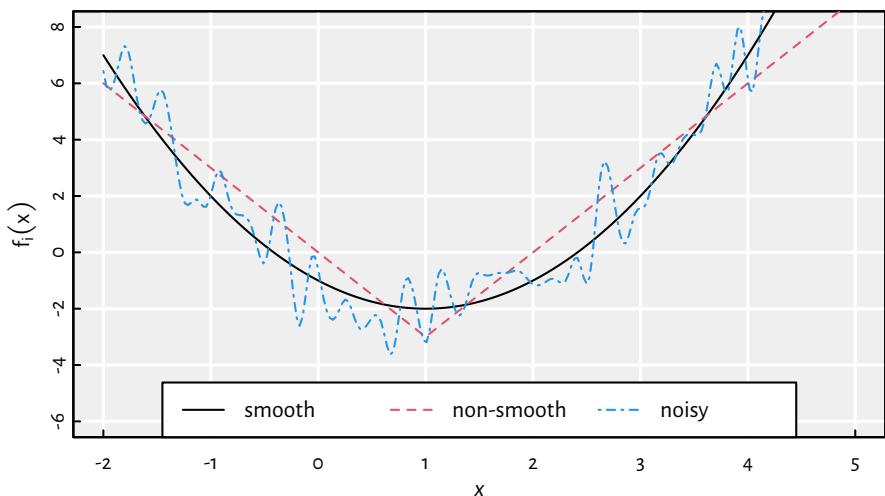


Figure 9.4: (#fig:smooth_vs_nonsmooth) Smooth vs. non-smooth vs. noisy objective functions

9.1.3 Example Objective over a 2D Domain

Of course, our objective function does not necessarily have to be defined over a one-dimensional domain.

For example, consider the following function:

$$g(x_1, x_2) = \log((x_1^2 + x_2 - 5)^2 + (x_1 + x_2^2 - 3)^2 + x_1^2 - 1.60644 \dots)$$

```

g <- function(x1, x2)
  log((x1^2+x2-5)^2+(x1+x2^2-3)^2+x1^2-1.60644366086443841)
x1 <- seq(-5, 5, length.out=100)
x2 <- seq(-5, 5, length.out=100)
# outer() expands two vectors to form a 2D grid
# and applies a given function on each point
y <- outer(x1, x2, g)

```

There are four local minima:

x1	x2	f(x1,x2)
2.2780	-0.61343	1.3564
-2.6123	-2.34546	1.7051
1.7988	1.19879	0.6955
-1.5423	2.15641	0.0000

The global minimum is at $\mathbf{x}^* = (x_1^*, x_2^*)$ as below:

```
g(-1.542255693195422641930153, 2.156405289793087261832605)
```

```
## [1] 0
```

Let's explore various ways of depicting f first. A contour plot and a heat map are given in Figure @ref(fig:contour_g).

```

par(mfrow=c(1,2)) # 2 in 1
# lefthand plot:
contour(x1, x2, y, nlevels=25)
points(-1.54226, 2.15641, col=2, pch=3)
# righthand plot:
image(x1, x2, y)
contour(x1, x2, y, add=TRUE)

```

Two perspective plots (views from different angles) are given in Figure @ref(fig:perspective_g).

```

par(mfrow=c(1,2)) # 2 in 1
persp(x1, x2, y, phi=30, theta=-5, shade=2, border=NA)
persp(x1, x2, y, phi=30, theta=75, shade=2, border=NA)

```

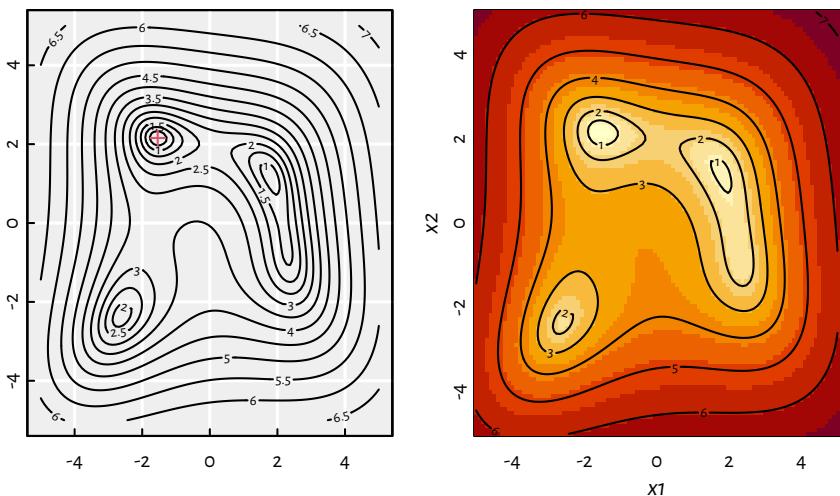


Figure 9.5: (#fig:contour_g) A contour plot and a heat map of $g(x_1, x_2)$

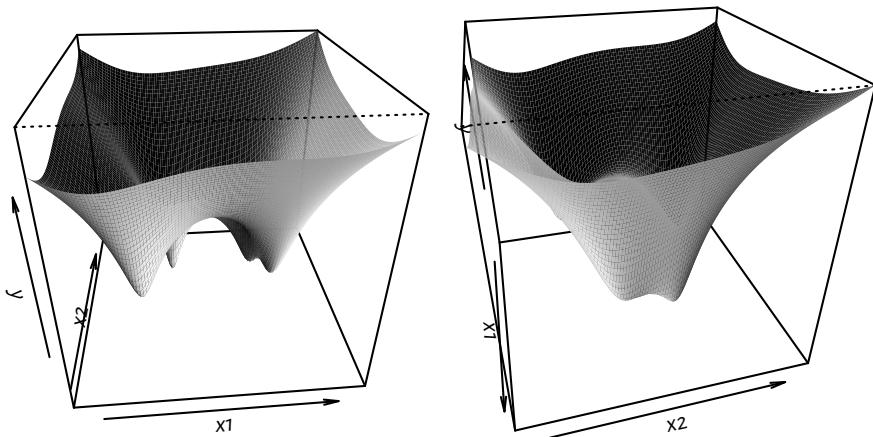


Figure 9.6: (#fig:perspective_g) Perspective plots of $g(x_1, x_2)$

Remark. As usual, depicting functions that are defined over high-dimensional (3D and higher) domains is... difficult. Usually 1D or 2D projections can give us some neat intuitions though.

9.1.4 Example Optimisation Problems in Machine Learning

In **multiple linear regression** we were minimising the sum of squared residuals

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^n (\beta_0 + \beta_1 x_{i,1} + \dots + \beta_p x_{i,p} - y_i)^2.$$

In **binary logistic regression** we were minimising the cross-entropy:

$$\min_{\beta \in \mathbb{R}^p} -\frac{1}{n} \sum_{i=1}^n \left(\begin{array}{l} y_i \log \left(\frac{1}{1+e^{-(\beta_0+\beta_1 x_{i,1}+\dots+\beta_p x_{i,p})}} \right) \\ +(1-y_i) \log \left(\frac{e^{-(\beta_0+\beta_1 x_{i,1}+\dots+\beta_p x_{i,p})}}{1+e^{-(\beta_0+\beta_1 x_{i,1}+\dots+\beta_p x_{i,p})}} \right) \end{array} \right).$$

9.2 Iterative Methods

9.2.1 Introduction

Many optimisation algorithms are built around the following scheme:

Starting from a random point, perform a walk, in each step deciding where to go based on the idea of where the location of the minimum might be.

Example. Imagine we're to cycle from Deakin University's Burwood Campus to the CBD not knowing the route and with GPS disabled – we'll have to ask many people along the way, but we'll eventually (because most people are good) get to some CBD (say, in Perth).

More formally, we are interested in iterative algorithms that operate in a greedy-like manner:

1. $\mathbf{x}^{(0)}$ – initial guess (e.g., generated at random)
2. for $i = 1, \dots, M$:
 - a. $\mathbf{x}^{(i)} = \mathbf{x}^{(i-1)} + [\text{guessed direction}]$
 - b. if $|f(\mathbf{x}^{(i)}) - f(\mathbf{x}^{(i-1)})| < \varepsilon$ break
3. return $\mathbf{x}^{(i)}$ as result

Note that there are two stopping criteria, based on:

- M = maximum number of iterations,
- ε = tolerance, e.g. 10^{-8} .

9.2.2 Example in R

R has a built-in function, `optim()`, that provides an implementation of (amongst others) the **BFGS method** (proposed by Broyden, Fletcher, Goldfarb and Shanno in 1970).

Remark. (*) BFGS uses the assumption that the objective function is smooth – the [guessed direction] is determined by computing the (partial) derivatives (or their finite-difference approximations). However, they might work well even if this is not the case. We'll be able to derive similar algorithms (called quasi-Newton ones) ourselves once we learn about Taylor series approximation by reading a book/taking a course on calculus.

Here, we shall use the BFGS as a *black-box* continuous optimisation method, i.e., without going into how it has been defined (in terms of our assumed math skills, it might be too early for this). Despite that, will still be able to point out a few interesting patterns.

```
optim(par, fn, method="BFGS")
```

where:

- `par` – an initial guess (a numeric vector of length p)
- `fn` – an objective function to minimise (takes a vector of length p on input, returns a single number)

Let us minimise the `g` function defined above (the one with the 2D domain):

```
# g needs to be rewritten to accept a 2-ary vector
g_vectorised <- function(x12) g(x12[1], x12[2])
# random starting point with coordinates in [-5, 5]
(x12_init <- runif(2, -5, 5))
```

```
## [1] -2.1242 2.8831
(res <- optim(x12_init, g_vectorised, method="BFGS"))
```

```
## $par
## [1] -1.5423 2.1564
##
## $value
## [1] 1.4131e-12
##
## $counts
## function gradient
##      101      21
##
## $convergence
## [1] 0
##
## $message
## NULL
```

Note that:

- `par` gives the location of the local minimum found,
- `value` gives the value of g at `par`,
- convergence of \circ is a successful one (we were able to satisfy the $|f(\mathbf{x}^{(i)}) - f(\mathbf{x}^{(i-1)})| < \varepsilon$ condition).

We can even depict the points that the algorithm is “visiting”, see Figure 9.7.

Remark. (*) Technically, the algorithm needs to evaluate a few more points in order to make the decision on where to go next (BFGS approximates the gradient and the Hessian matrix).

```
g_vectorised_plot <- function(x12) {
  points(x12[1], x12[2], col=2, pch=3) # draw
  g(x12[1], x12[2]) # return value
}
contour(x1, x2, y, nlevels=25)
res <- optim(x12_init, g_vectorised_plot, method="BFGS")
```

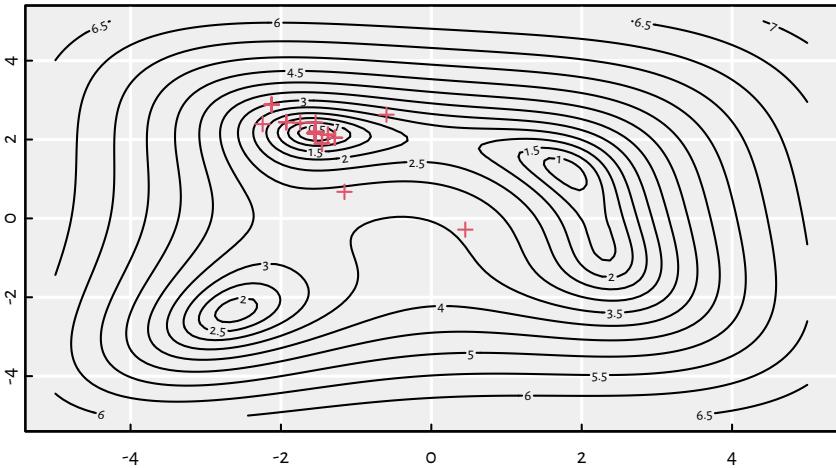


Figure 9.7: Each plotting symbol marks a point where the objective function was evaluated by the BFGS method

9.2.3 Convergence to Local Optima

We were lucky, because the local minimum that the algorithm has found coincides with the global minimum.

Let's see where does the BFGS algorithm converge if seek the minimum of the above

g starting from many randomly chosen points uniformly distributed over the square $[-5, 5] \times [-5, 5]$:

```
res_value <- replicate(1000, {
  # this will be iterated 100 times
  x12_init <- runif(2, -5, 5)
  res <- optim(x12_init, g_vectorised, method="BFGS")
  res$value # return value from each iteration
})
table(round(res_value,3))

##
##      0 0.695 1.356 1.705
##  273   352   156   219
```

Unfortunately, we find the global minimum only in $\sim 25\%$ cases, compare Figure @ref(fig:bfgs_multi_hist).

```
hist(res_value, col="white", breaks=100, main=NA); box()
```

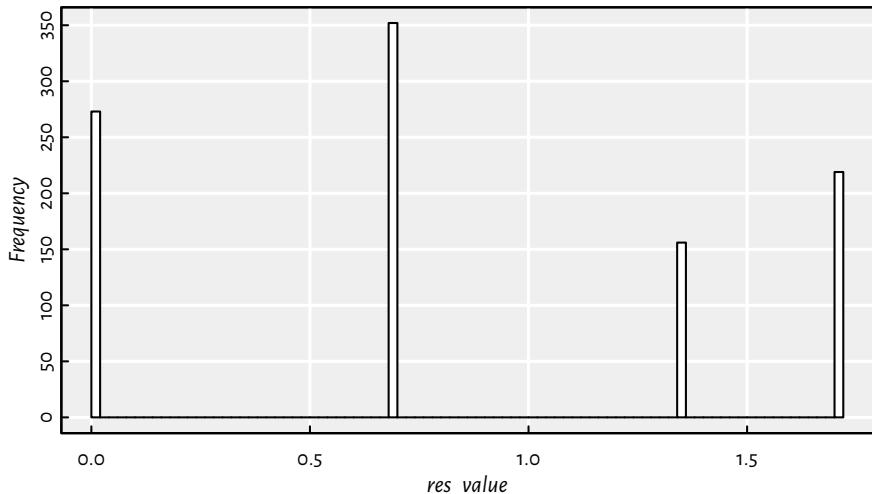


Figure 9.8:(#fig:bfgs_multi_hist) A histogram of the objective function's value at the local minimum found when using a random initial guess

Figure @ref(fig:bfgs_multi_where) depicts all the random starting points and where do we converge from them. Note that by starting in the neighbourhood of $(0, -4)$ we can actually end up in any of the 4 local minima. Moreover, even starting close to the global minimum (ca. $(-2, 2)$) does not guarantee that we will end up there.

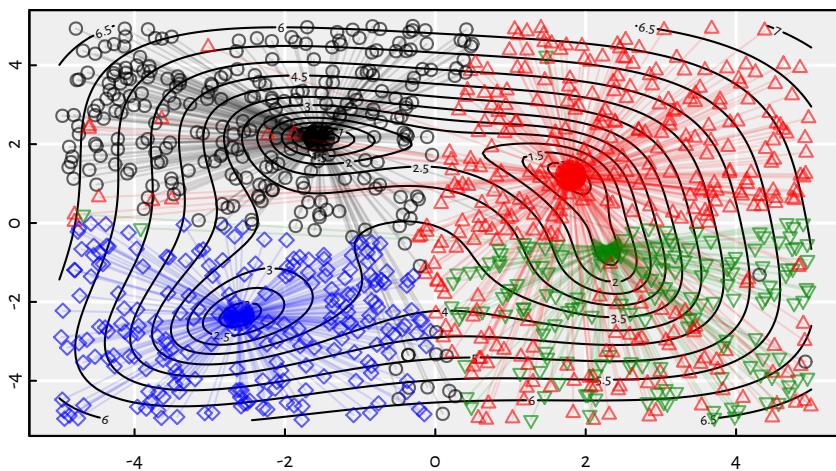


Figure 9.9: (#fig:bfgs_multi_where) Each line segment connects a starting point to the point of where the BFGS algorithm converged

9.2.4 Random Restarts

A kind of “remedy” for the above limitation could be provided by *repeated local search*: in order to robustify an optimisation procedure it is often advised to consider multiple random initial points and pick the best solution amongst the identified local optima.

```
# N           - number of restarts
# par_generator - a function generating initial guesses
# ...          - further arguments to optim()
optim_with_restarts <- function(par_generator, ..., N=10) {
  res_best <- list(value=Inf) # cannot be worse than this
  for (i in 1:N) {
    res <- optim(par_generator(), ...)
    if (res$value < res_best$value)
      res_best <- res # a better candidate found
  }
  res_best
}

optim_with_restarts(function() runif(2, -5, 5),
  g_vectorised, method="BFGS", N=10)

## $par
## [1] -1.5423  2.1564
##
## $value
```

```
## [1] 3.9702e-13
##
## $counts
## function gradient
##      48      17
##
## $convergence
## [1] 0
##
## $message
## NULL
```

Exercise 9.1 Food for thought: Can we really really guarantee that the global minimum will be found within N tries?

Solution.

Absolutely not.

9.3 Gradient Descent

9.3.1 Function Gradient (*)

How to choose the [guessed direction] in our iterative optimisation algorithm?

If we are minimising a smooth function, the simplest possible choice is to use the information included in the objective's **gradient**, which provides us with the information about the direction where the function decreases the fastest.

Definition. (*) Gradient of $f : \mathbb{R}^p \rightarrow \mathbb{R}$, denoted $\nabla f : \mathbb{R}^p \rightarrow \mathbb{R}^p$, is the vector of all its partial derivatives, (∇ – nabla symbol = differential operator)

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(\mathbf{x}) \\ \vdots \\ \frac{\partial f}{\partial x_p}(\mathbf{x}) \end{bmatrix}$$

If we have a function $f(x_1, \dots, x_p)$, the partial derivative w.r.t. the i -th variable, denoted $\frac{\partial f}{\partial x_i}$ is like an ordinary derivative w.r.t. x_i where $x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_p$ are assumed constant.

Remark. Function differentiation is an important concept – see how it's referred to in, e.g., the `keras` package manual at <https://keras.rstudio.com/reference/fit.html>.

Recall our g function defined above:

$$g(x_1, x_2) = \log((x_1^2 + x_2 - 5)^2 + (x_1 + x_2^2 - 3)^2 + x_1^2 - 1.60644 \dots)$$

It can be shown (*) that:

$$\begin{aligned}\frac{\partial g}{\partial x_1}(x_1, x_2) &= \frac{4x_1(x_1^2 + x_2 - 5) + 2(x_1 + x_2^2 - 3) + 2x_1}{(x_1^2 + x_2 - 5)^2 + (x_1 + x_2^2 - 3)^2 + x_1^2 - 1.60644 \dots} \\ \frac{\partial g}{\partial x_2}(x_1, x_2) &= \frac{2(x_1^2 + x_2 - 5) + 4x_2(x_1 + x_2^2 - 3)}{(x_1^2 + x_2 - 5)^2 + (x_1 + x_2^2 - 3)^2 + x_1^2 - 1.60644 \dots}\end{aligned}$$

```
grad_g_vectorised <- function(x) {
  c(
    4*x[1]*(x[1]^2+x[2]-5)+2*(x[1]+x[2]^2-3)+2*x[1],
    2*(x[1]^2+x[2]-5)+4*x[2]*(x[1]+x[2]^2-3)
  )/(
    (x[1]^2+x[2]-5)^2+(x[1]+x[2]^2-3)^2+x[1]^2-1.60644366086443841
  )
}
```

9.3.2 Three Facts on the Gradient

For now, we should emphasise three important facts:

Fact 1. If we are incapable of deriving the gradient analytically, we can rely on its finite differences approximation. Each partial derivative can be estimated by means of:

$$\frac{\partial f}{\partial x_i}(x_1, \dots, x_p) \approx \frac{f(x_1, \dots, x_i + \delta, \dots, x_p) - f(x_1, \dots, x_i, \dots, x_p)}{\delta}$$

for some small $\delta > 0$, say, $\delta = 10^{-6}$.

Remark. (*) Actually, a function's partial derivative, by definition, is the limit of the above as $\delta \rightarrow 0$.

Example implementation:

```
# gradient of f at x=c(x[1], ..., x[p])
grad_approx <- function(f, x, delta=1e-6) {
  p <- length(x)
  gf <- numeric(p) # vector of length p
  for (i in 1:p) {
    xi <- x
    xi[i] <- xi[i]+delta
```

```

    gf[i] <- f(xi)
}
(gf-f(x))/delta
}

```

Remark. (*) Interestingly, some modern vector/matrix algebra frameworks like TensorFlow (upon which `keras` is built) or PyTorch, feature methods to “derive” the gradient algorithmically (autodiff; automatic differentiation).

Sanity check:

```

grad_approx(g_vectorised, c(-2, 2))

## [1] -3.1865 -1.3656
grad_g_vectorised(c(-2, 2))

## [1] -3.1865 -1.3656
grad_approx(g_vectorised, c(-1.542255693, 2.15640528979))

## [1] 1.0588e-05 1.9817e-05
grad_g_vectorised(c(-1.542255693, 2.15640528979))

## [1] 4.1292e-09 3.5771e-10

```

By the way, there is also the `grad()` function in package `numDeriv` that might be a little more accurate (uses a different approximation formula).

Fact 2. The gradient of f at \mathbf{x} , $\nabla f(\mathbf{x})$, is a vector that points in the direction of the steepest slope. On the other hand, minus gradient, $-\nabla f(\mathbf{x})$, is the direction where the function decreases the fastest.

Remark. (*) This can be shown by considering a function’s first-order Taylor series approximation.

Each gradient is a vector, therefore it can be depicted as an arrow. Figure 9.10 illustrates a few scaled gradients of the g function at different points – each arrow connects a point \mathbf{x} to $\mathbf{x} \pm 0.1\nabla f(\mathbf{x})$.

Note that the blue arrows point more or less in the direction of the local minimum. Therefore, in our iterative algorithm, we may try taking the direction of the minus gradient! How far should we go in that direction? Well, a bit. We will refer to the desired step size as the **learning rate**, η .

This will be called the **gradient descent** method (GD; Cauchy, 1847).

Fact 3. If a function f has a local minimum at \mathbf{x}^* , then its gradient vanishes there, i.e., $\nabla f(\mathbf{x}^*) = [0, \dots, 0]$.

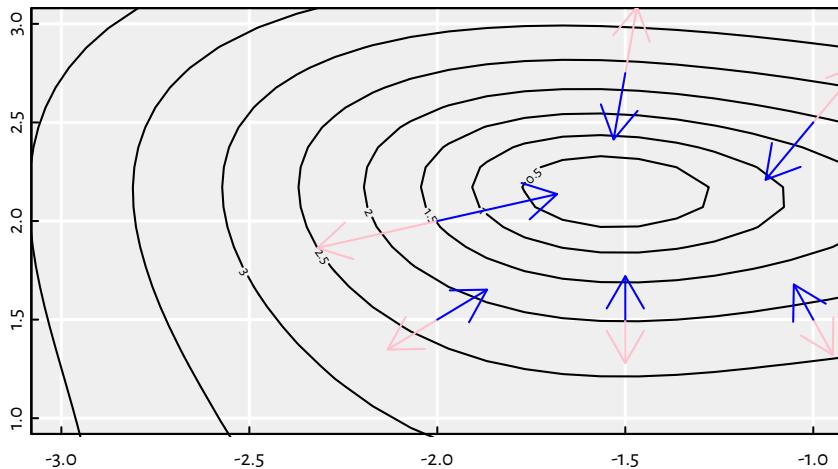


Figure 9.10: Scaled gradients (pink arrows) and minus gradients (blue arrows) of $g(x_1, x_2)$ at different points

Note that the above condition is a necessary, not sufficient one. For example, the gradient also vanishes at a maximum or at a saddle point. In fact, we have what follows.

Theorem. (****) More generally, a twice-differentiable function has a local minimum at \mathbf{x}^* if and only if its gradient vanishes there and $\nabla^2 f(\mathbf{x}^*)$ (Hessian matrix = matrix of all second-order derivatives) is positive-definite.

9.3.3 Gradient Descent Algorithm (GD)

Taking the above into account, we arrive at the gradient descent algorithm:

1. $\mathbf{x}^{(0)}$ – initial guess (e.g., generated at random)
2. for $i = 1, \dots, M$:
 - a. $\mathbf{x}^{(i)} = \mathbf{x}^{(i-1)} - \eta \nabla f(\mathbf{x}^{(i-1)})$
 - b. if $|f(\mathbf{x}^{(i)}) - f(\mathbf{x}^{(i-1)})| < \varepsilon$ break
3. return $\mathbf{x}^{(i)}$ as result

where $\eta > 0$ is a step size frequently referred to as the *learning rate*, because that's much more cool. We usually set η of small order of magnitude, say 0.01 or 0.1.

An implementation of the gradient descent algorithm is straightforward. In essence, it's the `par <- par - eta*grad_g_vectorised(par)` expression run in a loop, until convergence.

```
# par - initial guess
# fn - a function to be minimised
# gr - a function to return the gradient of fn
# eta - learning rate
# maxit - maximum number of iterations
# tol - convergence tolerance
optim_gd <- function(par, fn, gr, eta=0.01,
                      maxit=1000, tol=1e-8) {
  f_last <- fn(par)
  for (i in 1:maxit) {
    par <- par - eta*grad_g_vectorised(par) # update step
    f_cur <- fn(par)
    if (abs(f_cur-f_last) < tol) break
    f_last <- f_cur
  }
  list( # see ?optim, section `Value`
    par=par,
    value=g_vectorised(par),
    counts=i,
    convergence=as.integer(i==maxit)
  )
}
```

Tests of the g function. First, let's try $\eta = 0.01$. Figure 9.11 zooms in the contour plot so that we can see the actual path the algorithm has taken.

```
eta <- 0.01
res <- optim_gd(c(-3, 1), g_vectorised, grad_g_vectorised, eta=eta)
str(res)
```

```
## List of 4
## $ par      : num [1:2] -1.54 2.16
## $ value    : num 1.33e-08
## $ counts   : int 135
## $ convergence: int 0
```

Now let's try $\eta = 0.05$.

```
eta <- 0.05
res <- optim_gd(c(-3, 1), g_vectorised, grad_g_vectorised, eta=eta)
str(res)
```

```
## List of 4
## $ par      : num [1:2] -1.54 2.15
## $ value    : num 0.000203
## $ counts   : int 417
## $ convergence: int 0
```

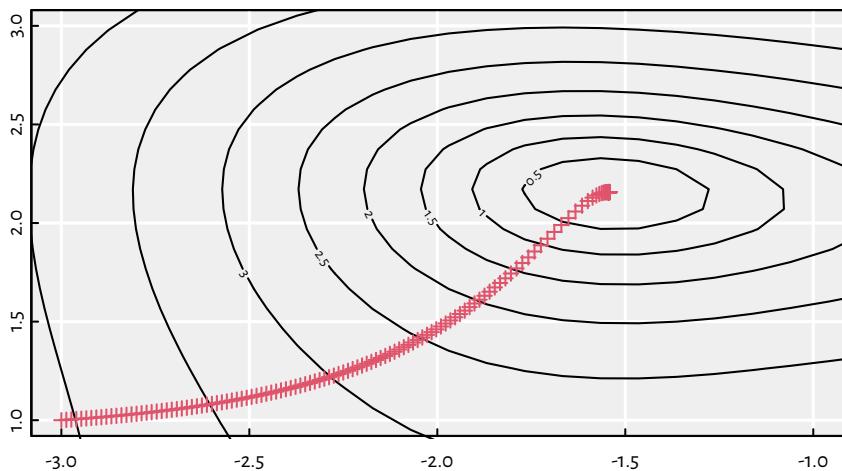


Figure 9.11: Path taken by the gradient descent algorithm with $\eta = 0.01$

With an increased step size, the algorithm needed many more iterations (3 times as many), see Figure 9.12 for the path taken.

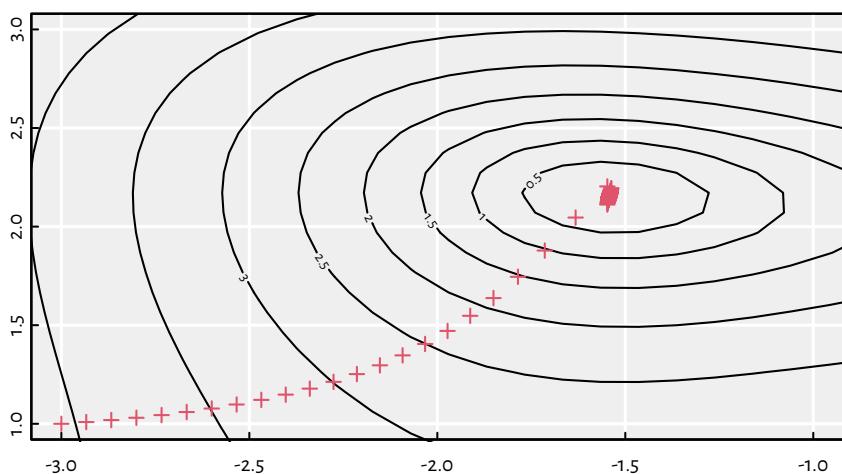


Figure 9.12: Path taken by the gradient descent algorithm with $\eta = 0.05$

And now for something completely different: $\eta = 0.1$, see Figure 9.13.

```

eta <- 0.1
res <- optim_gd(c(-3, 1), g_vectorised, grad_g_vectorised, eta=eta)
str(res)

## List of 4
## $ par      : num [1:2] -1.52 2.33
## $ value    : num 0.507
## $ counts   : int 1000
## $ convergence: int 1

```

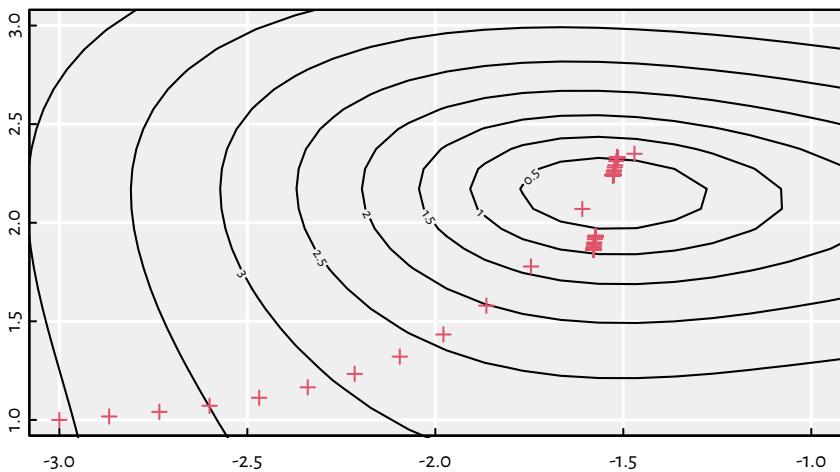


Figure 9.13: Path taken by the gradient descent algorithm with $\eta = 0.1$

The algorithm failed to converge.

If the learning rate η is too small, the convergence might be too slow or we might get stuck at a plateau. On the other hand, if η is too large, we might be overshooting and end up bouncing around the minimum.

This is why many optimisation libraries (including keras/TensorFlow) implement some of the following ideas:

- *learning rate decay* – start with large η , decreasing it in every iteration, say, by some percent;
- *line search* – determine optimal η in every step by solving a 1-dimensional optimisation problem w.r.t. $\eta \in [0, \eta_{\max}]$;
- *momentum* – the update step is based on a combination of the gradient direction and the previous change of the parameters, Δx ; can be used to accelerate search in the relevant direction and minimise oscillations.

Exercise 9.2 Try implementing at least the first of the above heuristics yourself. You can set `eta <- eta*0.95` in every iteration of the gradient descent procedure.

9.3.4 Example: MNIST (*)

In the previous chapter we've studied the MNIST dataset. Let us go back to the task of fitting a multiclass logistic regression model.

```
library("keras")
mnist <- dataset_mnist()

# get train/test images in greyscale
X_train <- mnist$train$x/255 # to [0,1]
X_test  <- mnist$test$x/255 # to [0,1]

# get the corresponding labels in {0,1,...,9}:
Y_train <- mnist$train$y
Y_test  <- mnist$test$y
```

The labels need to be one-hot encoded:

```
one_hot_encode <- function(Y) {
  stopifnot(is.numeric(Y))
  c1 <- min(Y) # first class label
  cK <- max(Y) # last class label
  K <- cK-c1+1 # number of classes
  Y2 <- matrix(0, nrow=length(Y), ncol=K)
  Y2[cbind(1:length(Y), Y-c1+1)] <- 1
  Y2
}

Y_train2 <- one_hot_encode(Y_train)
Y_test2 <- one_hot_encode(Y_test)
```

Our task is to find the parameters \mathbf{B} that minimise cross entropy $E(\mathbf{B})$ over the training set:

$$\min_{\mathbf{B} \in \mathbb{R}^{785 \times 10}} -\frac{1}{n^{\text{train}}} \sum_{i=1}^{n^{\text{train}}} \log \Pr(Y = y_i^{\text{train}} | \mathbf{x}_i^{\text{train}}, \mathbf{B}).$$

In the previous chapter, we've relied on the methods implemented in the `keras` package. Let's do that all by ourselves now.

In order to come up with a working version of the gradient descent procedure for classifying of MNIST digits, we will need to derive and implement `grad_cross_entropy()`. We do that below using matrix notation.

Remark. In the first reading, you can jump to the *Safe landing zone* below with no much loss in what we're trying to convey here (you will then treat `grad_cross_entropy()` as a black-box function). Nevertheless, keep in mind that this is the kind of maths you will need to master anyway sooner than later – this is inevitable. Perhaps you should go back to, e.g., the appendix on Matrix Computations with R or the chapter on Linear Regression? Learning maths is not a linear, step-by-step process. Everyone is different and will have a different path to success. The material needs to be frequently revisited, it will “click” someday, don’t you worry; good stuff isn’t built in a day or seven.

Recall that the output of the logistic regression model (1-layer neural network with softmax) can be written in the matrix form as:

$$\hat{\mathbf{Y}} = \text{softmax}(\dot{\mathbf{X}} \mathbf{B}),$$

where $\dot{\mathbf{X}} \in \mathbb{R}^{n \times 785}$ is a matrix representing n images of size 28×28 , augmented with a column of 1s, and $\mathbf{B} \in \mathbb{R}^{785 \times 10}$ is the coefficients matrix and softmax is applied on each matrix row separately.

Of course, by the definition of matrix multiplication, $\hat{\mathbf{Y}}$ will be a matrix of size $n \times 10$, where $\hat{y}_{i,k}$ represents the predicted probability that the i -th image depicts the k -th digit.

```
# convert to matrices of size n*784
# and add a column of 1s
X_train1 <- cbind(1.0, matrix(X_train, ncol=28*28))
X_test1 <- cbind(1.0, matrix(X_test, ncol=28*28))
```

The `nn_predict()` function implements the above formula for $\hat{\mathbf{Y}}$:

```
softmax <- function(T) {
  T <- exp(T)
  T/rowSums(T)
}

nn_predict <- function(B, X) {
  softmax(X %*% B)
}
```

Let’s define the functions to compute the cross-entropy (which we shall minimise) and accuracy (which we shall report to a user):

```
cross_entropy <- function(Y_true, Y_pred) {
  -sum(Y_true*log(Y_pred))/nrow(Y_true)
}

accuracy <- function(Y_true, Y_pred) {
  # both arguments are one-hot encoded
  Y_true_decoded <- apply(Y_true, 1, which.max)
  Y_pred_decoded <- apply(Y_pred, 1, which.max)
  # proportion of equal corresponding pairs:
```

```

    mean(Y_true_decoded == Y_pred_decoded)
}

```

It may be shown (***) that the gradient of cross-entropy (with respect to the parameter matrix \mathbf{B}) can be expressed in the matrix form as:

$$\frac{1}{n} \dot{\mathbf{X}}^T (\hat{\mathbf{Y}} - \mathbf{Y})$$

```

grad_cross_entropy <- function(X, Y_true, Y_pred) {
  t(X) %*% (Y_pred - Y_true) / nrow(Y_true)
}

```

Of course, we could always substitute the gradient with the finite difference approximation. Yet, this would be much slower).

The more mathematically inclined reader will surely notice that by expanding the formulas given in the previous chapter, we can write cross-entropy in the non-matrix form (n – number of samples, K – number of classes, $p + 1$ – number of model parameters; in our case $K = 10$ and $p = 784$) as:

$$\begin{aligned}
E(\mathbf{B}) &= -\frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K y_{i,k} \log \left(\frac{\exp \left(\sum_{j=0}^p \dot{x}_{i,j} \beta_{j,k} \right)}{\sum_{c=1}^K \exp \left(\sum_{j=0}^p \dot{x}_{i,j} \beta_{j,c} \right)} \right) \\
&= \frac{1}{n} \sum_{i=1}^n \left(\log \left(\sum_{k=1}^K \exp \left(\sum_{j=0}^p \dot{x}_{i,j} \beta_{j,k} \right) \right) - \sum_{k=1}^K y_{i,k} \sum_{j=0}^p \dot{x}_{i,j} \beta_{j,k} \right).
\end{aligned}$$

Partial derivatives of cross-entropy w.r.t. $\beta_{a,b}$ in non-matrix form can be derived (*** so as to get:

$$\begin{aligned}
\frac{\partial E}{\partial \beta_{a,b}}(\mathbf{B}) &= \frac{1}{n} \sum_{i=1}^n \dot{x}_{i,a} \left(\frac{\exp \left(\sum_{j=0}^p \dot{x}_{i,j} \beta_{j,b} \right)}{\sum_{k=1}^K \exp \left(\sum_{j=0}^p \dot{x}_{i,j} \beta_{j,k} \right)} - y_{i,b} \right) \\
&= \frac{1}{n} \sum_{i=1}^n \dot{x}_{i,a} (\hat{y}_{i,b} - y_{i,b}).
\end{aligned}$$

Safe landing zone. In case you're lost with the above, continue from here. However, in the near future, harden up and revisit the skipped material to get the most out of our discussion.

We now have all the building blocks to implement the gradient descent method. The algorithm below follows exactly the same scheme as the one in the g function example. This

time, however, we play with a parameter matrix **B** (not a parameter vector $[x_1, x_2]$) and we compute the gradient of cross-entropy (by means of `grad_cross_entropy()`), not the gradient of g .

Note that a call to `system.time(expr)` measures the time (in seconds) spent evaluating an expression `expr`.

```
# random matrix of size 785x10 - initial guess
B <- matrix(rnorm(ncol(X_train1)*ncol(Y_train2)),
             nrow=ncol(X_train1))
eta <- 0.1    # learning rate
maxit <- 100 # number of GD iterations
system.time({ # measure time spent
  # for simplicity, we stop only when we reach maxit
  for (i in 1:maxit) {
    B <- B - eta*grad_cross_entropy(
      X_train1, Y_train2, nn_predict(B, X_train1))
  }
}) # `user` - processing time in seconds:

##    user  system elapsed
## 102.932 40.194 36.602
```

Unfortunately, the method's convergence is really slow (we are optimising over 7850 parameters...) and the results after 100 iterations are disappointing:

```
accuracy(Y_train2, nn_predict(B, X_train1))
```

```
## [1] 0.46462
accuracy(Y_test2, nn_predict(B, X_test1))

## [1] 0.4735
```

Recall that in the previous chapter we obtained much better classification accuracy by using the `keras` package. What are we doing wrong then? Maybe `keras` implements some Super-Fancy Hyper Optimisation Framework (TM) (R) that we could get access to for only \$19.99 per month?

9.3.5 Stochastic Gradient Descent (SGD) (*)

It turns out that there's a very simple cure for the slow convergence of our method.

It might be shocking for some, but sometimes the true global minimum of cross-entropy for the whole training set is not exactly what we *really* want. In our predictive modelling task, we are minimising train error, but what we actually desire is to minimise the test error (which we cannot refer to while training = no cheating!).

It is therefore rational to assume that both the train and the test set consist of random digits independently sampled from the set of “all the possible digits out there in the world”.

Looking at the original objective (cross-entropy):

$$E(\mathbf{B}) = -\frac{1}{n^{\text{train}}} \sum_{i=1}^{n^{\text{train}}} \log \Pr(Y = y_i^{\text{train}} | \mathbf{x}_{i,:}^{\text{train}}, \mathbf{B}).$$

How about we try fitting to different random samples of the train set in each iteration of the gradient descent method instead of fitting to the whole train set?

$$E(\mathbf{B}) \simeq -\frac{1}{b} \sum_{i=1}^b \log \Pr(Y = y_{\text{random_index}_i}^{\text{train}} | \mathbf{x}_{\text{random_index}_i,:}^{\text{train}}, \mathbf{B}),$$

where b is some fixed batch size. Such an approach is often called **stochastic gradient descent**.

Remark. This scheme is sometimes referred to as **mini-batch** gradient descent in the literature; some researchers reserve the term “stochastic” only for batches of size 1.

Stochastic gradient descent can be implemented very easily:

```
B <- matrix(rnorm(ncol(X_train1)*ncol(Y_train2)),
             nrow=ncol(X_train1))
eta <- 0.1
maxit <- 100
batch_size <- 32
system.time({
  for (i in 1:maxit) {
    wh <- sample(nrow(X_train1), size=batch_size)
    B <- B - eta*grad_cross_entropy(
      X_train1[wh,], Y_train2[wh,],
      nn_predict(B, X_train1[wh,]))
  }
})
##     user   system elapsed
##   0.109   0.016   0.125
accuracy(Y_train2, nn_predict(B, X_train1))
## [1] 0.46198
accuracy(Y_test2, nn_predict(B, X_test1))
## [1] 0.4693
```

The errors are much worse but at least we got the (useless) solution very quickly. That's the “fail fast” rule in practice.

However, why don't we increase the number of iterations and see what happens? We've

allowed the classic gradient descent to scrabble around the MNIST dataset for almost 2 minutes.

```
B <- matrix(rnorm(ncol(X_train1)*ncol(Y_train2)),
            nrow=ncol(X_train1))
eta <- 0.1
maxit <- 10000
batch_size <- 32
system.time({
  for (i in 1:maxit) {
    wh <- sample(nrow(X_train1), size=batch_size)
    B <- B - eta*grad_cross_entropy(
      X_train1[wh,], Y_train2[wh,],
      nn_predict(B, X_train1[wh,]))
  }
})
##    user  system elapsed
##  7.936   0.270   8.206
accuracy(Y_train2, nn_predict(B, X_train1))

## [1] 0.89222
accuracy(Y_test2, nn_predict(B, X_test1))

## [1] 0.8935
```

Bingo! Let's take a closer look at how the train/test error behaves in each iteration for different batch sizes. Figures @ref(fig:mnist_sgd) and @ref(fig:mnist_sgd2b) depict the cases of `batch_size` of 32 and 128, respectively.

The time needed to go through 10000 iterations with batch size of 32 is:

```
##    user  system elapsed
##  89.921  23.053  35.112
```

What's more, batch size of 128 takes:

```
##    user  system elapsed
## 211.826  90.260  59.978
```

Exercise 9.3 *Draw conclusions.*

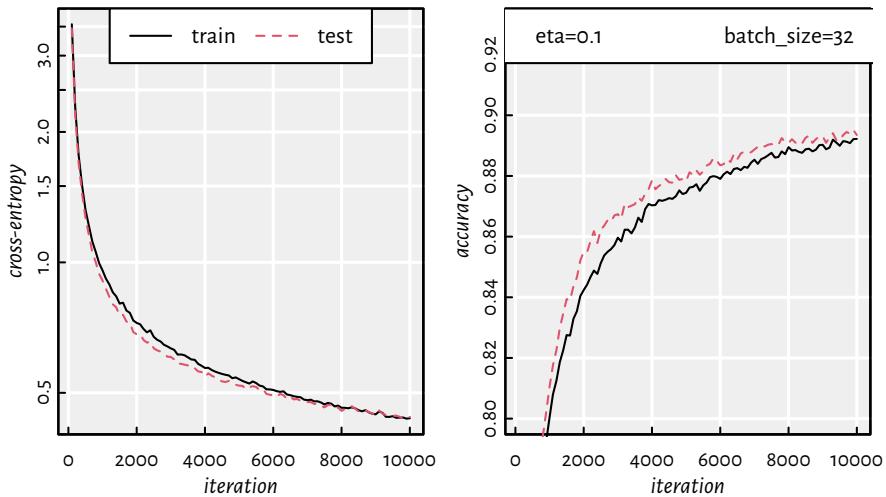


Figure 9.14: (#fig:mnist_sgd) Cross-entropy and accuracy on the train and test set in each iteration of SGD; batch size of 32

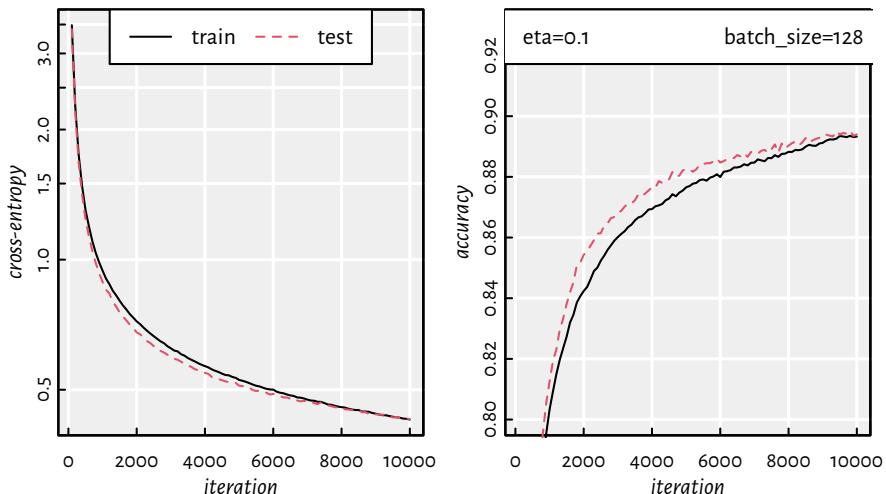


Figure 9.15: (#fig:mnist_sgd2b) Cross-entropy and accuracy on the train and test set in each iteration of SGD; batch size of 128

9.4 A Note on Convex Optimisation (*)

Are there any cases where we are sure that a local minimum is the global minimum? It turns out that the answer to this is positive; for example, when we minimise objective functions that fulfil a special property defined below.

First let's note that given two points $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^p$, by taking any $\theta \in [0, 1]$, the point defined as:

$$\mathbf{t} = \theta\mathbf{x}_1 + (1 - \theta)\mathbf{x}_2$$

lies on a (straight) line segment between \mathbf{x}_1 and \mathbf{x}_2 .

Definition. We say that a function $f : \mathbb{R}^p \rightarrow \mathbb{R}$ is *convex*, whenever:

$$(\forall \mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^p)(\forall \theta \in [0, 1]) \quad f(\theta\mathbf{x}_1 + (1 - \theta)\mathbf{x}_2) \leq \theta f(\mathbf{x}_1) + (1 - \theta)f(\mathbf{x}_2)$$

In other words, the function's value at any convex combination of two points is not greater than that combination of the function values at these two points. See Figure @ref(fig:convex_function) for a graphical illustration of the above.

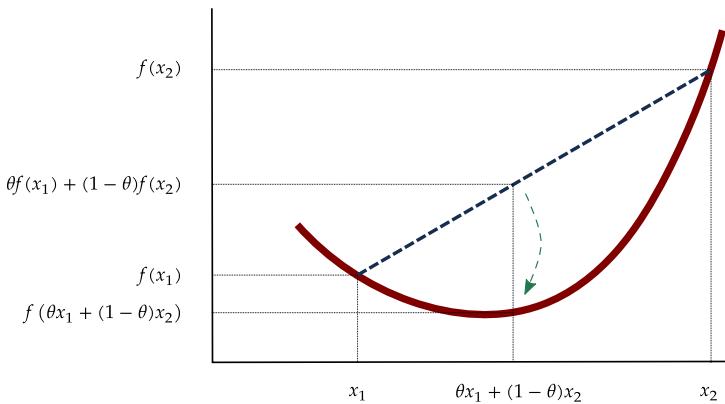


Figure 9.16: An illustration of the definition of a convex function

The following result addresses the question we posed at the beginning of this section.

Theorem. For any *convex* function f , if f has a local minimum at \mathbf{x}^+ then \mathbf{x}^+ is also its global minimum.

Convex functions are ubiquitous in machine learning, but of course not every objective function we are going to deal with will fulfil this property. Here are some basic examples

of convex functions and how they come into being, see, e.g., (Boyd & Vandenberghe 2004) for more:

- the functions mapping x to $x, x^2, |x|, e^x$ are all convex,
- $f(x) = |x|^p$ is convex for all $p \geq 1$,
- if f is convex, then $-f$ is concave,
- if f_1 and f_2 are convex, then $w_1 f_1 + w_2 f_2$ are convex for any $w_1, w_2 \geq 0$,
- if f_1 and f_2 are convex, then $\max\{f_1, f_2\}$ is convex,
- if f and g are convex and g is non-decreasing, then $g(f(x))$ is convex.

The above feature the building blocks of our error measures in supervised learning problems! In particular, sum of squared residuals in linear regression is a convex function of the underlying parameters. Also, cross-entropy in logistic regression is a convex function of the underlying parameters.

Theorem. (****) If a function is twice differentiable, then its convexity can be judged based on the positive-definiteness of its Hessian matrix.

Note that optimising convex functions is *relatively* easy, especially if they are differentiable. This is because they are quite well-behaving. However, it doesn't mean that we an analytic solution to the problem of their minimisation. Methods such as gradient descent or BFGS should work well (unless there are vast regions where a function is constant or the function's is defined over a large number of parameters).

Remark. (**) There is a special class of constrained optimisation problems called linear and, more generally, quadratic programming that involves convex functions. Moreover, the Karush–Kuhn–Tucker (KKT) conditions address the more general problem of minimisation with constraints (i.e., not over the whole \mathbb{R}^p set); see (Nocedal & Wright 2006, Fletcher 2008) for more details.

Remark. Not only functions, but also sets can be said to be convex. We say that $C \subseteq \mathbb{R}^p$ is a *convex set*, whenever the line segment joining any two points in C is fully included in C . More formally, for every $\mathbf{x}_1 \in C$ and $\mathbf{x}_2 \in C$, it holds $\theta \mathbf{x}_1 + (1 - \theta) \mathbf{x}_2 \in C$ for all $\theta \in [0, 1]$; see Figure @ref(fig:convex_set) for an illustration.

9.5 Outro

9.5.1 Remarks

Solving continuous problems with many variables (e.g., deep neural networks) is time consuming – the more variables to optimise over (e.g., model parameters, think the number of interconnections between all the neurons), the slower the optimisation process.

Moreover, it might be the case that the sole objective function takes long to compute (think of image classification with large training samples).

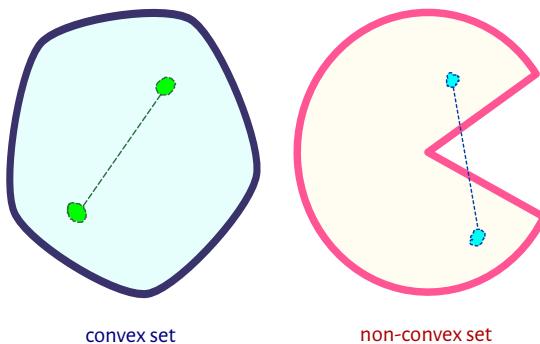


Figure 9.17: A convex and a non-convex set

Remark. (*) Although theoretically possible, good luck fitting a logistic regression model to MNIST with `optim()`'s BFGS – there are 7850 variables!

Training deep neural networks with SGD is even slower (more parameters), but there is a trick to propagate weight updates layer by layer, called *backpropagation* (actually used in every neural network library), see, e.g., (Sarle & others 2002) and (Goodfellow et al. 2016). Moreover, `keras` and similar libraries implement automatic differentiation procedures that make its user's life much easier (swiping some of the tedious math under the comfy carpet).

`keras` implements various optimisers that we can refer to in the `compile()` function, see <https://keras.rstudio.com/reference/compile.html> and <https://keras.io/optimizers/>:

- `SGD` – stochastic gradient descent supporting momentum and learning rate decay,
- `RMSprop` – divides the gradient by a running average of its recent magnitude,
- `Adam` – adaptive momentum,

and so on. These are all non-complicated variations of the pure stochastic GD. Some of them are just tricks that work well in some examples and destroy the convergence on many other ones. You can get into their details in a dedicated book/course aimed at covering neural networks (see, e.g., (Sarle & others 2002), (Goodfellow et al. 2016)), but we have already developed some good intuitions here.

Keep in mind that with methods such as GD or SGD, there is no guarantee we reach a minimum, but an approximate solution is better than no solution at all. Also sometimes (especially in ML applications) we don't really need the actual minimum (e.g., when optimising the error with respect to the train set). Those “mathematically pure” will find that a bit... unaesthetic, but here we are. Maybe the solution makes your boss or client happy, maybe it generates revenue. Maybe it helps solve some other problem. Some

claim that *a* solution is better than no solution at all, remember? But... is it really always the case though?

TODO

Recommended further reading: (Nocedal & Wright 2006), (Boyd & Vandenberghe 2004), (Fletcher 2008).

Next chapter...

10

Clustering with K-Means

TODO In this chapter, we will:

- ...
 - ...
-

10.1 Within-Cluster Sum of Squares

Prove that it holds:

$$\begin{aligned}\text{Var}(x_1, \dots, x_n) &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= \frac{1}{n} \sum_{i=1}^n \left(x_i - \frac{1}{n} \sum_{j=1}^n x_j \right)^2 \\ &= \frac{1}{n} \sum_{i=1}^n x_i^2 - \left(\frac{1}{n} \sum_{i=1}^n x_i \right)^2 \\ &= \frac{1}{2n^2} \sum_{i=1}^n \sum_{j=1}^n (x_i - x_j)^2 \\ &= \frac{1}{n^2} \sum_{i=1}^{n-1} \sum_{j=i+1}^n (x_i - x_j)^2\end{aligned}$$

In other words, the (biased) sample variance (average squared distance to the mean, \bar{x}) is equal to average of squared values minus square of the average which is equal to half of average distance between each pair of points (or just the average distance between each unique pair). Note that the former requires $\sim 2n$ arithmetic operations, while the latter $\sim n^2$

within-cluster sum of squares:

$$\begin{aligned}f(C_1, \dots, C_K) &= \sum_{k=1}^K \sum_{x_i \in C_k} \|x_i - \mu_k\|^2 \\ &= \sum_{k=1}^K \sum_{x_i \in C_k} \sum_{u=1}^d (x_{i,u} - \mu_{k,u})^2 \\ &= \sum_{k=1}^K \sum_{u=1}^d \sum_{x_i \in C_k} \left(x_{i,u} - \frac{1}{|C_k|} \sum_{x_j \in C_k} x_{j,u} \right)^2 \\ &= \sum_{k=1}^K \frac{2}{|C_k|} \sum_{x_i \in C_k} \sum_{x_j \in C_k} \sum_{u=1}^d (x_{i,u} - x_{j,u})^2 \\ &= \sum_{k=1}^K \frac{2}{|C_k|} \sum_{x_i \in C_k} \sum_{x_j \in C_k} \|x_i - x_j\|^2\end{aligned}$$

agglomerative strategy (Ward Jr. 1963)

divisive strategy (Edwards & Cavalli-Sforza 1965)

exhaustive search over MST - each $\binom{n-1}{k-1}$ possible splits of an MST (Caliński & Harabasz 1974)

??? - agglomerative over MST

??? - divisive over MST

k-means - fixed-point iteration algorithm ((MacQueen 1967) LLoyd, 1957)

TODO?? relation to Linear discriminant analysis and PCA??

10.2 K-means Clustering

10.2.1 Example in R

..hierarchical clustering is nice, because it outputs a whole hierarchy of nested partitions and it works in arbitrary spaces equipped with a distance, but most algorithms are slow for large datasets

Let's begin our clustering adventure by applying the K-means clustering method to find $K = 3$ groups in the famous Fisher's iris data set (variables Sepal.Width and Petal.Length variables only):

```
X <- as.matrix(iris[,c(3,2)])
# never forget to set nstart>>1!
km <- kmeans(X, centers=3, nstart=10)
km$cluster # labels assigned to each of 150 points:

## [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [35] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [69] 3 3 3 3 3 3 3 3 3 2 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
## [ reached getOption("max.print") -- omitted 51 entries ]
```

Remark. Later we'll see that nstart is responsible for random restarting the (local) optimisation procedure, just as we did in the previous chapter.

Let's draw a scatter plot that depicts the detected clusters:

```
plot(X, col=km$cluster)
```

The colours in Figure 10.1 indicate the detected clusters. The left group is clearly well-separated from the other two.

What can we do with this information? Well, if we were experts on plants (in the 1930s), that'd definitely be something ground-breaking. Figure 10.2 is a version of the aforementioned scatter plot now with the true iris species added.

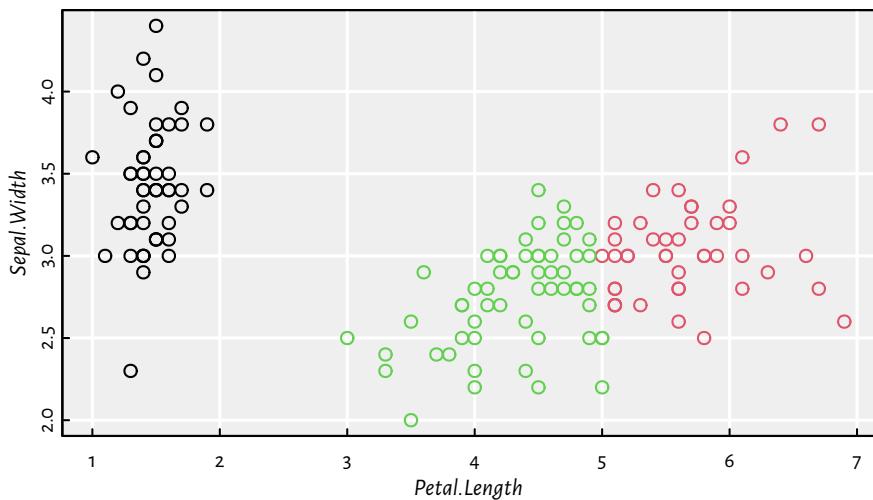


Figure 10.1: 3-means clustering on a projection of the Iris dataset

```
plot(X, col=km$cluster, pch=as.numeric(iris$Species))
```

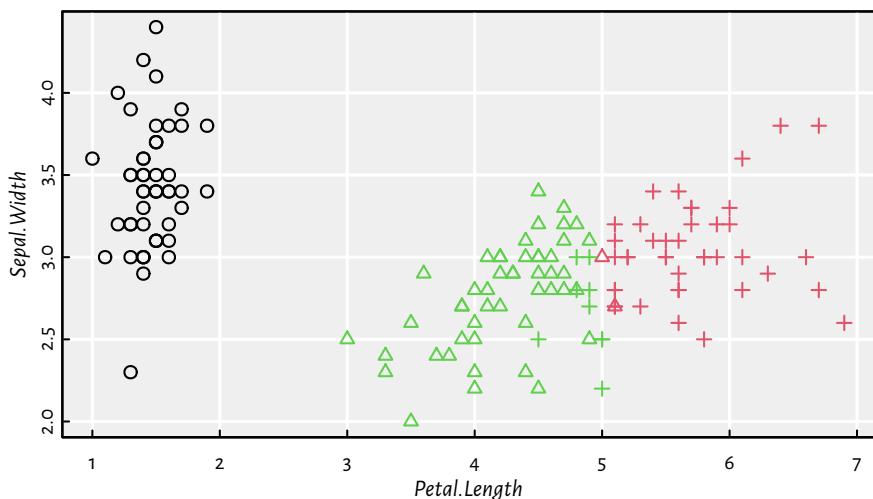


Figure 10.2: 3-means clustering (colours) vs true Iris species (shapes)

Here is a contingency table for detected clusters vs. true iris species:

```
(C <- table(km$cluster, iris$Species))
```

```
##          setosa versicolor virginica
## 1      50           0           0
## 2      0            2           41
## 3      0           48           9
```

It turns out that the discovered partition matches the original iris species very well. We have just made a “discovery” in the field of botany (actually some research fields classify their objects of study into families, genres etc. by means of such tools).

Were the actual Iris species what we had hoped to match? Was that our aim? Well, surely we have had begun our journey with “clear minds” (yet with hungry eyes). Note that the true class labels were not used during the clustering procedure – we’re dealing with an unsupervised learning problem here. The result turned useful, it’s a win.

Remark. (*) There are several indices that assess the similarity of two partitions, for example the Adjusted Rand Index (ARI) the Normalised Mutual Information Score (NMI) or set matching-based measures, see, e.g., (Hubert & Arabie 1985), (Rezaei & Fränti 2016).

10.2.2 Problem Statement

The aim of *K-means clustering* is to find K “good” cluster centres $\mu_{1,.}, \dots, \mu_{K,.}$

Then, a point $\mathbf{x}_{i,..}$ will be assigned to the cluster represented by the closest centre. Here, by *closest* we mean the *squared Euclidean distance*.

More formally, assuming all the points are in a p -dimensional space, \mathbb{R}^p , we define the distance between the i -th point and the k -th centre as:

$$d(\mathbf{x}_{i,.}, \mu_{k,.}) = \|\mathbf{x}_{i,.} - \mu_{k,.}\|^2 = \sum_{j=1}^p (x_{i,j} - \mu_{k,j})^2$$

Then the i -th point’s cluster is determined by:

$$C(i) = \arg \min_{k=1,\dots,K} d(\mathbf{x}_{i,.}, \mu_{k,.}),$$

where, as usual, $\arg \min$ (argument minimum) is the index k that minimises the given expression.

In the previous example, the three identified cluster centres in \mathbb{R}^2 are given by (see Figure @ref{fig:kmeans_problem1} for illustration):

```
km$centers
```

```
##   Petal.Length Sepal.Width
## 1      1.4620      3.4280
```

```

## 2      5.6721    3.0326
## 3      4.3281    2.7509

plot(X, col=km$cluster, asp=1) # asp=1 gives the same scale on both axes
points(km$centers, cex=2, col=4, pch=16)

```

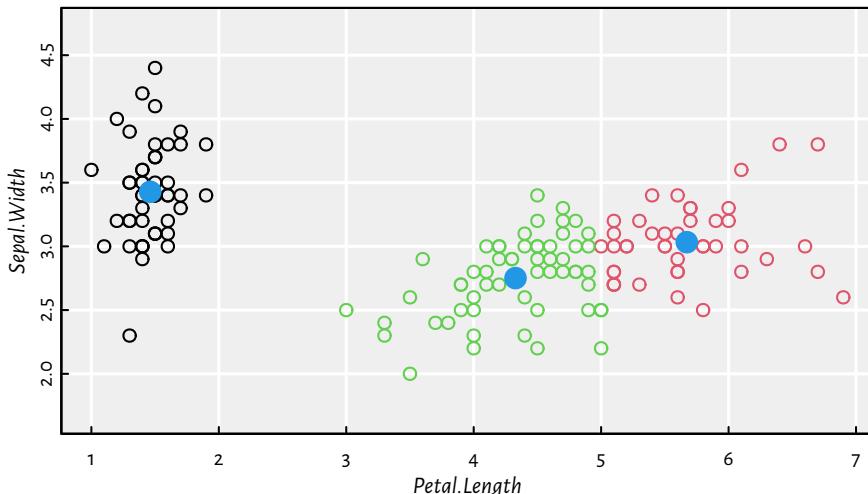


Figure 10.3: (#fig:kmeans_problem1) Cluster centres (blue dots) identified by the 3-means algorithm

Figure @ref(fig:kmeans_problem2) depicts the partition of the whole \mathbb{R}^2 space into clusters based on the closeness to the three cluster centres.

To compute the distances between all the points and the cluster centres, we may call `pdist::pdist()`:

```

library("pdist")
D <- as.matrix(pdist(X, km$centers))^2
head(D)

##          [,1]  [,2]  [,3]
## [1,] 0.009028 18.469 9.1348
## [2,] 0.187028 18.252 8.6357
## [3,] 0.078228 19.143 9.3709
## [4,] 0.109028 17.411 8.1199
## [5,] 0.033428 18.573 9.2946
## [6,] 0.279428 16.530 8.2272

```

where $D[i,k]$ gives the squared Euclidean distance between $x_{i,\cdot}$ and $\mu_{k,\cdot}$.

The cluster memberships the ($\arg \min$ s) can now be determined by:

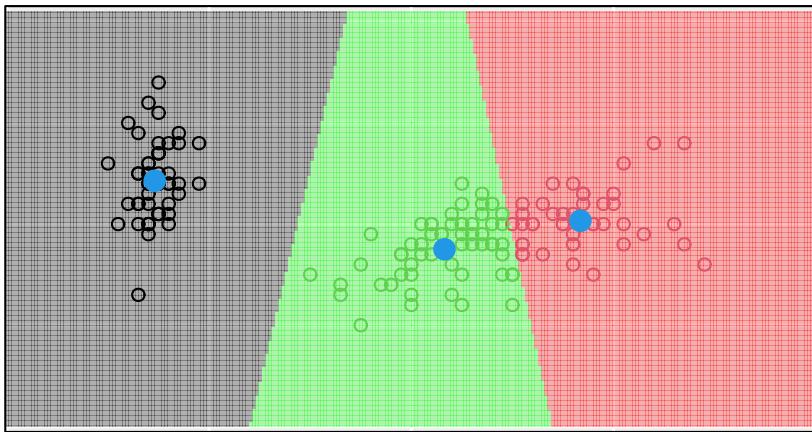


Figure 10.4: (#fig:kmeans_problem2) The division of the whole space into three sets based on the proximity to cluster centres (a so-called Voronoi diagram)

```
(idx <- apply(D, 1, which.min)) # for every row of D...
## [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [35] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
## [69] 3 3 3 3 3 3 3 3 3 2 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
## [ reached getOption("max.print") -- omitted 51 entries ]
all(km$cluster == idx) # sanity check
## [1] TRUE
```

10.2.3 Algorithms for the K-means Problem

All good, but how do we find “good” cluster centres? Good, better, best... yet again we are in a need for a goodness-of-fit metric. In the K-means clustering, we determine μ_1, \dots, μ_K , that minimise the total within-cluster distances (distances from each point to each own cluster centre):

$$\min_{\mu_1, \dots, \mu_K} \sum_{i=1}^n d(\mathbf{x}_{i,.}, \mu_{C(i),.}),$$

Note that the μ s are also “hidden” inside the point-to-cluster belongingness mapping, C. Expanding the above yields:

$$\min_{\mu_1, \dots, \mu_K \in \mathbb{R}^p} \sum_{i=1}^n \left(\min_{k=1, \dots, K} \sum_{j=1}^p (x_{i,j} - \mu_{k,j})^2 \right).$$

Unfortunately, the min operator in the objective function makes this optimisation problem not tractable with the methods discussed in the previous chapter.

The above problem is *hard* to solve (* more precisely, it is an NP-hard problem). Therefore, in practice we use various heuristics to solve it. The `kmeans()` function itself implements 3 of them: the Hartigan-Wong, Lloyd (a.k.a. Lloyd-Forgy) and MacQueen (MacQueen 1967) algorithms.

Remark. (*) Technically, there is no such thing as “*the* K-means algorithm” – all the aforementioned methods are particular heuristic approaches to solving the K-means clustering problem formalised as the above optimisation task. By setting `nstart = 10` above, we ask the (Hartigan-Wong, which is the default one in `kmeans()`) algorithm to find 10 solution candidates obtained by considering different random initial clusterings and choose the best one (with respect to the sum of within-cluster distances) amongst them. This does not guarantee finding the optimal solution, especially for very unbalanced datasets, but increases the likelihood of such.

Remark. The *squared* Euclidean distance was of course chosen to make computations easier. It turns out that for any given subset of input points $\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_m}$, the point $\mu_{k,\cdot}$ that minimises the total distances to all of them, i.e.,

$$\min_{\mu_{k,\cdot} \in \mathbb{R}^p} \sum_{\ell=1}^m \left(\sum_{j=1}^p (x_{i_\ell, j} - \mu_{k,j})^2 \right),$$

is exactly these points’ *centroid* – which is given by the componentwise arithmetic means of their coordinates.

For example:

```
colMeans(X[km$cluster == 1,]) # centroid of the points in the 1st cluster

## Petal.Length Sepal.Width
##      1.462       3.428

km$centers[1,] # the centre of the 1st cluster

## Petal.Length Sepal.Width
##      1.462       3.428
```

TODO: see Lloyd’s (and Forgy’s) original papers to verify these method defs

Among the various heuristics to solve the K-means problem, Lloyd’s algorithm (1957) is perhaps the simplest. This is probably the reason why it is sometimes referred to as “*the*” K-means algorithm:

1. Start with random cluster centres $\mu_{1,.}, \dots, \mu_{K,.}$
2. For each point $\mathbf{x}_{i,.}$, determine its closest centre $C(i) \in \{1, \dots, K\}$:

$$C(i) = \arg \min_{k=1,\dots,K} d(\mathbf{x}_{i,.}, \mu_{k,.}).$$

3. For each cluster $k \in \{1, \dots, K\}$, compute the new cluster centre $\mu_{k,.}$ as the centroid of all the point indices i such that $C(i) = k$.
4. If the cluster centres changed since the last iteration, go to step 2, otherwise stop and return the result.

(*) Here's an example implementation. As the initial cluster centres, let's pick some "noisy" versions of K randomly chosen points in \mathbf{X} .

```
set.seed(12345)
K <- 3

# Random initial cluster centres:
M <- jitter(X[sample(1:nrow(X), K),])
M

##      Petal.Length Sepal.Width
## [1,]      5.1004     3.0814
## [2,]      4.7091     3.1861
## [3,]      3.3196     2.4094
```

In what follows, we will be maintaining a matrix such that $D[i,k]$ is the distance between the i -th point and the k -th centre and a vector such that $idx[i]$ denotes the index of the cluster centre closest to the i -th point.

```
D <- as.matrix(pdist(X, M))^2
idx <- apply(D, 1, which.min)

repeat {
  # Determine the new cluster centres:
  M <- t(sapply(1:K, function(k) {
    # the centroid of all points in the k-th cluster:
    colMeans(X[idx==k,])
  }))

  # Store the previous cluster belongingness info:
  old_idx <- idx

  # Recompute D and idx:
  D <- as.matrix(pdist(X, M))^2
  idx <- apply(D, 1, which.min)
```

```

# Check if converged already:
if (all(idx == old_idx)) break
}

```

Let's compare the obtained cluster centres with the ones returned by `kmeans()`:

```

M # our result

##      Petal.Length Sepal.Width
## [1,]      5.6721     3.0326
## [2,]      4.3281     2.7509
## [3,]      1.4620     3.4280

km$center # result of kmeans()

##      Petal.Length Sepal.Width
## 1      1.4620     3.4280
## 2      5.6721     3.0326
## 3      4.3281     2.7509

```

These two represent exactly the same 3-partitions (note that the actual labels (the order of centres) are not important).

The value of the objective function (total within-cluster distances) at the identified candidate solution is equal to:

```
sum(D[cbind(1:nrow(X),idx)]) # indexing with a 2-column matrix!
```

```

## [1] 40.737
km$tot.withinss # as reported by kmeans()

## [1] 40.737

```

We would need it if we were to implement the `nstart` functionality, which is left as an:

Exercise 10.1 (*) Wrap the implementation of the Lloyd algorithm into a standalone R function, with a similar look-and-feel as the original `kmeans()`.

On a side note, our algorithm needed 4 iterations to identify the (locally optimal) cluster centres. Figure @ref(fig:kmeanimpl_plot) depicts its quest for the clustering grail.

10.2.4 K-means Revisited

In **K-means clustering** we are minimising the squared Euclidean distance to each point's cluster centre:

$$\min_{\mu_1, \dots, \mu_K, \in \mathbb{R}^p} \sum_{i=1}^n \left(\min_{k=1, \dots, K} \sum_{j=1}^p (x_{i,j} - \mu_{k,j})^2 \right).$$

This is an (NP-)hard problem! There is no efficient exact algorithm.

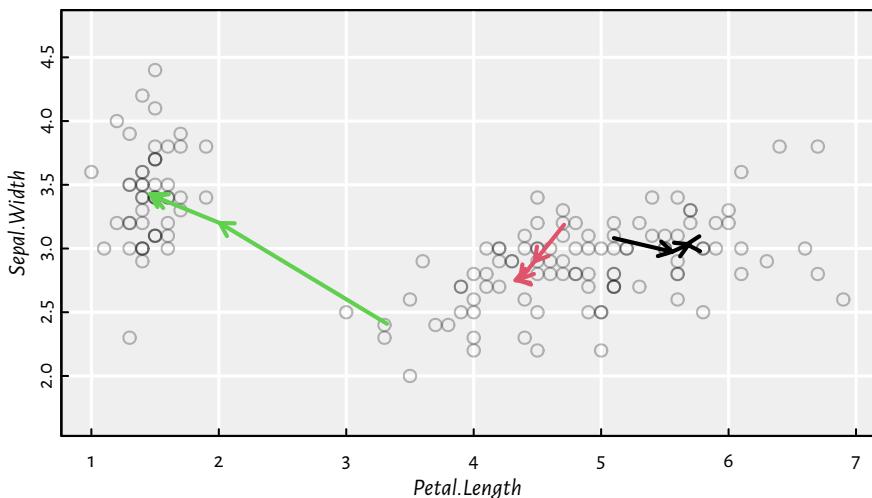


Figure 10.5: (#fig:kmeanimpl_plot) The arrows denote the cluster centres in each iteration of the Lloyd algorithm

We need approximations. In the last chapter, we have discussed the iterative Lloyd's algorithm (1957), which is amongst a few procedures implemented in the `kmeans()` function.

Lloyd, Stuart P. (1957). "Least square quantization in PCM". Bell Telephone Laboratories Paper. Published much later: Lloyd, Stuart P. (1982). "Least squares quantization in PCM" (PDF). IEEE Transactions on Information Theory. 28 (2): 129–137. doi:10.1109/TIT.1982.1056489.

MacQueen (MacQueen 1967) studies the “k-means process” (and its asymptotic behaviour) *Stated informally, the k-means procedure consists of simply starting with k groups each of which consists of a single random point, and thereafter adding each new point to the group whose mean the new point is nearest. After a point is added to a group, the mean of that group is adjusted in order to take account of the new point. Thus at each stage the k-means are, in fact, the means of the groups they represent (hence the term k-means).* (p.283)

Hartigan-Wong == ??

To recall, Lloyd's algorithm (1957) is sometimes referred to as “the” K-means algorithm:

1. Start with random cluster centres $\mu_{1,.}, \dots, \mu_{K,.}$
2. For each point $\mathbf{x}_{i,.}$, determine its closest centre $C(i) \in \{1, \dots, K\}$.
3. For each cluster $k \in \{1, \dots, K\}$, compute the new cluster centre $\mu_{k,.}$ as the componentwise arithmetic mean of the coordinates of all the point indices i such that $C(i) = k$.

4. If the cluster centres changed since last iteration, go to step 2, otherwise stop and return the result.

As the procedure might get stuck in a local minimum, a few restarts are recommended (as usual).

Hence, we are used to calling:

```
kmeans(X, centers=k, nstart=10)
```

10.2.5 optim() vs. kmeans()

Let us compare how a general-purpose optimiser such as the BFGS algorithm implemented in `optim()` compares with a customised, problem-specific solver.

We will need some benchmark data.

```
gen_cluster <- function(n, p, m, s) {
  vectors <- matrix(rnorm(n*p), nrow=n, ncol=p)
  unit_vectors <- vectors/sqrt(rowSums(vectors^2))
  unit_vectors*rnorm(n, 0, s)+rep(m, each=n)
}
```

The above function generates n points in \mathbb{R}^p from a distribution centred at $\mathbf{m} \in \mathbb{R}^p$, spread randomly in every possible direction with scale factor s .

Two example clusters in \mathbb{R}^2 :

```
# plot the "black" cluster
plot(gen_cluster(500, 2, c(0, 0), 1), col="#00000022", pch=16,
      xlim=c(-3, 4), ylim=c(-3, 4), asp=1, ann=FALSE)
# plot the "red" cluster
points(gen_cluster(250, 2, c(1.5, 1), 0.5), col="#ff000022", pch=16)
```

Let's generate the benchmark dataset \mathbf{X} that consists of three clusters in a high-dimensional space.

```
set.seed(123)
p <- 32
Ns <- c(50, 100, 20)
Ms <- c(0, 1, 2)
s <- 1.5*p
K <- length(Ns)

X <- lapply(1:K, function(k)
```

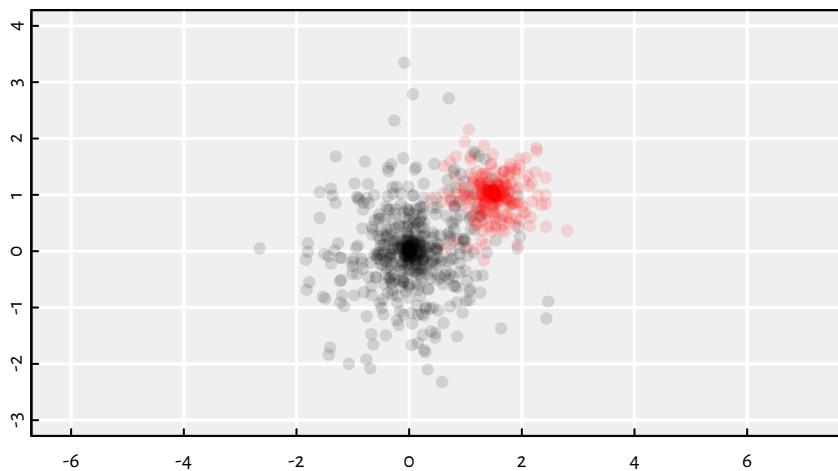


Figure 10.6: (#fig:gendata_example) plot of chunk gendata_example

```
gen_cluster(Ns[k], p, rep(Ms[k], p), s))
X <- do.call(rbind, X) # rbind(X[[1]], X[[2]], X[[3]])
```

The objective function for the K-means clustering problem:

```
library("FNN")
get_fitness <- function(mu, X) {
  # For each point in X,
  # get the index of the closest point in mu:
  memb <- FNN::get.knnx(mu, X, 1)$nn.index

  # compute the sum of squared distances
  # between each point and its closes cluster centre:
  sum((X-mu[memb,])^2)
}
```

Setting up the solvers:

```
min_HartiganWong <- function(mu0, X)
  get_fitness(
    # algorithm="Hartigan-Wong"
    kmeans(X, mu0, iter.max=100)$centers,
    X)
min_Lloyd <- function(mu0, X)
  get_fitness(
    kmeans(X, mu0, iter.max=100, algorithm="Lloyd")$centers,
```

```
X)
min_optim <- function(mu0, X)
  optim(mu0,
    function(mu, X) {
      get_fitness(matrix(mu, nrow=nrow(mu0)), X)
    }, X=X, method="BFGS", control=list(reltol=1e-16)
 )$val
```

Running the simulation:

```
nstart <- 100
set.seed(123)
res <- replicate(nstart, {
  mu0 <- X[sample(nrow(X), K),]
  c(
    HartiganWong=min_HartiganWong(mu0, X),
    Lloyd=min_Lloyd(mu0, X),
    optim=min_optim(mu0, X)
  )
})
```

Notice a considerable variability of the objective function at the local minima found:

```
boxplot(as.data.frame(t(res)), horizontal=TRUE, col="white")
```

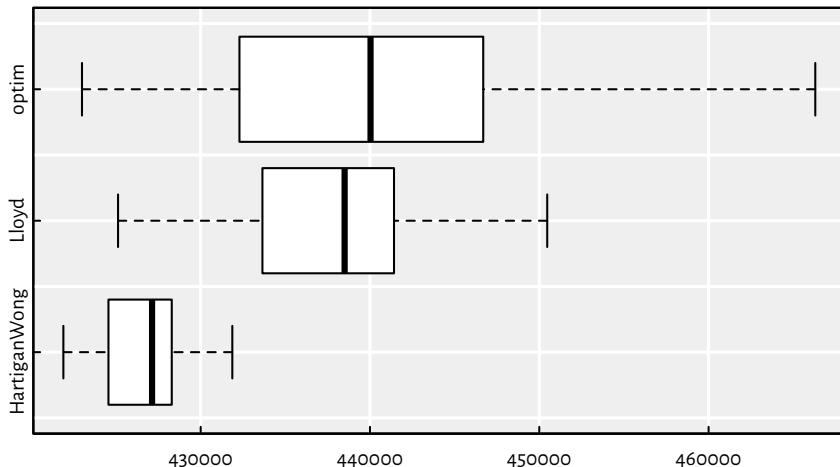


Figure 10.7: plot of chunk gendata5

```
print(apply(res, 1, function(x)
```

```
c(summary(x), sd=sd(x))
))

##      HartiganWong    Lloyd  optim
## Min.        421889 425119.5 422989
## 1st Qu.     424663 433669.3 432446
## Median     427129 438502.2 440033
## Mean       426557 438075.0 440635
## 3rd Qu.     428243 441381.3 446614
## Max.       431869 450469.7 466303
## sd          2301   5709.3  10888
```

Of course, we are interested in the smallest value of the objective, because we're trying to pinpoint the global minimum.

```
print(apply(res, 1, min))

## HartiganWong      Lloyd      optim
##        421889      425119      422989
```

The Hartigan-Wong algorithm (the default one in `kmeans()`) is the most reliable one of the three:

- it gives the best solution (low bias)
- the solutions have the lowest degree of variability (low variance)
- it is the fastest:

```
library("microbenchmark")
set.seed(123)
mu0 <- X[sample(nrow(X), K),]
summary(microbenchmark(
  HartiganWong=min_HartiganWong(mu0, X),
  Lloyd=min_Lloyd(mu0, X),
  optim=min_optim(mu0, X),
  times=10
), unit="relative")

##           expr      min       lq     mean     median       uq
## 1 HartiganWong  1.1362  1.1515  1.1326  1.1152  1.0734
## 2      Lloyd   1.0000  1.0000  1.0000  1.0000  1.0000
## 3      optim 1642.3472 1671.7683 1529.0246 1540.3921 1404.0872
##           max neval
## 1    1.2846     10
## 2    1.0000     10
## 3 1430.7911     10

print(min(res))

## [1] 421889
```

Is it the global minimum?

We don't know, we just didn't happen to find anything better (yet).

Did we put enough effort to find it?

Well, maybe. We can try more random restarts:

```
res_tried_very_hard <- kmeans(X, K, nstart=100000, iter.max=10000)$centers  
print(get_fitness(res_tried_very_hard, X))
```

```
## [1] 421889
```

Is it good enough?

It depends what we'd like to do with this. Does it make your boss happy? Does it generate revenue? Does it help solve any other problem? Is it useful anyhow? Are you really looking for the global minimum?

10.3 Exercises

10.3.1 Clustering of the World Factbook

Let's perform a cluster analysis of countries based on the information contained in the World Factbook dataset:

```
factbook <- read.csv("datasets/world_factbook_2020.csv",  
comment.char="#")
```

Exercise 10.2 Remove all the columns that consist of more than 40 missing values. Then remove all the rows with at least 1 missing value.

Solution.

To remove appropriate columns, we must first count the number of NAs in them.

```
count_na_in_columns <- sapply(factbook, function(x) sum(is.na(x)))
factbook <- factbook[count_na_in_columns <= 40] # column removal
```

Getting rid of the rows plagued by missing values is as simple as calling the `na.omit()` function:

```
factbook <- na.omit(factbook) # row removal
dim(factbook) # how many rows and cols remained
```

```
## [1] 203 23
```

Missing value removal is necessary for metric-based clustering methods, especially K-means. Otherwise, some of the computed distances would be not available.



Exercise 10.3 Standardise all the numeric columns.

Solution.

Distance-based methods are very sensitive to the order of magnitude of the variables, and our dataset is a mess with regards to this (population, GDP, birth rate, oil production etc.) – standardisation of variables is definitely a good idea:

```
for (i in 2:ncol(factbook)) # skip `country`
  factbook[[i]] <- (factbook[[i]] - mean(factbook[[i]]))/
    sd(factbook[[i]]))
```

Recall that Z-scores (values of the standardised variables) have a very intuitive interpretation: 0 is the value equal to the column mean, 1 is one standard deviation above the mean, -2 is two standard deviations below the mean etc.



Exercise 10.4 Apply the 2-means algorithm, i.e., K-means with $K = 2$. Analyse the results.

Solution.

Calling `kmeans()`:

```
km <- kmeans(factbook[-1], 2, nstart=10)
```

Let's split the country list w.r.t. the obtained cluster labels. It turns out that the obtained partition is heavily imbalanced, so we'll print only the contents of the first group:

```
km_countries <- split(factbook[[1]], km$cluster)
km_countries[[1]]
```

```
## [1] "China"           "India"            "United States"
```

With regards to which criteria has the K-means algorithm distinguished the countries? Let's inspect the cluster centres to check the average Z-scores of all the countries in each cluster:

```
t(km$centers) # transposed for readability
```

	1	2
## area	3.661581	-0.0549237
## population	6.987279	-0.1048092
## median_age	0.477991	-0.0071699
## population_growth_rate	-0.252774	0.0037916
## birth_rate	-0.501030	0.0075155
## death_rate	0.153915	-0.0023087
## net_migration_rate	0.236449	-0.0035467
## infant_mortality_rate	-0.139577	0.0020937
## life_expectancy_at_birth	0.251541	-0.0037731
## total_fertility_rate	-0.472716	0.0070907
## gdp_purchasing_power_parity	7.213681	-0.1082052
## gdp_real_growth_rate	0.369499	-0.0055425
## gdp_per_capita_ppp	0.298103	-0.0044715
## labor_force	6.914319	-0.1037148
## taxes_and_other_revenues	-0.922735	0.0138410
## budget_surplus_or_deficit	-0.012627	0.0001894
## inflation_rate_consumer_prices	-0.096626	0.0014494
## exports	5.341178	-0.0801177
## imports	5.956538	-0.0893481
## telephones_fixed_lines	5.989858	-0.0898479
## internet_users	6.997126	-0.1049569
## airports	4.551832	-0.0682775

Countries in Cluster 2 are... average (Z-scores ≈ 0). On the other hand, the three countries in Cluster 1 dominate the others w.r.t. area, population, GDP PPP, labour force etc.

Exercise 10.5 Apply the complete linkage agglomerative hierarchical clustering algorithm. ■

Solution.

Recall that the complete linkage-based method is implemented in the `hclust()` function:

```
d <- dist(factbook[-1]) # skip `country'
h <- hclust(d, method="complete")
```

A “nice” number of clusters to divide our dataset into can be read from the dendrogram, see Figure @ref{fig:clustering_factbook7}.

```
plot(h, labels=FALSE, ann=FALSE); box()
```

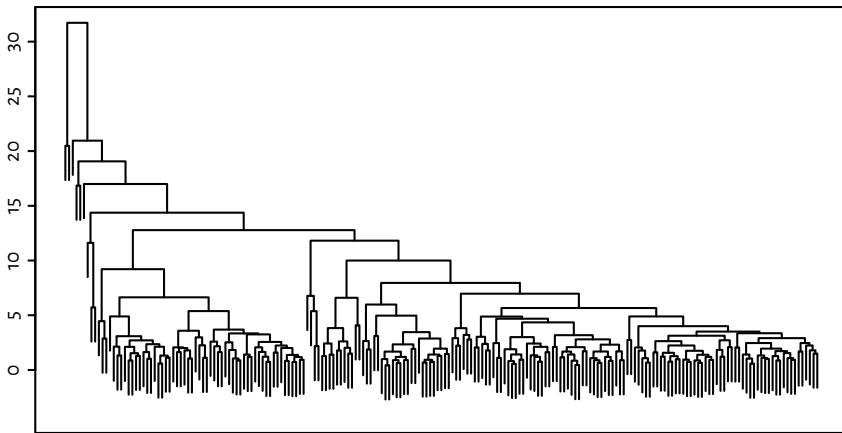


Figure 10.8: (#fig:clustering_factbook7) Cluster dendrogram for the World Factbook dataset – Complete linkage

It seems that a 9-partition might reveal something interesting, because it will distinguish two larger country groups. However, there will be many singletons if we do so either way.

```
y <- cutree(h, 9)
h_countries <- split(factbook[[1]], y)
sapply(h_countries, length) # number of elements in each cluster
```

```
##   1   2   3   4   5   6   7   8   9
## 138  56   1   1   3   1   1   1   1
```

Most likely this is not an interesting partitioning of this dataset, therefore we'll not be exploring it any further. ■

Exercise 10.6 Apply the Genie clustering algorithm (package `genieclust`).

Solution.

The Genie algorithm (Gagolewski et al. 2016) is a hierarchical clustering algorithm implemented in R package `genieclust`. Its interface is compatible with `hclust()`.

```
library("genieclust")
d <- dist(factbook[-1])
g <- gclus(g)
```

The cluster dendrogram in Figure @ref(fig:clustering_factbook10) reveals 3 evident clusters.

```
plot(g, labels=FALSE, ann=FALSE); box()
```

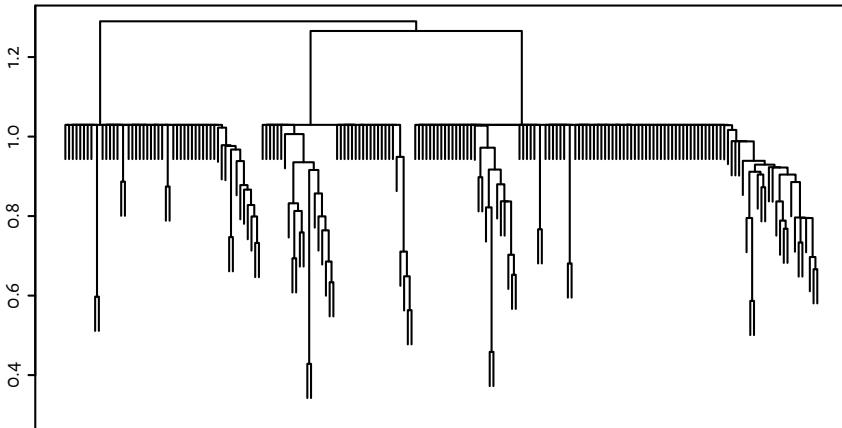


Figure 10.9: (#fig:clustering_factbook10) Cluster dendrogram for the World Factbook dataset – Genie algorithm

Let's determine the 3-partition of the data set.

```
y <- cutree(g, 3)
```

Here are few countries in each cluster:

```
y <- cutree(g, 3)
sapply(split(factbook$country, y), sample, 6)
```

```
##      1          2          3
## [1,] "Curacao"   "Uganda"   "Netherlands"
## [2,] "Puerto Rico" "Senegal"   "Bosnia and Herzegovina"
## [3,] "Israel"     "Niger"     "United Kingdom"
## [4,] "Barbados"   "Djibouti"  "Norway"
## [5,] "Mongolia"   "Haiti"    "Denmark"
## [6,] "Guam"       "Kenya"    "Palau"
```

We can draw the countries in each cluster on a map by using the *rworldmap* package (see its documentation for more details), see Figure @ref(fig:clustering_factbook12).

```

library("rworldmap")
mapdata <- data.frame(country=factbook$country, cluster=y)
# 3 country names must be adjusted to get a match
mapdata$country[mapdata$country == "Czechia"] <- "Czech Republic"
mapdata$country[mapdata$country == "Eswatini"] <- "Swaziland"
mapdata$country[mapdata$country == "Cabo Verde"] <- "Cape Verde"
mapdata <- joinCountryData2Map(mapdata, joinCode="NAME",
                                nameJoinColumn="country")

## 203 codes from your data successfully matched countries in the map
## 0 codes from your data failed to match with a country code in the map
## 40 codes from the map weren't represented in your data

par(mar=c(0,0,0,0))
mapCountryData(mapdata, nameColumnToPlot="cluster",
               catMethod="categorical", missingCountryCol="gray",
               colourPalette=palette()[2:4],
               mapTitle="", addLegend=TRUE)

```

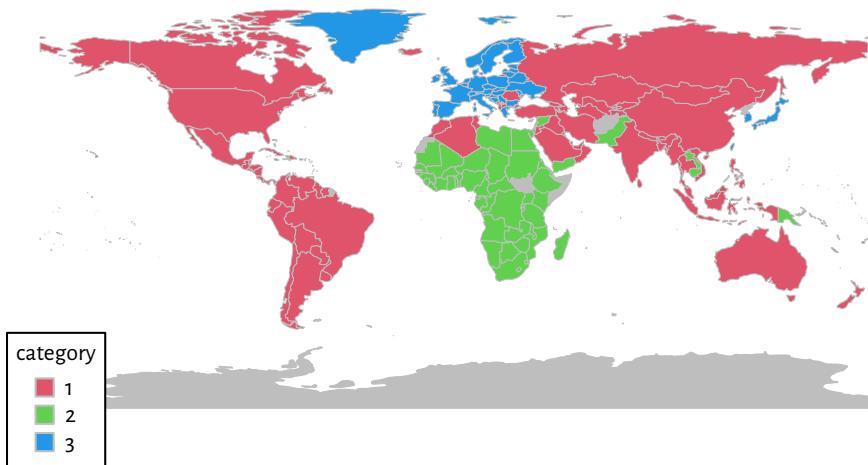


Figure 10.10: (#fig:clustering_factbook12) 3 clusters discovered by the Genie algorithm

Here are the average Z-scores in each cluster:

```
round(sapply(split(factbook[-1], y), colMeans), 3)
```

	1	2	3
## area	0.124	-0.068	-0.243
## population	0.077	-0.058	-0.130
## median_age	0.118	-1.219	1.261

```

## population_growth_rate      -0.227  1.052 -0.757
## birth_rate                  -0.316  1.370 -0.930
## death_rate                  -0.439  0.071  1.075
## net_migration_rate         -0.123  0.053  0.260
## infant_mortality_rate     -0.366  1.399 -0.835
## life_expectancy_at_birth   0.354 -1.356  0.812
## total_fertility_rate       -0.363  1.332 -0.758
## gdp_purchasing_power_parity 0.084 -0.213  0.052
## gdp_real_growth_rate      -0.062  0.126  0.002
## gdp_per_capita_ppp          0.021 -0.744  0.905
## labor_force                 0.087 -0.096 -0.107
## taxes_and_other_revenues   -0.095 -0.584  1.006
## budget_surplus_or_deficit  -0.113 -0.188  0.543
## inflation_rate_consumer_prices 0.044 -0.013 -0.099
## exports                     -0.013 -0.318  0.447
## imports                      0.007 -0.308  0.379
## telephones_fixed_lines      0.048 -0.244  0.186
## internet_users              0.093 -0.178 -0.016
## airports                     0.104 -0.131 -0.107

```

That is really interesting! The interpretation of the above is left to the reader.



10.3.2 Unbalance Dataset – K-Means Needs Multiple Starts

Let us consider a benchmark (artificial) dataset proposed in (Rezaei & Fräntti 2016):

```

unbalance <- as.matrix(read.csv("datasets/sipu_unbalance.csv",
  header=FALSE, sep=" ", comment.char="#"))
unbalance <- unbalance/10000-30 # a more user-friendly scale

```

According to its authors, this dataset is comprised of 8 clusters: there are 3 groups on the lefthand side (2000 points each) and 5 on the right side (100 each).

```
plot(unbalance, asp=1)
```

Exercise 10.7 Apply the K-means algorithm with $K = 8$.

Solution.

Of course, here by “the” K-means we mean the default method available in the `kmeans()` function. The clustering results are depicted in Figure @ref{fig:sipu_unbalance3a}.

```

km <- kmeans(unbalance, 8, nstart=10)
plot(unbalance, asp=1, col=km$cluster)

```

This is far from what we expected. The total within-cluster distances are equal to:

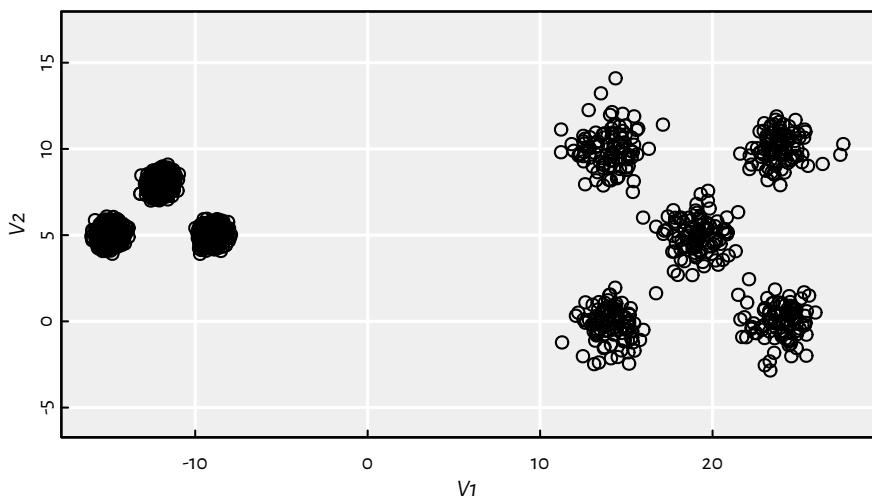


Figure 10.11: (#fig:sipu_unbalance2) `sipu_unbalance` dataset

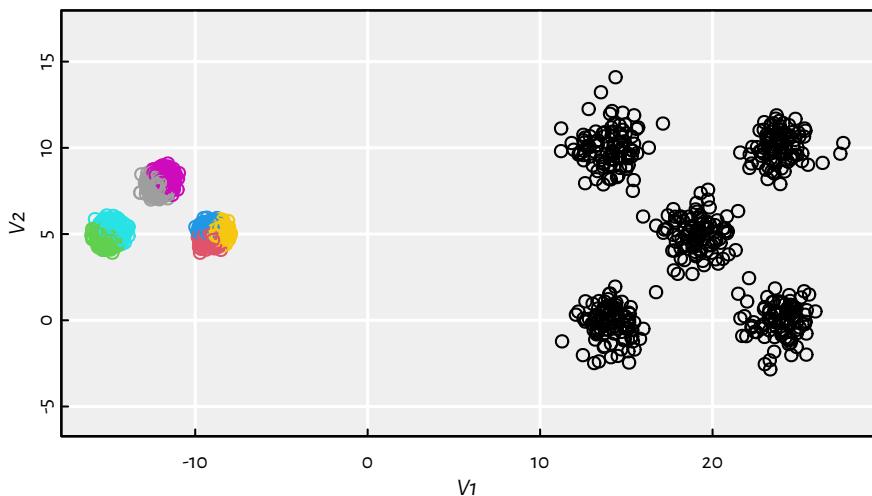


Figure 10.12: (#fig:sipu_unbalance3a) Results of K-means on the `sipu_unbalance` dataset

```
km$tot.withinss
```

```
## [1] 21713
```

Increasing the number of restarts even further improves the solution, but the local minimum is still far from the global one, compare Figure @ref(fig:sipu_unbalance3b).

```
km <- suppressWarnings(kmeans(unbalance, 8, nstart=1000, iter.max=1000))
plot(unbalance, asp=1, col=km$cluster)
```

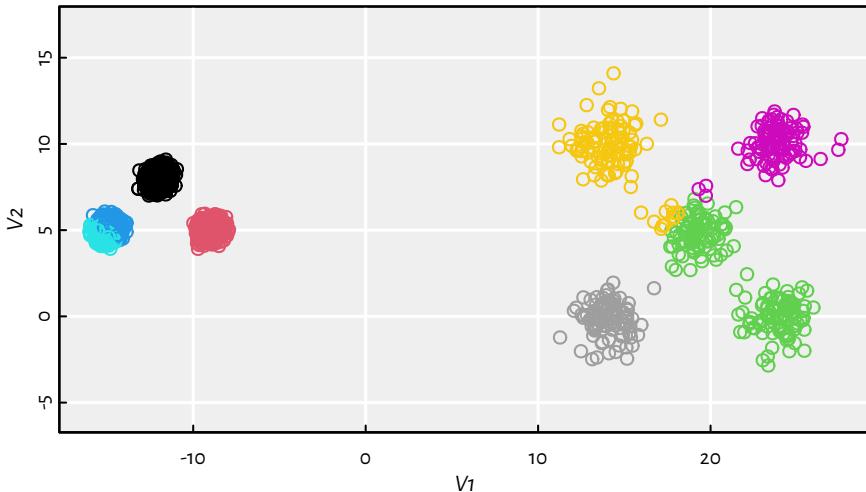


Figure 10.13: (#fig:sipu_unbalance3b) Results of K-means on the `sipu_unbalance` dataset – many more restarts

```
km$tot.withinss
```

```
## [1] 4378
```

Exercise 10.8 Apply the K-means algorithm starting from a “good” initial guess on the true cluster centres.

Solution.

Clustering is – in its essence – an unsupervised learning method, so what we’re going to do now could be called, let’s be blunt about it, cheating. Luckily, we have an oracle at our disposal – it has provided us with the following educated guesses (by looking at the scatter plot) about the localisation of the cluster centres:

```
cntr <- matrix(ncol=2, byrow=TRUE, c(
  -15, 5,
  -12, 10,
  -10, 5,
  15, 0,
  15, 10,
  20, 5,
  25, 0,
  25, 10))
```

Running `kmeans()` yields the clustering depicted in Figure @ref(fig:sipu_unbalance6).

```
km <- kmeans(unbalance, cntr)
plot(unbalance, asp=1, col=km$cluster)
```

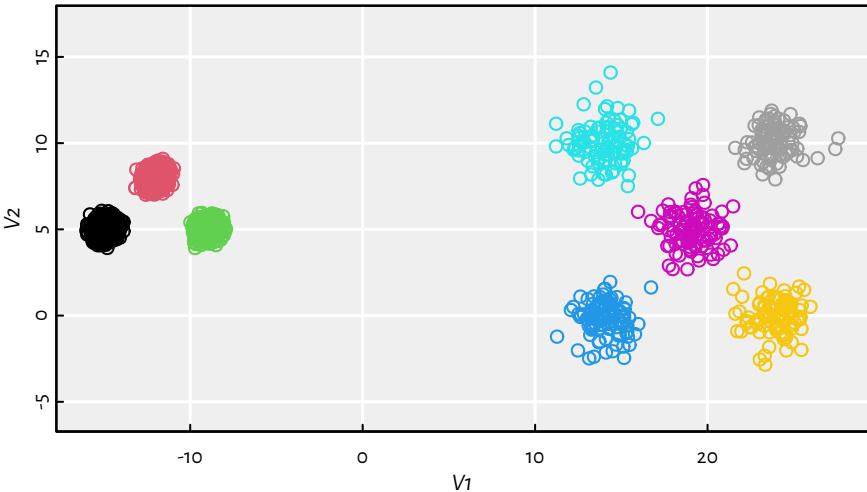


Figure 10.14: (#fig:sipu_unbalance6) Results of K-means on the `sipu_unbalance` dataset – an educated guess on the cluster centres' locations

The total within-cluster distances are now equal to:

```
km$tot.withinss
```

```
## [1] 2144.9
```

This is finally the globally optimal solution to the K-means problem we were asked to solve. Recall that the algorithms implemented in the `kmeans()` function are just fast heuristics that are supposed to find local optima of the K-means objective function, which is given by the within-cluster sum of squared Euclidean distances.

10.3.3 Clustering of Typical 2D Benchmark Datasets

Let us consider a few clustering benchmark datasets available at https://github.com/gagolews/clustering_benchmarks_v1 and <http://cs.joensuu.fi/sipu/datasets/>. Here is a list of file names together with the corresponding numbers of clusters (as given by datasets' authors):

```
files <- c("datasets/wut_isolation.csv",
         "datasets/wut_mk2.csv",
         "datasets/wut_z3.csv",
         "datasets/sipu_aggregation.csv",
         "datasets/sipu_pathbased.csv",
         "datasets/sipu_unbalance.csv")
Ks <- c(3, 2, 4, 7, 3, 8)
```

All the datasets are two-dimensional, hence we'll be able to visualise the obtained results and assess the sensibility of the obtained clusterings.

Exercise 10.9 *Apply the K-means, the single, average and complete linkage and the Genie algorithm (from package `genieclust`) on the aforementioned datasets and discuss the results.*

Solution.

Apart from a call to the Genie algorithm with the default parameters, we will also look at the results it generates when we set the `gini_threshold` parameter to 0.5 (default is 0.3; smaller thresholds lead to clusters of more balanced sizes as measured by the Gini index).

The following function is our workhorse that will perform all the computations and will draw all the figures for a single dataset:

```
clusterise <- function(file, K) {
  X <- read.csv(file,
                header=FALSE, sep=" ", comment.char="#")
  d <- dist(X)
  par(mfrow=c(2, 3))
  par(mar=c(0.5, 0.5, 2, 0.5))

  y <- kmeans(X, K, nstart=10)$cluster
  plot(X, asp=1, col=y, ann=FALSE, axes=FALSE)
  mtext("K-means", line=0.5)

  y <- cutree(hclust(d, "complete"), K)
  plot(X, asp=1, col=y, ann=FALSE, axes=FALSE)
  mtext("Complete Linkage", line=0.5)

  y <- cutree(hclust(d, "average"), K)
  plot(X, asp=1, col=y, ann=FALSE, axes=FALSE)
  mtext("Average Linkage", line=0.5)}
```

```

y <- cutree(hclust(d, "single"), K)
plot(X, asp=1, col=y, ann=FALSE, axes=FALSE)
mtext("Single Linkage", line=0.5)

y <- genie(d, K) # gini_threshold=0.3
plot(X, asp=1, col=y, ann=FALSE, axes=FALSE)
mtext("Genie (default)", line=0.5)

y <- genie(d, K, gini_threshold=0.5)
plot(X, asp=1, col=y, ann=FALSE, axes=FALSE)
mtext("Genie (g=0.5)", line=0.5)
}

```

Applying the above as `clusterise(files[i], Ks[i])` yields Figures @ref(fig:clustering_benchmarks_plot1)-@ref(fig:clustering_benchmarks_plot6).



Figure 10.15: (#fig:clustering_benchmarks_plot1) Clustering of the `wut_isolation` data-set

Note that, by definition, K-means is only able to detect clusters of convex shapes. The Genie algorithm, on the other hand, might fail to detect clusters of very small sizes amongst the more populous ones. Single linkage is very sensitive to outliers in data – it often outputs clusters of cardinality 1.



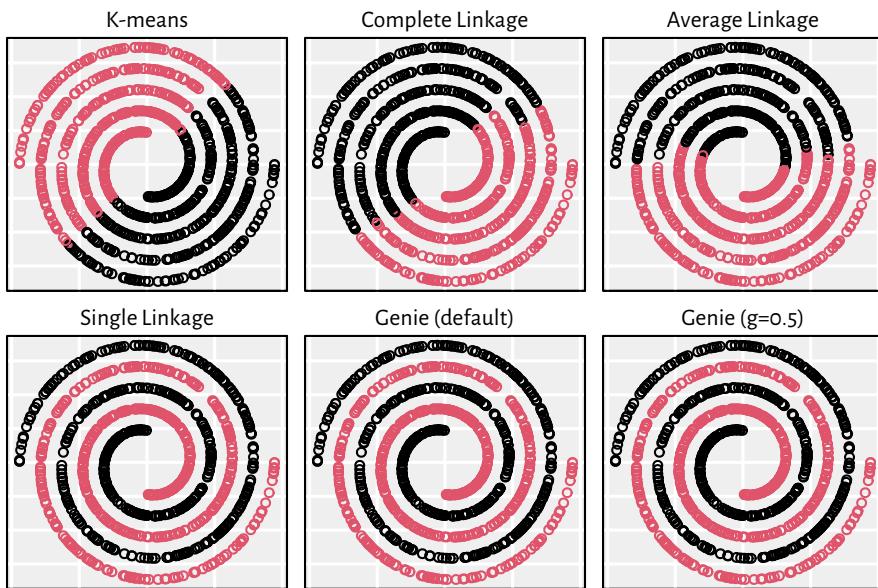


Figure 10.16: (#fig:clustering_benchmarks_plot2) Clustering of the `wut_mk2` dataset

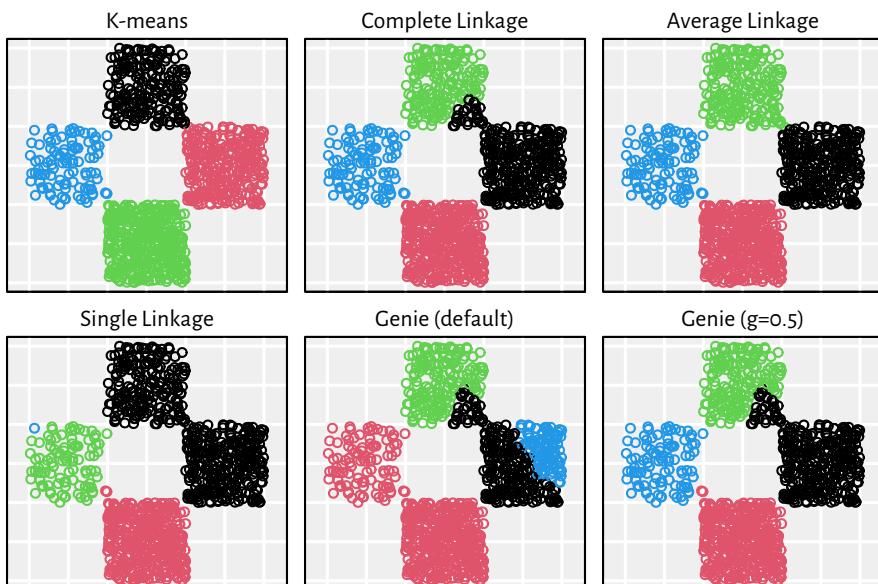


Figure 10.17: (#fig:clustering_benchmarks_plot3) Clustering of the `wut_z3` dataset

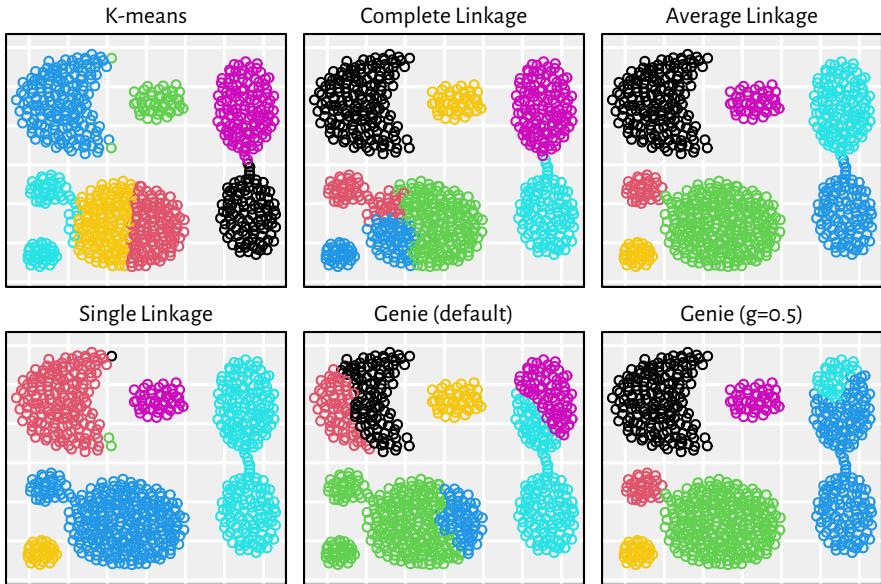


Figure 10.18: (#fig:clustering_benchmarks_plot4) Clustering of the `sipu_aggregation` dataset

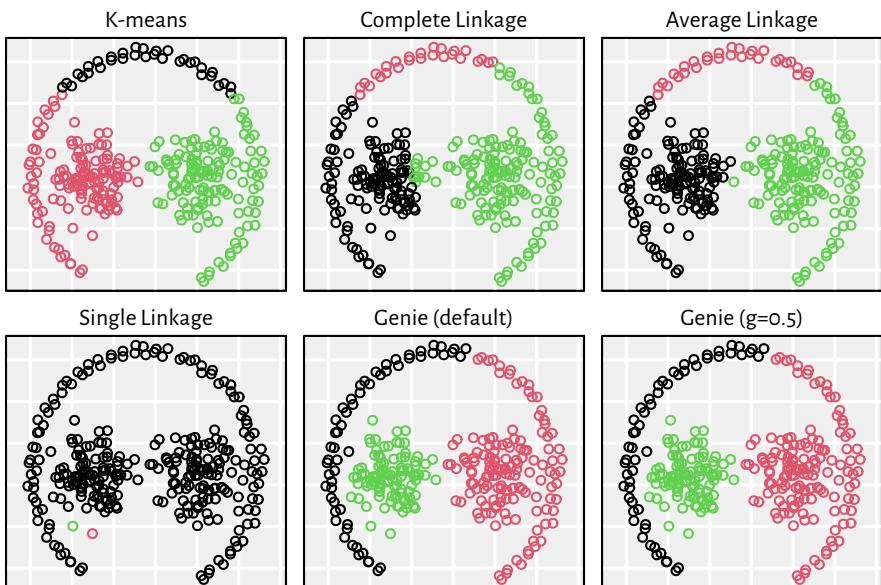


Figure 10.19: (#fig:clustering_benchmarks_plot5) Clustering of the `sipu_pathbased` dataset

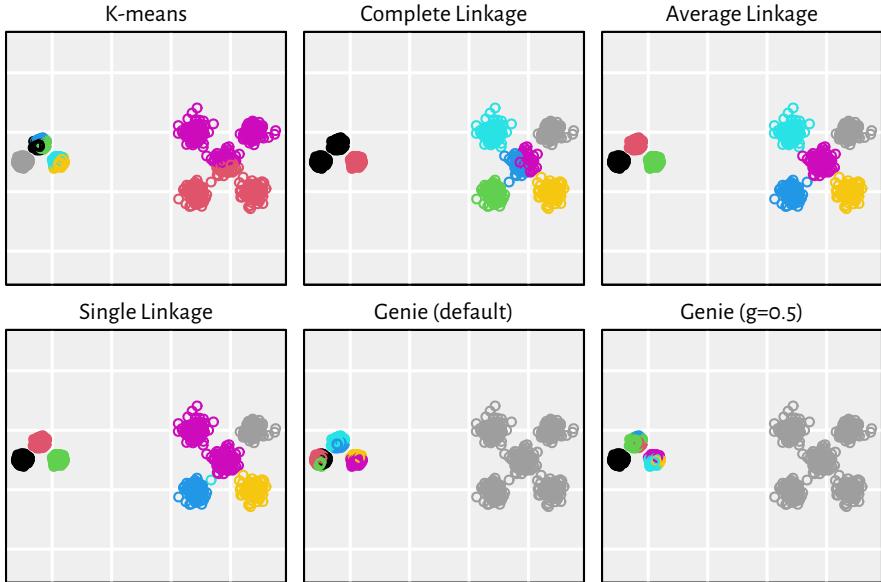


Figure 10.20: (#fig:clustering_benchmarks_plot6) Clustering of the `sipu_unbalance` dataset

10.3.4 Wine Quality – `volatile.acidity` and `sulphates`

Let's consider the Wine Quality dataset:

```
wine_quality <- read.csv("datasets/wine_quality_all.csv",
comment.char="#")
```

Exercise 10.10 Apply the 2-means clustering algorithm (i.e., K-means with $K = 2$) on a subset of `wine_quality` consisting only of the `volatile.acidity` and `sulphates` columns.

Exercise 10.11 Print the contingency table (see the `table()` function) for the discovered clusters vs. the wine colour (see the `color` column in `wine_quality`).

Exercise 10.12 Draw two scatter plots of `volatile.acidity` vs. `sulphates`:

- one where points' colours correspond to the discovered cluster labels (1st cluster = black symbols, 2nd cluster = red symbols),
- the other one such that points' colours correspond to the true wine colours (white wines = black symbols, red wines = red symbols).

Is there a good match between the discovered clusters and the true wine colours? Discuss.

10.3.5 Wine Quality – `chlorides` and `total.sulfur.dioxide`

The Wine Quality dataset again:

```
wine_quality <- read.csv("datasets/wine_quality_all.csv",
comment.char="#")
```

Exercise 10.13 Create a matrix X by extracting the `chlorides` and `total.sulfur.dioxide` columns from `wines`.

Exercise 10.14 Standardise both columns in X (i.e., from each column, subtract its mean and then divide by its standard deviation).

Exercise 10.15 Apply the 2-means clustering algorithm (i.e., K-means with $K = 2$) on X . Store the obtained cluster labels in a vector named `y_kmeans`.

Exercise 10.16 Apply the average linkage hierarchical clustering method. Cut the obtained hierarchy (see the `cutree()` function) so as to obtain 2 clusters and store the results in a vector named `y_average`.

Exercise 10.17 Apply the single linkage hierarchical clustering method. Cut the obtained hierarchy into two groups and store the results in a vector named `y_single`.

Exercise 10.18 Apply the complete linkage hierarchical clustering method. Cut the obtained hierarchy (see the `cutree()` function) into two groups and store the results in a vector named `y_complete`.

Exercise 10.19 Print the 4 contingency tables (see the `table()` function) for each of the 4 discovered partitions (`y_kmeans`, ..., `y_complete`) vs. the wine colour (see the `color` column in `wines`). Is there a good match between the discovered clusters and the true wine colours? Discuss.

10.4 Outro

10.4.1 Remarks

In K-means, we need to specify the number of clusters, K , in advance. What if we don't have any idea how to choose this parameter (which is often the case)?

Also, the problem with K-means is that there is no guarantee that a K -partition is any "similar" to the K' -one for $K \neq K'$, see Figure @ref(fig:kmeans_different_K).

```
X <- as.matrix(iris[,c(3,2)])
# never forget to set nstart>>1!
km <- kmeans(X, centers=3, nstart=10)
km$cluster # labels assigned to each of 150 points:

## [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [35] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
## [69] 3 3 3 3 3 3 3 3 2 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
## [ reached getOption("max.print") -- omitted 51 entries ]
```

```
km1 <- kmeans(X, 3, nstart=10)
km2 <- kmeans(X, 4, nstart=10)
plot(X, col=km1$cluster, pch=km2$cluster, asp=1)
```

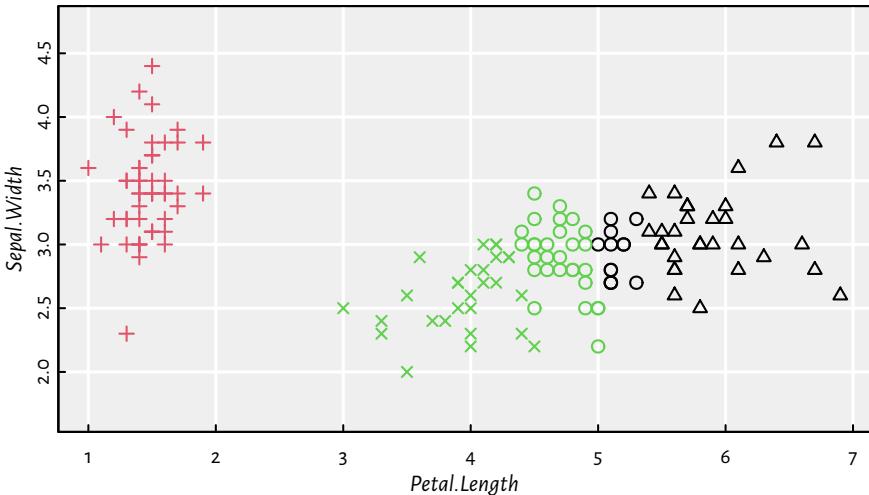


Figure 10.21: (#fig:kmeans_different_K) 3-means (colours) vs. 4-means (symbols) on example data; the “circle” cluster cannot decide if it likes the green or the black one more

Unsupervised learning is often performed during the data pre-processing and exploration stage. Assessing the quality of clustering is particularly challenging as, unlike in a supervised setting, we have no access to “ground truth” information.

In practice, we often apply different clustering algorithms and just see where they lead us. There’s no teacher that would tell us what we should do, so whatever we do is awesome, right? Well, not precisely. Most frequently, you, my dear reader, will work for some party that’s genuinely interested in your explaining why did you spent the last month coming up with nothing useful at all. Thus, the main body of work related to proving the usefull/less-ness will be on you.

Clustering methods can aid us in supervised tasks – instead of fitting a single “large model”, it might be useful to fit separate models to each cluster.

To sum up, the aim of K-means is to find K clusters based on the notion of the points’ closeness to the cluster centres. Remember that K must be set in advance. By definition (* via its relation to Voronoi diagrams), all clusters will be of convex shapes.

However, we may try applying K' -means for $K' \gg K$ to obtain a “fine grained” compressed representation of data and then combine the (sub)clusters into more meaningful groups using other methods (such as the hierarchical ones).

Iterative K-means algorithms are very fast (e.g., a mini-batch version of the algorithm can be implemented to speed up the optimisation process) even for large data sets, but they may fail to find a desirable solution, especially if clusters are unbalanced.

Hierarchical methods, on the other hand, output a whole family of mutually nested partitions, which may provide us with insight into the underlying structure of data data. Unfortunately, there is no easy way to assign new points to existing clusters; yet, you can always build a classifier (e.g., a decision tree or a neural network) that learns the discovered labels.

A linkage scheme must be chosen with care, for instance, single linkage can be sensitive to outliers. However, it is generally the fastest. The methods implemented in `hclust()` are generally slow; they have time complexity between $O(n^2)$ and $O(n^3)$.

Remark. Note that the `fastcluster` package provides a more efficient and memory-saving implementation of some methods available via a call to `hclust()`. See also the `genieclust` package for a super-robust version of the single linkage algorithm based on the datasets's Euclidean minimum spanning tree, which can be computed quite quickly.

Finally, note that all the discussed clustering methods are based on the notion of pairwise distances. These of course tend to behave weirdly in high-dimensional spaces (“the curse of dimensionality”). Moreover, some hardcore feature engineering might be needed to obtain meaningful results.

TODO

Recommended further reading: (James et al. 2017: Section 10.3)

Other: (Hastie et al. 2017: Section 14.3)

Additionally, check out other noteworthy clustering approaches:

- Genie (see R package `genieclust`) (Gagolewski et al. 2016)
- ITM (Müller et al. 2012)
- DBSCAN, HDBSCAN* (Ling 1973, Ester et al. 1996, Campello et al. 2015)
- K-medoids, K-medians
- Fuzzy C-means (a.k.a. weighted K-means) (Bezdek et al. 1984)
- Spectral clustering; e.g., (Ng et al. 2001)
- BIRCH (Zhang et al. 1996)

Next chapter...

11

Discrete Optimisation

TODO In this chapter, we will:

- ...
 - ...
-

11.1 Introduction

11.1.1 Recap

Recall that an **optimisation task** deals with finding an element \mathbf{x} in a **search space** \mathbb{D} , that minimises or maximises an **objective function** $f : \mathbb{D} \rightarrow \mathbb{R}$:

$$\min_{\mathbf{x} \in \mathbb{D}} f(\mathbf{x}) \quad \text{or} \quad \max_{\mathbf{x} \in \mathbb{D}} f(\mathbf{x}),$$

In one of the previous chapters, we were dealing with **unconstrained continuous optimisation**, i.e., we assumed the search space is $\mathbb{D} = \mathbb{R}^p$ for some p .

Example problems of this kind: minimising mean squared error in linear regression or minimising cross-entropy in logistic regression.

The class of general-purpose iterative algorithms we've previously studied fit into the following scheme:

1. $\mathbf{x}^{(0)}$ – initial guess (e.g., generated at random)
2. for $i = 1, \dots, M$:
 - a. $\mathbf{x}^{(i)} = \mathbf{x}^{(i-1)} + [\text{guessed direction, e.g., } -\eta \nabla f(\mathbf{x})]$
 - b. if $|f(\mathbf{x}^{(i)}) - f(\mathbf{x}^{(i-1)})| < \varepsilon$ break
3. return $\mathbf{x}^{(i)}$ as result

where:

- M = maximum number of iterations
- ε = tolerance, e.g., 10^{-8}
- $\eta > 0$ = learning rate

The algorithms such as gradient descent and BFGS (see `optim()`) give satisfactory results in the case of **smooth and well-behaving objective functions**.

However, if an objective has, e.g., many plateaus (regions where it is almost constant), those methods might easily get stuck in local minima.

The K-means clustering's objective function is a not particularly pleasant one – it involves a nested search for the closest cluster, with the use of the `min` operator.

11.2 Outro

11.2.1 Remarks

For any $p \geq 1$, the search space type determines the problem class:

- $\mathbb{D} \subseteq \mathbb{R}^p$ – **continuous optimisation**

In particular:

- $\mathbb{D} = \mathbb{R}^p$ – continuous unconstrained
- $\mathbb{D} = [a_1, b_1] \times \dots \times [a_n, b_n]$ – continuous with box constraints
- constrained with k linear inequality constraints

$$\begin{cases} a_{1,1}x_1 + \dots + a_{1,p}x_p \leq b_1 \\ \vdots \\ a_{k,1}x_1 + \dots + a_{k,p}x_p \leq b_k \end{cases}$$

However, there are other possibilities as well:

- $\mathbb{D} \subseteq \mathbb{Z}^p$ (\mathbb{Z} – the set of integers) – **discrete optimisation**

In particular:

- $\mathbb{D} = \{0, 1\}^p$ – 0-1 optimisation (hard!)
- \mathbb{D} is finite (but perhaps large, its objects can be enumerated) – **combination optimisation**

For example:

- \mathbb{D} = all possible routes between two points on a map.

These optimisation tasks tend to be much harder than the continuous ones.

Genetic algorithms might come in handy in such cases.

Specialised methods, customised to solve a specific problem (like Lloyd's algorithm) will often outperform generic ones (like SGD, genetic algorithms) in terms of speed and reliability.

All in all, we prefer a suboptimal solution obtained by means of heuristics to no solution at all.

Problems that you could try solving with GAs include variable selection in multiple regression – finding the subset of features optimising the AIC (this is a hard problem to and forward selection was just a simple greed heuristic).

TODO

further reading ...

next chapter ...

12

Feature Selection

TODO In this chapter, we will:

- ...
 - ...
-

12.1 Introduction

12.1.1 Recap

Recall that an **optimisation task** deals with finding an element \mathbf{x} in a **search space** \mathbb{D} , that minimises or maximises an **objective function** $f : \mathbb{D} \rightarrow \mathbb{R}$:

$$\min_{\mathbf{x} \in \mathbb{D}} f(\mathbf{x}) \quad \text{or} \quad \max_{\mathbf{x} \in \mathbb{D}} f(\mathbf{x}),$$

In one of the previous chapters, we were dealing with **unconstrained continuous optimisation**, i.e., we assumed the search space is $\mathbb{D} = \mathbb{R}^p$ for some p .

Example problems of this kind: minimising mean squared error in linear regression or minimising cross-entropy in logistic regression.

The class of general-purpose iterative algorithms we've previously studied fit into the following scheme:

1. $\mathbf{x}^{(0)}$ – initial guess (e.g., generated at random)
2. for $i = 1, \dots, M$:
 - a. $\mathbf{x}^{(i)} = \mathbf{x}^{(i-1)} + [\text{guessed direction, e.g., } -\eta \nabla f(\mathbf{x})]$
 - b. if $|f(\mathbf{x}^{(i)}) - f(\mathbf{x}^{(i-1)})| < \varepsilon$ break
3. return $\mathbf{x}^{(i)}$ as result

where:

- M = maximum number of iterations
- ε = tolerance, e.g., 10^{-8}
- $\eta > 0$ = learning rate

The algorithms such as gradient descent and BFGS (see `optim()`) give satisfactory results in the case of **smooth and well-behaving objective functions**.

However, if an objective has, e.g., many plateaus (regions where it is almost constant), those methods might easily get stuck in local minima.

The K-means clustering's objective function is a not particularly pleasant one – it involves a nested search for the closest cluster, with the use of the `min` operator.

12.2 Outro

12.2.1 Remarks

For any $p \geq 1$, the search space type determines the problem class:

- $\mathbb{D} \subseteq \mathbb{R}^p$ – **continuous optimisation**

In particular:

- $\mathbb{D} = \mathbb{R}^p$ – continuous unconstrained
- $\mathbb{D} = [a_1, b_1] \times \dots \times [a_n, b_n]$ – continuous with box constraints
- constrained with k linear inequality constraints

$$\begin{cases} a_{1,1}x_1 + \dots + a_{1,p}x_p \leq b_1 \\ \vdots \\ a_{k,1}x_1 + \dots + a_{k,p}x_p \leq b_k \end{cases}$$

However, there are other possibilities as well:

- $\mathbb{D} \subseteq \mathbb{Z}^p$ (\mathbb{Z} – the set of integers) – **discrete optimisation**

In particular:

- $\mathbb{D} = \{0, 1\}^p$ – 0-1 optimisation (hard!)
- \mathbb{D} is finite (but perhaps large, its objects can be enumerated) – **combination optimisation**

For example:

- \mathbb{D} = all possible routes between two points on a map.

These optimisation tasks tend to be much harder than the continuous ones.

Genetic algorithms might come in handy in such cases.

Specialised methods, customised to solve a specific problem (like Lloyd's algorithm) will often outperform generic ones (like SGD, genetic algorithms) in terms of speed and reliability.

All in all, we prefer a suboptimal solution obtained by means of heuristics to no solution at all.

Problems that you could try solving with GAs include variable selection in multiple regression – finding the subset of features optimising the AIC (this is a hard problem to and forward selection was just a simple greed heuristic).

TODO

further reading ...

next chapter ...

13

Shallow and Deep Neural Networks

TODO In this chapter, we will:

- ...
 - ...
-

13.1 Introduction

13.1.1 Binary Logistic Regression: Recap

Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be an input matrix that consists of n points in a p -dimensional space.

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{bmatrix}$$

In other words, we have a database on n objects. Each object is described by means of p numerical features.

With each input \mathbf{x}_i , we associate the desired output y_i which is a categorical label – hence we will be dealing with **classification** tasks again.

To recall, in **binary logistic regression** we model the probabilities that a given input belongs to either of the two classes:

$$\begin{aligned} \Pr(Y = 1|\mathbf{X}, \boldsymbol{\beta}) &= \phi(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p) \\ \Pr(Y = 0|\mathbf{X}, \boldsymbol{\beta}) &= 1 - \phi(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p) \end{aligned}$$

where $\phi(t) = \frac{1}{1+e^{-t}} = \frac{e^t}{1+e^t}$ is the logistic sigmoid function.

It holds:

$$\begin{aligned} \Pr(Y = 1|\mathbf{X}, \boldsymbol{\beta}) &= \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}}, \\ \Pr(Y = 0|\mathbf{X}, \boldsymbol{\beta}) &= \frac{1}{1 + e^{+(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}} = \frac{e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}}. \end{aligned}$$

The fitting of the model was performed by minimising the cross-entropy (log-loss):

$$\min_{\beta \in \mathbb{R}^{p+1}} -\frac{1}{n} \sum_{i=1}^n \left(y_i \log \hat{y}_i + (1 - y_i) \log (1 - \hat{y}_i) \right).$$

where $\hat{y}_i = \Pr(Y = 1 | \mathbf{x}_{i,.}, \beta)$.

This is equivalent to:

$$\min_{\beta \in \mathbb{R}^{p+1}} -\frac{1}{n} \sum_{i=1}^n \left(y_i \log \Pr(Y = 1 | \mathbf{x}_{i,.}, \beta) + (1 - y_i) \log \Pr(Y = 0 | \mathbf{x}_{i,.}, \beta) \right).$$

Note that for each i , either the left or the right term (in the bracketed expression) vanishes.

Hence, we may also write the above as:

$$\min_{\beta \in \mathbb{R}^{p+1}} -\frac{1}{n} \sum_{i=1}^n \log \Pr(Y = y_i | \mathbf{x}_{i,.}, \beta).$$

In this chapter we will generalise the binary logistic regression model:

- First we will consider the case of many classes (multiclass classification). This will lead to the multinomial logistic regression model.
- Then we will note that the multinomial logistic regression is a special case of a feed-forward neural network.

13.1.2 Data

We will study the famous classic – the MNIST image classification dataset (Modified National Institute of Standards and Technology database), see <http://yann.lecun.com/exdb/mnist/>

It consists of 28×28 pixel images of handwritten digits:

- `train`: 60,000 training images,
- `t10k`: 10,000 testing images.

A few image instances from each class are depicted in Figure @ref(fig:mnist_demo).

There are 10 unique digits, so this is a multiclass classification problem.

Remark. The dataset is already “too easy” for testing of the state-of-the-art classifiers (see the notes below), but it’s a great educational example.

Accessing MNIST via the `keras` package (which we will use throughout this chapter anyway) is easy:



Figure 13.1: (#fig:mnist_demo) Example images in the MNIST database

```
library("keras")
mnist <- dataset_mnist()
X_train <- mnist$train$x
Y_train <- mnist$train$y
X_test <- mnist$test$x
Y_test <- mnist$test$y
```

X_train and X_test consist of 28×28 pixel images.

```
dim(X_train)

## [1] 60000    28    28

dim(X_test)

## [1] 10000    28    28
```

X_train and X_test are 3-dimensional arrays, think of them as vectors of 60000 and 10000 matrices of size 28×28 , respectively.

These are grey-scale images, with 0 = black, ..., 255 = white:

```
range(X_train)

## [1]  0 255
```

Numerically, it's more convenient to work with colour values converted to 0.0 = black, ..., 1.0 = white:

```
X_train <- X_train/255
X_test <- X_test/255
```

`Y_train` and `Y_test` are the corresponding integer labels:

```
length(Y_train)
## [1] 60000
length(Y_test)
## [1] 10000
table(Y_train) # label distribution in the training sample

## Y_train
##   0   1   2   3   4   5   6   7   8   9
## 5923 6742 5958 6131 5842 5421 5918 6265 5851 5949
table(Y_test) # label distribution in the test sample

## Y_test
##   0   1   2   3   4   5   6   7   8   9
## 980 1135 1032 1010 982 892 958 1028 974 1009
```

Here is how we can plot one of the digits (see Figure @ref(fig:mnist_info2b)):

```
id <- 123 # image ID to show
image(z=t(X_train[id,,]), col=grey.colors(256, 0, 1),
      axes=FALSE, asp=1, ylim=c(1, 0))
legend("topleft", bg="white",
       legend=sprintf("True label=%d", Y_train[id]))
```

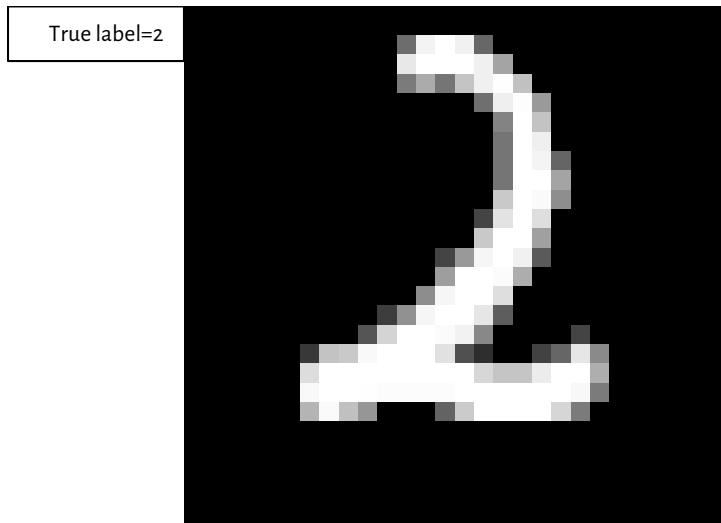


Figure 13.2: (#fig:mnist_info2b) Example image from the MNIST dataset

13.2 Multinomial Logistic Regression

13.2.1 A Note on Data Representation

So... you may now be wondering “how do we construct an image classifier, this seems so complicated!”.

For a computer, (almost) everything is just numbers.

Instead of playing with n matrices, each of size 28×28 , we may “flatten” the images so as to get n “long” vectors of length $p = 784$.

```
X_train2 <- matrix(X_train, ncol=28*28)
X_test2 <- matrix(X_test, ncol=28*28)
```

The classifiers studied here do not take the “spatial” positioning of the pixels into account anyway. Hence, now we’re back to our “comfort zone”.

Remark. (*) See, however, convolutional neural networks (CNNs), e.g., in (Goodfellow et al. 2016).

13.2.2 Extending Logistic Regression

Let us generalise the binary logistic regression model to a 10-class one (or, more generally, K -class one).

This time we will be modelling ten probabilities, with $\Pr(Y = k|\mathbf{X}, \mathbf{B})$ denoting the *confidence* that a given image \mathbf{X} is in fact the k -th digit:

$$\begin{aligned}\Pr(Y = 0|\mathbf{X}, \mathbf{B}) &= \dots \\ \Pr(Y = 1|\mathbf{X}, \mathbf{B}) &= \dots \\ &\vdots \\ \Pr(Y = 9|\mathbf{X}, \mathbf{B}) &= \dots\end{aligned}$$

where \mathbf{B} is the set of underlying model parameters (to be determined soon).

In binary logistic regression, the class probabilities are obtained by “cleverly normalising” (by means of the logistic sigmoid) the outputs of a linear model (so that we obtain a value in $[0, 1]$).

$$\Pr(Y = 1|\mathbf{X}, \boldsymbol{\beta}) = \phi(\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p)}}$$

In the multinomial case, we can use a separate linear model for each digit so that each $\Pr(Y = k|\mathbf{X}, \mathbf{B})$, $k = 0, 1, \dots, 9$, is given as a function of:

$$\beta_{0,k} + \beta_{1,k} X_1 + \dots + \beta_{p,k} X_p.$$

Therefore, instead of a parameter vector of length $(p + 1)$, we will need a parameter matrix of size $(p + 1) \times 10$ representing the model's definition.

Side note. The upper case of β is B .

Then, these 10 numbers will have to be normalised so as to they are all greater than 0 and sum to 1.

To maintain the spirit of the original model, we can apply $e^{-(\beta_{0,k} + \beta_{1,k}X_1 + \dots + \beta_{p,k}X_p)}$ to get a positive value, because the co-domain of the exponential function $t \mapsto e^t$ is $(0, \infty)$.

Then, dividing each output by the sum of all the outputs will guarantee that the total sum equals 1.

This leads to:

$$\begin{aligned}\Pr(Y = 0|\mathbf{X}, \mathbf{B}) &= \frac{e^{-(\beta_{0,0} + \beta_{1,0}X_1 + \dots + \beta_{p,0}X_p)}}{\sum_{k=0}^9 e^{-(\beta_{0,k} + \beta_{1,k}X_1 + \dots + \beta_{p,k}X_p)}}, \\ \Pr(Y = 1|\mathbf{X}, \mathbf{B}) &= \frac{e^{-(\beta_{0,1} + \beta_{1,1}X_1 + \dots + \beta_{p,1}X_p)}}{\sum_{k=0}^9 e^{-(\beta_{0,k} + \beta_{1,k}X_1 + \dots + \beta_{p,k}X_p)}}, \\ &\vdots \\ \Pr(Y = 9|\mathbf{X}, \mathbf{B}) &= \frac{e^{-(\beta_{0,9} + \beta_{1,9}X_1 + \dots + \beta_{p,9}X_p)}}{\sum_{k=0}^9 e^{-(\beta_{0,k} + \beta_{1,k}X_1 + \dots + \beta_{p,k}X_p)}}.\end{aligned}$$

This reduces to the binary logistic regression if we consider only the classes 0 and 1 and fix $\beta_{0,0} = \beta_{1,0} = \dots = \beta_{p,0} = 0$ (as $e^0 = 1$).

13.2.3 Softmax Function

The above transformation (that maps 10 arbitrary real numbers to positive ones that sum to 1) is called the **softmax** function (or *softargmax*).

```
softmax <- function(T) {
  T2 <- exp(T) # ignore the minus sign above
  T2/sum(T2)
}
round(rbind(
  softmax(c(0, 0, 10, 0, 0, 0, 0, 0, 0, 0)),
  softmax(c(0, 0, 10, 0, 0, 0, 10, 0, 0, 0)),
  softmax(c(0, 0, 10, 0, 0, 0, 9, 0, 0, 0)),
  softmax(c(0, 0, 10, 0, 0, 0, 9, 0, 0, 8))), 2)

##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## [1,]     0    0 1.00    0    0    0 0.00    0    0 0.00
## [2,]     0    0 0.50    0    0    0 0.50    0    0 0.00
## [3,]     0    0 0.73    0    0    0 0.27    0    0 0.00
## [4,]     0    0 0.67    0    0    0 0.24    0    0 0.09
```

13.2.4 One-Hot Encoding and Decoding

The ten class-belongingness-degrees can be decoded to obtain a single label by simply choosing the class that is assigned the highest probability.

```
y_pred <- softmax(c(0, 0, 10, 0, 0, 0, 9, 0, 0, 8))
round(y_pred, 2) # probabilities of Y=0, 1, 2, ..., 9

## [1] 0.00 0.00 0.67 0.00 0.00 0.00 0.24 0.00 0.00 0.09

which.max(y_pred)-1 # 1..10 -> 0..9

## [1] 2
```

Remark. `which.max(y)` returns an index k such that $y[k] == \max(y)$ (recall that in R the first element in a vector is at index 1). Mathematically, we denote this operation as $\arg \max_{k=1,\dots,K} y_k$.

To make processing the outputs of a logistic regression model more convenient, we will apply the so-called **one-hot-encoding** of the labels.

Here, each label will be represented as a 0-1 vector of 10 probabilities – with probability 1 corresponding to the true class only.

For instance:

```
y <- 2 # true class (this is just an example)
y2 <- rep(0, 10)
y2[y+1] <- 1 # +1 because we need 0..9 -> 1..10
y2 # one-hot-encoded y

## [1] 0 0 1 0 0 0 0 0 0 0
```

To one-hot encode *all* the reference outputs in R, we start with a matrix of size $n \times 10$ populated with “0”s:

```
Y_train2 <- matrix(0, nrow=length(Y_train), ncol=10)
```

Next, for every i , we insert a “1” in the i -th row and the $(Y_{\text{train}}[i]+1)$ -th column:

```
# Note the "+1" 0..9 -> 1..10
Y_train2[cbind(1:length(Y_train), Y_train+1)] <- 1
```

Remark. In R, indexing a matrix A with a 2-column matrix B, i.e., $A[B]$, allows for an easy access to $A[B[1,1], B[1,2]], A[B[2,1], B[2,2]], A[B[3,1], B[3,2]], \dots$

Sanity check:

```
head(Y_train)

## [1] 5 0 4 1 9 2

head(Y_train2)
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## [1,]    0    0    0    0    0    1    0    0    0    0
## [2,]    1    0    0    0    0    0    0    0    0    0
## [3,]    0    0    0    0    1    0    0    0    0    0
## [4,]    0    1    0    0    0    0    0    0    0    0
## [5,]    0    0    0    0    0    0    0    0    0    1
## [6,]    0    0    1    0    0    0    0    0    0    0
```

Let us generalise the above idea and write a function that can one-hot-encode any vector of integer labels:

```
one_hot_encode <- function(Y) {
  stopifnot(is.numeric(Y))
  c1 <- min(Y) # first class label
  cK <- max(Y) # last class label
  K <- cK-c1+1 # number of classes

  Y2 <- matrix(0, nrow=length(Y), ncol=K)
  Y2[cbind(1:length(Y), Y-c1+1)] <- 1
  Y2
}
```

Encode Y_{train} and Y_{test} :

```
Y_train2 <- one_hot_encode(Y_train)
Y_test2 <- one_hot_encode(Y_test)
```

13.2.5 Cross-entropy Revisited

Our classifier will be outputting $K = 10$ probabilities.

The true class labels are not one-hot-encoded so that they are represented as vectors of $K - 1$ zeros and a single one.

How to measure the “agreement” between these two?

In essence, we will be comparing the probability vectors as generated by a classifier, \hat{Y} :

```
round(y_pred, 2)
```

```
## [1] 0.00 0.00 0.67 0.00 0.00 0.00 0.24 0.00 0.00 0.09
```

with the one-hot-encoded true probabilities, Y :

```
y2
```

```
## [1] 0 0 1 0 0 0 0 0 0 0
```

It turns out that one of the definitions of cross-entropy introduced above already handles

the case of multiclass classification:

$$E(\mathbf{B}) = -\frac{1}{n} \sum_{i=1}^n \log \Pr(Y = y_i | \mathbf{x}_{i,\cdot}, \mathbf{B}).$$

The smaller the probability corresponding to the ground-truth class outputted by the classifier, the higher the penalty, see Figure @ref(fig:cross_entropy_revisited_example3).

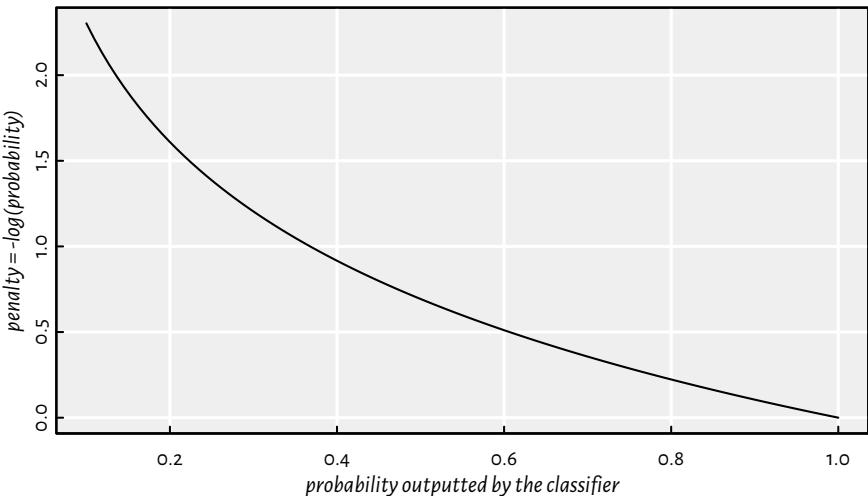


Figure 13.3: (#fig:cross_entropy_revisited_example3) The less the classifier is confident about the prediction of the actually true label, the greater the penalty

To sum up, we will be solving the optimisation problem:

$$\min_{\mathbf{B} \in \mathbb{R}^{(p+1) \times 10}} -\frac{1}{n} \sum_{i=1}^n \log \Pr(Y = y_i | \mathbf{x}_{i,\cdot}, \mathbf{B}).$$

This has no analytical solution, but can be solved using iterative methods (see the chapter on optimisation).

(*) Side note: A single term in the above formula,

$$\log \Pr(Y = y_i | \mathbf{x}_{i,\cdot}, \mathbf{B})$$

given:

- y_pred – a vector of 10 probabilities generated by the model:

$$[\Pr(Y = 0 | \mathbf{x}_{i,\cdot}, \mathbf{B}) \ \Pr(Y = 1 | \mathbf{x}_{i,\cdot}, \mathbf{B}) \ \dots \ \Pr(Y = 9 | \mathbf{x}_{i,\cdot}, \mathbf{B})]$$

- $y2$ – a one-hot-encoded version of the true label, y_i , of the form:

$$[0 \ 0 \ \dots \ 0 \ 1 \ 0 \ \dots \ 0]$$

can be computed as:

```
sum(y2*log(y_pred))
## [1] -0.40782
```

13.2.6 Problem Formulation in Matrix Form (**)

The definition of a multinomial logistic regression model for a multiclass classification task involving classes $\{1, 2, \dots, K\}$ is slightly bloated.

Assuming that $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the input matrix, to compute the K predicted probabilities for the i -th input,

$$[\hat{y}_{i,1} \ \hat{y}_{i,2} \ \dots \ \hat{y}_{i,K}],$$

given a parameter matrix $\mathbf{B}^{(p+1) \times K}$, we apply:

$$\begin{aligned}\hat{y}_{i,1} = \Pr(Y = 1 | \mathbf{x}_{i,\cdot}, \mathbf{B}) &= \frac{e^{\beta_{0,1} + \beta_{1,1}x_{i,1} + \dots + \beta_{p,1}x_{i,p}}}{\sum_{k=1}^K e^{\beta_{0,k} + \beta_{1,k}x_{i,1} + \dots + \beta_{p,k}x_{i,p}}}, \\ &\vdots \\ \hat{y}_{i,K} = \Pr(Y = K | \mathbf{x}_{i,\cdot}, \mathbf{B}) &= \frac{e^{\beta_{0,K} + \beta_{1,K}x_{i,1} + \dots + \beta_{p,K}x_{i,p}}}{\sum_{k=1}^K e^{\beta_{0,k} + \beta_{1,k}x_{i,1} + \dots + \beta_{p,k}x_{i,p}}}.\end{aligned}$$

Remark. We have dropped the minus sign in the exponentiation for brevity of notation.

Note that we can always map $b'_{j,k} = -b_{j,k}$.

It turns out we can make use of matrix notation to tidy the above formulas.

Denote the linear combinations prior to computing the softmax function with:

$$\begin{aligned}t_{i,1} &= \beta_{0,1} + \beta_{1,1}x_{i,1} + \dots + \beta_{p,1}x_{i,p}, \\ &\vdots \\ t_{i,K} &= \beta_{0,K} + \beta_{1,K}x_{i,1} + \dots + \beta_{p,K}x_{i,p}.\end{aligned}$$

We have:

- $x_{i,j}$ – the i -th observation, the j -th feature;
- $\hat{y}_{i,k}$ – the i -th observation, the k -th class probability;
- $\beta_{j,k}$ – the coefficient for the j -th feature when computing the k -th class.

Note that by augmenting $\dot{\mathbf{X}} = [1 \ \mathbf{X}] \in \mathbb{R}^{n \times (p+1)}$ by adding a column of 1s, i.e., where $\dot{x}_{i,0} = 1$ and $\dot{x}_{i,j} = x_{i,j}$ for all $j \geq 1$ and all i , we can write the above as:

$$\begin{aligned}t_{i,1} &= \sum_{j=0}^p \dot{x}_{i,j} \beta_{j,1} = \dot{\mathbf{x}}_{i,\cdot} \boldsymbol{\beta}_{\cdot,1}, \\ &\vdots \\ t_{i,K} &= \sum_{j=0}^p \dot{x}_{i,j} \beta_{j,K} = \dot{\mathbf{x}}_{i,\cdot} \boldsymbol{\beta}_{\cdot,K}.\end{aligned}$$

We can get the K linear combinations all at once in the form of a row vector by writing:

$$[t_{i,1} \ t_{i,2} \ \dots \ t_{i,K}] = \dot{\mathbf{x}}_{i,\cdot} \ \mathbf{B}.$$

Moreover, we can do that for all the n inputs by writing:

$$\mathbf{T} = \hat{\mathbf{X}} \mathbf{B}.$$

Yes yes yes! This is a single matrix multiplication, we have $\mathbf{T} \in \mathbb{R}^{n \times K}$.

To obtain $\hat{\mathbf{Y}}$, we have to apply the softmax function on every row of \mathbf{T} :

$$\hat{\mathbf{Y}} = \text{softmax}(\hat{\mathbf{X}} \mathbf{B}).$$

That's it. Take some time to appreciate the elegance of this notation.

Methods for minimising cross-entropy expressed in matrix form will be discussed in the next chapter.

13.3 Artificial Neural Networks

13.3.1 Artificial Neuron

A neuron can be thought of as a mathematical function, see Figure 13.4, which has its specific inputs and an output.

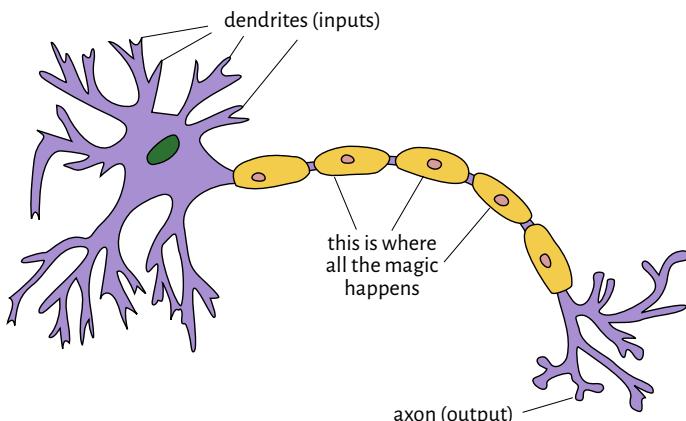


Figure 13.4: Neuron as a mathematical (black box) function; image based on: <https://en.wikipedia.org/wiki/File:Neuron3.png> by Egm4313.s12 at English Wikipedia, licensed under the Creative Commons Attribution-Share Alike 3.0 Unported license

The Linear Threshold Unit (McCulloch and Pitts, 1940s), the Perceptron (Rosenblatt, 1958) and the Adaptive Linear Neuron (Widrow and Hoff, 1960) were amongst the first models of an artificial neuron that could be used for the purpose of pattern recognition, see Figure 13.5. They can be thought of as processing units that compute a weighted sum of the inputs, which is then transformed by means of a nonlinear “activation” function.

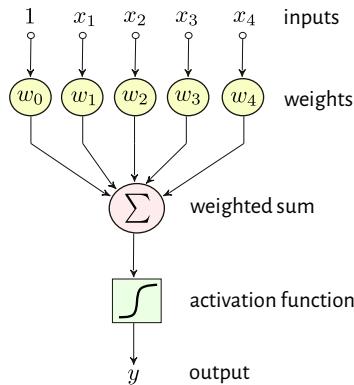


Figure 13.5: A simple model of an artificial neuron

13.3.2 Logistic Regression as a Neural Network

The above resembles our binary logistic regression model, where we determine a linear combination (a weighted sum) of p inputs and then transform it using the logistic sigmoid “activation” function. We can easily depict it in the Figure 13.4-style, see Figure @ref(fig:logistic_regression_binary).

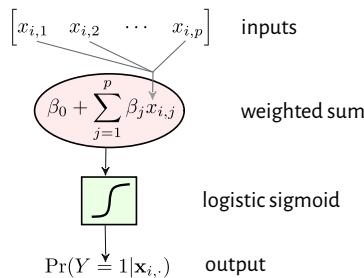


Figure 13.6: Binary logistic regression

On the other hand, a multiclass logistic regression can be depicted as in Figure @ref(fig:logistic_regression_multiclass). In fact, we can consider it as an instance of a:

- **single layer** (there is only one processing step that consists of 10 units),
- **densely connected** (all the inputs are connected to all the components below),
- **feed-forward** (the outputs are generated by processing the inputs from “top” to “bottom”, there are no loops in the graph etc.)

artificial neural network that uses the softmax as the activation function.

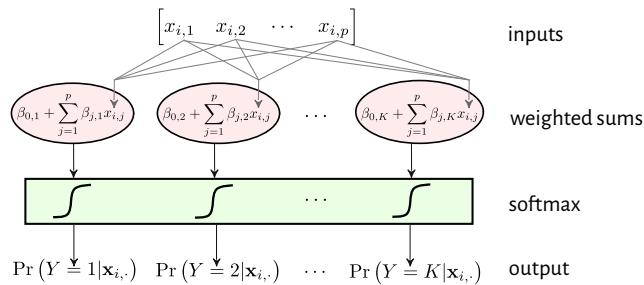


Figure 13.7: Multinomial logistic regression

13.3.3 Example in R

To train such a neural network (i.e., fit a multinomial logistic regression model), we will use the `keras` package, a wrapper around the (GPU-enabled) TensorFlow library.

The training of the model takes a few minutes (for more complex models and bigger datasets – it could take hours/days). Thus, it is wise to store the computed model (the **B** coefficient matrix and the accompanying `keras`'s auxiliary data) for further reference:

```
file_name <- "datasets/mnist_keras_model1.h5"
if (!file.exists(file_name)) { # File doesn't exist -> compute
  set.seed(123)
  # Start with an empty model
  model1 <- keras_model_sequential()
  # Add a single layer with 10 units and softmax activation
  layer_dense(model1, units=10, activation="softmax")
  # We will be minimising the cross-entropy,
  # sgd == stochastic gradient descent, see the next chapter
  compile(model1, optimizer="sgd",
          loss="categorical_crossentropy")
  # Fit the model (sloooooow!)
  fit(model1, X_train2, Y_train2, epochs=10)
  # Save the model for future reference
  save_model_hdf5(model1, file_name)
} else { # File exists -> reload the model
  model1 <- load_model_hdf5(file_name)
}
```

Let's make predictions over the test set:

```

Y_pred2 <- predict(model1, X_test2)
round(head(Y_pred2), 2) # predicted class probabilities

##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## [1,] 0.00 0.00 0.00 0.00 0.00 0.00 0.00 1.00 0.00 0.00
## [2,] 0.01 0.00 0.93 0.01 0.00 0.01 0.04 0.00 0.00 0.00
## [3,] 0.00 0.96 0.02 0.00 0.00 0.00 0.00 0.00 0.01 0.00
## [4,] 1.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
## [5,] 0.00 0.00 0.01 0.00 0.91 0.00 0.01 0.01 0.01 0.05
## [6,] 0.00 0.98 0.00 0.00 0.00 0.00 0.00 0.00 0.01 0.00

```

Then, we can one-hot-decode the output probabilities:

```

Y_pred <- apply(Y_pred2, 1, which.max)-1 # 1..10 -> 0..9
head(Y_pred, 20) # predicted outputs

```

```

## [1] 7 2 1 0 4 1 4 9 6 9 0 6 9 0 1 5 9 7 3 4
head(Y_test, 20) # true outputs

```

```

## [1] 7 2 1 0 4 1 4 9 5 9 0 6 9 0 1 5 9 7 3 4

```

Accuracy on the test set:

```
mean(Y_test == Y_pred)
```

```
## [1] 0.9169
```

Performance metrics for each digit separately (see also Figure 13.8):

i	Acc	Prec	Rec	F	TN	FN	FP	TP
0	0.9924	0.94664	0.97755	0.96185	8966	22	54	958
1	0.9923	0.95920	0.97357	0.96633	8818	30	47	1105
2	0.9803	0.92214	0.88372	0.90252	8891	120	77	912
3	0.9802	0.89417	0.91188	0.90294	8881	89	109	921
4	0.9833	0.90148	0.93177	0.91637	8918	67	100	915
5	0.9793	0.91415	0.84753	0.87958	9037	136	71	756
6	0.9885	0.93142	0.94990	0.94057	8975	48	67	910
7	0.9834	0.92843	0.90856	0.91839	8900	94	72	934
8	0.9754	0.86473	0.88604	0.87525	8891	111	135	863
9	0.9787	0.90040	0.88702	0.89366	8892	114	99	895

Note how misleading the individual accuracies are! Averaging over the above table's columns gives:

```

##   Acc    Prec    Rec     F
## 0.98338 0.91628 0.91575 0.91575

```

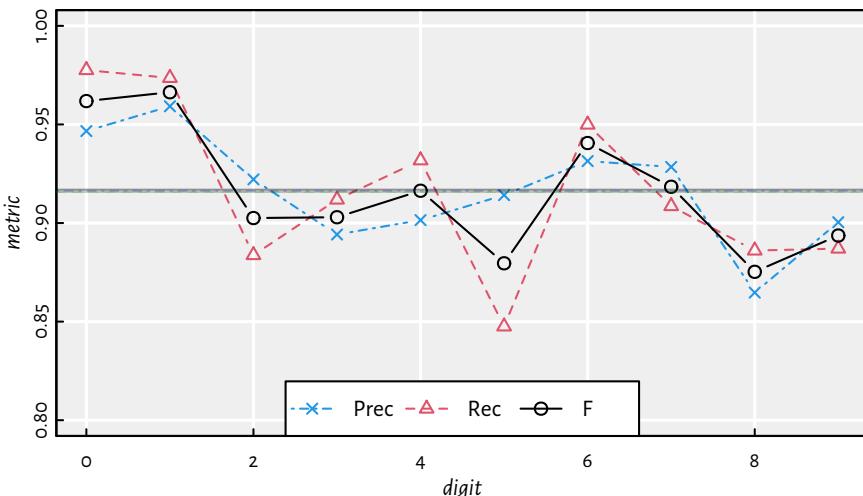


Figure 13.8: Performance metrics for multinomial logistic regression on MNIST

13.4 Deep Neural Networks

13.4.1 Introduction

In a brain, a neuron's output is an input to a bunch of other neurons. We could try aligning neurons into many interconnected layers. This leads to a structure like the one in Figure 13.9.

13.4.2 Activation Functions

Each layer's outputs should be transformed by some non-linear activation function. Otherwise, we'd end up with linear combinations of linear combinations, which are linear combinations themselves.

Example activation functions that can be used in hidden (inner) layers:

- `relu` – The rectified linear unit:

$$\psi(t) = \max(t, 0),$$

- `sigmoid` – The logistic sigmoid:

$$\phi(t) = \frac{1}{1 + \exp(-t)},$$

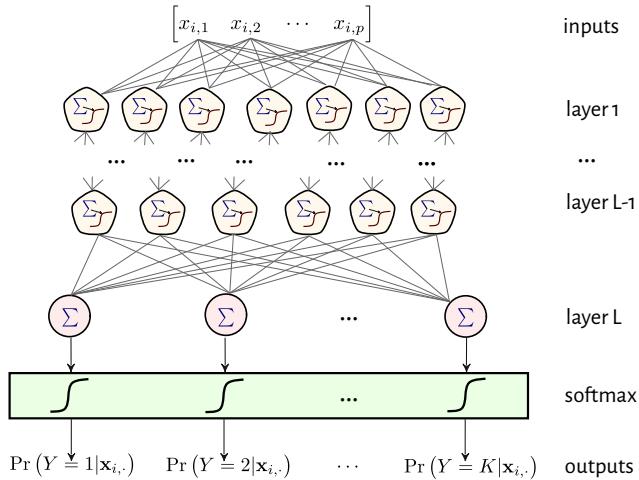


Figure 13.9: A multi-layer neural network

- \tanh – The hyperbolic function:

$$\tanh(t) = \frac{\exp(t) - \exp(-t)}{\exp(t) + \exp(-t)}.$$

There is not much difference between them, but some might be more convenient to handle numerically than the others, depending on the implementation.

13.4.3 Example in R - 2 Layers

Let's construct a 2-layer Neural Network of the type 784-800-10:

```
file_name <- "datasets/mnist_keras_model2.h5"
if (!file.exists(file_name)) {
  set.seed(123)
  model2 <- keras_model_sequential()
  layer_dense(model2, units=800, activation="relu")
  layer_dense(model2, units=10, activation="softmax")
  compile(model2, optimizer="sgd",
          loss="categorical_crossentropy")
  fit(model2, X_train2, Y_train2, epochs=10)
  save_model_hdf5(model2, file_name)
} else {
  model2 <- load_model_hdf5(file_name)
```

```

}

Y_pred2 <- predict(model2, X_test2)
Y_pred <- apply(Y_pred2, 1, which.max)-1 # 1..10 -> 0..9
mean(Y_test == Y_pred) # accuracy on the test set

## [1] 0.9583

```

Performance metrics for each digit separately, see also Figure 13.10:

i	Acc	Prec	Rec	F	TN	FN	FP	TP
0	0.9948	0.96215	0.98571	0.97379	8982	14	38	966
1	0.9962	0.98156	0.98502	0.98329	8844	17	21	1118
2	0.9911	0.96000	0.95349	0.95673	8927	48	41	984
3	0.9898	0.94773	0.95149	0.94960	8937	49	53	961
4	0.9919	0.95829	0.95927	0.95878	8977	40	41	942
5	0.9911	0.95470	0.94507	0.94986	9068	49	40	843
6	0.9920	0.94888	0.96868	0.95868	8992	30	50	928
7	0.9906	0.95517	0.95331	0.95424	8926	48	46	980
8	0.9899	0.95421	0.94148	0.94780	8982	57	44	917
9	0.9892	0.95643	0.93558	0.94589	8948	65	43	944

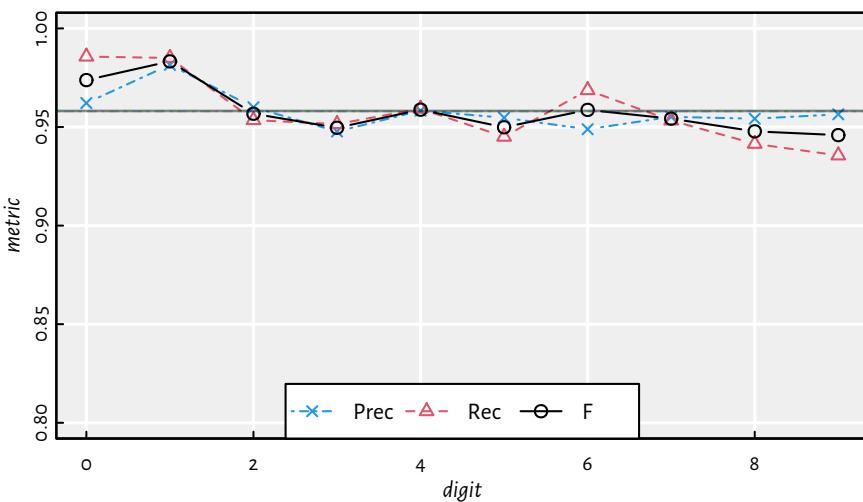


Figure 13.10: Performance metrics for a 2-layer net 784-800-10 [relu] on MNIST

13.4.4 Example in R - 6 Layers

How about a 6-layer Deep Neural Network like 784-2500-2000-1500-1000-500-10? Here you are:

```
file_name <- "datasets/mnist_keras_model3.h5"
if (!file.exists(file_name)) {
  set.seed(123)
  model3 <- keras_model_sequential()
  layer_dense(model3, units=2500, activation="relu")
  layer_dense(model3, units=2000, activation="relu")
  layer_dense(model3, units=1500, activation="relu")
  layer_dense(model3, units=1000, activation="relu")
  layer_dense(model3, units=500, activation="relu")
  layer_dense(model3, units=10, activation="softmax")
  compile(model3, optimizer="sgd",
          loss="categorical_crossentropy")
  fit(model3, X_train2, Y_train2, epochs=10)
  save_model_hdf5(model3, file_name)
} else {
  model3 <- load_model_hdf5(file_name)
}

Y_pred2 <- predict(model3, X_test2)
Y_pred <- apply(Y_pred2, 1, which.max)-1 # 1..10 -> 0..9
mean(Y_test == Y_pred) # accuracy on the test set
```

```
## [1] 0.9797
```

Performance metrics for each digit separately, see also Figure 13.11.

i	Acc	Prec	Rec	F	TN	FN	FP	TP
0	0.9966	0.97395	0.99184	0.98281	8994	8	26	972
1	0.9975	0.98856	0.98943	0.98899	8852	12	13	1123
2	0.9951	0.98615	0.96609	0.97602	8954	35	14	997
3	0.9964	0.99093	0.97327	0.98202	8981	27	9	983
4	0.9960	0.97006	0.98982	0.97984	8988	10	30	972
5	0.9962	0.98856	0.96861	0.97848	9098	28	10	864
6	0.9963	0.98019	0.98121	0.98070	9023	18	19	940
7	0.9961	0.98338	0.97860	0.98098	8955	22	17	1006
8	0.9939	0.96065	0.97741	0.96896	8987	22	39	952
9	0.9953	0.97436	0.97919	0.97677	8965	21	26	988

Exercise 13.1 Test the performance of different neural network architectures (different number of layers, different number of neurons in each layer etc.). Yes, it's more art than science! Many tried to

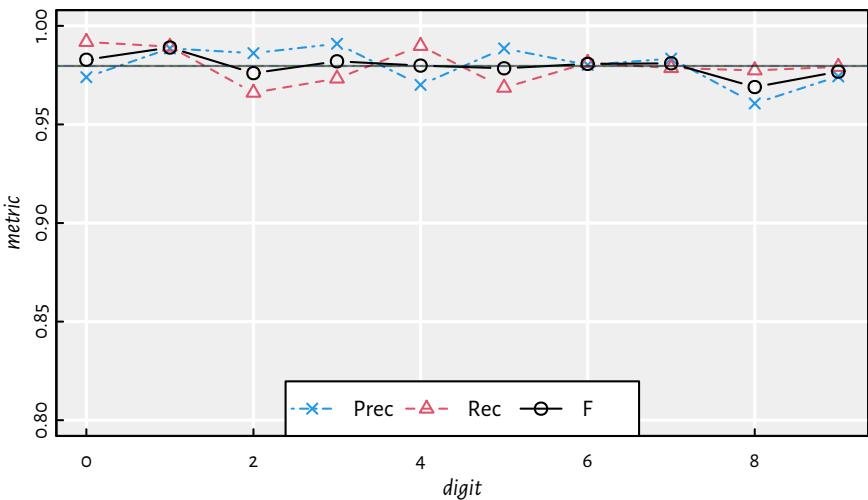


Figure 13.11: Performance metrics for a 6-layer net 784-2500-2000-1500-1000-500-10 [relu] on MNIST

come up with various “rules of thumb”, see, for example, the `comp.ai.neural-nets` FAQ (Sarle & others 2002) at <http://www.faqs.org/faqs/ai-faq/neural-nets/part3/preamble.html>, but what works well in one problem might not be generalisable to another one.

13.5 Preprocessing of Data

13.5.1 Introduction

Do not underestimate the power of appropriate data preprocessing — deep neural networks are not a universal replacement for a data engineer’s hard work!

On top of that, they are not interpretable – these are merely black-boxes.

Among the typical transformations of the input images we can find:

- normalisation of colours (setting brightness, stretching contrast, etc.),
- repositioning of the image (centring),
- deskewing (see below),
- denoising (e.g., by blurring).

Another frequently applied technique concerns an expansion of the training data — we can add “artificially contaminated” images to the training set (e.g., slightly rotated digits) so as to be more ready to whatever will be provided in the test test.

13.5.2 Image Deskewing

Deskewing of images (“straightening” of the digits) is amongst the most typical transformations that can be applied on MNIST.

Unfortunately, we don't have (yet) the necessary mathematical background to discuss this operation in very detail.

Luckily, we can apply it on each image anyway.

See the GitHub repository at <https://github.com/gagolews/Playground.R> for an example notebook and the `deskew.R` script.

```
# See https://github.com/gagolews/Playground.R
source("~/R/Playground.R/deskew.R")
# new_image <- deskew(old_image)
```

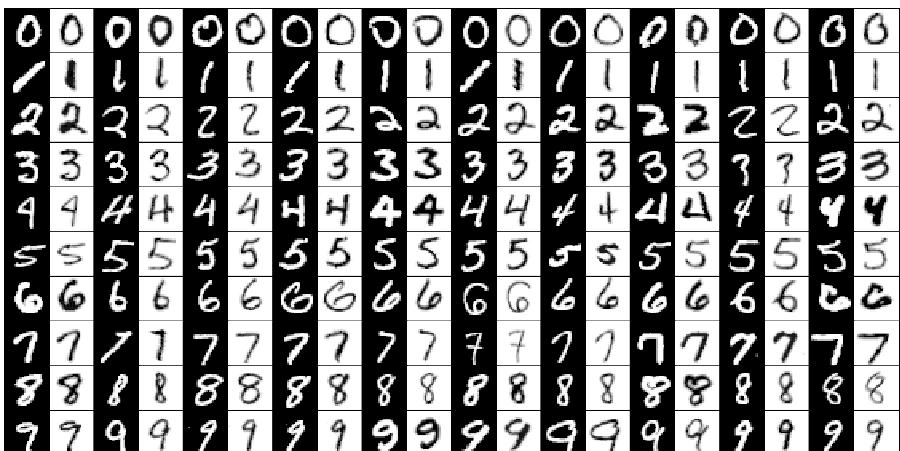


Figure 13.12: Deskewing of the MNIST digits

Let's take a look at Figure 13.12. In each pair, the left image (black background) is the original one, and the right image (palette inverted for purely dramatic effects) is its deskewed version.

Below we deskew each image in the training as well as in the test sample. This also takes a long time, so let's store the resulting objects for further reference:

```
file_name <- "datasets/mnist_deskewed_train.rds"
if (!file.exists(file_name)) {
  Z_train <- X_train
  for (i in 1:dim(Z_train)[1]) {
    Z_train[i,,] <- deskew(Z_train[i,,])
  }
  Z_train2 <- matrix(Z_train, ncol=28*28)
  saveRDS(Z_train2, file_name)
```

```

} else {
  Z_train2 <- readRDS(file_name)
}

file_name <- "datasets/mnist_deskewed_test.rds"
if (!file.exists(file_name)) {
  Z_test <- X_test
  for (i in 1:dim(Z_test)[1]) {
    Z_test[i,,] <- deskew(Z_test[i,,])
  }
  Z_test2 <- matrix(Z_test, ncol=28*28)
  saveRDS(Z_test2, file_name)
} else {
  Z_test2 <- readRDS(file_name)
}

```

Remark. RDS is a compressed file format used by R for object serialisation (quickly storing its verbatim copies so that they can be reloaded at any time).

Multinomial logistic regression model (1-layer NN):

```

file_name <- "datasets/mnist_keras_model1d.h5"
if (!file.exists(file_name)) {
  set.seed(123)
  model1d <- keras_model_sequential()
  layer_dense(model1d, units=10, activation="softmax")
  compile(model1d, optimizer="sgd",
          loss="categorical_crossentropy")
  fit(model1d, Z_train2, Y_train2, epochs=10)
  save_model_hdf5(model1d, file_name)
} else {
  model1d <- load_model_hdf5(file_name)
}

Y_pred2 <- predict(model1d, Z_test2)
Y_pred <- apply(Y_pred2, 1, which.max)-1 # 1..10 -> 0..9
mean(Y_test == Y_pred) # accuracy on the test set

```

```
## [1] 0.9488
```

Performance metrics for each digit separately, see also Figure 13.13.

i	Acc	Prec	Rec	F	TN	FN	FP	TP
0	0.9939	0.95450	0.98469	0.96936	8974	15	46	965
1	0.9959	0.98236	0.98150	0.98193	8845	21	20	1114
2	0.9878	0.95409	0.92636	0.94002	8922	76	46	956

i	Acc	Prec	Rec	F	TN	FN	FP	TP
3	0.9904	0.95069	0.95446	0.95257	8940	46	50	964
4	0.9888	0.94118	0.94501	0.94309	8960	54	58	928
5	0.9905	0.94426	0.94955	0.94690	9058	45	50	847
6	0.9905	0.95565	0.94468	0.95013	9000	53	42	905
7	0.9892	0.96000	0.93385	0.94675	8932	68	40	960
8	0.9855	0.91162	0.94251	0.92680	8937	56	89	918
9	0.9851	0.92914	0.92270	0.92591	8920	78	71	931

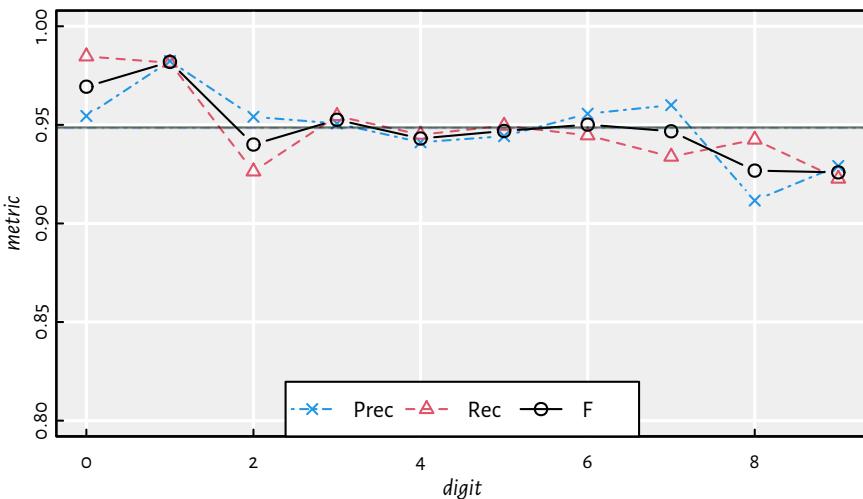


Figure 13.13: Performance of Multinomial Logistic Regression on the deskewed MNIST

13.5.3 Summary of All the Models Considered

Let's summarise the quality of all the considered classifiers. Figure 13.14 gives the F-measures, for each digit separately.

Note that the applied preprocessing of data increased the prediction accuracy.

The same information can also be included on a heat map which is depicted in Figure 13.15 (see the `image()` function in R).

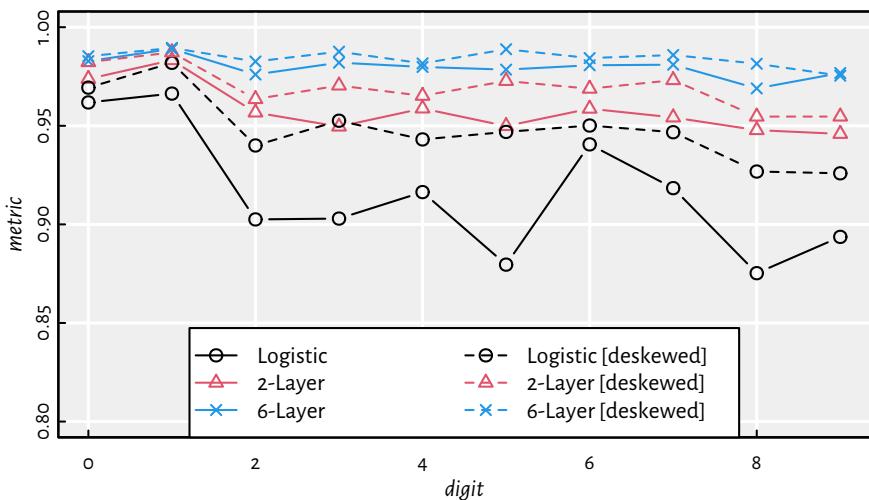


Figure 13.14: Summary of F-measures for each classified digit and every method

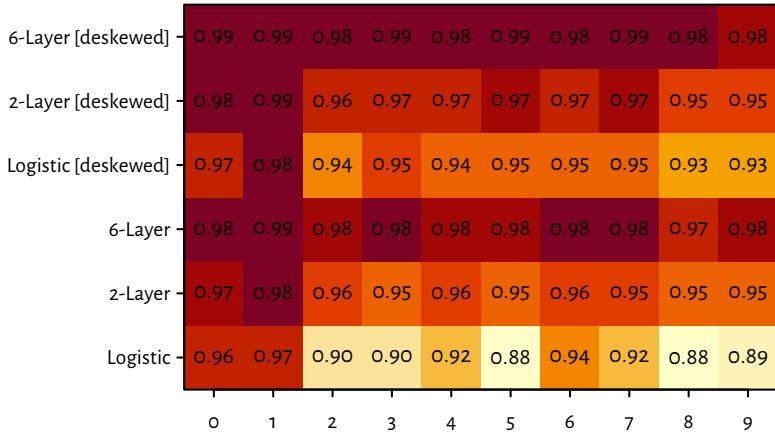


Figure 13.15: A heat map of F-measures for each classified digit and each method

13.6 Outro

13.6.1 Remarks

We have discussed a multinomial logistic regression model as a generalisation of the binary one.

This in turn is a special case of feed-forward neural networks.

There's a lot of hype (again...) for deep neural networks in many applications, including vision, self-driving cars, natural language processing, speech recognition etc.

Many different architectures of neural networks and types of units are being considered in theory and in practice, e.g.:

- convolutional neural networks apply a series of signal (e.g., image) transformations in first layers, they might actually “discover” deskewing automatically etc.;
- recurrent neural networks can imitate long/short-term memory that can be used for speech synthesis and time series prediction.

Main drawbacks of deep neural networks:

- learning is very slow, especially with very deep architectures (days, weeks);
- models are not explainable (black boxes) and hard to debug;
- finding good architectures is more art than science (maybe: more of a craftsmanship even);
- sometimes using deep neural network is just an excuse for being too lazy to do proper data cleansing and pre-processing.

There are many issues and challenges that are tackled in more advanced AI/ML courses and books, such as (Goodfellow et al. 2016).

13.6.2 Beyond MNIST

The MNIST dataset is a classic, although its use in deep learning research is nowadays discouraged – the dataset is not considered challenging anymore – state of the art classifiers can reach 99.8% accuracy.

See Zalando's Fashion-MNIST (by Kashif Rasul & Han Xiao) at <https://github.com/zalandoresearch/fashion-mnist> for a modern replacement.

Alternatively, take a look at CIFAR-10 and CIFAR-100 (<https://www.cs.toronto.edu/~kriz/cifar.html>) by A. Krizhevsky et al. or at ImageNet (<http://image-net.org/index>) for an even greater challenge.

TODO

Recommended further reading: (James et al. 2017: Chapter 11), (Sarle & others 2002) and (Goodfellow et al. 2016)

See also the `keras` package tutorials available at: <https://cran.r-project.org/web/packages/keras/index.html> and <https://keras.rstudio.com>.

next chapter...

14

Recommender Systems

TODO In this chapter, we will:

- ...
 - ...
-

14.1 Introduction

Recommender (recommendation) systems aim to predict the rating a *user* would give to an *item*.

For example:

- playlist generators (Spotify, YouTube, Netflix, ...),
- content recommendations (Facebook, Instagram, Twitter, Apple News, ...),
- product recommendations (Amazon, Alibaba, ...).

Implementing recommender systems, according to (Ricci et al. 2011), might:

- increase the number of items sold,
- increase users' satisfaction,
- increase users' fidelity,
- allow a company to sell more diverse items,
- allow to better understand what users want.

Exercise 14.1 *Think of the last time you found some recommendation useful.*

They can also increase the users' frustration.

Exercise 14.2 *Think of the last time you found a recommendation useless and irritating. What might be the reasons why you have been provided with such a suggestion?*

14.1.1 The Netflix Prize

In 2006 Netflix (back then a DVD rental company) released one of the most famous benchmark sets for recommender systems, which helped boost the research on algorithms in this field.

See <https://www.kaggle.com/netflix-inc/netflix-prize-data>; data archived at <https://web.archive.org/web/20090925184737/http://archive.ics.uci.edu/ml/datasets/Netflix+Prize> and https://archive.org/details/nf_prize_dataset.tar

The dataset consists of:

- 480,189 users
- 17,770 movies
- 100,480,507 ratings in the training sample:
 - MovieID
 - CustomerID
 - Rating from 1 to 5
 - Title
 - YearOfRelease from 1890 to 2005
 - Date of rating in the range 1998-11-01 to 2005-12-31

The *quiz set* consists of 1,408,342 ratings and it was used by the competitors to assess the quality of their algorithms and compute the leaderboard scores.

Root mean squared error (RMSE) of predicted vs. true rankings was chosen as a performance metric.

The *test set* of 1,408,789 ratings (not made publicly available) was used to determine the winner.

On 21 September 2009, the grand prize of US\$1,000,000 was given to the BellKor's Pragmatic Chaos team which improved over the Netflix's *Cinematch* algorithm by 10.06%, achieving the winning RMSE of 0.8567 on the test subset.

14.1.2 Main Approaches

Current recommender systems are quite complex and use a fusion of various approaches, also those based on external knowledge bases.

However, we may distinguish at least two core approaches, see (Ricci et al. 2011) for more:

- *Collaborative Filtering* is based on the assumption that if two people interact with the same product, they're likely to have other interests in common as well.

John and Mary both like bananas and apples and dislike spinach. John likes sushi. Mary hasn't tried sushi yet. It seems they might have similar tastes, so we recommend that Mary should give sushi a try.

- *Content-based Filtering* builds users' profiles that represent information about what kind of products they like.
-

We have discovered that John likes fruit but dislikes vegetables. An orange is a fruit (an item similar to those he liked in the past) with which John is yet to interact. Thus, it is suggested that John should give it a try.

Jim Bennett, at that time the vice president of recommendation systems at Netflix on the idea behind the original Cinematch algorithm (see <https://www.technologyreview.com/s/406637/the-1-million-netflix-challenge/> and <https://web.archive.org/web/20070821194257/http://www.netflixprize.com/faq>):

First, you collect 100 million user ratings for about 18,000 movies. Take any two movies and find the people who have rated both of them. Then look to see if the people who rate one of the movies highly rate the other one highly, if they liked one and not the other, or if they didn't like either movie. Based on their ratings, Cinematch sees whether there's a correlation between those people. Now, do this for all possible pairs of 65,000 movies.

Exercise 14.3 Is the above an example of the collaborative or context-based filtering?

14.1.3 Formalism

Let $\mathcal{U} = \{U_1, \dots, U_n\}$ denote the set of n users.

Let $\mathcal{I} = \{I_1, \dots, I_p\}$ denote the set of p items.

Let $\mathbf{R} \in \mathbb{R}^{n \times p}$ be a user-item matrix such that:

$$r_{u,i} = \begin{cases} r & \text{if the } u\text{-th user ranked the } i\text{-th item as } r > 0 \\ 0 & \text{if the } u\text{-th user hasn't interacted with the } i\text{-th item yet} \end{cases}$$

Remark. Note that Θ is used to denote a missing value (NA) here.

In particular, we can assume:

- $r_{u,i} \in \{0, 1, \dots, 5\}$ (ratings on the scale 1–5 or no interaction)

- $r_{u,i} \in \{0, 1\}$ (“Like” or no interaction)

The aim of an recommender system is to predict the rating $\hat{r}_{u,i}$ that the u -th user would give to the i -th item provided that currently $r_{u,i} = 0$.

14.2 Collaborative Filtering

14.2.1 Example

.	Apple	Banana	Sushi	Spinach	Orange
Anne	1	5	5		1
Beth	1	1	5	5	1
John	5	5		1	
Kate	1	1	5	5	1
Mark	5	5	1	1	5
Sara	?	5		?	5

In **user-based collaborative filtering**, we seek users with similar preference profiles/rating patters.

“User A has similar behavioural patterns as user B, so A should suggested with what B likes.”

In **item-based collaborative filtering**, we seek items with similar (dis)likeability structure.

“Users who (dis)liked X also (dis)liked Y”.

Exercise 14.4 Will Sara enjoy her spinach? Will Sara enjoy her apple?

An example **R** matrix in R:

```
R <- matrix(
  c(
    1, 5, 5, 0, 1,
    1, 1, 5, 5, 1,
    5, 5, 0, 1, 0,
    1, 1, 5, 5, 1,
    5, 5, 1, 1, 5,
    0, 5, 0, 0, 5
  ), byrow=TRUE, nrow=6, ncol=5,
  dimnames=list(
    c("Anne", "Beth", "John", "Kate", "Mark", "Sara"),
    c("Apple", "Banana", "Sushi", "Spinach", "Orange")
  )
)

R

##      Apple Banana Sushi Spinach Orange
## Anne     1       5     5      0      1
## Beth     1       1     5      5      1
## John     5       5     0      1      0
## Kate     1       1     5      5      1
## Mark     5       5     1      1      5
## Sara     0       5     0      0      5
```

14.2.2 Similarity Measures

Assuming \mathbf{a}, \mathbf{b} are two sequences of length k (in our setting, k is equal to either n or p), let S be the following similarity measure between two rating vectors:

$$S(\mathbf{a}, \mathbf{b}) = \frac{\sum_{i=1}^k a_i b_i}{\sqrt{\sum_{i=1}^k a_i^2} \sqrt{\sum_{i=1}^k b_i^2}}$$

```
cosim <- function(a, b) sum(a*b)/sqrt(sum(a^2)*sum(b^2))
```

We call it the **cosine similarity**. We have $S(\mathbf{a}, \mathbf{b}) \in [-1, 1]$, where we get 1 for two identical elements. Similarity of 0 is obtained for two unrelated (“orthogonal”) vectors. For nonnegative sequences, negative similarities are not generated.

(*) Another frequently considered similarity measure is a version of the Pearson correlation coefficient that ignores all 0-valued observations, see also the `use` argument to the `cor()` function.

14.2.3 User-Based Collaborative Filtering

User-based approaches involve comparing each user against every other user (pairwise comparisons of the rows in \mathbf{R}). This yields a similarity degree between the i -th and the j -th user:

```

 $s_{i,j}^U = S(\mathbf{r}_{i,\cdot}, \mathbf{r}_{j,\cdot}).$ 

SU <- matrix(0, nrow=nrow(R), ncol=nrow(R),
             dimnames=dimnames(R)[c(1,1)]) # and empty n*n matrix
for (i in 1:nrow(R)) {
  for (j in 1:nrow(R)) {
    SU[i,j] <- cosim(R[i,], R[j,])
  }
}
round(SU, 2)

##      Anne Beth John Kate Mark Sara
## Anne 1.00 0.61 0.58 0.61 0.63 0.59
## Beth 0.61 1.00 0.29 1.00 0.39 0.19
## John 0.58 0.29 1.00 0.29 0.81 0.50
## Kate 0.61 1.00 0.29 1.00 0.39 0.19
## Mark 0.63 0.39 0.81 0.39 1.00 0.81
## Sara 0.59 0.19 0.50 0.19 0.81 1.00

```

In order to obtain the previously unobserved rating $\hat{r}_{u,i}$ using the user-based approach, we typically look for the K most similar users and aggregate their corresponding scores (for some $K \geq 1$).

More formally, let $\{U_{v_1}, \dots, U_{v_K}\} \subseteq \mathcal{U} \setminus \{U_u\}$ be the set of users maximising $s_{u,v_1}^U, \dots, s_{u,v_K}^U$ and having $r_{v_1,i}, \dots, r_{v_K,i} > 0$. Then:

$$\hat{r}_{u,i} = \frac{1}{K} \sum_{\ell=1}^K r_{v_\ell,i}.$$

Remark. The arithmetic mean can be replaced with, e.g., the more or a weighted arithmetic mean where weights are proportional to s_{u,v_ℓ}^U .

This is very similar to the K -nearest neighbour heuristic!

```

K <- 2
(sim <- order(SU["Sara",], decreasing=TRUE))

## [1] 6 5 1 3 2 4

```

```

# sim gives the indices of people in decreasing order
# of similarity to Sara:
dimnames(R)[[1]][sim] # the corresponding names

## [1] "Sara" "Mark" "Anne" "John" "Beth" "Kate"

# Remove those who haven't tried Spinach yet (including Sara):
sim <- sim[ R[sim, "Spinach"]>0 ]
dimnames(R)[[1]][sim]

## [1] "Mark" "John" "Beth" "Kate"

# aggregate the Spinach ratings of the K most similar people:
mean(R[sim[1:K], "Spinach"])

## [1] 1

```

14.2.4 Item-Based Collaborative Filtering

Item-based schemes rely on pairwise comparisons between the items (columns in **R**). Hence, a similarity degree between the i -th and the j -th item is given by:

```

SI <- matrix(0, nrow=ncol(R), ncol=ncol(R),
             dimnames=dimnames(R)[c(2,2)]) # an empty p*p matrix
for (i in 1:ncol(R)) {
  for (j in 1:ncol(R)) {
    SI[i,j] <- cosim(R[,i], R[,j])
  }
}
round(SI, 2)

##          Apple Banana Sushi Spinach Orange
## Apple     1.00   0.78  0.32   0.38   0.53
## Banana    0.78   1.00  0.45   0.27   0.78
## Sushi     0.32   0.45  1.00   0.81   0.32
## Spinach   0.38   0.27  0.81   1.00   0.29
## Orange    0.53   0.78  0.32   0.29   1.00

```

In order to obtain the previously unobserved rating $\hat{r}_{u,i}$ using the item-based approach, we typically look for the K most similar items and aggregate their corresponding scores (for some $K \geq 1$)

More formally, let $\{I_{j_1}, \dots, I_{j_K}\} \in \mathcal{I}$ $\{I_i\}$ be the set of items maximising $s_{i,j_1}^I, \dots, s_{i,j_K}^I$ and having $r_{u,j_1}, \dots, r_{u,j_K} > 0$. Then:

$$\hat{r}_{u,i} = \frac{1}{K} \sum_{\ell=1}^K r_{u,j_\ell}.$$

Remark. Similarly to the previous case, the arithmetic mean can be replaced with, e.g., the mode or a weighted arithmetic mean where weights are proportional to s_{i,j_ℓ}^l .

```
K <- 2
(sim <- order(SI["Apple",], decreasing=TRUE))

## [1] 1 2 5 4 3
# sim gives the indices of items in decreasing order
# of similarity to Apple:
dimnames(R)[[2]][sim] # the corresponding item types

## [1] "Apple"    "Banana"   "Orange"   "Spinach"  "Sushi"
# Remove these which Sara haven't tried yet (e.g., Apples):
sim <- sim[ R["Sara", sim]>0 ]
dimnames(R)[[2]][sim]

## [1] "Banana" "Orange"
# aggregate Sara's ratings of the K most similar items:
mean(R["Sara", sim[1:K]])

## [1] 5
```

14.3 Exercise: The MovieLens Dataset (*)

14.3.1 Dataset

Let's make a few recommendations based on the MovieLens-9/2018-Small dataset available at <https://grouplens.org/datasets/movielens/latest/>, see also <https://movielens.org/> and (Harper & Konstan 2015).

The dataset consists of ca. 100,000 ratings to 9,000 movies by 600 users. It was last updated on September 2018.

This is already a pretty large dataset! We might run into problems with memory usage and high run-time.

```
movies <- read.csv("datasets/ml-9-2018-small/movies.csv",
  comment.char="#")
head(movies, 4)

##   movieId          title
```

```

## 1      1      Toy Story (1995)
## 2      2      Jumanji (1995)
## 3      3  Grumpier Old Men (1995)
## 4      4 Waiting to Exhale (1995)

##                                genres
## 1 Adventure|Animation|Children|Comedy|Fantasy
## 2           Adventure|Children|Fantasy
## 3           Comedy|Romance
## 4 Comedy|Drama|Romance

nrow(movies)

## [1] 9742

ratings <- read.csv("datasets/ml-9-2018-small/ratings.csv",
  comment.char="#")
head(ratings, 4)

##   userId movieId rating timestamp
## 1      1       1     4 964982703
## 2      1       3     4 964981247
## 3      1       6     4 964982224
## 4      1      47     5 964983815

nrow(ratings)

## [1] 100836

table(ratings$rating)

## 
##   0.5     1    1.5     2    2.5     3    3.5     4    4.5     5
## 1370  2811  1791  7551  5550  20047 13136  26818  8551  13211

```

14.3.2 Data Cleansing

movieIds should be re-encoded, as not every film is mentioned/rated in the database. We will re-map the movieIds to consecutive integers.

```

# the list of all rated movieIds:
movieId2 <- unique(ratings$movieId)
# max user Id (these could've been cleaned up too):
(n <- max(ratings$userId))

## [1] 610

# number of unique movies:
(p <- length(movieId2))

## [1] 9724

```

```
# remove unrated movies:
movies <- movies[movies$movieId %in% movieId2, ]

# we'll map movieId2[i] to i for each i=1,...,p:
movies$movieId <- match(movies$movieId, movieId2)
ratings$movieId <- match(ratings$movieId, movieId2)
# order the movies by the new movieId so that
# the movie with Id==i is in the i-th row:
movies <- movies[order(movies$movieId),]
stopifnot(all(movies$movieId == 1:p)) # sanity check
```

We will use a sparse matrix data type (from R package `Matrix`) to store the ratings data, $\mathbf{R} \in \mathbb{R}^{n \times p}$.

Remark. *Sparse* matrices contain many zeros. Instead of storing all the $np = 5931640$ elements, only the lists of non-zero ones are saved, 100836 values in total. This way, we might save a lot of memory. The drawback is that, amongst others, random access to the elements in a sparse matrix takes more time.

```
library("Matrix")
R <- Matrix(0.0, sparse=TRUE, nrow=n, ncol=p)
# This is a vectorised operation;
# it is faster than a for loop over each row
# in the ratings matrix:
R[cbind(ratings$userID, ratings$movieId)] <- ratings$rating
```

Let's preview a few first rows and columns:

```
R[1:6, 1:18]
```

```
## 6 x 18 sparse Matrix of class "dgCMatrix"
##
## [1,] 4 4 4 5 5 3 5 4 5 5 5 5 3 5 4 5 3 3
## [2,] .
## [3,] .
## [4,] .
## [5,] 4 . . 4 . 4 . . . . . . 5 2
## [6,] . 5 4 4 1 . 5 4 . 3 4 . 3 . 2 5
```

14.3.3 Item-Item Similarities

To recall, the cosine similarity between $\mathbf{r}_{\cdot,i}, \mathbf{r}_{\cdot,j} \in \mathbb{R}^n$ is given by:

$$s_{i,j}^I = S_C(\mathbf{r}_{\cdot,i}, \mathbf{r}_{\cdot,j}) = \frac{\sum_{k=1}^n r_{k,i} r_{k,j}}{\sqrt{\sum_{k=1}^n r_{k,i}^2} \sqrt{\sum_{k=1}^n r_{k,j}^2}}$$

In vector/matrix algebra notation, this is:

$$s_{i,j}^I = S_C(\mathbf{r}_{\cdot,i}, \mathbf{r}_{\cdot,j}) = \frac{\mathbf{r}_{\cdot,i}^T \mathbf{r}_{\cdot,j}}{\sqrt{\mathbf{r}_{\cdot,i}^T \mathbf{r}_{\cdot,i}} \sqrt{\mathbf{r}_{\cdot,j}^T \mathbf{r}_{\cdot,j}}}$$

As $\mathbf{R} \in \mathbb{R}^{n \times p}$, we can compute all the $p \times p$ cosine similarities at once by applying:

$$\mathbf{S}^I = \frac{\mathbf{R}^T \mathbf{R}}{\mathbf{l}^T}$$

where $\mathbf{l} \in \mathbb{R}^{p \times 1}$ is a column vector such that with $l_i = \sqrt{\mathbf{r}_{\cdot,i}^T \mathbf{r}_{\cdot,i}}$ and by \div we mean elementwise division.

Cosine similarities for item-item comparisons:

```
norms <- as.matrix(sqrt(colSums(R^2)))
Rx <- as.matrix(crossprod(R, R))
SI <- Rx/tcrossprod(norms)
SI[is.nan(SI)] <- 0 # there were some divisions by zero
```

Remark. `crossprod(A,B)` gives $\mathbf{A}^T \mathbf{B}$ and `tcrossprod(A,B)` gives $\mathbf{A} \mathbf{B}^T$.

14.3.4 Example Recommendations

```
recommend <- function(i, K, SI, movies) {
  # get K most similar movies to the i-th one
  ms <- order(SI[i,], decreasing=TRUE)
  data.frame(
    Title=movies$title[ms[1:K]],
    SIC=SI[i,ms[1:K]]
  )
}

movies$title[1215]

## [1] "Monty Python's The Meaning of Life (1983)"
recommend(1215, 10, SI, movies)

##                                     Title      SIC
## 1 Monty Python's The Meaning of Life (1983) 1.000000
```

```

## 2           Monty Python's Life of Brian (1979) 0.61097
## 3           Monty Python and the Holy Grail (1975) 0.51415
## 4 House of Flying Daggers (Shi mian mai fu) (2004) 0.49322
## 5   Hitchhiker's Guide to the Galaxy, The (2005) 0.45482
## 6           Bowling for Columbine (2002) 0.45051
## 7           Shaun of the Dead (2004) 0.44566
## 8 O Brother, Where Art Thou? (2000) 0.44541
## 9           Ghost World (2001) 0.44416
## 10          Full Metal Jacket (1987) 0.44285

movies$title[1]

## [1] "Toy Story (1995)"

recommend(1, 10, SI, movies)

##                                     Title      SIC
## 1           Toy Story (1995) 1.00000
## 2           Toy Story 2 (1999) 0.57260
## 3           Jurassic Park (1993) 0.56564
## 4 Independence Day (a.k.a. ID4) (1996) 0.56426
## 5 Star Wars: Episode IV - A New Hope (1977) 0.55739
## 6           Forrest Gump (1994) 0.54710
## 7 Lion King, The (1994) 0.54115
## 8 Star Wars: Episode VI - Return of the Jedi (1983) 0.54109
## 9           Mission: Impossible (1996) 0.53891
## 10          Groundhog Day (1993) 0.53417

```

...and so on.

14.3.5 Clustering

All our ratings are $r_{i,j} \geq 0$, therefore the cosine similarity is $s_{i,j}^I \geq 0$. Moreover, it holds $s_{i,j}^I \leq 1$. Thus, a cosine similarity matrix can be turned into a dissimilarity matrix:

```

DI <- 1.0-SI
DI[DI<0] <- 0.0 # account for numeric inaccuracies
DI <- as.dist(DI)

```

This enables us to perform, e.g., the cluster analysis of items:

```

library("genie")

## Loading required package: genieclust
h <- hclust2(DI)
plot(h, labels=FALSE, ann=FALSE); box()

```

A 14-partition might look nice, let's give it a try:

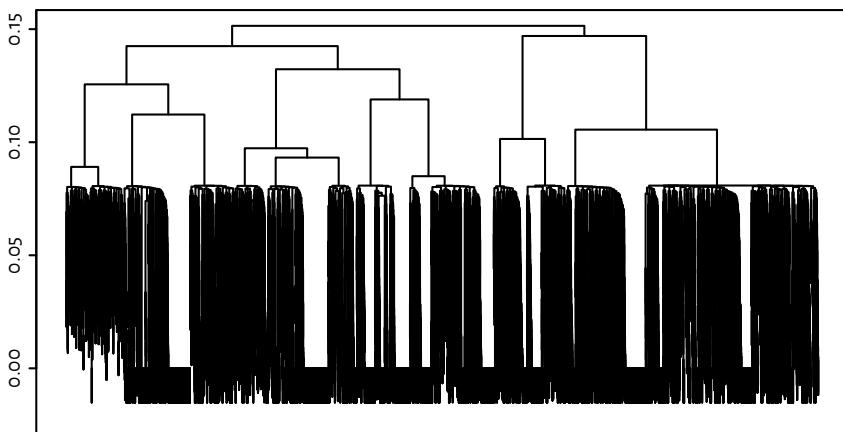


Figure 14.1: Cluster dendrogram for the movies

```
c <- cutree(h, k=14)
```

Example movies in the 3rd cluster:

Bottle Rocket (1996), Clerks (1994), Star Wars: Episode IV - A New Hope (1977), Swingers (1996), Monty Python's Life of Brian (1979), E.T. the Extra-Terrestrial (1982), Monty Python and the Holy Grail (1975), Star Wars: Episode V - The Empire Strikes Back (1980), Princess Bride, The (1987), Raiders of the Lost Ark (Indiana Jones and the Raiders of the Lost Ark) (1981), Star Wars: Episode VI - Return of the Jedi (1983), Blues Brothers, The (1980), Duck Soup (1933), Groundhog Day (1993), Back to the Future (1985), Young Frankenstein (1974), Indiana Jones and the Last Crusade (1989), Grosse Pointe Blank (1997), Austin Powers: International Man of Mystery (1997), Men in Black (a.k.a. MIB) (1997)

The definitely have something in common!

Example movies in the 1st cluster:

Toy Story (1995), Heat (1995), Seven (a.k.a. Se7en) (1995), Usual Suspects, The (1995), From Dusk Till Dawn (1996), Braveheart (1995), Rob Roy (1995), Desperado (1995), Billy Madison (1995), Dumb & Dumber (Dumb and Dumber) (1994), Ed Wood (1994), Pulp Fiction (1994), Stargate (1994), Tommy Boy (1995), Clear and Present Danger (1994), Forrest Gump (1994), Jungle Book, The (1994), Mask, The (1994), Fugitive, The (1993), Jurassic Park (1993)

... and so forth.

14.4 Outro

14.4.1 Remarks

Good recommender systems are perfect tools to increase the revenue of any user-centric enterprise.

Not a single algorithm, but an ensemble (a proper combination) of different approaches is often used in practice, see the Further Reading section below for the detailed information of the Netflix Prize winners.

Recommender systems are an interesting fusion of the techniques we have already studied – linear models, K-nearest neighbours etc.

Building recommender systems is challenging, because data is large yet often sparse. For instance, the ratio of available ratings vs. all possible user-item valuations for the Netflix Prize (obviously, it is just a sample of the complete dataset that Netflix has) is equal to:

```
100480507/(480189*17770)
```

```
## [1] 0.011776
```

A *sparse matrix* (see R package `Matrix`) data structure is often used for storing of and computing over such data effectively.

Note that some users are *biased* in the sense that they are more critical or enthusiastic than average users.

Exercise 14.5 Is 3 stars a “bad”, “fair enough” or “good” rating for you? Would you go to a bar/restaurant ranked 3.0 by your favourite Maps app community?

It is particularly challenging to predict the preferences of users that cast few ratings, e.g., those who just signed up (*the cold start problem*).

“Hill et al. [1995] have shown that users provide inconsistent ratings when asked to rate the same movie at different times. They suggest that an algorithm cannot be more accurate than the variance in a user’s ratings for the same item.” (Herlocker et al. 2004: p. 6)

It is good to take into account the temporal (time-based) characteristics of data as well as external knowledge (e.g., how long ago a rating was cast, what is a film’s genre).

The presented approaches are vulnerable to attacks – bots may be used to promote or inhibit selected items.

TODO

Recommended further reading: (Herlocker et al. 2004), (Ricci et al. 2011), (Lü & others 2012), (Harper & Konstan 2015). See also the Netflix prize winners: (Koren 2009), (Töscher et al. 2009), (Piotte & Chabbert 2009). Also take a look at the R package `recommenderlab` (amongst others).

next chapter....

15

Natural Language Processing

TODO In this chapter, we will:

- ...
 - ...
-

15.1 TO DO

TO DO

..

TODO

further reading...

A

Notation Convention

Abbreviations

a.k.a. == also known as

w.r.t. == with respect to

s.t. == such that

iff == if and only if

e.g. == for example (Latin: *exempli gratia*)

i.e. == that is (Latin: *id est*)

etc. == and so forth (Latin: *et cetera*)

AI == artificial intelligence

API == application programming interface

GA == genetic algorithm

GD == gradient descent

GLM == generalised linear model

ML == machine learning

NN == neural network

SGD == stochastic gradient descent

IDE = integrated development environment

1D, 2D, 3D, ... == 1-, 2-, 3-dimensional etc.

Notation Convention – Logic and Set Theory

\in – “is in”

By writing $x \in \{a, b, c\}$ we mean that “ x is in a set that consists of a, b and c ” or “ x is either a, b or c ”

$A \subseteq B$ – set A is a subset of set B (every element in A belongs to B , $x \in A$ implies that $x \in B$)

\forall – for all

\exists – exists

$A \cup B$ – union (sum) of two sets, $x \in A \cup B$ iff $x \in A$ or $x \in B$ ($\cup = \text{cup}$)

$A \cap B$ – intersection (sum) of two sets, $x \in A \cap B$ iff $x \in A$ and $x \in B$ ($\cap = \text{cap}$)

$A \setminus B$ – difference of two sets, $x \in A \setminus B$ iff $x \in A$ and $x \notin B$

$A \times B$ – Cartesian product of two sets, $A \times B = \{(a, b) : a \in A, b \in B\}$

$A^p = A \times A \times \dots \times A$ (p times) for any p

$\bigcup_{i=1}^n C_i = C_1 \cup C_2 \cup \dots \cup C_n$ – the union of n indexed sets

$\bigcap_{i=1}^n C_i = C_1 \cap C_2 \cap \dots \cap C_n$ – the intersection of n indexed sets

Notation Convention – Symbols

X, Y, A, I, C – bold (I use it for denoting vectors and matrices)

$\mathbb{X}, \mathbb{Y}, \mathbb{A}, \mathbb{I}, \mathbb{C}$ – blackboard bold (I sometimes use it for sets)

$\mathcal{X}, \mathcal{Y}, \mathcal{A}, \mathcal{I}, \mathcal{C}$ – calligraphic (I use it for set families = sets of sets)

$X, x, \mathbf{X}, \mathbf{x}$ – inputs (usually)

$Y, y, \mathbf{Y}, \mathbf{y}$ – outputs

$\hat{Y}, \hat{y}, \hat{\mathbf{Y}}, \hat{\mathbf{y}}$ – predicted outputs (usually)

- X – independent/explanatory/predictor variable

- Y – dependent/response/predicted variable

\mathbb{R} – the set of real numbers, $\mathbb{R} = (-\infty, \infty)$

\mathbb{N} – the set of natural numbers, $\mathbb{N} = \{1, 2, 3, \dots\}$

\mathbb{N}_0 – the set of natural numbers including zero, $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$

\mathbb{Z} – the set of integer numbers, $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$

$[0, 1]$ – the unit interval

(a, b) – an open interval; $x \in (a, b)$ iff $a < x < b$ for some $a < b$

$[a, b]$ – a closed interval; $x \in [a, b]$ iff $a \leq x \leq b$ for some $a \leq b$

Notation Convention – Vectors and Matrices

$\mathbf{x} = (x_1, \dots, x_n)$ – a sequence of n elements (n -ary sequence/vector)

if it consists of real numbers, we write $\mathbf{x} \in \mathbb{R}^n$

$\mathbf{x} = [x_1 \ x_2 \ \dots \ x_p]$ – a row vector, $\mathbf{x} \in \mathbb{R}^{1 \times p}$ (a matrix with 1 row)

$\mathbf{x} = [x_1 \ x_2 \ \dots \ x_n]^T$ – a column vector, $\mathbf{x} \in \mathbb{R}^{n \times 1}$ (a matrix with 1 column)

$\mathbf{X} \in \mathbb{R}^{n \times p}$ – matrix with n rows and p columns

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{bmatrix}$$

$x_{i,j}$ – element in the i -th row, j -th column

$\mathbf{x}_{i,\cdot}$ – the i -th row of \mathbf{X}

$\mathbf{x}_{\cdot,j}$ – the j -th column of \mathbf{X}

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_{1,\cdot} \\ \mathbf{x}_{2,\cdot} \\ \vdots \\ \mathbf{x}_{n,\cdot} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_{\cdot,1} & \mathbf{x}_{\cdot,2} & \cdots & \mathbf{x}_{\cdot,p} \end{bmatrix}.$$

$$\mathbf{x}_{i,\cdot} = \begin{bmatrix} x_{i,1} & x_{i,2} & \cdots & x_{i,p} \end{bmatrix}.$$

$$\mathbf{x}_{\cdot,j} = \begin{bmatrix} x_{1,j} & x_{2,j} & \cdots & x_{n,j} \end{bmatrix}^T = \begin{bmatrix} x_{1,j} \\ x_{2,j} \\ \vdots \\ x_{n,j} \end{bmatrix},$$

T denotes the matrix transpose; $\mathbf{B} = \mathbf{A}^T$ is a matrix such that $b_{i,j} = a_{j,i}$.

$\|\mathbf{x}\| = \|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$ – the Euclidean norm

Notation Convention – Functions

$f : X \rightarrow Y$ means that f is a function mapping inputs from set X (the domain of f) to elements of Y (the codomain)

$x \mapsto x^2$ denotes a (inline) function mapping x to x^2 , equivalent to `function(x) x^2` in R

$\exp x = e^x$ – exponential function with base $e \approx 2.718$

$\log x$ – natural logarithm (base e)

it holds $e^x = y$ iff $\log y = x$

$\log ab = \log a + \log b$

$\log a^c = c \log a$

$\log a/b = \log a - \log b$

$$\log 1 = 0$$

$$\log e = 1$$

hence $\log e^x = x$

Notation Convention – Sums and Products

$$\sum_{i=1}^n x_i = x_1 + x_2 + \dots + x_n$$

$\sum_{i=1,\dots,n} x_i$ – the same

$\sum_{i \in \{1,\dots,n\}} x_i$ – the same

note display (stand-alone) $\sum_{i=1}^n x_i$ vs text (in-line) $\sum_{i=1}^n x_i$ style

$$\prod_{i=1}^n x_i = x_1 x_2 \dots x_n$$

For example, n-factorial, $n!$ is given by $n! = \prod_{i=1}^n i = 1 \cdot 2 \cdot 3 \cdots \cdot n$ (with the convention that $0! = 1$).

B

Setting Up the R Environment

TODO In this chapter, we will:

- ...
 - ...
-

B.1 Installing R

R and Python are *the* languages of modern data science. The former is slightly more oriented towards data modelling, analysis and visualisation as well as statistical computing. It has a gentle learning curve, which makes it very suitable even for beginners – just like us!

R is available for Windows as well as MacOS, Linux and other Unix-like operating systems. It can be downloaded from the R project website, see <https://www.r-project.org/> (or installed through system-specific package repositories).

Remark. From now on we assume that you have installed the newest version of the R environment.

B.2 Installing an IDE

As we wish to make our first steps with the R language as stress- and hassle-free as possible, let's stick to a very user-friendly development environment called RStudio, which can be downloaded from <https://rstudio.com/products/rstudio/> (choose RStudio Desktop Open Source Edition).

Remark. There are of course many other options for working with R, both interactive and non-interactive, including Jupyter Notebooks (see <https://irkernel.github.io/>), dynamically generated reports (see <https://yihui.org/knitr/options/>) and plain shell scripts executed from a terminal. However, for now let's leave that to more advanced users.

B.3 Installing Recommended Packages

Once we get the above up and running, from within RStudio, we need to install a few packages which we're going to use during the course of this course. Execute the following commands in the R console (bottom-left RStudio pane):

```
pkgs <- c("Cairo", "DEoptim", "fastcluster", "FNN", "genie",
  "genieclust", "glmnet", "gsl", "hydroPSO", "ISLR",
  "keras", "MASS", "Matrix", "microbenchmark", "pdist",
  "randomForest", "RColorBrewer", "recommenderlab",
  "rpart", "rpart.plot", "rworldmap", "scatterplot3d",
  "stringi", "tensorflow", "tidyverse", "titanic", "vioplot",
  "xgboost")
install.packages(pkgs)
```

What is more, in order to be able to play with neural networks, we will need some Python environment, for example the Anaconda Distribution Python 3.x, see <https://www.anaconda.com/distribution/>.

Remark. Do **not** download Python 2.7.

Installation instructions can be found at <https://docs.anaconda.com/anaconda/install/>. This is required for the R packages tensorflow and keras, see <https://tensorflow.rstudio.com/installation/>. Once this is installed, execute the following R commands in the console:

```
library("tensorflow")
install_tensorflow()
```

B.4 First R Script in RStudio

Let's open RStudio and perform the following steps:

1. Create a New Project where we will store all the scripts related to this book. Click *File* → *New Project* and then choose to start in a brand new working directory, in any location you like. Choose *New Project* as the project type.

From now on, we are assuming that the project name is *LMLCR* and the project has been opened. All source files we create will be relative to the project directory.

2. Create a new R source file, *File* → *New File* → *R Script*. Save the file as, for example, *sandbox_01.R*.

The source editor (top left pane) behaves just like any other text editor. Standard keyboard shortcuts are available, such as CTRL+C and CTRL+V (Cmd+C and Cmd+V on MacOS) for copy and paste, respectively.

A list of keyboard shortcuts is available at <https://support.rstudio.com/hc/en-us/articles/200711853-Keyboard-Shortcuts>

3. Input the following R code into the editor:

```
# My first R script
# This is a comment

# Another comment

# Everything from '#' to the end of the line
#     is ignored by the R interpreter
print("Hello world") # prints a given character string
print(2+2) # evaluates the expression and prints the result
x <- seq(0, 10, length.out=100) # a new numeric vector
y <- x^2 # squares every element in x
plot(x, y, las=1, type="l") # plots y as a function of x
```

4. Execute the 5 above commands, line by line, by positioning the keyboard cursor accordingly and pressing Ctrl+Enter (Cmd+Return on MacOS).

Each time, the command will be copied to the console (bottom-left pane) and evaluated.

The last line generates a nice plot which will appear in the bottom-right pane.

While you learn, we recommend that you get used to writing your code in an R script and executing it just as we did above.

On a side note, you can execute (source) the whole script by pressing Ctrl+Shift+S (Cmd+Shift+S on MacOS).

B.5 Exercises

B.5.1 First Steps with Vectors

Exercise B.1 Print the contents of the *x* and *y* vectors that we have created above by issuing the `print(x)` and `print(y)` commands. Here, by a vector we mean a sequence of numbers.

Exercise B.2 Check the length of the vectors by applying `length(x)` and `length(y)`. Read the results.

Exercise B.3 Inspect a few elements in both x and y , for example $y[1]$ gives the first element and $y[length(y)]$ – the last one.

B.5.2 Basic Plotting

Exercise B.4 Read the manual page on the `seq()` function by executing the `?seq` command.

Exercise B.5 Using the `seq()` function with appropriate parameters, create a vector $x2$ with elements 0, 0.1, 0.2, ..., 1.0.

Remark. In one of the above examples, `<-` denotes the assignment operator. By writing `x <- ...` we create a new named object (variable) called x . Use a similar notation to create $x2$.

Exercise B.6 Then create a vector $y2$ that gives the square root of each element in $x2$. Use the `sqrt()` function.

Exercise B.7 Plot $y2$ as a function of $x2$.

Remark. You can access the list of graphical parameters that the `plot()` function takes by executing `?plot.default` and `?plot`. For example, `plot(..., col="red")` draws a red curve. See `colours()` for the list of available colours. Check out the meaning of the `lty` and `lwd` parameters. Moreover, change `type="l"` to `type="p"`. Check out the meaning of the `pch` and `cex` parameters.

TODO

Next chapter...

C

Vector Algebra in R

TODO In this chapter, we will:

- ...
- ...

This chapter is a step-by-step guide to vector computations in R. It also explains the basic mathematical notation around vectors.

You're encouraged to not only simply *read* the chapter, but also to execute yourself the R code provided. Play with it, do some experiments, get curious about how R works. Read the documentation on the functions you are calling, e.g., `?seq`, `?sample` and so on.

Technical and mathematical literature isn't belletristic. It requires *active (pro-active even)* thinking. Sometimes going through a single page can take an hour. Or a day. If you don't understand something, keep thinking, go back, ask yourself questions, take a look at other sources. This is not a *linear* process. This is what makes it fun and creative. To become a good programmer you need a lot of practice, there are no shortcuts. But the whole endeavour is worth the hassle!

C.1 Motivation

Vector and matrix algebra provides us with a convenient language for expressing computations on sequential and tabular data.

Vector and matrix algebra operations are supported by every major programming language – either natively (e.g., R, Matlab, GNU Octave, Mathematica) or via an additional library/package (e.g, Python with numpy, tensorflow or pytorch; C++ with Eigen/Armadillo; C, C++ or Fortran with LAPACK).

By using matrix notation, we generate more concise and readable code.

For instance, given two vectors $\mathbf{x} = (x_1, \dots, x_n)$ and $\mathbf{y} = (y_1, \dots, y_n)$ like:

```
x <- c(1.5, 3.5, 2.3, -6.5)
y <- c(2.9, 8.2, -0.1, 0.8)
```

Instead of writing:

```
s <- 0
n <- length(x)
for (i in 1:n)
  s <- s + (x[i]-y[i])^2
sqrt(s)

## [1] 9.1159
```

to mean:

$$\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^n (x_i - y_i)^2},$$

which denotes the (Euclidean) distance between the two vectors (the square root of the sum of squared differences between the corresponding elements in x and y), we shall soon become used to writing:

```
sqrt(sum((x-y)^2))
```

```
## [1] 9.1159
```

or:

$$\sqrt{(\mathbf{x} - \mathbf{y})^T (\mathbf{x} - \mathbf{y})}$$

or even:

$$\|\mathbf{x} - \mathbf{y}\|_2$$

In order to be able to read this notation, we only have to get to know the most common “building blocks”. There are just a few of them, but it’ll take some time until we become comfortable with their use.

What’s more, we should note that vectorised code might be much faster than the `for` loop-based one (a.k.a. “iterative” style):

```
library("microbenchmark")
n <- 10000
x <- runif(n) # n random numbers in [0, 1]
y <- runif(n)
print(microbenchmark(
  t1={
    # "iterative" style
    s <- 0
    n <- length(x)
```

```

for (i in 1:n)
  s <- s + (x[i]-y[i])^2
sqrt(s)
},
t2={
  # "vectorised" style
  sqrt(sum((x-y)^2))
}
), signif=3, unit='relative')

## Unit: relative
##   expr min  lq  median  uq  max neval
##     t1 125 118    111    114 111  99.3   100
##     t2     1     1      1       1     1  1.0    100

```

C.2 Numeric Vectors

C.2.1 Creating Numeric Vectors

First let's introduce a few ways with which we can create numeric vectors.

C.2.1.1 c()

The `c()` function combines a given list of values to form a sequence:

```
c(1, 2, 3)
```

```
## [1] 1 2 3
c(1, 2, 3, c(4, 5), c(6, c(7, 8)))
```

```
## [1] 1 2 3 4 5 6 7 8
```

Note that when we use the assignment operator, `<-` or `=` (both are equivalent), printing of the output is suppressed:

```
x <- c(1, 2, 3) # doesn't print anything
print(x)
```

```
## [1] 1 2 3
```

However, we can enforce it by parenthesising the whole expression:

```
(x <- c(1, 2, 3))
```

```
## [1] 1 2 3
```

In order to determine that `x` is indeed a numeric vector, we call:

```
mode(x)
## [1] "numeric"
class(x)
## [1] "numeric"
```

Remark. These two functions might return different results. For instance, in the next chapter we note that a numeric matrix will yield `mode()` of `numeric` and `class()` of `matrix`.

What is more, we can get the number of elements in `x` by calling:

```
length(x)
## [1] 3
```

C.2.1.2 seq()

To create an arithmetic progression, i.e., a sequence of equally-spaced numbers, we can call the `seq()` function

```
seq(1, 9, 2)
## [1] 1 3 5 7 9
```

If we access the function's documentation (by executing `?seq` in the console), we'll note that the function takes a couple of parameters: `from`, `to`, `by`, `length.out` etc.

The above call is equivalent to:

```
seq(from=1, to=9, by=2)
## [1] 1 3 5 7 9
```

The `by` argument can be replaced with `length.out`, which gives the desired size:

```
seq(0, 1, length.out=5)
## [1] 0.00 0.25 0.50 0.75 1.00
```

Note that R supports partial matching of argument names:

```
seq(0, 1, len=5)
## [1] 0.00 0.25 0.50 0.75 1.00
```

Quite often we need progressions with step equal to 1 or -1. Such vectors can be generated by applying the `:` operator.

```
1:10      # from:to (inclusive)
## [1] 1 2 3 4 5 6 7 8 9 10
```

```
-1:-10  
## [1] -1 -2 -3 -4 -5 -6 -7 -8 -9 -10
```

C.2.1.3 rep()

Moreover, `rep()` replicates a given vector. Check out the function's documentation (see `?rep`) for the meaning of the arguments provided below.

```
rep(1, 5)  
  
## [1] 1 1 1 1 1  
rep(1:3, 4)  
  
## [1] 1 2 3 1 2 3 1 2 3 1 2 3  
rep(1:3, c(2, 4, 3))  
  
## [1] 1 1 2 2 2 2 3 3 3 3  
rep(1:3, each=4)  
  
## [1] 1 1 1 1 2 2 2 2 3 3 3 3
```

C.2.1.4 Pseudo-Random Vectors

We can also generate vectors of pseudo-random values. For instance, the following generates 5 deviates from the uniform distribution (every number has the same probability) on the unit (i.e., $[0, 1]$) interval:

```
runif(5, 0, 1)  
  
## [1] 0.56490 0.55881 0.44148 0.20764 0.66964
```

We call such numbers pseudo-random, because they are generated arithmetically. In fact, by setting the random number generator's state (also called the *seed*), we can obtain *reproducible* results.

```
set.seed(123)  
runif(5, 0, 1) # a,b,c,d,e  
  
## [1] 0.28758 0.78831 0.40898 0.88302 0.94047  
runif(5, 0, 1) # f,g,h,i,j  
  
## [1] 0.045556 0.528105 0.892419 0.551435 0.456615  
set.seed(123)  
runif(5, 0, 1) # a,b,c,d,e again!  
  
## [1] 0.28758 0.78831 0.40898 0.88302 0.94047
```

Note the difference between the uniform distribution on $[0, 1]$ and the normal distribution with expected value of 0 and standard deviation of 1 (also called the standard normal distribution), see Figure C.1.

```
par(mfrow=c(1, 2)) # align plots in one row and two columns
hist(runif(10000, 0, 1), col="white", ylim=c(0, 2500)); box()
hist(rnorm(10000, 0, 1), col="white", ylim=c(0, 2500)); box()
```

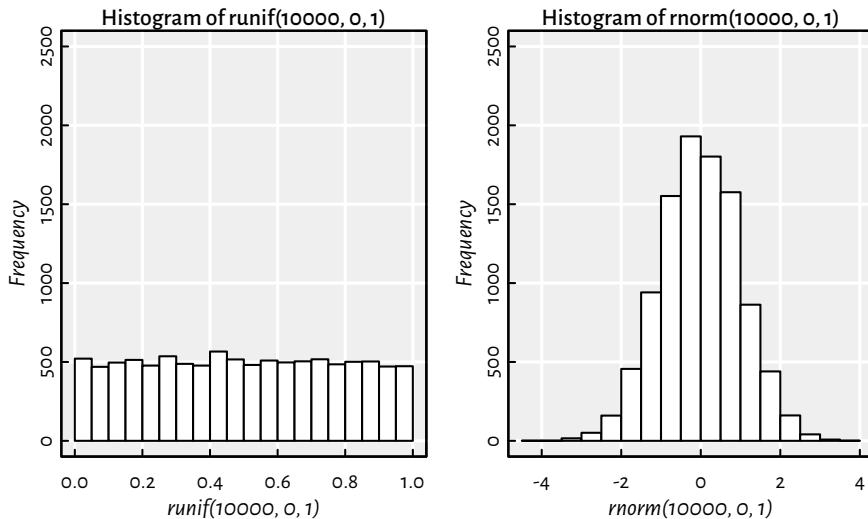


Figure C.1: Uniformly vs. normally distributed random variables

Another useful function samples a number of values from a given vector, either with or without replacement:

```
sample(1:10, 8, replace=TRUE) # with replacement
## [1] 3 3 10 2 6 5 4 6
sample(1:10, 8, replace=FALSE) # without replacement
## [1] 9 5 3 8 1 4 6 10
```

Note that if n is a single number, `sample(n, ...)` is equivalent to `sample(1:n, ...)`. This is a dangerous behaviour than may lead to bugs in our code. Read more at `?sample`.

C.2.2 Vector-Scalar Operations

Mathematically, we sometimes refer to a vector that is reduced to a single component as a *scalar*. We are used to denoting such objects with lowercase letters such as $a, b, i, s, x \in \mathbb{R}$.

Remark. Note that some programming languages distinguish between atomic numer-

ical entities and length-one vectors, e.g., 7 vs. [7] in Python. This is not the case in R, where `length(7)` returns 1.

Vector-scalar arithmetic operations such as `s*x` (multiplication of a vector $\mathbf{x} = (x_1, \dots, x_n)$ by a scalar s) result in a vector \mathbf{y} such that $y_i = sx_i, i = 1, \dots, n$.

The same rule holds for, e.g., $s + \mathbf{x}, \mathbf{x} - s, \mathbf{x}/s$.

```
0.5 * c(1, 10, 100)
```

```
## [1] 0.5 5.0 50.0
```

```
10 + 1:5
```

```
## [1] 11 12 13 14 15
```

```
seq(0, 10, by=2)/10
```

```
## [1] 0.0 0.2 0.4 0.6 0.8 1.0
```

By $-\mathbf{x}$ we will mean $(-1)\mathbf{x}$:

```
-seq(0, 1, length.out=5)
```

```
## [1] 0.00 -0.25 -0.50 -0.75 -1.00
```

Note that in R the same rule applies for exponentiation:

```
(0:5)^2 # synonym: (1:5)**2
```

```
## [1] 0 1 4 9 16 25
```

```
2^(0:5)
```

```
## [1] 1 2 4 8 16 32
```

However, in mathematics, we are **not** used to writing $2^{\mathbf{x}}$ or \mathbf{x}^2 .

C.2.3 Vector-Vector Operations

Let $\mathbf{x} = (x_1, \dots, x_n)$ and $\mathbf{y} = (y_1, \dots, y_n)$ be two vectors of identical lengths.

Arithmetic operations $\mathbf{x} + \mathbf{y}$ and $\mathbf{x} - \mathbf{y}$ are performed *elementwise*, i.e., they result in a vector \mathbf{z} such that $z_i = x_i + y_i$ and $z_i = x_i - y_i$, respectively, $i = 1, \dots, n$.

```
x <- c(1, 2, 3, 4)
```

```
y <- c(1, 10, 100, 1000)
```

```
x+y
```

```
## [1] 2 12 103 1004
```

```
x-y
```

```
## [1] 0 -8 -97 -996
```

Although in mathematics we are **not** used to using any special notation for elementwise multiplication, division and exponentiation, this is available in R.

```
x*y
## [1] 1 20 300 4000
x/y
## [1] 1.000 0.200 0.030 0.004
y^x
## [1] 1e+00 1e+02 1e+06 1e+12
```

Remark. $1e+12$ is a number written in the *scientific notation*. It means “1 times 10 to the power of 12”, i.e., 1×10^{12} . Physicists love this notation, because they are used to dealing with very small (think sizes of quarks) and very large (think distances between galaxies) entities.

Moreover, in R the **recycling rule** is applied if we perform elementwise operations on vectors of *different* lengths – the shorter vector is recycled as many times as needed to match the length of the longer vector, just as if we were performing:

```
rep(1:3, length.out=12) # recycle 1,2,3 to get 12 values
```

```
## [1] 1 2 3 1 2 3 1 2 3 1 2 3
```

Therefore:

```
1:6 * c(1)
## [1] 1 2 3 4 5 6
1:6 * c(1,10)
## [1] 1 20 3 40 5 60
1:6 * c(1,10,100)
## [1] 1 20 300 4 50 600
1:6 * c(1,10,100,1000)
```

```
## Warning in 1:6 * c(1, 10, 100, 1000): longer object length is not a
## multiple of shorter object length
```

```
## [1] 1 20 300 4000 5 60
```

Note that a warning is not an error – we still get a result that makes sense.

C.2.4 Aggregation Functions

R implements a couple of *aggregation* functions:

- `sum(x) = sum $\sum_{i=1}^n x_i = x_1 + x_2 + \dots + x_n$`
- `prod(x) = prod $\prod_{i=1}^n x_i = x_1 x_2 \dots x_n$`
- `mean(x) = mean $\frac{1}{n} \sum_{i=1}^n x_i$` – arithmetic mean
- `var(x) = var $((x - mean(x))^2) / (length(x) - 1) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \frac{1}{n} \sum_{j=1}^n x_j)^2$` – variance
- `sd(x) = sqrt(var(x))` – standard deviation

see also: `min()`, `max()`, `median()`, `quantile()`.

Remark. Remember that you can always access the R manual by typing `?functionname`, e.g., `?quantile`.

Remark. Note that $\sum_{i=1}^n x_i$ can also be written as $\sum_{i=1}^n x_i$ or even $\sum_{i=1,\dots,n} x_i$. These all mean the sum of x_i for i from 1 to n , that is, the sum of x_1, x_2, \dots, x_n , i.e., $x_1 + x_2 + \dots + x_n$.

```
x <- runif(1000)
mean(x)
```

```
## [1] 0.49728
median(x)
```

```
## [1] 0.48995
min(x)

## [1] 0.00046535
max(x)
```

```
## [1] 0.9994
```

C.2.5 Special Functions

Furthermore, R supports numerous mathematical functions, e.g., `sqrt()`, `abs()`, `round()`, `log()`, `exp()`, `cos()`, `sin()`. All of them are vectorised – when applied on a vector of length n , they yield a vector of length n in result.

For example, here is how we can compute the square roots of all the integers between 1 and 9:

```
sqrt(1:9)
```

```
## [1] 1.0000 1.4142 1.7321 2.0000 2.2361 2.4495 2.6458 2.8284 3.0000
```

Vectorisation is super-convenient when it comes to, for instance, plotting (see Figure C.2).

```
x <- seq(-2*pi, 6*pi, length.out=51)
```

```
plot(x, sin(x), type="l")
lines(x, cos(x), col="red") # add a curve to the current plot
```

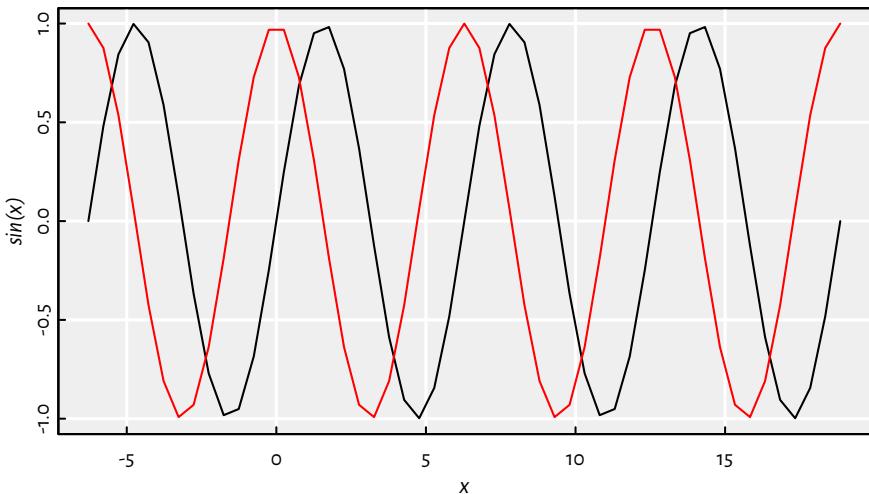


Figure C.2: An example plot of the sine and cosine functions

Exercise C.1 Try increasing the `length.out` argument to make the curves smoother.

C.2.6 Norms and Distances

Norms are used to measure the size of an object. Mathematically, we will also be interested in the following norms:

- Euclidean norm:

$$\|\mathbf{x}\| = \|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$$

this is nothing else than the `length` of the vector \mathbf{x}

- Manhattan (taxicab) norm:

$$\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|$$

- Chebyshev (maximum) norm:

$$\|\mathbf{x}\|_\infty = \max_{i=1,\dots,n} |x_i| = \max\{|x_1|, |x_2|, \dots, |x_n|\}$$

The above norms can be easily implemented by means of the building blocks introduced above. This is super easy:

```

z <- c(1, 2)
sqrt(sum(z^2)) # or norm(z, "2"); Euclidean

## [1] 2.2361

sum(abs(z))    # Manhattan

## [1] 3

max(abs(z))    # Chebyshev

## [1] 2

```

Also note that all the norms easily generate the corresponding *distances* (metrics) between two given points. In particular:

$$\|\mathbf{x} - \mathbf{y}\| = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

gives the *Euclidean distance* (metric) between the two vectors.

```

u <- c(1, 0)
v <- c(1, 1)
sqrt(sum((u-v)^2))

## [1] 1

```

This is the “normal” distance that you learned at school.

C.2.7 Dot Product (*)

What is more, given two vectors of identical lengths, \mathbf{x} and \mathbf{y} , we define their *dot product* (a.k.a. *scalar* or *inner product*) as:

$$\mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^n x_i y_i.$$

Let’s stress that this is not the same as the elementwise vector multiplication in R – the result is a single number.

```

u <- c(1, 0)
v <- c(1, 1)
sum(u*v)

## [1] 1

```

Remark. (*) Note that the squared Euclidean norm of a vector is equal to the dot product of the vector and itself, $\|\mathbf{x}\|^2 = \mathbf{x} \cdot \mathbf{x}$.

(*) Interestingly, a dot product has a nice geometrical interpretation:

$$\mathbf{x} \cdot \mathbf{y} = \|\mathbf{x}\| \|\mathbf{y}\| \cos \alpha$$

where α is the angle between the two vectors. In other words, it is the product of the lengths of the two vectors and the cosine of the angle between them. Note that we can get the cosine part by computing the dot product of the *normalised* vectors, i.e., such that their lengths are equal to 1.

For example, the two vectors \mathbf{u} and \mathbf{v} defined above can be depicted as in Figure C.3.

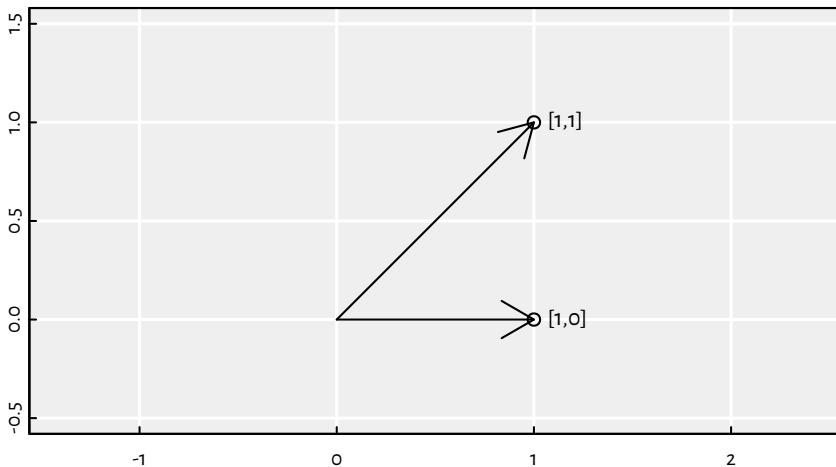


Figure C.3: Example vectors in 2D

We can compute the angle between them by calling:

```
(len_u <- sqrt(sum(u^2))) # length == Euclidean norm
## [1] 1
(len_v <- sqrt(sum(v^2)))
## [1] 1.4142
(cos_angle_uv <- (sum(u*v)/(len_u*len_v))) # cosine of the angle
## [1] 0.70711
acos(cos_angle_uv)*180/pi # angle in degs
## [1] 45
```

C.2.8 Missing and Other Special Values

R has a notion of a missing (not-available) value. It is very useful in data analysis, as we sometimes don't have an information on an object's feature. For instance, we might not know a patient's age if he was admitted to the hospital unconscious.

```
x <- c(1, 2, NA, 4, 5)
```

Operations on missing values generally result in missing values – that makes a lot sense.

```
x + 11:15
```

```
## [1] 12 14 NA 18 20
```

```
mean(x)
```

```
## [1] NA
```

If we wish to compute a vector's aggregate after all, we can get rid of the missing values by calling `na.omit()`:

```
mean(na.omit(x)) # mean of non-missing values
```

```
## [1] 3
```

We can also make use of the `na.rm` parameter of the `mean()` function (however, not every aggregation function has it – always refer to documentation).

```
mean(x, na.rm=TRUE)
```

```
## [1] 3
```

Remark. Note that in R, a dot has no special meaning. `na.omit` is as good of a function's name or variable identifier as `na_omit`, `naOmit`, `NAOMIT`, `naomit` and `NaOmit`. Note that, however, R is a case-sensitive language – these are all different symbols. Read more in the *Details* section of `?make.names`.

Moreover, some arithmetic operations can result in infinities ($\pm\infty$):

```
log(0)
```

```
## [1] -Inf
```

```
10^1000 # too large
```

```
## [1] Inf
```

Also, sometimes we'll get a *not-a-number*, `NaN`. This is not a missing value, but a “invalid” result.

```
sqrt(-1)
```

```
## Warning in sqrt(-1): NaNs produced
```

```
## [1] NaN
```

```
log(-1)

## Warning in log(-1): NaNs produced
## [1] NaN
Inf-Inf
## [1] NaN
```

C.3 Logical Vectors

C.3.1 Creating Logical Vectors

In R there are 3 (!) logical values: TRUE, FALSE and geez, I don't know, NA maybe?

```
c(TRUE, FALSE, TRUE, NA, FALSE, FALSE, TRUE)
```

```
## [1] TRUE FALSE TRUE NA FALSE FALSE TRUE
(x <- rep(c(TRUE, FALSE, NA), 2))
```

```
## [1] TRUE FALSE NA TRUE FALSE NA
mode(x)
```

```
## [1] "logical"
```

```
class(x)
```

```
## [1] "logical"
```

```
length(x)
```

```
## [1] 6
```

Remark. By default, T is a synonym for TRUE and F for FALSE. This may be changed though so it's better not to rely on these.

C.3.2 Logical Operations

Logical operators such as & (and) and | (or) are performed in the same manner as arithmetic ones, i.e.:

- they are elementwise operations and
- recycling rule is applied if necessary.

For example,

```
TRUE & TRUE
## [1] TRUE
TRUE & c(TRUE, FALSE)

## [1] TRUE FALSE
c(FALSE, FALSE, TRUE, TRUE) | c(TRUE, FALSE, TRUE, FALSE)

## [1] TRUE FALSE TRUE TRUE
```

The `!` operator stands for logical elementwise negation:

```
!c(TRUE, FALSE)
```

```
## [1] FALSE TRUE
```

Generally, operations on NAs yield NA unless other solution makes sense.

```
u <- c(TRUE, FALSE, NA)
v <- c(TRUE, TRUE, TRUE, FALSE, FALSE, NA, NA, NA)
u & v # elementwise AND (conjunction)

## [1] TRUE FALSE NA FALSE FALSE FALSE NA FALSE NA
u | v # elementwise OR (disjunction)

## [1] TRUE TRUE TRUE TRUE FALSE NA TRUE NA NA
!u # elementwise NOT (negation)

## [1] FALSE TRUE NA
```

C.3.3 Comparison Operations

We can compare the corresponding elements of two numeric vectors and get a logical vector in result. Operators such as `<` (less than), `<=` (less than or equal), `==` (equal), `!=` (not equal), `>` (greater than) and `>=` (greater than or equal) are again elementwise and use the recycling rule if necessary.

```
3 < 1:5 # c(3, 3, 3, 3, 3) < c(1, 2, 3, 4, 5)

## [1] FALSE FALSE FALSE TRUE TRUE

1:2 == 1:4 # c(1,2,1,2) == c(1,2,3,4)

## [1] TRUE TRUE FALSE FALSE

z <- c(0, 3, -1, 1, 0.5)
(z >= 0) & (z <= 1)

## [1] TRUE FALSE FALSE TRUE TRUE
```

C.3.4 Aggregation Functions

Also note the following operations on *logical* vectors:

```
z <- 1:10
all(z >= 5) # are all values TRUE?
```

```
## [1] FALSE
any(z >= 5) # is there any value TRUE?
```

```
## [1] TRUE
```

Moreover:

```
sum(z >= 5) # how many TRUE values are there?
```

```
## [1] 6
mean(z >= 5) # what is the proportion of TRUE values?
```

```
## [1] 0.6
```

The behaviour of `sum()` and `mean()` is dictated by the fact that, when interpreted in numeric terms, `TRUE` is interpreted as numeric 1 and `FALSE` as 0.

```
as.numeric(c(FALSE, TRUE))
```

```
## [1] 0 1
```

Therefore in the example above we have:

```
z >= 5
## [1] FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
as.numeric(z >= 5)

## [1] 0 0 0 0 1 1 1 1 1
sum(as.numeric(z >= 5)) # the same as sum(z >= 5)

## [1] 6
```

Yes, there are 6 values equal to `TRUE` (or 6 ones after conversion), the sum of zeros and ones gives the number of ones.

C.4 Character Vectors

C.4.1 Creating Character Vectors

Individual character strings can be created using double quotes or apostrophes. These are the elements of character vectors

```
(x <- "a string")  
  
## [1] "a string"  
mode(x)  
  
## [1] "character"  
class(x)  
  
## [1] "character"  
length(x)  
  
## [1] 1  
rep(c("aaa", 'bb', "c"), 2)  
  
## [1] "aaa" "bb" "c" "aaa" "bb" "c"
```

C.4.2 Concatenating Character Vectors

To join (concatenate) the corresponding elements of two or more character vectors, we call the `paste()` function:

```
paste(c("a", "b", "c"), c("1", "2", "3"))  
  
## [1] "a 1" "b 2" "c 3"  
paste(c("a", "b", "c"), c("1", "2", "3"), sep="")  
  
## [1] "a1" "b2" "c3"
```

Also note:

```
paste(c("a", "b", "c"), 1:3) # the same as as.character(1:3)  
  
## [1] "a 1" "b 2" "c 3"  
paste(c("a", "b", "c"), 1:6) # recycling  
  
## [1] "a 1" "b 2" "c 3" "a 4" "b 5" "c 6"  
paste(c("a", "b", "c"), 1:6, c("!", "?"))
```

```
## [1] "a 1 !" "b 2 ?" "c 3 !" "a 4 ?" "b 5 !" "c 6 ?"
```

C.4.3 Collapsing Character Vectors

We can also collapse a sequence of strings to a single string:

```
paste(c("a", "b", "c", "d"), collapse="")
```

```
## [1] "abcd"
```

```
paste(c("a", "b", "c", "d"), collapse=",")
```

```
## [1] "a,b,c,d"
```

C.5 Vector Subsetting

C.5.1 Subsetting with Positive Indices

In order to extract subsets (parts) of vectors, we use the square brackets:

```
(x <- seq(10, 100, 10))
```

```
## [1] 10 20 30 40 50 60 70 80 90 100
```

```
x[1] # the first element
```

```
## [1] 10
```

```
x[length(x)] # the last element
```

```
## [1] 100
```

More than one element at a time can also be extracted:

```
x[1:3] # the first three
```

```
## [1] 10 20 30
```

```
x[c(1, length(x))] # the first and the last
```

```
## [1] 10 100
```

For example, the `order()` function returns the indices of the smallest, 2nd smallest, 3rd smallest, ..., the largest element in a given vector. We will use this function when implementing our first classifier.

```
y <- c(50, 30, 10, 20, 40)
```

```
(o <- order(y))
```

```
## [1] 3 4 2 5 1
```

Hence, we see that the smallest element in y is at index 3 and the largest at index 1:

```
y[o[1]]  
## [1] 10  
y[o[length(y)]]  
## [1] 50
```

Therefore, to get a sorted version of y , we call:

```
y[o] # see also sort(y)  
## [1] 10 20 30 40 50  
We can also obtain the 3 largest elements by calling:  
y[order(y, decreasing=TRUE)[1:3]]  
## [1] 50 40 30
```

C.5.2 Subsetting with Negative Indices

Subsetting with a vector of negative indices, *excludes* the elements at given positions:

```
x[-1] # all but the first  
## [1] 20 30 40 50 60 70 80 90 100  
x[-(1:3)]  
## [1] 40 50 60 70 80 90 100  
x[-c(1:3, 5, 8)]  
## [1] 40 60 70 90 100
```

C.5.3 Subsetting with Logical Vectors

We may also subset a vector x of length n with a logical vector l also of length n . The i -th element, x_i , will be extracted if and only if the corresponding l_i is true.

```
x[c(TRUE, FALSE, FALSE, FALSE, TRUE, FALSE, TRUE, TRUE, FALSE)]  
## [1] 10 50 70 80 100  
This gets along nicely with comparison operators that yield logical vectors on output.  
x>50  
## [1] FALSE FALSE FALSE FALSE FALSE TRUE TRUE TRUE TRUE TRUE
```

```
x[x>50] # select elements in x that are greater than 50
## [1] 60 70 80 90 100
x[x<30 | x>70]
## [1] 10 20 80 90 100
x[x<max(x)] # getting rid of the greatest element
## [1] 10 20 30 40 50 60 70 80 90
x[x > min(x) & x < max(x)] # return all but the smallest and greatest one
## [1] 20 30 40 50 60 70 80 90
```

Of course, e.g., `x[x<max(x)]` returns a new, independent object. In order to remove the greatest element in `x` permanently, we can write `x <- x[x<max(x)]`.

C.5.4 Replacing Elements

Note that the three above vector indexing schemes (positive, negative, logical indices) allow for replacing specific elements with new values.

```
x[-1] <- 10000
x
## [1] 10 10000 10000 10000 10000 10000 10000 10000 10000 10000
x[(-(1:7))] <- c(1, 2, 3)
x
## [1] 10 10000 10000 10000 10000 10000 10000 10000 10000 10000
## [1] 1 2 3
```

C.5.5 Other Functions

`head()` and `tail()` return, respectively, a few (6 by default) first and last elements of a vector.

```
head(x) # head(x, 6)
## [1] 10 10000 10000 10000 10000 10000
tail(x, 3)
## [1] 1 2 3
```

Sometimes the `which()` function can come in handy. For a given logical vector, it returns all the indices where TRUE elements are stored.

```
which(c(TRUE, FALSE, TRUE, TRUE, FALSE, FALSE, TRUE))
## [1] 1 3 4 7
```

```
print(y) # recall

## [1] 50 30 10 20 40
which(y>30)

## [1] 1 5
```

Note that `y[y>70]` gives the same result as `y[which(y>70)]` but is faster (because it involves less operations).

`which.min()` and `which.max()` return the index of the smallest and the largest element, respectively:

```
which.min(y) # where is the minimum?
```

```
## [1] 3
which.max(y)

## [1] 1
y[which.min(y)] # min(y)

## [1] 10
```

`is.na()` indicates which elements are missing values (NAs):

```
z <- c(1, 2, NA, 4, NA, 6)
is.na(z)
```

```
## [1] FALSE FALSE TRUE FALSE TRUE FALSE
```

Therefore, to remove them from `z` permanently, we can write (compare `na.omit()`, see also `is.finite()`):

```
(z <- z[!is.na(z)])
```

```
## [1] 1 2 4 6
```

C.6 Named Vectors

C.6.1 Creating Named Vectors

Vectors in R can be *named* – each element can be assigned a string label.

```
x <- c(20, 40, 99, 30, 10)
names(x) <- c("a", "b", "c", "d", "e")
x # a named vector
```

```
## a b c d e
## 20 40 99 30 10
```

Other ways to create named vectors include:

```
c(a=1, b=2, c=3)

## a b c
## 1 2 3

structure(1:3, names=c("a", "b", "c"))

## a b c
## 1 2 3
```

For instance, the `summary()` function returns a named vector:

```
summary(x) # NAMED vector, we don't want this here yet
```

```
##      Min. 1st Qu. Median     Mean 3rd Qu.    Max.
##      10.0    20.0   30.0    39.8   40.0    99.0
```

This gives the minimum, 1st quartile (25%-quantile), Median (50%-quantile), arithmetic mean, 3rd quartile (75%-quantile) and maximum.

Note that `x` is still a numeric vector, we can perform various operations on it as usual:

```
sum(x)
```

```
## [1] 199
```

```
x[x>3]
```

```
## a b c d e
## 20 40 99 30 10
```

Names can be dropped by calling:

```
unname(x)
```

```
## [1] 20 40 99 30 10
```

```
as.numeric(x) # we need to know the type of x though
```

```
## [1] 20 40 99 30 10
```

C.6.2 Subsetting Named Vectors with Character String Indices

It turns out that extracting elements from a named vector can *also* be performed by means of a vector of character string indices:

```
x[c("a", "d", "b")]
```

```
## a d b
```

```
## 20 30 40  
summary(x)[c("Median", "Mean")]  
  
## Median Mean  
## 30.0 39.8
```

C.7 Factors

Factors are *special* kinds of vectors that are frequently used to store qualitative data, e.g., species, groups, types. Factors are convenient in situations where we have many observations, but the number of distinct (unique) values is relatively small.

C.7.1 Creating Factors

For example, the following character vector:

```
(col <- sample(c("blue", "red", "green"), replace=TRUE, 10))  
  
## [1] "green" "green" "green" "red"    "green" "red"    "red"    "red"  
## [9] "green" "blue"
```

can be converted to a factor by calling:

```
(fcol <- factor(col))  
  
## [1] green green green red    green red    red    red    green blue  
## Levels: blue green red
```

Note how different is the way factors are printed out on the console.

C.7.2 Levels

We can easily obtain the list unique labels:

```
levels(fcol)  
  
## [1] "blue"  "green" "red"
```

Those can be re-encoded by calling:

```
levels(fcol) <- c("b", "g", "r")  
fcol  
  
## [1] g g g r g r r g b  
## Levels: b g r
```

To create a contingency table (in the form of a named numeric vector, giving how many values are at each factor level), we call:

```
table(fcol)
```

```
## fcol
## b g r
## 1 5 4
```

C.7.3 Internal Representation (*)

Factors have a look-and-feel of character vectors, however, internally they are represented as integer sequences.

```
class(fcol)
```

```
## [1] "factor"
```

```
mode(fcol)
```

```
## [1] "numeric"
```

```
as.numeric(fcol)
```

```
## [1] 2 2 2 3 2 3 3 3 2 1
```

These are always integers from 1 to M inclusive, where M is the number of levels. Their meaning is given by the `levels()` function: in the example above, the meaning of the codes 1, 2, 3 is, respectively, b, g, r.

If we wished to generate a factor with a specific order of labels, we could call:

```
factor(col, levels=c("red", "green", "blue"))
```

```
## [1] green green green red   green red   red   red   green blue
## Levels: red green blue
```

We can also assign different labels upon creation of a factor:

```
factor(col, levels=c("red", "green", "blue"), labels=c("r", "g", "b"))
```

```
## [1] g g g r g r r g b
## Levels: r g b
```

Knowing how factors are represented is important when we deal with factors that are built around data that *look like* numeric. This is because their conversion to numeric gives the internal codes, not the actual values:

```
(f <- factor(c(1, 3, 0, 1, 4, 0, 0, 1, 4)))
```

```
## [1] 1 3 0 1 4 0 0 1 4
```

```
## Levels: 0 1 3 4
```

```
as.numeric(f) # not necessarily what we want here

## [1] 2 3 1 2 4 1 1 2 4

as.numeric(as.character(f)) # much better

## [1] 1 3 0 1 4 0 0 1 4
```

Moreover, that idea is labour-saving in contexts such as plotting of data that are grouped into different classes. For instance, here is a scatter plot for the Sepal.Length and Petal.Width variables in the `iris` dataset (which is an object of type `data.frame`, see below). Flowers are of different Species, and we wish to indicate which point belongs to which class:

```
which_preview <- c(1, 11, 51, 69, 101) # indexes we show below
iris$Sepal.Length[which_preview]
```

```
## [1] 5.1 5.4 7.0 6.2 6.3
```

```
iris$Petal.Width[which_preview]
```

```
## [1] 0.2 0.2 1.4 1.5 2.5
```

```
iris$Species[which_preview]
```

```
## [1] setosa      setosa      versicolor versicolor virginica
```

```
## Levels: setosa versicolor virginica
```

```
as.numeric(iris$Species)[which_preview]
```

```
## [1] 1 1 2 2 3
```

```
plot(iris$Sepal.Length, # x (it's a vector)
     iris$Petal.Width, # y (it's a vector)
     col=as.numeric(iris$Species), # colours
     pch=as.numeric(iris$Species))
)
```

The above (see Figure C.4) was possible because the Species column is a factor object with:

```
levels(iris$Species)
```

```
## [1] "setosa"      "versicolor"   "virginica"
```

and the meaning of `pch` of 1, 2, 3, ... is “circle”, “triangle”, “plus”, ..., respectively. What’s more, there’s a default palette that maps consecutive integers to different colours:

```
palette()
```

```
## [1] "black"       "#DF536B"     "#61D04F"     "#2297E6"     "#28E2E5"     "#CD0BBC"
## [7] "#F5C710"    "gray62"
```

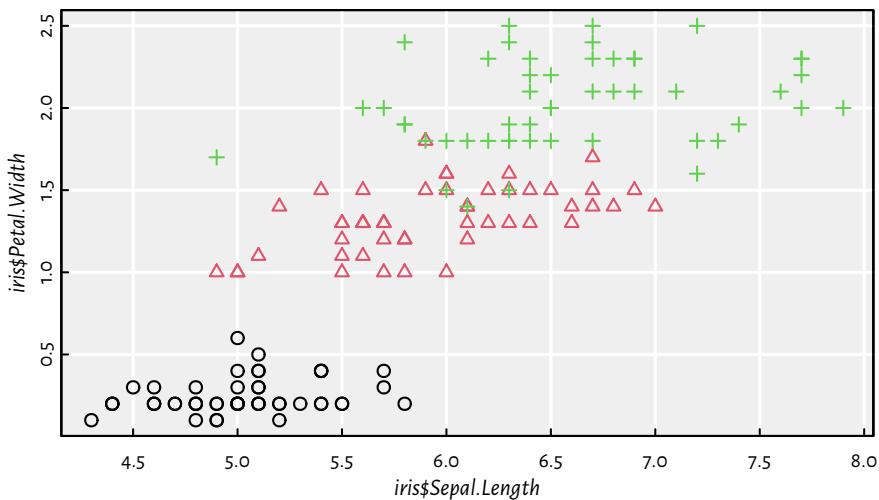


Figure C.4: `as.numeric()` on factors can be used to create different plotting styles

Hence, black circles mark irises from the 1st class, i.e., “setosa”.

C.8 Lists

Numeric, logical and character vectors are *atomic* objects – each component is of the same type. Let’s take a look at what happens when we create an atomic vector out of objects of different types:

```
c("nine", FALSE, 7, TRUE)
```

```
## [1] "nine"   "FALSE"   "7"      "TRUE"
```

```
c(FALSE, 7, TRUE, 7)
```

```
## [1] 0 7 1 7
```

In each case, we get an object of the most “general” type which is able to represent our data.

On the other hand, R *lists* are *generalised* vectors. They can consist of arbitrary R objects, possibly of mixed types – also other lists.

C.8.1 Creating Lists

Most commonly, we create a generalised vector by calling the `list()` function.

```
(l <- list(1:5, letters, runif(3)))  
## [[1]]  
## [1] 1 2 3 4 5  
##  
## [[2]]  
## [1] "a" "b" "c" "d" "e" "f" "g" "h" "i" "j" "k" "l" "m" "n" "o" "p" "q"  
## [18] "r" "s" "t" "u" "v" "w" "x" "y" "z"  
##  
## [[3]]  
## [1] 0.95683 0.45333 0.67757  
mode(l)  
## [1] "list"  
class(l)  
## [1] "list"  
length(l)  
## [1] 3
```

There's a more compact way to print a list on the console:

```
str(l)  
## List of 3  
## $ : int [1:5] 1 2 3 4 5  
## $ : chr [1:26] "a" "b" "c" "d" ...  
## $ : num [1:3] 0.957 0.453 0.678
```

We can also convert an atomic vector to a list by calling:

```
as.list(1:3)  
## [[1]]  
## [1] 1  
##  
## [[2]]  
## [1] 2  
##  
## [[3]]  
## [1] 3
```

C.8.2 Named Lists

List, like other vectors, may be assigned a `names` attribute.

```

names(l) <- c("a", "b", "c")
l

## $a
## [1] 1 2 3 4 5
##
## $b
## [1] "a" "b" "c" "d" "e" "f" "g" "h" "i" "j" "k" "l" "m" "n" "o" "p" "q"
## [18] "r" "s" "t" "u" "v" "w" "x" "y" "z"
##
## $c
## [1] 0.95683 0.45333 0.67757

```

C.8.3 Subsetting and Extracting From Lists

Applying a square brackets operator creates a sub-list, which is of type list as well.

```

l[-1]

## $b
## [1] "a" "b" "c" "d" "e" "f" "g" "h" "i" "j" "k" "l" "m" "n" "o" "p" "q"
## [18] "r" "s" "t" "u" "v" "w" "x" "y" "z"
##
## $c
## [1] 0.95683 0.45333 0.67757

l[c("a", "c")]

## $a
## [1] 1 2 3 4 5
##
## $c
## [1] 0.95683 0.45333 0.67757

l[1]

## $a
## [1] 1 2 3 4 5

Note in the 3rd case we deal with a list of length one, not a numeric vector.

To extract (dig into) a particular (single) element, we use double square brackets:

l[[1]]

## [1] 1 2 3 4 5

l[["c"]]

## [1] 0.95683 0.45333 0.67757

```

The latter can equivalently be written as:

```
l$c
```

```
## [1] 0.95683 0.45333 0.67757
```

C.8.4 Common Operations

Lists, because of their generality (they can store any kind of object), have few dedicated operations. In particular, it neither makes sense to add, multiply, ... two lists together nor to aggregate them.

However, if we wish to run some operation on each element, we can call list-apply:

```
(k <- list(x=runif(5), y=runif(6), z=runif(3))) # a named list
```

```
## $x
## [1] 0.57263 0.10292 0.89982 0.24609 0.04206
##
## $y
## [1] 0.32792 0.95450 0.88954 0.69280 0.64051 0.99427
##
## $z
## [1] 0.65571 0.70853 0.54407

lapply(k, mean)

## $x
## [1] 0.37271
##
## $y
## [1] 0.74992
##
## $z
## [1] 0.6361
```

The above computes the mean of each of the three numeric vectors stored inside list k. Moreover:

```
lapply(k, range)
```

```
## $x
## [1] 0.04206 0.89982
##
## $y
## [1] 0.32792 0.99427
##
## $z
## [1] 0.54407 0.70853
```

The built-in function `range(x)` returns `c(min(x), max(x))`.

`unlist()` tries (it might not always be possible) to unwind a list to a simpler, atomic form:

```
unlist(lapply(k, mean))
```

```
##      x      y      z
## 0.37271 0.74992 0.63610
```

Moreover, `split(x, f)` classifies elements in a vector `x` into subgroups defined by a factor (or an object coercible to) of the same length.

```
x <- c( 1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
f <- c("a", "b", "a", "a", "c", "b", "b", "a", "a", "b")
split(x, f)
```

```
## $a
## [1] 1 3 4 8 9
##
## $b
## [1] 2 6 7 10
##
## $c
## [1] 5
```

This is very useful when combined with `lapply()` and `unlist()`. For instance, here are the mean sepal lengths for each of the three flower species in the famous `iris` dataset.

```
unlist(lapply(split(iris$Sepal.Length, iris$Species), mean))
```

```
##    setosa versicolor  virginica
##      5.006      5.936      6.588
```

By the way, if we take a look at the documentation of `?lapply`, we will note that that this function is defined as `lapply(X, FUN, ...)`. Here `...` denotes the optional arguments that will be passed to `FUN`.

In other words, `lapply(X, FUN, ...)` returns a list `Y` of length `length(X)` such that `Y[[i]] <- FUN(X[[i]], ...)` for each `i`. For example, `mean()` has an additional argument `na.rm` that aims to remove missing values from the input vector. Compare the following:

```
t <- list(1:10, c(1, 2, NA, 4, 5))
unlist(lapply(t, mean))
```

```
## [1] 5.5  NA
unlist(lapply(t, mean, na.rm=TRUE))
```

```
## [1] 5.5 3.0
```

Of course, we can always pass a custom (self-made) function object as well:

```

min_mean_max <- function(x) {
  # the last expression evaluated in the function's body
  # gives its return value:
  c(min(x), mean(x), max(x))
}
lapply(k, min_mean_max)

## $x
## [1] 0.04206 0.37271 0.89982
##
## $y
## [1] 0.32792 0.74992 0.99427
##
## $z
## [1] 0.54407 0.63610 0.70853

```

or, more concisely (we can skip the curly braces here – they are normally used to group many expressions into one; also, if we don't plan to re-use the function again, there's no need to give it a name):

```

lapply(k, function(x) c(min(x), mean(x), max(x)))

## $x
## [1] 0.04206 0.37271 0.89982
##
## $y
## [1] 0.32792 0.74992 0.99427
##
## $z
## [1] 0.54407 0.63610 0.70853

```

C.9 Exercises

C.9.1 AUD/EUR Exchange Rates

Let's load the dataset on historical daily currency exchange rates of EUR/AUD, i.e., the Euro as base currency quoted relative to the Australian dollar between 4 January 1999 and 19 May 2020:

```

x <- scan("datasets/currency_exchange_single.csv")
head(x)

## [1] 1.9100 1.8944 1.8820 1.8474 1.8406 1.8134

```

The dataset is described in more detail in Appendix F and is available for download from the book's homepage.

Exercise C.2 Plot the daily prices by calling `plot(x)`. Set the `xlab` and `ylab` arguments to "day" and "USD to AUD", respectively.

Exercise C.3 Load the `pracma` package by calling `library("pracma")`. If you haven't installed the package yet, call `install.packages("pracma")` from the R console.

Exercise C.4 A call to `movavg(x, n)` returns the n -moving average of a numeric vector x . This transformation is frequently used to smoothen noisy data, especially time series such as the one we are dealing with here. Create a plot of the 5-moving average (`movavg(x, 5)`).

Exercise C.5 Create a plot of the 5, 15 and 90-moving averages, all in a single figure. The `plot()` function can be used to draw the first curve, and then new ones can be added by calling `lines()`.

Exercise C.6 The dataset contains missing values (NAs). Remove all missing observations before computing the moving averages and visually assess how this changes the corresponding figures.

Exercise C.7 Compute the minimum, maximum, median and the arithmetic mean of x for (a) the whole series, (b) the last 30 days only, (c) the first 30 days only.

Exercise C.8 A call to `d <- diff(x)` computes a vector of relative daily changes in exchange rates ($d[i] = x[i+1] - x[i]$). Draw the box and whisker plot and histogram by calling `boxplot(d)` and `hist(d)`, respectively. In your own words, explain what information can be read from these figures.

Exercise C.9 Perform a similar data analysis for the AUD/EUR rates.

C.10 Further Reading

Remember to read about all the functions mentioned in this chapter in their respective manual pages, e.g., `?seq`, `?c` etc.

Recommended further reading: (Venables et al. 2020)

Other: (Deisenroth et al. 2020), (Peng 2019), (Wickham & Grolemund 2017)

TODO

Next chapter...

D

Matrix Algebra in R

TODO In this chapter, we will:

- ...
- ...

Vectors are one-dimensional objects – they represent “flat” sequences of values. Matrices, on the other hand, are two-dimensional – they represent tabular data, where values aligned into rows and columns. Matrices (and their extensions – data frames, which we’ll cover in the next chapter) are predominant in data science, where objects are typically represented by means of feature vectors.

Below are some examples of structured datasets in matrix forms.

```
head(as.matrix(iris[,1:4]))
```

```
##      Sepal.Length Sepal.Width Petal.Length Petal.Width
## [1,]         5.1       3.5        1.4       0.2
## [2,]         4.9       3.0        1.4       0.2
## [3,]         4.7       3.2        1.3       0.2
## [4,]         4.6       3.1        1.5       0.2
## [5,]         5.0       3.6        1.4       0.2
## [6,]         5.4       3.9        1.7       0.4
```

```
WorldPhones
```

```
##      N.Amer Europe Asia S.Amer Oceania Africa Mid.Amer
## 1951  45939 21574 2876   1815    1646     89     555
## 1956  60423 29990 4708   2568    2366   1411     733
## 1957  64721 32510 5230   2695    2526   1546     773
## 1958  68484 35218 6662   2845    2691   1663     836
## 1959  71799 37598 6856   3000    2868   1769     911
## 1960  76036 40341 8220   3145    3054   1905    1008
## 1961  79831 43173 9053   3338    3224   2005    1076
```

The aim of this chapter is to cover the most essential matrix operations, both from the computational perspective and the mathematical one.

D.1 Creating Matrices

D.1.1 `matrix()`

A matrix can be created – amongst others – with a call to the `matrix()` function.

```
(A <- matrix(c(1, 2, 3, 4, 5, 6), byrow=TRUE, nrow=2))
```

```
##      [,1] [,2] [,3]
## [1,]    1    2    3
## [2,]    4    5    6
class(A)
```

```
## [1] "matrix" "array"
```

Given a numeric vector of length 6, we've asked R to convert to a numeric matrix with 2 rows (the `nrow` argument). The number of columns has been deduced automatically (otherwise, we would additionally have to pass `ncol=3` to the function).

Using mathematical notation, above we have defined $\mathbf{A} \in \mathbb{R}^{2 \times 3}$:

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$$

We can fetch the size of the matrix by calling:

```
dim(A) # number of rows, number of columns
```

```
## [1] 2 3
```

We can also “promote” a “flat” vector to a column vector, i.e., a matrix with one column by calling:

```
as.matrix(1:3)
```

```
##      [,1]
## [1,]    1
## [2,]    2
## [3,]    3
```

D.1.2 Stacking Vectors

Other ways to create a matrix involve stacking a couple of vectors of equal lengths along each other:

```
rbind(1:3, 4:6, 7:9) # row bind
```

```
##      [,1] [,2] [,3]
```

```
## [1,]    1    2    3
## [2,]    4    5    6
## [3,]    7    8    9
cbind(1:3, 4:6, 7:9) # column bind
```

```
##      [,1] [,2] [,3]
## [1,]    1    4    7
## [2,]    2    5    8
## [3,]    3    6    9
```

These functions also allow for adding new rows/columns to existing matrices:

```
rbind(A, c(-1, -2, -3))
```

```
##      [,1] [,2] [,3]
## [1,]    1    2    3
## [2,]    4    5    6
## [3,]   -1   -2   -3
```

```
cbind(A, c(-1, -2))
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    1    2    3   -1
## [2,]    4    5    6   -2
```

D.1.3 Beyond Numeric Matrices

Note that logical matrices are possible as well. For instance, knowing that comparison such as `<` and `==` are performed elementwise also in the case of matrices, we can obtain:

```
A >= 3
```

```
##      [,1] [,2] [,3]
## [1,] FALSE FALSE TRUE
## [2,] TRUE  TRUE TRUE
```

Moreover, although much more rarely used, we can define character matrices:

```
matrix(LETTERS[1:12], ncol=6)
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,] "A"  "C"  "E"  "G"  "I"  "K"
## [2,] "B"  "D"  "F"  "H"  "J"  "L"
```

D.1.4 Naming Rows and Columns

Just like vectors could be equipped with `names` attribute:

```
c(a=1, b=2, c=3)
```

```
## a b c
## 1 2 3
```

matrices can be assigned row and column labels in the form of a list of two character vectors:

```
dimnames(A) <- list(
  c("a", "b"),      # row labels
  c("x", "y", "z") # column labels
)
A

##   x y z
## a 1 2 3
## b 4 5 6
```

D.1.5 Other Methods

The `read.table()` (and its special case, `read.csv()`), can be used to read a matrix from a text file. We will cover it in the next chapter, because technically it returns a data frame object (which we can convert to a matrix with a call to `as.matrix()`).

`outer()` applies a given (vectorised) function on each pair of elements from two vectors, forming a two-dimensional “grid”. More precisely `outer(x, y, f, ...)` returns a matrix Z with `length(x)` rows and `length(y)` columns such that $z_{i,j} = f(x_i, y_j, \dots)$, where \dots are optional further arguments to `f`.

```
outer(c(1, 10, 100), 1:5, "*") # apply the multiplication operator
```

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]    1    2    3    4    5
## [2,]   10   20   30   40   50
## [3,]  100  200  300  400  500

outer(c("A", "B"), 1:8, paste, sep="-") # concatenate strings
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
## [1,] "A-1" "A-2" "A-3" "A-4" "A-5" "A-6" "A-7" "A-8"
## [2,] "B-1" "B-2" "B-3" "B-4" "B-5" "B-6" "B-7" "B-8"
```

`simplify2array()` is an extension of the `unlist()` function. Given a list of vectors, each of length one, it will return an “unlisted” vector. However, if a list of equisized vectors of greater lengths is given, these will be converted to a matrix.

```
simplify2array(list(1, 11, 21))
```

```
## [1] 1 11 21
```

```

simplify2array(list(1:3, 11:13, 21:23))

##      [,1] [,2] [,3]
## [1,]    1   11   21
## [2,]    2   12   22
## [3,]    3   13   23

simplify2array(list(1, 11:12, 21:23)) # no can do

## [[1]]
## [1] 1
##
## [[2]]
## [1] 11 12
##
## [[3]]
## [1] 21 22 23

sapply(...) is a nice application of the above, meaning simplify2array(lapply(...)).  

sapply(split(iris$Sepal.Length, iris$Species), mean)

##      setosa versicolor  virginica
##      5.006      5.936      6.588

sapply(split(iris$Sepal.Length, iris$Species), summary)

##      setosa versicolor  virginica
## Min.   4.300   4.900   4.900
## 1st Qu. 4.800   5.600   6.225
## Median 5.000   5.900   6.500
## Mean   5.006   5.936   6.588
## 3rd Qu. 5.200   6.300   6.900
## Max.   5.800   7.000   7.900

```

Of course, custom functions can also be applied:

```

min_mean_max <- function(x) {
  # returns a named vector with three elements
  # (note that the last expression in a function's body
  # is its return value)
  c(min=min(x), mean=mean(x), max=max(x))
}

sapply(split(iris$Sepal.Length, iris$Species), min_mean_max)

##      setosa versicolor  virginica
## min   4.300   4.900   4.900
## mean  5.006   5.936   6.588
## max   5.800   7.000   7.900

```

Lastly, `table(x, y)` creates a contingency matrix that counts the number of unique pairs of corresponding elements from two vectors of equal lengths.

```
library("titanic") # data on the passengers of the RMS Titanic
table(titanic_train$Survived)

##
##    0    1
## 549 342

table(titanic_train$Sex)

##
## female male
##   314   577

table(titanic_train$Survived, titanic_train$Sex)

##
##      female male
##    0     81  468
##    1    233  109
```

D.1.6 Internal Representation (*)

Note that by setting `byrow=TRUE` in a call to the `matrix()` function above, we are reading the elements of the input vector in the row-wise (row-major) fashion. The default is the column-major order, which might be a little unintuitive for some of us.

```
A <- matrix(c(1, 2, 3, 4, 5, 6), ncol=3, byrow=TRUE)
B <- matrix(c(1, 2, 3, 4, 5, 6), ncol=3) # byrow=FALSE
```

It turns out that is exactly the order in which the matrix is stored internally. Under the hood, it is an ordinary numeric vector:

```
mode(B)      # == mode(A)

## [1] "numeric"
length(B)    # == length(A)

## [1] 6
as.numeric(A)

## [1] 1 4 2 5 3 6
as.numeric(B)

## [1] 1 2 3 4 5 6
```

Also note that we can create a different view on the same underlying data vector:

```
dim(A) <- c(3, 2) # 3 rows, 2 columns
A

##      [,1] [,2]
## [1,]    1    5
## [2,]    4    3
## [3,]    2    6

dim(B) <- c(3, 2) # 3 rows, 2 columns
B

##      [,1] [,2]
## [1,]    1    4
## [2,]    2    5
## [3,]    3    6
```

D.2 Common Operations

D.2.1 Matrix Transpose

The matrix *transpose* is denoted with \mathbf{A}^T :

```
t(A)
```

```
##      [,1] [,2] [,3]
## [1,]    1    4    2
## [2,]    5    3    6
```

Hence, $\mathbf{B} = \mathbf{A}^T$ is a matrix such that $b_{i,j} = a_{j,i}$.

In other words, in the transposed matrix, rows become columns and columns become rows. For example:

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \end{bmatrix} \quad \mathbf{A}^T = \begin{bmatrix} a_{1,1} & a_{2,1} \\ a_{1,2} & a_{2,2} \\ a_{1,3} & a_{2,3} \end{bmatrix}$$

D.2.2 Matrix-Scalar Operations

Operations such as $s\mathbf{A}$ (multiplication of a matrix by a scalar), $-\mathbf{A}$, $s + \mathbf{A}$ etc. are applied on each element of the input matrix:

```
(A <- matrix(c(1, 2, 3, 4, 5, 6), byrow=TRUE, nrow=2))
```

```
##      [,1] [,2] [,3]
## [1,]    1    2    3
```

```
## [2,]    4    5    6
(-1)*A
```

```
## [,1] [,2] [,3]
## [1,] -1   -2   -3
## [2,] -4   -5   -6
```

In R, the same rule holds when we compute other operations (despite the fact that, mathematically, e.g., A^2 or $A \geq 0$ might have a different meaning):

```
A^2 # this is not A-matrix-multiply-A, see below
```

```
## [,1] [,2] [,3]
## [1,]    1    4    9
## [2,]   16   25   36
```

```
A>=3
```

```
## [,1] [,2] [,3]
## [1,] FALSE FALSE TRUE
## [2,] TRUE  TRUE TRUE
```

D.2.3 Matrix-Matrix Operations

If $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times p}$ are two matrices of identical sizes, then $\mathbf{A} + \mathbf{B}$ and $\mathbf{A} - \mathbf{B}$ are understood elementwise, i.e., they result in $\mathbf{C} \in \mathbb{R}^{n \times p}$ such that $c_{i,j} = a_{i,j} \pm b_{i,j}$.

```
A-A
```

```
## [,1] [,2] [,3]
## [1,]    0    0    0
## [2,]    0    0    0
```

In R (but not when we use mathematical notation), all other arithmetic, logical and comparison operators are also applied in an elementwise fashion.

```
A*A
```

```
## [,1] [,2] [,3]
## [1,]    1    4    9
## [2,]   16   25   36
```

```
(A>2) & (A<=5)
```

```
## [,1] [,2] [,3]
## [1,] FALSE FALSE TRUE
## [2,] TRUE  TRUE FALSE
```

D.2.4 Matrix Multiplication (*)

Mathematically, \mathbf{AB} denotes the **matrix multiplication**. It is a very different operation to the elementwise multiplication.

```
(A <- rbind(c(1, 2), c(3, 4)))

##      [,1] [,2]
## [1,]    1    2
## [2,]    3    4

(I <- rbind(c(1, 0), c(0, 1)))

##      [,1] [,2]
## [1,]    1    0
## [2,]    0    1

A %*% I # matrix multiplication

##      [,1] [,2]
## [1,]    1    2
## [2,]    3    4
```

This is not the same as the elementwise $\mathbf{A} * \mathbf{I}$.

Matrix multiplication can only be performed on two matrices of *compatible sizes* – the number of columns in the left matrix must match the number of rows in the right operand.

Given $\mathbf{A} \in \mathbb{R}^{n \times p}$ and $\mathbf{B} \in \mathbb{R}^{p \times m}$, their multiply is a matrix $\mathbf{C} = \mathbf{AB} \in \mathbb{R}^{n \times m}$ such that c_{ij} is the dot product of the i -th row in \mathbf{A} and the j -th column in \mathbf{B} :

$$c_{ij} = \mathbf{a}_{i,\cdot} \cdot \mathbf{b}_{\cdot,j} = \sum_{k=1}^p a_{i,k} b_{k,j}$$

for $i = 1, \dots, n$ and $j = 1, \dots, m$.

Exercise D.1 Multiply a few simple matrices of sizes 2×2 , 2×3 , 3×2 etc. using pen and paper and checking the results in R.

Also remember that, mathematically, *squaring* a matrix is done in terms of matrix multiplication, i.e., $\mathbf{A}^2 = \mathbf{AA}$. It can only be performed on *square* matrices, i.e., ones with the same number of rows and columns. This is again different than R's elementwise \mathbf{A}^2 .

Note that $\mathbf{A}^T \mathbf{A}$ gives the matrix that consists of the dot products of all the pairs of columns in \mathbf{A} .

```
crossprod(A) # same as t(A) %*% A
```

```
##      [,1] [,2]
## [1,]    10   14
## [2,]    14   20
```

In one of the chapters on Regression, we note that the Pearson linear correlation coefficient can be beautifully expressed this way.

D.2.5 Aggregation of Rows and Columns

The `apply()` function may be used to transform or summarise individual rows or columns in a matrix. More precisely:

- `apply(A, 1, f)` applies a given function f on each *row* of A .
- `apply(A, 2, f)` applies a given function f on each *column* of A .

Usually, either f returns a single value (when we wish to aggregate all the elements in a row/column) or returns the same number of values (when we wish to transform a row/column). The latter case is covered in the next subsection.

Let's compute the mean of each row and column in A :

```
(A <- matrix(1:18, byrow=TRUE, nrow=3))

##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]    1    2    3    4    5    6
## [2,]    7    8    9   10   11   12
## [3,]   13   14   15   16   17   18

apply(A, 1, mean) # synonym: rowMeans(A)

## [1] 3.5 9.5 15.5

apply(A, 2, mean) # synonym: colMeans(A)

## [1] 7 8 9 10 11 12
```

We can also fetch the minimal and maximal value by means of the `range()` function:

```
apply(A, 1, range)

##      [,1] [,2] [,3]
## [1,]    1    7   13
## [2,]    6   12   18

apply(A, 2, range)

##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]    1    2    3    4    5    6
## [2,]   13   14   15   16   17   18
```

Of course, a custom function can be provided as well. Here we compute the minimum, average and maximum of each row:

```
apply(A, 1, function(row) c(min(row), mean(row), max(row)))

##      [,1] [,2] [,3]
```

```
## [1,] 1.0 7.0 13.0
## [2,] 3.5 9.5 15.5
## [3,] 6.0 12.0 18.0
```

D.2.6 Vectorised Special Functions

The special functions mentioned in the previous chapter, e.g., `sqrt()`, `abs()`, `round()`, `log()`, `exp()`, `cos()`, `sin()`, are also performed in an elementwise manner when applied on a matrix object.

```
round(1/A, 2) # rounds every element in 1/A
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,] 1.00 0.50 0.33 0.25 0.20 0.17
## [2,] 0.14 0.12 0.11 0.10 0.09 0.08
## [3,] 0.08 0.07 0.07 0.06 0.06 0.06
```

An example plot of the absolute values of sine and cosine functions depicted using the `matplot()` function (see Figure D.1).

```
x <- seq(-2*pi, 6*pi, by=pi/100)
Y <- cbind(sin(x), cos(x)) # a matrix with two columns
Y <- abs(Y) # take the absolute value of every element in Y
matplot(x, Y, type="l")
```

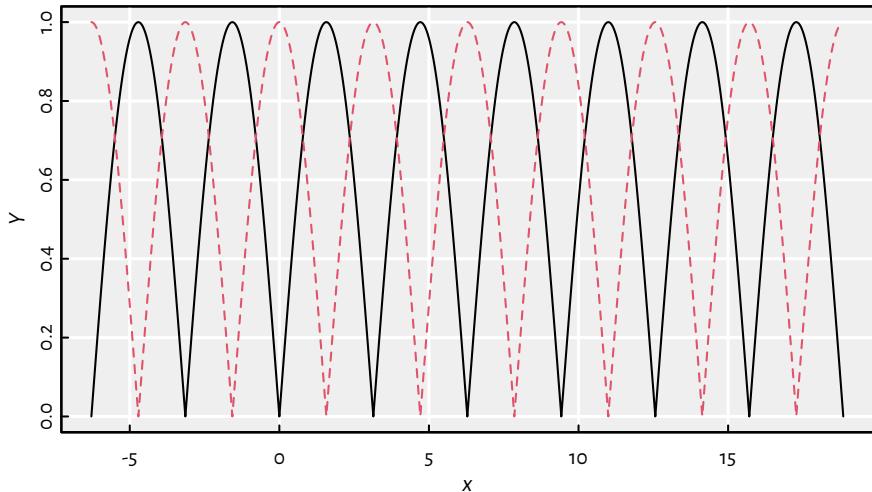


Figure D.1: Example plot with `matplot()`

D.2.7 Matrix-Vector Operations

Mathematically, there is no generally agreed upon convention defining arithmetic operations between matrices and vectors.

(*) The only exception is the matrix – vector multiplication in the case where an argument is a column or a row vector, i.e., in fact, a matrix. Hence, given $\mathbf{A} \in \mathbb{R}^{n \times p}$ we may write \mathbf{Ax} only if $\mathbf{x} \in \mathbb{R}^{p \times 1}$ is a column vector. Similarly, \mathbf{yA} makes only sense whenever $\mathbf{y} \in \mathbb{R}^{1 \times n}$ is a row vector.

Remark. Please take notice of the fact that we consistently discriminate between different bold math fonts and letter cases: \mathbf{X} is a matrix, \mathbf{x} is a row or column vector (still a matrix, but a sequence-like one) and x is an ordinary vector (one-dimensional sequence).

However, in R, we might sometimes wish to vectorise an arithmetic operation between a matrix and a vector in a row- or column-wise fashion. For example, if $\mathbf{A} \in \mathbb{R}^{n \times p}$ is a matrix and $\mathbf{m} \in \mathbb{R}^{1 \times p}$ is a row vector, we might want to subtract m_i from each element in the i -th column. Here, the `apply()` function comes in handy again.

Example: to create a *centred* version of a given matrix, we need to subtract from each element the arithmetic mean of its column.

```
(A <- cbind(c(1, 2), c(2, 4), c(5, 8)))
```

```
##      [,1] [,2] [,3]
## [1,]     1     2     5
## [2,]     2     4     8
(m <- apply(A, 2, mean)) # same as colMeans(A)
## [1] 1.5 3.0 6.5
t(apply(A, 1, function(r) r-m)) # note the transpose here
```

```
##      [,1] [,2] [,3]
## [1,] -0.5   -1  -1.5
## [2,]  0.5    1   1.5
```

The above is equivalent to:

```
apply(A, 2, function(c) c-mean(c))
```

```
##      [,1] [,2] [,3]
## [1,] -0.5   -1  -1.5
```

```
## [2,] 0.5 1 1.5
```

D.3 Matrix Subsetting

Example matrices:

```
(A <- matrix(1:12, byrow=TRUE, nrow=3))

##      [,1] [,2] [,3] [,4]
## [1,]     1     2     3     4
## [2,]     5     6     7     8
## [3,]     9    10    11    12

B <- A
dimnames(B) <- list(
  c("a", "b", "c"),      # row labels
  c("x", "y", "z", "w") # column labels
)
B

##   x  y  z  w
## a 1  2  3  4
## b 5  6  7  8
## c 9 10 11 12
```

D.3.1 Selecting Individual Elements

Matrices are two-dimensional structures: items are aligned in rows and columns. Hence, to extract an element from a matrix, we will need two indices. Mathematically, given a matrix \mathbf{A} , $a_{i,j}$ stands for the element in the i -th row and the j -th column. The same in R:

```
A[1, 2] # 1st row, 2nd columns
```

```
## [1] 2
B["a", "y"] # using dimnames == B[1,2]

## [1] 2
```

D.3.2 Selecting Rows and Columns

We will sometimes use the following notation to emphasise that a matrix \mathbf{A} consists of n rows or p columns:

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_{1,\cdot} \\ \mathbf{a}_{2,\cdot} \\ \vdots \\ \mathbf{a}_{n,\cdot} \end{bmatrix} = \begin{bmatrix} \mathbf{a}_{\cdot,1} & \mathbf{a}_{\cdot,2} & \cdots & \mathbf{a}_{\cdot,p} \end{bmatrix}.$$

Here, $\mathbf{a}_{i,\cdot}$ is a *row vector* of length p , i.e., a $(1 \times p)$ -matrix:

$$\mathbf{a}_{i,\cdot} = \begin{bmatrix} a_{i,1} & a_{i,2} & \cdots & a_{i,p} \end{bmatrix}.$$

Moreover, $\mathbf{a}_{\cdot,j}$ is a *column vector* of length n , i.e., an $(n \times 1)$ -matrix:

$$\mathbf{a}_{\cdot,j} = \begin{bmatrix} a_{1,j} & a_{2,j} & \cdots & a_{n,j} \end{bmatrix}^T = \begin{bmatrix} a_{1,j} \\ a_{2,j} \\ \vdots \\ a_{n,j} \end{bmatrix},$$

We can extract individual rows and columns from a matrix by using the following notation:

```
A[1,] # 1st row
## [1] 1 2 3 4
A[,2] # 2nd column
## [1] 2 6 10
B["a",] # of course, B[1,] is still legal
## x y z w
## 1 2 3 4
B[, "y"]
## a b c
## 2 6 10
```

Note that by extracting a single row/column, we get an atomic (one-dimensional) vector. However, we can preserve the dimensionality of the output object by passing `drop=FALSE`:

```
A[ 1, , drop=FALSE] # 1st row
## [,1] [,2] [,3] [,4]
## [1,] 1 2 3 4
A[ , 2, drop=FALSE] # 2nd column
```

```

##      [,1]
## [1,]    2
## [2,]    6
## [3,]   10
B["a",     , drop=FALSE]

##  x y z w
## a 1 2 3 4
B[     , "y", drop=FALSE]

##      y
## a  2
## b  6
## c 10

```

Now this is what we call proper row and column vectors!

D.3.3 Selecting Submatrices

To create a sub-block of a given matrix we pass two indexers, possibly of length greater than one:

```

A[1:2, c(1, 2, 4)] # rows 1,2 columns 1,2,4

##      [,1] [,2] [,3]
## [1,]    1    2    4
## [2,]    5    6    8
B[c("a", "b"), c(1, 2, 4)]

##  x y w
## a 1 2 4
## b 5 6 8
A[c(1, 3), 3]

## [1] 3 11
A[c(1, 3), 3, drop=FALSE]

##      [,1]
## [1,]    3
## [2,]   11

```

D.3.4 Selecting Based on Logical Vectors and Matrices

We can also subset a matrix with a logical matrix of the same size. This always yields a (flat) vector in return.

```
A[A>8]
```

```
## [1] 9 10 11 12
```

Logical vectors can also be used as indexers:

```
A[c(TRUE, FALSE, TRUE),] # select 1st and 3rd row
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    1    2    3    4
## [2,]    9   10   11   12
```

```
A[,colMeans(A)>6] # columns with means > 6
```

```
##      [,1] [,2]
## [1,]    3    4
## [2,]    7    8
## [3,]   11   12
```

```
B[B[,"x"]>1 & B[,"x"]<=9,] # All rows where x is in (1, 9]
```

```
##   x  y  z  w
## b 5  6  7  8
## c 9 10 11 12
```

D.3.5 Selecting Based on Two-Column Matrices

Lastly, note that we can also index a matrix A with a 2-column matrix I, i.e., A[I]. This allows for an easy access to A[I[1,1], I[1,2]], A[I[2,1], I[2,2]], A[I[3,1], I[3,2]], ...

```
I <- cbind(c(1, 3, 2, 1, 2),
            c(2, 3, 2, 1, 4)
)
A[I]
```

```
## [1] 2 11 6 1 8
```

This is exactly A[1, 2], A[3, 3], A[2, 2], A[1, 1], A[2, 4].

D.4 Exercises

D.4.1 Currency Exchange Rates

Let's load the dataset on historical daily EUR to AUD/NZD/GBP/PLN/etc. currency exchange rates:

```

currency_exchange <- read.csv("datasets/currency_exchange_all.csv.gz",
  comment.char="#")
# Convert Day to numeric:
currency_exchange$Day <- as.numeric(as.Date(currency_exchange$Day))
rates <- as.matrix(currency_exchange)
head(rates, 3)

##          Day      AUD      BGN      BRL      CAD      CHF      CNY      CZK      DKK
## [1,] 18401 1.6751 1.9558 6.2357 1.5251 1.0633 7.7816 27.490 7.4562
## [2,] 18400 1.6736 1.9558 6.2701 1.5202 1.0521 7.7068 27.610 7.4548
## [3,] 18397 1.6805 1.9558 6.3172 1.5231 1.0513 7.6759 27.589 7.4576
##          GBP      HKD      HRK      HUF      IDR      ILS      INR      ISK      JPY
## [1,] 0.89535 8.4870 7.5713 350.35 16178 3.8601 82.853 156.5 118.00
## [2,] 0.89153 8.3964 7.5580 353.39 16077 3.8321 82.144 157.1 116.31
## [3,] 0.88738 8.3693 7.5693 354.70 16128 3.8178 81.989 157.5 115.53
##          KRW      MXN      MYR      NOK      NZD      PHP      PLN      RON      RUB
## [1,] 1341.2 25.857 4.7583 10.915 1.8004 55.567 4.5510 4.8428 79.384
## [2,] 1332.4 25.634 4.7320 10.966 1.8096 55.102 4.5596 4.8388 78.908
## [3,] 1332.6 25.897 4.6982 11.057 1.8145 54.778 4.5650 4.8400 79.621
##          SEK      SGD      THB      TRY      USD      ZAR
## [1,] 10.569 1.5513 34.947 7.4448 1.0950 19.965
## [2,] 10.610 1.5426 34.684 7.4276 1.0832 19.891
## [3,] 10.669 1.5397 34.656 7.4689 1.0798 20.076

```

The dataset originally has columns of mixed types, therefore, in order to represent it as a matrix, we converted Day to numeric. Here, 10595 means 1999-01-04 and 18401 denotes 2020-05-19 — this is the so-called UNIX date, the number of days since 1970-01-01.

Remark. We will discuss how to deal with tabular data of possibly mixed column types in the next chapter.

Exercise D.2 Draw a plot (see `?plot`) of EUR/AUD exchange rates between 2020-01-01 and 2020-05-10, i.e., between days 18262 and 18392, respectively. Use the Day column on the x-axis.

Remark. Note that the most recent exchange rate is listed in the first row of `rates`.

Exercise D.3 Draw a plot of the NZD/AUD exchange rates (by appropriately transforming EUR/AUD and EUD/NZD columns).

Exercise D.4 Compute the minimal, average and maximal exchange rates for all the 32 currencies in the dataset.

D.4.2 Currency Exchange Rates Relative to 1999

Let's get back to the dataset from the previous exercise:

```
head(rates, 3)
```

```
##          Day      AUD      BGN      BRL      CAD      CHF      CNY      CZK      DKK
```

```

## [1,] 18401 1.6751 1.9558 6.2357 1.5251 1.0633 7.7816 27.490 7.4562
## [2,] 18400 1.6736 1.9558 6.2701 1.5202 1.0521 7.7068 27.610 7.4548
## [3,] 18397 1.6805 1.9558 6.3172 1.5231 1.0513 7.6759 27.589 7.4576
##          GBP      HKD      HRK      HUF      IDR      ILS      INR      ISK      JPY
## [1,] 0.89535 8.4870 7.5713 350.35 16178 3.8601 82.853 156.5 118.00
## [2,] 0.89153 8.3964 7.5580 353.39 16077 3.8321 82.144 157.1 116.31
## [3,] 0.88738 8.3693 7.5693 354.70 16128 3.8178 81.989 157.5 115.53
##          KRW      MXN      MYR      NOK      NZD      PHP      PLN      RON      RUB
## [1,] 1341.2 25.857 4.7583 10.915 1.8004 55.567 4.5510 4.8428 79.384
## [2,] 1332.4 25.634 4.7320 10.966 1.8096 55.102 4.5596 4.8388 78.908
## [3,] 1332.6 25.897 4.6982 11.057 1.8145 54.778 4.5650 4.8400 79.621
##          SEK      SGD      THB      TRY      USD      ZAR
## [1,] 10.569 1.5513 34.947 7.4448 1.0950 19.965
## [2,] 10.610 1.5426 34.684 7.4276 1.0832 19.891
## [3,] 10.669 1.5397 34.656 7.4689 1.0798 20.076

```

Exercise D.5 From now on, we will be only interested in the following currencies: AUD, CHF, GBP and USD. Remove the irrelevant columns from `rates`. Also, reverse the order of rows in the data-frame so that the oldest record is listed first.

Exercise D.6 Create a matrix `Y` with exchange rates relative to 1999-01-04 (the first record).

Remark. To do the above, for each currency separately, divide every exchange rate by the rate found the first row. For example, relative rate of EUR/GBP equal to ca. 1.26 on the last day means that you could buy, with the same amount of EUR, 26% more of pounds on 2020-05-19 (where 1 EUR = 0.89535 GBP) than on 1999-01-04 (where 1 EUR = 0.7111 GBP).

After the transformation, the first and the last two rows in `Y` should look like:

```
# Y <- ... # to do
head(round(Y, 4), 2) # rounded to 4 fractional digits
```

```

##          AUD      CHF      GBP      USD
## [1,] 1.0000 1.0000 1.0000 1.0000
## [2,] 0.9918 0.9972 1.0015 1.0001
tail(round(Y, 4), 2)
```

```

##          AUD      CHF      GBP      USD
## [5532,] 0.8762 0.6507 1.2537 0.9188
## [5533,] 0.8770 0.6577 1.2591 0.9288
```

Exercise D.7 Explain in your own words (consider different scenarios):

- If you were a German investor in Scotland would it be better to have the relative EUR/GBP exchange rates going up or down?
- What if you were a Welsh tourist in Estonia?
- What if you were a producer of goods made from imported as well as locally sourced parts/ingredients?

Exercise D.8 Draw a single plot of the relative exchange rates, similar to the one in Figure D.2. Use `legend()` to add a legend and `abline()` to draw a horizontal line at $y = 1$.

Remark. You can either call `plot()` and then `lines()` to add new curves to the existing figure or call `matplot()` to draw all curves all at once. Note that you can set the range of values on the y-axis by setting the `ylim` argument in `plot()` or `matplot()`.

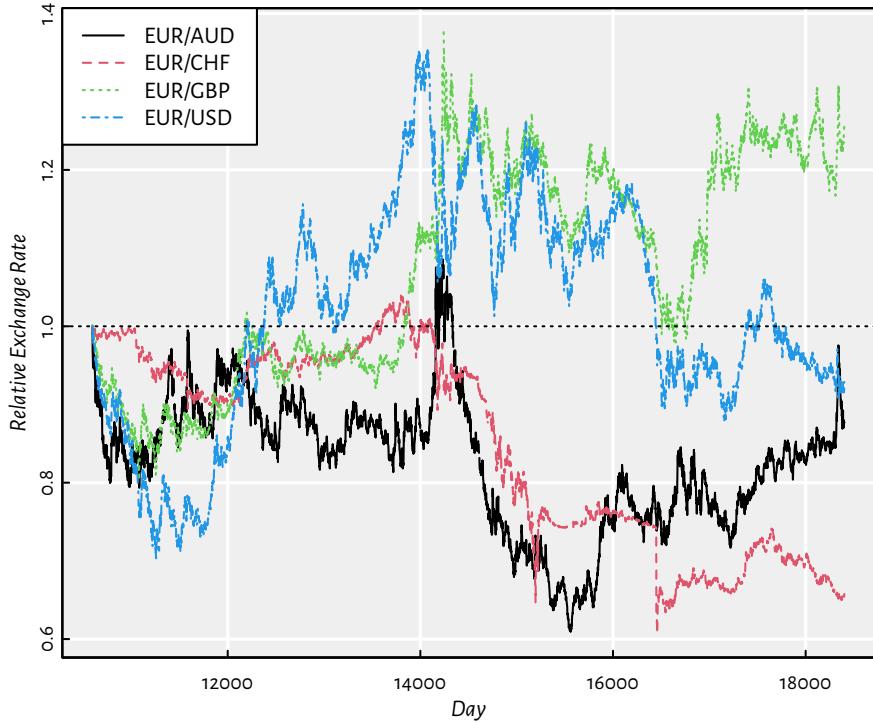


Figure D.2: Currency exchange rates relative to 1999-01-04 (with EUR as the base currency)

Exercise D.9 Imagine that you are asked to build a machine learning algorithm that predicts the exchange rate on the next day (which we in fact do in some other chapter). How would you use it to invest your money?

What could happen if the algorithm was 100% accurate but everyone in the world had access to it?

Do you think it is possible to predict the prices accurately based only on past exchange rates?

D.5 Further Reading

Recommended further reading: (Venables et al. 2020)

Other: (Deisenroth et al. 2020), (Peng 2019), (Wickham & Grolemund 2017)

TODO

Next chapter...

E

Data Frame Wrangling in R

TODO In this chapter, we will:

- ...
- ...

R `data.frames` are similar to matrices in the sense that we use them to store tabular data. However, in data frames each column can be of different type:

```
head(iris)

##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa

head(rpart::car90, 2)

##           Country Disp Disp2 Eng.Rev Front.Hd Frt.Leg.Room Frt.Shld
## Acura Integra  Japan  112    1.8    2935      3.5        41.5     53.0
## Acura Legend   Japan  163    2.7    2505      2.0        41.5     55.5
##             Gear.Ratio Gear2 HP HP.revs Height Length Luggage
## Acura Integra     3.26  3.21 130    6000    47.5     177      16
## Acura Legend      2.95  3.02 160    5900    50.0     191      14
##             Mileage Model2 Price Rear.Hd Rear.Seating RearShld
## Acura Integra     NA    11950   1.5     26.5     52.0
## Acura Legend      20    24760   2.0     28.5     55.5
##             Reliability Rim Sratio.m Sratio.p Steering Tank Tires
## Acura Integra Much better R14       NA     0.86    power 13.2 195/60
## Acura Legend  Much better R15       NA     0.96    power 18.0 205/60
##             Trans1 Trans2 Turning Type Weight Wheel.base Width
## Acura Integra man.5 auto.4       37 Small   2700      102     67
## Acura Legend   man.5 auto.4       42 Medium  3265      109     69
```

E.1 Creating Data Frames

Most frequently, we will be creating data frames based on a series of numeric, logical, characters vectors of identical lengths.

```
print(x)
```

```
##           u      v w
## 1 0.181517 TRUE A
## 2 0.919723 FALSE B
## 3 0.311723 FALSE C
## 4 0.064152 TRUE D
## 5 0.396422 FALSE E
```

Some objects, such as matrices, can easily be coerced to data frames:

```
(A <- matrix(1:12, byrow=TRUE, nrow=3,
             dimnames=list(
               NULL,       # row labels
               c("x", "y", "z", "w") # column labels
             )))

##      x  y  z  w
## [1,] 1  2  3  4
## [2,] 5  6  7  8
## [3,] 9 10 11 12
as.data.frame(A)

##      x  y  z  w
## 1  1  2  3  4
## 2  5  6  7  8
## 3  9 10 11 12
```

Named lists are amongst other candidates for a meaningful conversion:

```
(l <- lapply(split(iris$Sepal.Length, iris$Species),
            function(x) {
              c(min=min(x), median=median(x), mean=mean(x), max=max(x))
            }))

## $setosa
##   min median   mean   max
## 4.300  5.000  5.006  5.800
##
## $versicolor
##   min median   mean   max
```

```

##  4.900  5.900  5.936  7.000
##
## $virginica
##   min median   mean   max
##  4.900  6.500  6.588  7.900
as.data.frame(l)

##      setosa versicolor virginica
## min    4.300     4.900     4.900
## median 5.000     5.900     6.500
## mean   5.006     5.936     6.588
## max    5.800     7.000     7.900

```

E.2 Importing Data Frames

Many interesting data frames come from external sources, such as csv files, web APIs, SQL databases and so on.

In particular, `read.csv()` (see `?read.table` for a long list of tunable parameters) imports data from plain text files organised in a tabular manner (such as comma-separated lists of values):

```

f <- tempfile() # temporary file name
write.csv(x, f, row.names=FALSE) # save data frame to file
cat(readLines(f), sep="\n") # print file contents

## "u","v","w"
## 0.181517061544582,TRUE,"A"
## 0.919722604798153, FALSE,"B"
## 0.31172346835956, FALSE,"C"
## 0.0641516039613634, TRUE,"D"
## 0.396421572659165, FALSE,"E"

read.csv(f)

##          u      v      w
## 1 0.181517  TRUE A
## 2 0.919723 FALSE B
## 3 0.311723 FALSE C
## 4 0.064152  TRUE D
## 5 0.396422 FALSE E

```

Note that CSV is by far the most portable format for exchanging matrix-like objects between different programs (statistical or numeric computing environments, spreadsheets etc.).

E.3 Data Frame Subsetting

E.3.1 Each Data Frame is a List

First of all, we should note that each data frame is in fact represented as an ordinary named list:

```
class(x)  
  
## [1] "data.frame"  
  
typeof(x)  
  
## [1] "list"
```

Each column is stored as a separate list item. Having said that, we shouldn't be surprised that we already know how to perform quite a few operations on data frames:

```
length(x) # number of columns  
  
## [1] 3  
  
names(x) # column labels  
  
## [1] "u" "v" "w"  
x$u # accessing column `u` (synonym: x[["u"]])  
  
## [1] 0.181517 0.919723 0.311723 0.064152 0.396422  
x[[2]] # 2nd column  
  
## [1] TRUE FALSE FALSE TRUE FALSE  
x[c(1,3)] # a sub-data.frame  
  
##          u   w  
## 1 0.181517 A  
## 2 0.919723 B  
## 3 0.311723 C  
## 4 0.064152 D  
## 5 0.396422 E  
sapply(x, class) # apply class() on each column  
  
##          u           v           w  
## "numeric"  "logical"  "character"
```

E.3.2 Each Data Frame is Matrix-like

Data frames can be considered as “generalised” matrices. Therefore, operations such as subsetting will work in the same manner.

```
dim(x) # number of rows and columns

## [1] 5 3

x[1:2,] # first two rows

##           u      v w
## 1 0.18152 TRUE A
## 2 0.91972 FALSE B

x[,c(1,3)] # 1st and 3rd column, synonym: x[c(1,3)]

##           u w
## 1 0.181517 A
## 2 0.919723 B
## 3 0.311723 C
## 4 0.064152 D
## 5 0.396422 E

x[,1] # synonym: x[[1]]

## [1] 0.181517 0.919723 0.311723 0.064152 0.396422

x[,1,drop=FALSE] # synonym: x[1]

##           u
## 1 0.181517
## 2 0.919723
## 3 0.311723
## 4 0.064152
## 5 0.396422
```

Take a special note of selecting rows based on logical vectors. For instance, let’s extract all the rows from x where the values in the column named u are between 0.3 and 0.6:

```
x[x$u>=0.3 & x$u<=0.6, ]

##           u      v w
## 3 0.31172 FALSE C
## 5 0.39642 FALSE E

x[!(x[,"u"]<0.3 | x[,"u"]>0.6), ] # equivalent

##           u      v w
## 3 0.31172 FALSE C
## 5 0.39642 FALSE E
```

Moreover, subsetting based on integer vectors can be used to change the order of rows. Here is how we can sort the rows in `x` with respect to the values in column `u`:

```
(x_sorted <- x[order(x$u),])
```

```
##           u      v w
## 4 0.064152 TRUE D
## 1 0.181517 TRUE A
## 3 0.311723 FALSE C
## 5 0.396422 FALSE E
## 2 0.919723 FALSE B
```

Let's stress that the programming style we emphasise on here is very transparent. If we don't understand how a complex operation is being executed, we can always decompose it into smaller chunks that can be studied separately. For instance, as far as the last example is concerned, we can take a look at the manual of `?order` and then inspect the result of calling `order(x$u)`.

On a side note, we can re-set the row names by referring to:

```
row.names(x_sorted) <- NULL
x_sorted
```

```
##           u      v w
## 1 0.064152 TRUE D
## 2 0.181517 TRUE A
## 3 0.311723 FALSE C
## 4 0.396422 FALSE E
## 5 0.919723 FALSE B
```

E.4 Common Operations

We already know how to filter rows based on logical conditions, e.g.:

```
iris[iris$Petal.Width >= 1.2 & iris$Petal.Width <= 1.3,
  c("Petal.Width", "Species")]
```

```
##       Petal.Width   Species
## 54        1.3 versicolor
## 56        1.3 versicolor
## 59        1.3 versicolor
## 65        1.3 versicolor
## 72        1.3 versicolor
## 74        1.2 versicolor
## 75        1.3 versicolor
## 83        1.2 versicolor
```

```

## 88      1.3 versicolor
## 89      1.3 versicolor
## 90      1.3 versicolor
## 91      1.2 versicolor
## 93      1.2 versicolor
## 95      1.3 versicolor
## 96      1.2 versicolor
## 97      1.3 versicolor
## 98      1.3 versicolor
## 100     1.3 versicolor

iris[iris$Sepal.Length > 6.5 & iris$Species == "versicolor", ]

##   Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
## 51      7.0      3.2       4.7      1.4 versicolor
## 53      6.9      3.1       4.9      1.5 versicolor
## 59      6.6      2.9       4.6      1.3 versicolor
## 66      6.7      3.1       4.4      1.4 versicolor
## 76      6.6      3.0       4.4      1.4 versicolor
## 77      6.8      2.8       4.8      1.4 versicolor
## 78      6.7      3.0       5.0      1.7 versicolor
## 87      6.7      3.1       4.7      1.5 versicolor

```

and aggregate information in individual columns:

```
sapply(iris[1:4], summary)
```

```

##   Sepal.Length Sepal.Width Petal.Length Petal.Width
## Min.        4.3000    2.0000     1.000     0.1000
## 1st Qu.     5.1000    2.8000     1.600     0.3000
## Median     5.8000    3.0000     4.350     1.3000
## Mean        5.8433    3.0573     3.758     1.1993
## 3rd Qu.     6.4000    3.3000     5.100     1.8000
## Max.        7.9000    4.4000     6.900     2.5000

```

Quite frequently, we will be interested in summarising data within subgroups generated by a list of factor-like variables.

```
aggregate(iris[1:4], iris[5], mean)
```

```

##   Species Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1 setosa     5.006     3.428     1.462     0.246
## 2 versicolor  5.936     2.770     4.260     1.326
## 3 virginica   6.588     2.974     5.552     2.026

```

```
ToothGrowth[sample(nrow(ToothGrowth), 5), ] # 5 random rows
```

```

##   len supp dose
## 12 16.5  VC  1.0

```

```

## 10 7.0  VC 0.5
## 17 13.6  VC 1.0
## 50 27.3  OJ 1.0
## 60 23.0  OJ 2.0

aggregate(ToothGrowth[ "len" ], ToothGrowth[ c("supp", "dose") ], median)

##   supp dose   len
## 1  OJ  0.5 12.25
## 2  VC  0.5  7.15
## 3  OJ  1.0 23.45
## 4  VC  1.0 16.50
## 5  OJ  2.0 25.95
## 6  VC  2.0 25.95

```

Taking into account that `split()` accepts a data frame input as well, we can perform what follows:

```

sapply(
  # split iris into 3 sub-data.frames:
  split(iris, iris[5]),
  # on each sub-data.frame, apply the following function
  function(df) {
    # compute the mean of first four columns:
    sapply(df[1:4], mean)
  })

##           setosa versicolor virginica
## Sepal.Length  5.006      5.936     6.588
## Sepal.Width   3.428      2.770     2.974
## Petal.Length  1.462      4.260     5.552
## Petal.Width   0.246      1.326     2.026

sapply(split(iris, iris[5]), function(df) {
  c(Sepal.Length=summary(iris$Sepal.Length),
    Petal.Length=summary(iris$Petal.Length)
  )
})

##           setosa versicolor virginica
## Sepal.Length.Min.  4.3000     4.3000     4.3000
## Sepal.Length.1st Qu. 5.1000     5.1000     5.1000
## Sepal.Length.Median 5.8000     5.8000     5.8000
## Sepal.Length.Mean   5.8433     5.8433     5.8433
## Sepal.Length.3rd Qu. 6.4000     6.4000     6.4000
## Sepal.Length.Max.   7.9000     7.9000     7.9000
## Petal.Length.Min.  1.0000     1.0000     1.0000
## Petal.Length.1st Qu. 1.6000     1.6000     1.6000

```

```
## Petal.Length.Median 4.3500    4.3500    4.3500
## Petal.Length.Mean   3.7580    3.7580    3.7580
## Petal.Length.3rd Qu. 5.1000    5.1000    5.1000
## Petal.Length.Max.   6.9000    6.9000    6.9000
```

The above syntax is not super-convenient, but it only uses the building blocks that we have already mastered! That should be very appealing to the minimalists. Note that R packages such as `data.table` and `dplyr` offer more convenient substitutes – you can always learn them on your own (which takes time, but it's worth the hassle). They simplify the most common data wrangling tasks. Moreover, they have been optimised for speed – they can handle much larger data sets efficiently.

E.5 Metaprogramming and Formulas (*)

R (together with a few other programming languages such as Lisp and Scheme, that heavily inspired R's semantics) allows its programmers to apply some *metaprogramming* techniques, that is, to write programs that manipulate unevaluated R expressions.

For instance, take a close look at the following plot:

```
z <- seq(-2*pi, 2*pi, length.out=101)
plot(z, sin(z), type="l")
```

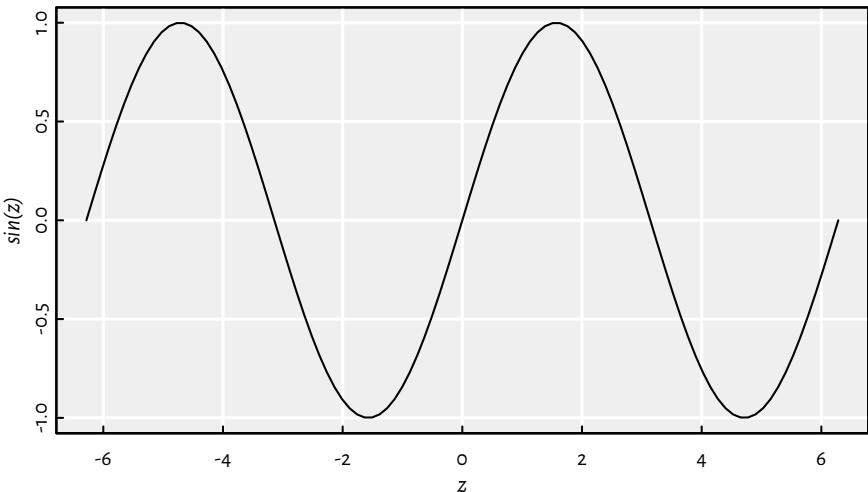


Figure E.1: Metaprogramming in action: Take a look at the Y axis label

How did the `plot()` function know that we are plotting `sin` of `z` (see Figure E.1)? It turns

out that, at any time, we not only have access to the value of an object (such as the result of evaluating `sin(z)`, which is a vector of 101 reals) but also to the expression that was passed as a function's argument itself.

```
test_meta <- function(x) {
  cat("x equals to ", x, "\n") # \n == newline
  cat("x stemmed from ", deparse(substitute(x)), "\n")
}
test_meta(2+7)

## x equals to  9
## x stemmed from  2 + 7
```

This is very powerful and yet potentially very confusing to the users, because we can write functions that don't compute the arguments provided in a way we expect them to (i.e., following the R language specification). Each function can constitute a new micro-verse, where with its own rules – we should always refer to the documentation.

For instance, consider the `subset()` function:

```
head(iris)

##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1       3.5        1.4       0.2    setosa
## 2          4.9       3.0        1.4       0.2    setosa
## 3          4.7       3.2        1.3       0.2    setosa
## 4          4.6       3.1        1.5       0.2    setosa
## 5          5.0       3.6        1.4       0.2    setosa
## 6          5.4       3.9        1.7       0.4    setosa

subset(iris, Sepal.Length>7.5, select=-(Sepal.Width:Petal.Width))

##      Sepal.Length  Species
## 106          7.6 virginica
## 118          7.7 virginica
## 119          7.7 virginica
## 123          7.7 virginica
## 132          7.9 virginica
## 136          7.7 virginica
```

Neither `Sepal.Length>6` nor `-(Sepal.Width:Petal.Width)` make sense as standalone R expressions! However, according to the `subset()` function's own rules, the former expression is considered as a row selector (here, `Sepal.Length` refers to a particular column *within* the `iris` data frame). The latter plays the role of a column filter (select everything but all the columns between...).

The `data.table` and `dplyr` packages (which are very popular) rely on this language feature all the time, so we shouldn't be surprised when we see them.

There is one more interesting language feature that is possible thanks to metaprogramming. *Formulas* are special R objects that consist of two unevaluated R expressions separated by a tilde (~). For example:

```
len ~ supp+dose
```

```
## len ~ supp + dose
```

A formula on its own has no meaning. However, many R functions accept formulas as arguments and can interpret them in various different ways.

For example, the `lm()` function that fits a linear regression model, uses formulas to specify the output and input variables:

```
lm(Sepal.Length~Petal.Length+Sepal.Width, data=iris)
```

```
##
## Call:
## lm(formula = Sepal.Length ~ Petal.Length + Sepal.Width, data = iris)
##
## Coefficients:
## (Intercept) Petal.Length Sepal.Width
##         2.249          0.472          0.596
```

On the other hand, `boxplot()` (see Figure E.2) allows for creating separate box-and-whisker plots for each subgroup given by a combination of factors.

```
boxplot(len~supp+dose, data=ToothGrowth,
horizontal=TRUE, col="white")
```

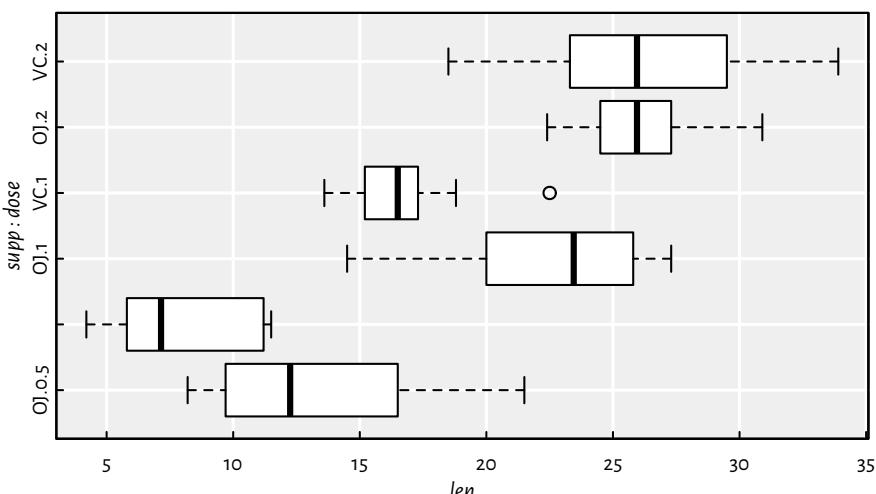


Figure E.2: Example box plot created via the formula interface

The `aggregate()` function supports formulas too:

```
aggregate(cbind(Sepal.Length, Sepal.Width)~Species, data=iris, mean)
```

```
##      Species Sepal.Length Sepal.Width
## 1      setosa      5.006     3.428
## 2 versicolor      5.936     2.770
## 3 virginica      6.588     2.974
```

We should therefore make sure that we know how every function interacts with a formula – information on that can be found in `?lm`, `?boxplot`, `?aggregate` and so forth.

E.6 Exercises

E.6.1 Urban Forest

In this task we will be working with the data on more than 70,000 trees in Melbourne, VIC, Australia. Before proceeding any further, read the dataset's description in Appendix F.

```
urban_forest <- read.csv("datasets/urban_forest.csv.gz",
  comment.char="#")
head(urban_forest, 3)

##      CoM.ID Common.Name           Scientific.Name   Genus Family
## 1 1036574    English Elm        Ulmus procera    Ulmus  Ulmaceae
## 2 1052946     Red Box Eucalyptus polyanthemos Eucalyptus Myrtaceae
## 3 1043012 River red gum Eucalyptus camaldulensis Eucalyptus Myrtaceae
##   Diameter.Breast.Height Year.Planted Date.Planted Age.Description
## 1                      90       1998 27/10/1998            Mature
## 2                      NA       1999 01/08/1999
## 3                      NA       1998 23/11/1998
##   Useful.Life.Expectancy Useful.Life.Expectancy.Value Precinct
## 1                 11-20 years                   20      NA
## 2
## 3
##   Located.In Upload.Date           Coordinate.Location
## 1      Park 10/03/2020 (-37.801406980432084, 144.97359902247373)
## 2      Park 10/03/2020 (-37.78262124666557, 144.95417032927972)
## 3      Park 10/03/2020 (-37.792607608607256, 144.95134248026756)
##   Latitude Longitude Easting Northing
## 1 -37.801     144.97  321599  5814285
## 2 -37.783     144.95  319843  5816332
## 3 -37.793     144.95  319618  5815219
```

Exercise E.1 Write R code that answers the following questions:

- How many trees have been planted between 2010 and 2015 (inclusive)?
- Are more trees located within a public park or along a street?
- What is the average age (in years, the current one is 2020) of the London Planes? (Give an approximate answer based on the `Year.Planted` column).

Exercise E.2 Fetch the IDs (`CoM.ID`), common names (`Common.Name`) and trunk diameters (`Diameter.Breast.Height`) of the 5 trees with the largest diameters at breast height. The output data frame must be sorted with respect to `Diameter.Breast.Height`, decreasingly.

```
widest_trees <- ...
print(widest_trees)
```

	CoM.ID	Common.Name	Diameter.Breast.Height
##	9776	1036933 Moreton Bay Fig	347
##	27297	1035852 Moreton Bay Fig	310
##	7836	1040042 Moreton Bay Fig	283
##	47577	1040378 Moreton Bay Fig	273
##	33671	1289230 Moreton Bay Fig	268

Exercise E.3 Create a new data frame that gives the number of trees planted in each year. Recall (from the dataset's description) that pre-2003 data might be inaccurate.

```
yearly_trees <- ...
head(yearly_trees)
```

	Year.Planted	Count
## 1	1899	48
## 2	1900	5399
## 3	1977	1
## 4	1997	6771
## 5	1998	12288
## 6	1999	3068

Exercise E.4 Depict the post-2003 tree count data on a plot similar to the one in Figure E.3. Is there any tendency in the data?

Exercise E.5 The City of Melbourne released the aforementioned dataset as part of the Open Data Platform, see <https://data.melbourne.vic.gov.au>. In your own words answer the following questions.

- What are the benefits of sharing such data with the general public?
- Are there any concerns?
- What can a machine learning engineer do with them?

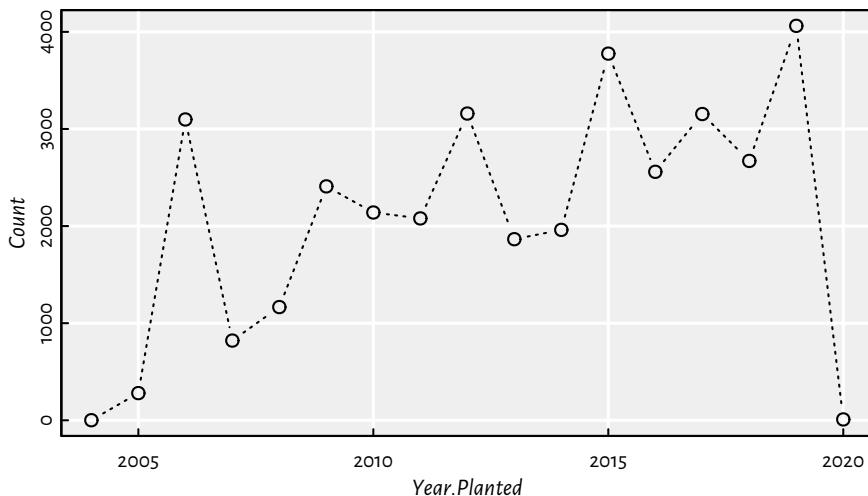


Figure E.3: The number of trees planted in the City of Melbourne each year

E.7 Air Quality

In this task, we will be working with the 2018 air quality data in the state of Victoria, Australia. Read the dataset's description in Appendix F.

```
air_quality <- read.csv("datasets/air_quality.csv.gz",
  comment.char="#")
head(air_quality, 3)
```

```
##   sample_point_id    sp_name latitude longitude      sample_datetime
## 1           10001 Alphington -37.778   145.03 01/01/2018 00:00:00
## 2           10001 Alphington -37.778   145.03 01/01/2018 00:00:00
## 3           10001 Alphington -37.778   145.03 01/01/2018 00:00:00
##   time_basis_id param_id          param_name value
## 1       1HR_AV     API    Airborne particle index  0.51
## 2       1HR_AV  BPM2.5 BAM Particles < 2.5 micron  3.70
## 3       1HR_AV      CO        Carbon Monoxide  0.10
##   param_std_unit_of_measure      param_short_name
## 1                           none Visibility Reduction
## 2                     ug/m3                  PM2.5
## 3                      ppm                   CO
```

Exercise E.6 Create a new dataframe name *gee* that consists of air quality data for Geelong South only. Omit all columns but *sample_datetime*, *param_id* and *value*.

```
gee <- ..
head(gee, 14)

##      sample_datetime param_id value
## 17 01/01/2018 00:00:00     API  0.40
## 18 01/01/2018 00:00:00   BPM2.5  5.10
## 19 01/01/2018 00:00:00      CO  0.10
## 20 01/01/2018 00:00:00    NO2  3.00
## 21 01/01/2018 00:00:00     O3 13.00
## 22 01/01/2018 00:00:00   PM10 10.20
## 23 01/01/2018 00:00:00     SO2  0.00
## 55 01/01/2018 01:00:00   BPM2.5  5.70
## 56 01/01/2018 01:00:00   PM10  7.40
## 80 01/01/2018 02:00:00     API  0.32
## 81 01/01/2018 02:00:00   BPM2.5  4.30
## 82 01/01/2018 02:00:00      CO  0.10
## 83 01/01/2018 02:00:00    NO2  1.00
## 84 01/01/2018 02:00:00     O3 13.00
```

Exercise E.7 There are 7 air quality parameters: API, BPM2.5, CO, NO2, O₃, PM10, SO₂. However, currently the data frame is in the “long” format – each measurement is represented in a separate row and there are many rows corresponding to the same dates. Convert the data frame to the “wide” format so as to obtain:

```
gee_wide <- ..
head(gee_wide, 3)
```

```
##      sample_datetime API BPM2.5 CO NO2 O3 PM10 SO2
## 1 01/01/2018 00:00:00 0.40    5.1 0.1  3 13 10.2  0
## 2 01/01/2018 01:00:00 NA     5.7 NA  NA  7.4 NA
## 3 01/01/2018 02:00:00 0.32    4.3 0.1  1 13  7.2  0
```

Missing measurements should be represented as NAs. By the way, conversion from “long” to “wide” format is also called “spreading” or “casting”.

Exercise E.8 Note that the `sample_datetime` column consists of ordinary strings. Using the `strptime()` function (see `?strptime`), convert it to proper date-time data:

```
class(gee_wide$sample_datetime) # before

## [1] "character"
gee_wide$sample_datetime <- ..
class(gee_wide$sample_datetime) # after

## [1] "POSIXlt" "POSIXt"
head(gee_wide, 3)

##      sample_datetime API BPM2.5 CO NO2 O3 PM10 SO2
```

```
## 1 2018-01-01 00:00:00 0.40    5.1 0.1    3 13 10.2    0
## 2 2018-01-01 01:00:00    NA    5.7  NA NA NA 7.4  NA
## 3 2018-01-01 02:00:00 0.32    4.3 0.1    1 13  7.2    0
```

Technical note. POSIXlt means “local time”, a special format returned by `strptime()` to represent date-time information, see `?POSIXlt`.

Exercise E.9 Below we extract month numbers from each date and add them as a new column in `gee_wide` (see `?strftime`):

```
gee_wide$month <- strftime(gee_wide$sample_datetime, "%m")
```

Compute the monthly averages of the 7 parameters.

```
gee_ave <- ..
gee_ave
```

```
##   month   API BPM2.5     CO NO2    O3 PM10   SO2
## 1    01 0.595  7.50 0.1373 3.74 18.8 23.4 0.187
## 2    02 0.458  5.64 0.1317 4.51 17.1 23.3 0.433
## 3    03 0.422  5.55 0.1228 4.85 17.6 24.1 0.519
## 4    04 0.664  8.67 0.1756 7.38 16.2 21.6 0.328
## 5    05 0.552  7.19 0.1758 7.46 17.9 22.3 0.411
## 6    06 0.638  9.23 0.2408 8.64 15.8 15.6 0.537
## 7    07 0.434  6.20 0.1626 6.23 19.1 16.0 0.639
## 8    08 0.408  5.99 0.1976 6.51 21.6 13.7 0.367
## 9    09 0.385  4.64 0.1155 5.87 20.9 16.0 0.344
## 10   10 0.442  5.65 0.1104 5.93 21.3 18.9 0.412
## 11   11 0.408  5.48 0.0776 4.63 19.5 17.7 0.211
## 12   12 0.455  6.38 0.0900 3.64 19.2 21.7 0.363
```

Exercise E.10 Draw a plot of the monthly averages of BPM2.5, NO2, O₃, PM10. Add a legend by calling `legend()`, compare Figure E.4.

Exercise E.11 In your own words answer the following questions.

- What can be deduced from the above plot?
- How machine learning algorithms could help different government bodies predict the air quality?
- What can they do with this information?

E.8 Further Reading

Recommended further reading: (Venables et al. 2020)

Other: (Peng 2019), (Wickham & Gromelund 2017)

R packages `dplyr` and `data.table` implement the most common data frame wrangling pro-

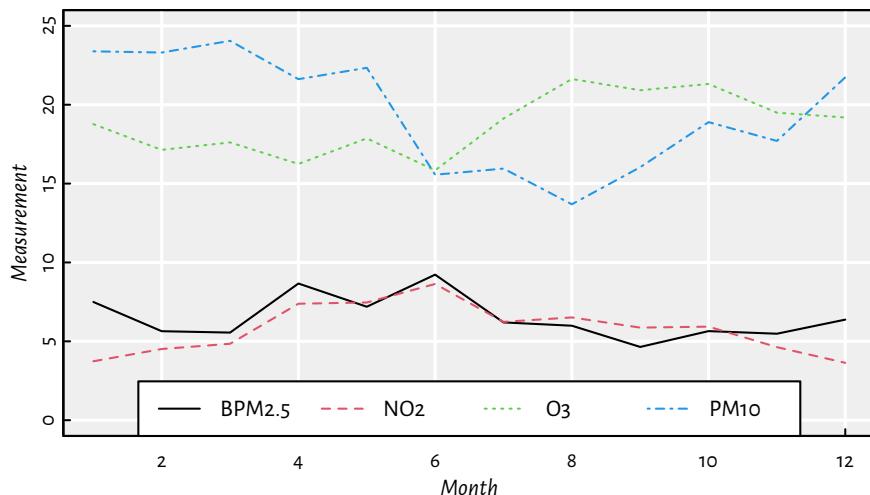


Figure E.4: Monthly averages of the air quality parameters in Geelong, VIC, Australia

cedures. You may find them very useful. Moreover, they are very fast even for large data sets. Additionally, the `magrittr` package provides a pipe operator, `%>%`, that simplifies the writing of complex, nested function calls. Do note that not everyone is a big fan of these, however.

TODO

Next chapter...

F

Datasets

TODO In this chapter, we will:

- ...
- ...

Below we briefly describe the most noteworthy datasets considered in this book. They are available for download at <https://github.com/gagolews/lmlcr/tree/master/datasets>. Many of them are real-world data, and need some proper cleaning, outlier removal etc. Moreover, for some of them, the machine learning models studied will not be of the highest quality. That's life.

"Good" datasets are actually hard to find! Many books analyse datasets that are "too easy"; they give their readers the false impression that anything can be modelled well with machine learning algorithms. Most often, you will be dealing with datasets that are incapable of telling any interesting story. Fear not, this is not personal, you are not stupid if you cannot find any useful pattern in your data! You might be forced (by your thesis supervisor, manager, client, you-the-utterly-ambitious-yourself that wants to prove your point, get that paper published, show off your superb data science skills in front of your workmates etc.) to squeeze something out of it anyway. If but few tear drops are really there, communicating this openly should be part of your work ethics.

Remark. The files with the `.csv.gz` extension are CSVs compressed with `gzip` (a widespread, open source format). R is able to decompress them on-the-fly when we use, e.g., `read.csv()`.

Compatibility note. We recommend that you use R version 4.0 or later. If this is not the case, please issue `options(stringsAsFactors=FALSE)` at the beginning of each R session or add the `stringsAsFactors=FALSE` argument to each call to `read.csv()`, `data.frame()` etc. Otherwise,

F.1 Sustainable Society Indices

Van de Kerk and Manuel in (Van de Kerk & Manuel 2008), by extending the famous definition from the "Brundtland Report" (World Commission on Environment and Development 1987), proclaim that:

A sustainable society is a society that:

- meets the needs of the present generation,
 - does not compromise the ability of future generations to meet their own needs,
 - in which each human being has the opportunity to develop itself in freedom, within a well-balanced society and in harmony with its surroundings.
-

In 2006 they have established the Sustainable Society Foundation, headquartered in the Netherlands, with the objective to develop the Sustainable Society Indices whose aim is to aid in answering the question of “How sustainable is your country?”.

ssi_2016_*

Last updated	December 2016
Provider	Sustainable Society Foundation
License	Public Domain
Source	http://www.ssfindex.com/

The most recent, 2016, edition of SSI includes 21 indicators that are based on publicly available data (UNESCO, FAO etc.). Raw scores (such as CO₂ emissions per capita, life expectancy at birth or the ratio of income of the richest 10% to the poorest 10% citizens) have been transformed onto the 0-10 scale by means of the formulas explained in detail in <http://www.ssfindex.com/ssi2016/wp-content/uploads/pdf/calculation-formulas-2016.pdf>

The 21 fine-grained indicators can be aggregated (by means of the geometric mean) into 7 categories, which in turn form 3 dimensions:

- Human Wellbeing
 - a. Basic Needs
 1. Sufficient Food
 2. Sufficient to Drink
 3. Safe Sanitation
 - b. Personal Development and Health
 4. Education
 5. Healthy Life
 6. Gender Equality
 - c. Well-balanced Society
- 7. Income Distribution

- 8. Population Growth
- 9. Good Governance
- Environmental Wellbeing
 - d. Natural Resources
 - 10. Biodiversity
 - 11. Renewable Water Resources
 - 12. Consumption
 - e. Climate and Energy
 - 13. Energy Use
 - 14. Energy Savings
 - 15. Greenhouse Gases
 - 16. Renewable Energy
- Economic Wellbeing
 - f. Transition
 - 17. Organic Farming
 - 18. Genuine Savings
 - g. Economy
 - 19. GDP
 - 20. Employment
 - 21. Public Debt

All the indices are explained in very detail in the whitepapers and reports available at <http://www.ssfindex.com/ssi/calculation-methodology/>.

```
ssi_indicators <- read.csv("datasets/ssi_2016_indicators.csv",
  comment.char="#")
head(ssi_indicators, 3)

##   Country SufficientFood SufficientWater SafeSanitation Education
## 1 Albania        10.00       9.51       9.32     8.8448
## 2 Algeria        10.00       8.40       8.74     8.0969
## 3 Angola         8.58       4.86       5.11     6.7095
##   HealthyLife GenderEquality IncomeDistribution PopulationGrowth
## 1     8.1333        7.01       8.3820      8.2310
## 2     7.7167        6.32       5.5454      4.0871
## 3     4.3167        6.37       3.2927      1.0000
##   GoodGovernance Biodiversity RenewableWaterResources Consumption
## 1        4.9560       5.5093       9.5659      5.5405
## 2        3.2780       6.5953       3.3080      6.7361
```

```

## 3      2.9769      4.1013      9.9525      7.5660
## EnergyUse EnergySavings GreenhouseGases RenewableEnergy
## 1      8.386       2.3251      8.5757      2.7286
## 2      7.346       1.0000      6.8426      1.0000
## 3      8.788       3.5602      9.2035      5.1720
## OrganicFarming GenuineSavings GDP Employment PublicDebt
## 1      1.0957      4.0346 5.4625      1.9989      2.6151
## 2      1.0038      9.4145 6.2929      3.8674      9.5776
## 3      1.0096      3.5712 3.9106      5.0662      2.6692

ssi_categories <- read.csv("datasets/ssi_2016_categories.csv",
  comment.char="#")
head(ssi_categories, 3)

## Country BasicNeeds PersonalDevelopmentAndHealth WellBalancedSociety
## 1 Albania      9.6058      7.9596      6.9926
## 2 Algeria      9.0212      7.3365      4.2039
## 3 Angola       5.9728      5.6928      2.1401
## NaturalResources ClimateAndEnergy Transition Economy
## 1      6.6343      4.6217      2.1025 3.0565
## 2      5.2772      2.6627      3.0741 6.1543
## 3      6.7594      6.2122      1.8988 3.7535

ssi_dimensions <- read.csv("datasets/ssi_2016_dimensions.csv",
  comment.char="#")
head(ssi_dimensions, 3)

## Country HumanWellbeing EnvironmentalWellbeing EconomicWellbeing
## 1 Albania      8.1162      5.3961      2.6317
## 2 Algeria      6.5283      3.5698      4.6622
## 3 Angola       4.1749      6.4411      2.8579

```

F.2 Air Quality

Environment Protection Authority of the state of Victoria, Australia published the 2018 air quality data on its Open Data platform. Various parameters are probed at different spots on an hourly basis.

air_quality

Last updated	22 May 2019
Provider	Environment Protection Authority Victoria
License	Creative Commons Attribution 4.0 International License

air_quality

Source <https://discover.data.vic.gov.au/dataset/epa-air-watch-all-sites-air-quality-hourly-averages-yearly>

```
air_quality <- read.csv("datasets/air_quality.csv.gz",
  comment.char="#")
head(air_quality, 3)

##   sample_point_id    sp_name latitude longitude   sample_datetime
## 1           10001 Alphington -37.778     145.03 01/01/2018 00:00:00
## 2           10001 Alphington -37.778     145.03 01/01/2018 00:00:00
## 3           10001 Alphington -37.778     145.03 01/01/2018 00:00:00
##   time_basis_id param_id          param_name value
## 1       1HR_AV      API Airborne particle index  0.51
## 2       1HR_AV    BPM2.5 BAM Particles < 2.5 micron  3.70
## 3       1HR_AV       CO  Carbon Monoxide  0.10
##   param_std_unit_of_measure param_short_name
## 1                           none Visibility Reduction
## 2                      ug/m3                  PM2.5
## 3                         ppm                   CO
```

Column	Description
sample_point_id	Integer identifier of the sample point
sp_name	Sample point name (city, suburb etc.)
latitude	
longitude	
sample_datetime	Date and time of the measurement; DD/MM/YYYY HH:MM:SS local time
time_basis_id	Denotes whether a 1-hour or 24-hour average is provided (1HR_AV, 24HR_AV)
param_id	Identifier of the air quality parameter: API, BPM2.5, CO, NO ₂ , O ₃ , PM ₁₀ , SO ₂ , PPM2.5, HPM ₁₀
param_name	Name of the parameter: Airborne particle index, BAM Particles < 2.5 micron, Carbon Monoxide, Nitrogen Dioxide, Ozone, TEOM Particles <10micron, Sulfur Dioxide, Partisol PM _{2.5} , Hivol PM ₁₀
value	Measured parameter value
param_std_unit_of_measure	Unit of measure
param_short_name	Short version of the parameter name: Visibility Reduction, PM _{2.5} , CO, NO ₂ , O ₃ , PM ₁₀ , SO ₂ , NA

F.3 Currency Exchange Rates

Statistical Data Warehouse of the European Central Bank System publishes currency exchange rates (quoted against EUR) that are updated once per working day.

The column names give the traditional ISO 4217 currency designators (see Table below), for instance JPY stands for the Japanese (JP) yen.

In each column like YYY, the quotation (EUR/YYY) is listed, giving the price of the Euro (here: the base currency) expressed in the counter/quote currency YYY, e.g., the value of 118.0 in JPY means that 1 euro = 118 yens.

currency_exchange	
Last updated	19 May 2020
Provider	European Central Bank
License	Provided free of charge to enhance public access to information about the activities of the European Central Bank System and the activities of the European System of Central Banks
Source	http://sdw.ecb.europa.eu/browse.do?node=9691296

```
currency_exchange <- read.csv("datasets/currency_exchange_all.csv.gz",
  comment.char="#")
# convert Day to Date-type object:
currency_exchange$Day <- as.Date(currency_exchange$Day)
head(currency_exchange, 2)

##           Day      AUD      BGN      BRL      CAD      CHF      CNY      CZK      DKK
## 1 2020-05-19 1.6751 1.9558 6.2357 1.5251 1.0633 7.7816 27.49 7.4562
## 2 2020-05-18 1.6736 1.9558 6.2701 1.5202 1.0521 7.7068 27.61 7.4548
##           GBP      HKD      HRK      HUF      IDR      ILS      INR      ISK      JPY      KRW
## 1 0.89535 8.4870 7.5713 350.35 16178 3.8601 82.853 156.5 118.00 1341.2
## 2 0.89153 8.3964 7.5580 353.39 16077 3.8321 82.144 157.1 116.31 1332.4
##           MXN      MYR      NOK      NZD      PHP      PLN      RON      RUB      SEK      SGD
## 1 25.857 4.7583 10.915 1.8004 55.567 4.5510 4.8428 79.384 10.569 1.5513
## 2 25.634 4.7320 10.966 1.8096 55.102 4.5596 4.8388 78.908 10.610 1.5426
##           THB      TRY      USD      ZAR
## 1 34.947 7.4448 1.0950 19.965
## 2 34.684 7.4276 1.0832 19.891
```

Column	Description
Day	Date in YYYY-MM-DD format, from 1999-01-04 to 2020-05-19, ordered decreasingly
AUD	Australian dollar
BGN	Bulgarian lev
BRL	Brazilian real
CAD	Canadian dollar
CHF	Swiss franc
CNY	Chinese yuan renminbi
CZK	Czech koruna
DKK	Danish krone
GBP	UK pound sterling
HKD	Hong Kong dollar
HRK	Croatian kuna
HUF	Hungarian forint
IDR	Indonesian rupiah
ILS	Israeli shekel
INR	Indian rupee
ISK	Iceland krona
JPY	Japanese yen
KRW	Korean won
MXN	Mexican peso
MYR	Malaysian ringgit
NOK	Norwegian krone
NZD	New Zealand dollar
PHP	Philippine peso
PLN	Polish zloty
RON	Romanian leu
RUB	Russian rouble
SEK	Swedish krona
SGD	Singapore dollar
THB	Thai baht
TRY	Turkish lira
USD	US dollar
ZAR	South African rand

From this dataset we derive two following ones.

currency_exchange_single. The EUR/AUD rates only, from the least to the most recent:

```
write.table(rev(currency_exchange$AUD),
           "datasets/currency_exchange_single.csv",
           col.names=FALSE, row.names=FALSE, quote=FALSE)
```

currency_exchange_diff. Relative daily changes in the EUR/AUD rates (we use it for building example exchange rates prediction algorithms):

```
n <- nrow(currency_exchange)
pred <- currency_exchange[n:1, c("Day", "AUD")] # from oldest to newest
row.names(pred) <- NULL # reset the row.names attribute
pred$Change <- NA
pred$Change[2:n] <- diff(pred$AUD) # iterative difference:
  # pred$Change[2] == pred$AUD[2]-pred$AUD[1], ...
  # pred$Change[n] == pred$AUD[n]-pred$AUD[n-1]
pred$Dir <- as.character(cut(pred$Change, c(-Inf, 0, Inf),
  c("dec", "inc"))) # inc if Change is positive, dec otherwise
for (k in 1:10) {
  name <- paste0("Lag", k)
  pred[[name]] <- NA
  pred[[name]][(k+1):n] <- pred$Change[1:(n-k)]
}
tail(pred)
```

```
##           Day     AUD  Change Dir    Lag1    Lag2    Lag3    Lag4
## 5528 2020-05-12 1.6625 -0.0084 dec  0.0096 -0.0091 -0.0342  0.0221
## 5529 2020-05-13 1.6687  0.0062 inc -0.0084  0.0096 -0.0091 -0.0342
## 5530 2020-05-14 1.6805  0.0118 inc  0.0062 -0.0084  0.0096 -0.0091
## 5531 2020-05-15 1.6805  0.0000 dec  0.0118  0.0062 -0.0084  0.0096
## 5532 2020-05-18 1.6736 -0.0069 dec  0.0000  0.0118  0.0062 -0.0084
## 5533 2020-05-19 1.6751  0.0015 inc -0.0069  0.0000  0.0118  0.0062
##           Lag5    Lag6    Lag7    Lag8    Lag9    Lag10
## 5528 -0.0197  0.0424 -0.0057 -0.0079 -0.0061 -0.0148
## 5529  0.0221 -0.0197  0.0424 -0.0057 -0.0079 -0.0061
## 5530 -0.0342  0.0221 -0.0197  0.0424 -0.0057 -0.0079
## 5531 -0.0091 -0.0342  0.0221 -0.0197  0.0424 -0.0057
## 5532  0.0096 -0.0091 -0.0342  0.0221 -0.0197  0.0424
## 5533 -0.0084  0.0096 -0.0091 -0.0342  0.0221 -0.0197
```

Note that `pred$Change[i]` gives by how much the exchange rate changed as compared to the previous day, i.e., it is equal to `pred$AUD[i]-pred$AUD[i-1]`. `Lag1` is the change from the preceding working day, `Lag2` is the one from 2 working days ago, etc.

```
write.csv(pred, "datasets/currency_exchange_diff.csv.gz",
  row.names=FALSE)
```

F.4 Urban Forest

The Urban Forest dataset gives location, species and lifespan of ca. 72,500 trees in Melbourne, VIC, Australia.

urban_forest

Last updated	10 March 2020
Provider	City of Melbourne
License	Creative Commons Attribution 4.0 International Public License
Source	https://data.melbourne.vic.gov.au/Environment/Trees-with-species-and-dimensions-Urban-Forest-/fp38-wiyy

```
urban_forest <- read.csv("datasets/urban_forest.csv.gz",
  comment.char="#")
head(urban_forest, 3)
```

```
##   CoM.ID Common.Name      Scientific.Name   Genus Family
## 1 1036574 English Elm       Ulmus procera    Ulmus Ulmaceae
## 2 1052946 Red Box   Eucalyptus polyanthemos Eucalyptus Myrtaceae
## 3 1043012 River red gum Eucalyptus camaldulensis Eucalyptus Myrtaceae
##   Diameter.Breast.Height Year.Planted Date.Planted Age.Description
## 1                      90     1998 27/10/1998        Mature
## 2                      NA     1999 01/08/1999
## 3                      NA     1998 23/11/1998
##   Useful.Life.Expectancy Useful.Life.Expectancy.Value Precinct
## 1           11-20 years                  20        NA
## 2
## 3
##   Located.In Upload.Date          Coordinate.Location
## 1       Park 10/03/2020 (-37.801406980432084, 144.97359902247373)
## 2       Park 10/03/2020 (-37.78262124666557, 144.95417032927972)
## 3       Park 10/03/2020 (-37.792607608607256, 144.95134248026756)
##   Latitude Longitude Easting Northing
## 1 -37.801    144.97  321599  5814285
## 2 -37.783    144.95  319843  5816332
## 3 -37.793    144.95  319618  5815219
```

Column	Description
CoM.ID	City of Melbourne's unique asset ID
Common.Name	
Scientific.Name	
Genus	

Column	Description
Family	
Diameter.Breast.Height	Diameter at breast height is a standard method of expressing the diameter of the trunk or bole of a standing tree. City of Melbourne Measures this 1.4m from ground level. 0.00 = Not yet assessed
Year.Planted	This is generally accurate for trees planted from 2003 onward, prior to 2003 this indicate the date the trees are added to the inventory.
Date.Planted	See above.
Age.Description	Based on date planted/data entry, therefore sometimes inaccurate; Null value = Not yet assessed
Useful.Life.Expectancy	The trees useful life expectancy. Last updated in 2009; Null value = Not yet assessed
Useful.Life.Expectancy.Value	Useful Life Expectancy Value, derived from the Useful Life Expectancy column, displayed as an integer
Precinct	The neighbourhood boundary defined for tree planning purposes
Located.In	This describes whether a tree is located within a public park or along a street
Upload.Date	
Coordinate.Location	
Latitude	
Longitude	
Easting	Easting Coordinate, please use MGA94-55
Northing	Northing Coordinate, please use MGA94-55

F.5 Wine Quality

The Wine Quality dataset (Cortez et al. 2009) describes 11 physicochemical features of Vinho Verde wine samples from the Minho region in northwest Portugal. Moreover, there are wine ratings given by wine experts and their colour.

wine_quality	
Last updated	2009
Provider	Paulo Cortez, Antonio Cerdeira, Fernando Almeida, Telmo Matos and Jose Reis via UCI (Dua et al. 2020)
License	Public Domain

wine_quality

Source <http://archive.ics.uci.edu/ml/datasets/Wine+Quality>

```
wine_quality <- read.csv("datasets/wine_quality_all.csv",
  comment.char="#")
head(wine_quality, 3)

##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1          7.4            0.70     0.00           1.9      0.076
## 2          7.8            0.88     0.00           2.6      0.098
## 3          7.8            0.76     0.04           2.3      0.092
##   free.sulfur.dioxide total.sulfur.dioxide density    pH sulphates
## 1                  11            34  0.9978 3.51      0.56
## 2                  25            67  0.9968 3.20      0.68
## 3                  15            54  0.9970 3.26      0.65
##   alcohol response color
## 1     9.4      5  red
## 2     9.8      5  red
## 3     9.8      5  red
```

Column	Description
fixed.acidity	(g(tartaric acid)/dm ³)
volatile.acidity	(g(acetic acid)/dm ³)
citric.acid	(g/dm ³)
residual.sugar	(g/dm ³)
chlorides	(g(sodium chloride)/dm ³)
free.sulfur.dioxide	(mg/dm ³)
total.sulfur.dioxide	(mg/dm ³)
density	(g/cm ³)
pH	
sulphates	(g(potassium sulphate)/dm ³)
alcohol	(vol.%)
response	Wine rating on the scale 0 (bad) to 10 (excellent): median of at least 3 valuations by sensory assessors
color	Wine colour: red (1599) or white (4898)

wine_train, wine_validate, wine_test. In the first chapters we want to set up a working classifier as fast as possible, without going into details on why and how to perform a train-test or train-validate-test split etc. Therefore, let's prepare a tailored subset of the wine quality database.

First, we select 600 white wines at random:

```
set.seed(444) # reproducibility matters
wine_subset <- wine_quality[wine_quality$color=="white",]
wine_subset <- wine_subset[sample(nrow(wine_subset), 600),]
```

Next we create a new column that tells us whether a given wine is bad (class 1, quality is not greater than 5) or not-bad:

```
wine_subset[, "bad"] <- as.numeric(wine_subset$response <= 5)
```

For simplicity, we are only interested in 3 following physicochemical features:

```
wine_subset <- wine_subset[, c(
  "chlorides", "density", "volatile.acidity", "bad"
)]
```

Then we perform a 50/25/25% split. The rows in the data frame are already randomised, therefore the first 300 can constitute the training sample, the next 150 can define the validation sample, and the 150 remaining ones – the test one.

```
write.csv(wine_subset[1:300, ],
          "datasets/wine_train.csv", row.names=FALSE)
write.csv(wine_subset[301:450, ],
          "datasets/wine_validate.csv", row.names=FALSE)
write.csv(wine_subset[451:600, ],
          "datasets/wine_test.csv", row.names=FALSE)
write.csv(wine_subset[301:450, -4],
          "datasets/wine_validate_X.csv", row.names=FALSE)
write.csv(wine_subset[451:600, -4],
          "datasets/wine_test_X.csv", row.names=FALSE)
write.csv(wine_subset[301:450, 4, drop=FALSE],
          "datasets/wine_validate_y.csv", row.names=FALSE)
write.csv(wine_subset[451:600, 4, drop=FALSE],
          "datasets/wine_test_y.csv", row.names=FALSE)
```

F.6 The World Factbook (Countries of the World)

The World Factbook 2020 dataset consists of country names, their population, area, GDP, mortality rates etc.

world_factbook

Last updated	3 April 2020
Provider	Central Intelligence Agency
License	Public Domain

world_factbook

Source <https://www.cia.gov/library/publications/resources/the-world-factbook/index.html>

```
factbook <- read.csv("datasets/world_factbook_2020.csv",
  comment.char="#")
factbook[1:6, 1:5] # preview
```

	country	area	population	median_age	population_growth_rate
## 1	Afghanistan	652230	36643815	19.5	2.38
## 2	Akrotiri	123	NA	NA	NA
## 3	Albania	28748	3074579	34.3	0.28
## 4	Algeria	2381740	42972878	28.9	1.52
## 5	American Samoa	224	49437	27.2	-1.40
## 6	Andorra	468	77000	46.2	-0.06

Column	Description
country	Country name
area	Area (km ²) as the sum of all land and water areas delimited by international boundaries and/or coastlines.
population	Population
median_age	Median age of population in years.
population_growth_rate	Compares the average annual % change in populations, resulting from a surplus (or deficit) of births over deaths and the balance of migrants entering and leaving a country.
birth_rate	Annual number of births/1,000 population
death_rate	Annual number of deaths/1,000 population
net_migration_rate	Difference between the annual number of persons entering and leaving a country/1,000 population
maternal_mortality_rate	Annual number of female deaths/100,000 live births from any cause related to or aggravated by pregnancy or its management (excluding accidental or incidental causes)

Column	Description
infant_mortality_rate	Annual number of deaths of infants under one year old/1,000 live births
life_expectancy_at_birth	Average number of years to be lived by a group of people born in the same year
total_fertility_rate	Average number of children per woman (if all women lived to the end of their childbearing years)
hiv_aids_adult_prevalence_rate	% adults (aged 15-49) living with HIV/AIDS
hiv_aids_people_living_with	Number of people living with HIV/AIDS
hiv_aids_deaths	The number of adults and children who died of AIDS during a calendar year
obesity_adult_prevalence_rate	% population considered to be obese
children_under_age_5_underweight	% children under 5 years considered to be underweight
education_expenditures	Public expenditure on education as % GDP
unemployment_youth_ages_15_to_24	% total labour force aged 15-24 that are unemployed
gdp_purchasing_power_parity	Gross domestic product (GDP) on a purchasing power parity (PPP) basis, i.e., the sum value of all goods and services produced in the country valued at prices prevailing in the United States (USD)
gdp_real_growth_rate	Annual % GDP growth adjusted for inflation
gdp_per_capita_ppp	GDP on PPP divided by population
gross_national_saving	Gross national disposable income minus final consumption expenditure as % GDP
industrial_production_growth_rate	Annual % increase in industrial production (includes manufacturing, mining and construction)
labor_force	Total labour force
unemployment_rate	% labour force that is without jobs
taxes_and_other_revenues	Total taxes and other revenues received by the national government as % GDP

Column	Description
<code>budget_surplus_or_deficit</code>	Difference between national government revenues and expenditures as % GDP
<code>public_debt</code>	Cumulative total of all government borrowings minus spending as % GDP
<code>inflation_rate_consumer_prices</code>	Annual % change in consumer prices
<code>current_account_balance</code>	Net trade in goods and services plus net earnings and net transfer payments to/from the rest of the world (USD)
<code>exports</code>	Total amount of merchandise exports (USD)
<code>imports</code>	Total amount of merchandise imports (USD)
<code>reserves_of_foreign_exchange_and_gold</code>	Financial assets that are available to the central monetary authority (USD)
<code>debt_external</code>	Total public and private debt owed to nonresidents (USD)
<code>electricity_production,_consumption,_exports,_imports</code>	Electricity produced, consumed, exported and imported (kWh)
<code>electricity_installed_generating_capacity</code>	Total capacity of currently installed generators to produce electricity (kW)
<code>electricity_from_fossil_fuels</code>	Capacity of plants that generate electricity by burning fossil fuels (coal, petroleum products and natural gas) as % total installed capacity
<code>electricity_from_nuclear_fuels</code>	Capacity of plants that generate electricity through radioactive decay of nuclear fuel as % total installed capacity
<code>electricity_from_hydroelectric_plants</code>	Capacity of plants that generate electricity by water-driven turbines as % total installed capacity
<code>electricity_from_other_renewable_sources</code>	Capacity of plants that generate electricity by using renewable energy sources other than hydroelectric (wind, waves, solar, geothermal) as % total installed capacity
<code>crude_oil_production,_exports,_imports</code>	Crude oil produced, exported and imported (barrels per day)

Column	Description
crude_oil_proved_reserves	Stock of proved reserves of crude oil (barrels)
refined_petroleum_products_production,_consumption,_exports,_imports	Refined petroleum products produced, consumed, exported and imported (barrels per day)
natural_gas_production,_consumption,_exports,_imports	Natural gas produced, consumed, exported and imported (m^3)
natural_gas_proved_reserves	Proved reserves of natural gas (m^3)
carbon_dioxide_emissions_from_consumption_of_energy	Total amount of carbon dioxide (in metric tons) released by burning fossil fuels in the process of producing and consuming energy
telephones_fixed_lines	Total number of main telephone lines in use
telephones_mobile_cellular	Total number of mobile cellular telephone subscribers
internet_users	Number of users that access the Internet
broadband_fixed_subscriptions	Total number of broadband internet subscribers
military_expenditures	Spending on defence programs as % GDP
airports	Total number of airports or airfields recognisable from the air
railways	Total route length of the railway network and of its component parts (km)
roadways	Total length of the road network (km)
waterways	Total length of navigable rivers, canals and other inland bodies of water (km)
merchant_marine	Number of ships engaged in the carriage of goods

F.7 EdStats (Country-Level Education Statistics)

The file `edstats_2019.csv.gz` provides us with many country-level Education Statistics extracted from the World Bank's Databank. Databank aggregates information from such sources as the UNESCO Institute for Statistics, OECD Programme for International Student Assessment (PISA) etc. The official description reads:

"The World Bank EdStats Query holds around 2,500 internationally comparable education indicators for access, progression, completion, literacy, teachers, population, and expenditures. The indicators cover the education cycle from pre-primary to tertiary education. The query also holds learning outcome data from international learning assessments (PISA, TIMSS, etc.), equity data from household surveys, and projection data to 2050."

edstats

Last updated	24 April 2020
Provider	The World Bank's Databank
License	Creative Commons Attribution 4.0 International License (CC BY 4.0)
Source	https://databank.worldbank.org/

```
edstats_2019 <- read.csv("datasets/edstats_2019.csv.gz",
  comment.char="#")
head(edstats_2019)

##   CountryName CountryCode
## 1 Afghanistan      AFG
## 2 Afghanistan      AFG
## 3 Afghanistan      AFG
## 4 Afghanistan      AFG
## 5 Afghanistan      AFG
## 6 Afghanistan      AFG
##
##                                     Series
## 1 Government expenditure on education as % of GDP (%)
## 2 Gross enrolment ratio, primary, female (%)
## 3 Net enrolment rate, primary, female (%)
## 4 Primary completion rate, both sexes (%)
## 5 PISA: Mean performance on the mathematics scale
## 6 PISA: Mean performance on the mathematics scale. Female
##           Code    Y2010   Y2011   Y2012   Y2013   Y2014   Y2015
## 1 SE.XPD.TOTL.GD.ZS 3.4794  3.462  2.6042  3.4545  3.6952  3.2558
## 2 SE.PRM.ENRR.FE 80.6355 80.937 86.3288 85.9021 86.7296 83.5044
## 3 SE.PRM.NENR.FE     NA     NA     NA     NA     NA     NA
## 4 SE.PRM.CMPT.ZS     NA     NA     NA     NA     NA     NA
## 5 LO.PISA.MAT       NA     NA     NA     NA     NA     NA
```

```
## 6 LO.PISA.MAT.FE NA NA NA NA NA
## Y2016 Y2017 Y2018 Y2019
## 1 4.2284 4.0589 NA NA
## 2 82.5584 82.0803 82.850 NA
## 3 NA NA NA NA
## 4 79.9346 84.4150 85.625 NA
## 5 NA NA NA NA
## 6 NA NA NA NA
```

This data frame is in a “long” format, where each indicator for each country is given in its own row. Note that we extracted the indicators reported between 2010 and 2019 and that some of them are not surveyed/updated every year.

edstats_2019_wide. Let’s convert this dataset to the “wide” format (one row per country, each indicator in its own column) based on the most recent indicators.

First we need a function that returns the last non-missing value in a given numeric vector. To recall, `na.omit()`, removes all missing values and `tail()` can be used to access the last observation easily. Unfortunately, if the vector consists of missing values only, the removal of NAs leads to an empty sequence. However, the trick we can use is that by extracting the first element from an empty vector by using `[...]`, we get a NA.

```
last_non_na <- function(x) tail(na.omit(x), 1)[1]
last_non_na(c(1, 2, NA, 3, NA, NA)) # example 1

## [1] 3
last_non_na(c(NA, NA, NA, NA, NA, NA)) # example 2

## [1] NA
```

Let’s extract the most recent indicator from each row in `edstats_2019`.

```
values <- apply(edstats_2019[-(1:4)], 1, last_non_na)
head(values)

## [1] 4.0589 82.8503 NA 85.6253 NA NA
```

Now, we shall create a data frame with 3 columns: name of the country, indicator code, indicator value. Let’s order it with respect to the first two columns.

```
edstats_2019 <- edstats_2019[c("CountryName", "Code")]
# add a new column at the righthand end:
edstats_2019["Value"] <- values
edstats_2019 <- edstats_2019[
  order(edstats_2019$CountryName, edstats_2019$Code), ]
head(edstats_2019)

##   CountryName      Code Value
## 59 Afghanistan HD.HCI.AMRT 0.7797
## 57 Afghanistan HD.HCI.AMRT.FE 0.8018
```

```
## 58 Afghanistan HD.HCI.AMRT.MA 0.7597
## 53 Afghanistan HD.HCI.EYRS 8.5800
## 51 Afghanistan HD.HCI.EYRS.FE 6.7300
## 52 Afghanistan HD.HCI.EYRS.MA 9.2100
```

To convert the data frame to a “wide” format, many readers would choose the `pivot_wider()` function from the `tidyverse` package (amongst others).

```
library("tidyverse")
edstats <- as.data.frame(
  pivot_wider(edstats_2019, names_from="Code", values_from="Value")
)
edstats[1, 1:7]

##   CountryName HD.HCI.AMRT HD.HCI.AMRT.FE HD.HCI.AMRT.MA HD.HCI.EYRS
## 1 Afghanistan     0.7797      0.8018      0.7597      8.58
##   HD.HCI.EYRS.FE HD.HCI.EYRS.MA
## 1             6.73          9.21
```

Side note (*). The above solution is of course perfectly fine and we can now live long and prosper. Nevertheless, we are here to learn new skills, so let’s note that it has the drawback that it required us to search for the answer on the internet (and go through many “answers” that actually don’t work). If we are not converting between the long and the wide formats on a daily basis, this might not be worth the hassle (moreover, there’s no guarantee that this function will work the same way in the future, that the package we relied on will provide the same API etc.).

Instead, by relying on a bit deeper knowledge of R programming (which we already have gained, see Appendices C, D and E), we could implement the relevant procedure manually. The downside is that this requires us to get out of our comfort zone and... think.

First, let’s generate the list of all countries and indicators:

```
countries <- unique(edstats_2019$CountryName)
head(countries)

## [1] "Afghanistan"    "Albania"        "Algeria"        "American Samoa"
## [5] "Andorra"         "Angola"

indicators <- unique(edstats_2019$Code)
head(indicators)

## [1] "HD.HCI.AMRT"    "HD.HCI.AMRT.FE" "HD.HCI.AMRT.MA" "HD.HCI.EYRS"
## [5] "HD.HCI.EYRS.FE" "HD.HCI.EYRS.MA"
```

Second, note that `edstats_2019` gives all the possible combinations (pairs) of the indexes and countries:

```
nrow(edstats_2019) # number of rows in edstats_2019
```

```
## [1] 23852
```

```
length(countries)*length(indicators) # number of pairs
```

```
## [1] 23852
```

Looking at the numbers in the `Value` column of `edstats_2019`, this will exactly provide us with our desired “wide” data matrix, if we read it in a rowwise manner. Hence, we can use `matrix(..., byrow=TRUE)` to generate it:

```
# edstats_2019 is already sorted w.r.t. CountryName and Code
```

```
edstats2 <- cbind(
```

```
CountryName=countries, # first column
```

```
as.data.frame(
```

```
matrix(edstats_2019$Value,
      byrow=TRUE,
      ncol=length(indicators),
      dimnames=list(NULL, indicators))
```

```
))
```

```
identical(edstats, edstats2)
```

```
## [1] TRUE
```

Now we can export `edstats` to a CSV file.

```
write.csv(edstats, "datasets/edstats_2019_wide.csv", row.names=FALSE)
```

We didn’t export the row names, because they’re useless in our case.

The table below lists the indicators included in `edstats_2019_wide.csv` based on their description included in `Series` column of the accompanying `edstats_meta.csv` file. For more details, refer to the `Definition`, `Source`, `Topic` columns therein.

Column	Description
<code>CountryName</code>	Country
<code>HD.HCI.AMRT</code>	Human Capital Index (HCI): Survival Rate from Age 15-60, Total
<code>HD.HCI.AMRT.FE</code>	Human Capital Index (HCI): Survival Rate from Age 15-60, Female
<code>HD.HCI.AMRT.MA</code>	Human Capital Index (HCI): Survival Rate from Age 15-60, Male
<code>HD.HCI.EYRS</code>	Human Capital Index (HCI): Expected Years of School, Total
<code>HD.HCI.EYRS.FE</code>	Human Capital Index (HCI): Expected Years of School, Female

Column	Description
HD.HCI.EYRS.MA	Human Capital Index (HCI): Expected Years of School, Male
HD.HCI.HLOS	Harmonised Test Scores, Total
HD.HCI.HLOS.FE	Harmonised Test Scores, Female
HD.HCI.HLOS.MA	Harmonised Test Scores, Male
HD.HCI.MORT	Human Capital Index (HCI): Probability of Survival to Age 5, Total
HD.HCI.MORT.FE	Human Capital Index (HCI): Probability of Survival to Age 5, Female
HD.HCI.MORT.MA	Human Capital Index (HCI): Probability of Survival to Age 5, Male
HD.HCI.OVRL	Human Capital Index (HCI) Score: Total (Scale 0-1)
HD.HCI.OVRL.FE	Human Capital Index (HCI) Score: Female (Scale 0-1)
HD.HCI.OVRL.MA	Human Capital Index (HCI) Score: Male (Scale 0-1)
IT.CMP.PCMP.P2	Personal computers/100 people
IT.NET.USER.P2	Internet users/100 people
LO.PISA.MAT	PISA: Mean performance on the mathematics scale
LO.PISA.MAT.FE	PISA: Mean performance on the mathematics scale. Female
LO.PISA.MAT.MA	PISA: Mean performance on the mathematics scale. Male
LO.PISA.REA	PISA: Mean performance on the reading scale
LO.PISA.REA.FE	PISA: Mean performance on the reading scale. Female
LO.PISA.REA.MA	PISA: Mean performance on the reading scale. Male
LO.PISA.SCI	PISA: Mean performance on the science scale
LO.PISA.SCI.FE	PISA: Mean performance on the science scale. Female
LO.PISA.SCI.MA	PISA: Mean performance on the science scale. Male
NY.GDP.MKTP.CD	GDP (current USD)
NY.GDP.PCAP.CD	GDP per capita (current USD)
NY.GDP.PCAP.PP.CD	GDP per capita, PPP (current international dollars)
NY.GNP.PCAP.CD	GNP per capita, Atlas method (current USD)
NY.GNP.PCAP.PP.CD	GNP per capita, PPP (current international dollars)
SE.COM.DURS	Duration of compulsory education (years)

Column	Description
SE.PRM.CMPT.ZS	Primary completion rate, both sexes (%)
SE.PRM.CMPT.FE.ZS	Primary completion rate, female (%)
SE.PRM.CMPT.MA.ZS	Primary completion rate, male (%)
SE.PRM.ENRL.TC.ZS	Pupil-teacher ratio in primary education (headcount basis)
SE.PRM.ENRR	Gross enrolment ratio, primary, both sexes (%)
SE.PRM.ENRR.FE	Gross enrolment ratio, primary, female (%)
SE.PRM.ENRR.MA	Gross enrolment ratio, primary, male (%)
SE.PRM.NENR	Net enrolment rate, primary, both sexes (%)
SE.PRM.NENR.FE	Net enrolment rate, primary, female (%)
SE.PRM.NENR.MA	Net enrolment rate, primary, male (%)
SE.PRM.PRIV.ZS	Percentage of enrolment in primary education in private institutions (%)
SE.SEC.ENRL.TC.ZS	Pupil-teacher ratio in secondary education (headcount basis)
SE.SEC.ENRR	Gross enrolment ratio, secondary, both sexes (%)
SE.SEC.ENRR.FE	Gross enrolment ratio, secondary, female (%)
SE.SEC.ENRR.MA	Gross enrolment ratio, secondary, male (%)
SE.SEC.NENR	Net enrolment rate, secondary, both sexes (%)
SE.SEC.NENR.MA	Net enrolment rate, secondary, male (%)
SE.SEC.PRIV.ZS	Percentage of enrolment in secondary education in private institutions (%)
SE.TER.ENRR	Gross enrolment ratio, tertiary, both sexes (%)
SE.TER.ENRR.FE	Gross enrolment ratio, tertiary, female (%)
SE.TER.ENRR.MA	Gross enrolment ratio, tertiary, male (%)
SE.TER.PRIV.ZS	Percentage of enrolment in tertiary education in private institutions (%)
SE.XPD.TOTL.GD.ZS	Government expenditure on education as % GDP
SL.TLF.ADVN.ZS	Labor force with advanced education (% total working-age population with advanced education)
SL.TLF.ADVN.FE.ZS	Labor force with advanced education, female (% female working-age population with advanced education)
SL.TLF.ADVN.MA.ZS	Labor force with advanced education, male (% male working-age population with advanced education)
SP.POP.TOTL	Population, total
SP.POP.TOTL.FE.IN	Population, female
SP.POP.TOTL.MA.IN	Population, male

Column	Description
SP.PRM.TOTL.IN	School age population, primary education, both sexes
SP.PRM.TOTL.FE.IN	School age population, primary education, female
SP.PRM.TOTL.MA.IN	School age population, primary education, male
SP.SEC.TOTL.IN	School age population, secondary education, both sexes
SP.SEC.TOTL.FE.IN	School age population, secondary education, female
SP.SEC.TOTL.MA.IN	School age population, secondary education, male
UIS.PTRHC.56	Pupil-teacher ratio in tertiary education (headcount basis)
UIS.SAP.CE	Population of compulsory school age, both sexes
UIS.SAP.CE.F	Population of compulsory school age, female
UIS.SAP.CE.M	Population of compulsory school age, male
UIS.XGDP.1.FSGOV	Government expenditure on primary education as % GDP
UIS.XGDP.23.FSGOV	Government expenditure on secondary education as % GDP
UIS.XGDP.56.FSGOV	Government expenditure on tertiary education as % GDP
UIS.X.PPP.1.FSGOV	Government expenditure on primary education, millions PPP USD
UIS.X.PPP.2T3.FSGOV	Government expenditure on secondary education, millions PPP USD
UIS.X.PPP.5T8.FSGOV	Government expenditure on tertiary education, millions PPP USD
UIS.XUNIT.GDPCAP.1.FSGOV	Initial government funding per primary student as % GDP per capita
UIS.XUNIT.GDPCAP.23.FSGOV	Initial government funding per secondary student as % GDP per capita
UIS.XUNIT.GDPCAP.5T8.FSGOV	Initial government funding per tertiary student as % GDP per capita
UIS.XUNIT.PPP.1.FSGOV.FFNTR	Initial government funding per primary student, PPP USD
UIS.XUNIT.PPP.2T3.FSGOV.FFNTR	Initial government funding per secondary student, PPP USD
UIS.XUNIT.PPP.5T8.FSGOV.FFNTR	Initial government funding per tertiary student, PPP USD
UIS.XUNIT.US.1.FSGOV.FFNTR	Initial government funding per primary student, USD

Column	Description
UIS.XUNIT.US.23.FSGOV.FFNTR	Initial government funding per secondary student, USD
UIS.XUNIT.US.5T8.FSGOV.FFNTR	Initial government funding per tertiary student, USD
UIS.X.US.1.FSGOV	Government expenditure on primary education, millions USD
UIS.X.US.2T3.FSGOV	Government expenditure on secondary education, millions USD
UIS.X.US.5T8.FSGOV	Government expenditure on tertiary education, millions USD

F.8 Food and Nutrient Database for Dietary Studies (FNDDS)

The Food Surveys Research Group affiliated with the US Department of Agriculture has published the Food and Nutrient Database for Dietary Studies 2015-2016. The database list nutrients in 8690 foods and beverages. See https://www.ars.usda.gov/ARSUserFiles/80400530/pdf/fndds/2015_2016_FNDDS_Doc.pdf for detailed documentation.

fndds_nutrients	
Last updated	2016
Provider	US Department of Agriculture, Agricultural Research Service
License	Public Domain
Source	http://www.ars.usda.gov/nea/bhnrc/fsrg

```
fndds_nutrients <- read.csv("datasets/fndds_nutrients_2016.csv.gz",
  comment.char="#")
fndds_nutrients[1:6, c(1:7, 10)] # preview

##      Code          Description WWEIA_Code
## 1 11000000 Milk, human        9602
## 2 11100000 Milk, NFS           1004
## 3 11111000 Milk, whole         1002
## 4 11111100 Milk, low sodium, whole 1002
## 5 11111150 Milk, calcium fortified, whole 1002
## 6 11111160 Milk, calcium fortified, low fat (1%) 1006
##   WWEIA_Description Energy Protein Carbohydrate Fat_total
## 1 Human milk     70    1.03      6.89     4.38
## 2 Milk, reduced fat 51    3.27      4.85     2.04
## 3 Milk, whole    61    3.15      4.80     3.25
```

```
## 4      Milk, whole    61    3.10      4.46    3.46
## 5      Milk, whole    61    3.15      4.80    3.25
## 6      Milk, lowfat   42    3.37      4.99    0.97
```

All nutrient amounts are given as per 100 g edible portion.

Column	Description
Code	Unique 8-digit identification code
Description	Primary description for a food/beverage
WWEIA_Code	Unique 4-digit identification code
WWEIA_Description	Description for a WWEIA category
Energy	Energy (kcal)
Protein	Protein (g)
Carbohydrate	Carbohydrate (g)
Sugars_total	Sugars, total (g)
Fiber_total_dietary	Fiber, total dietary (g)
Fat_total	Total Fat (g)
Fatty_acids_total_saturated	Fatty acids, total saturated (g)
Fatty_acids_total_monounsaturated	Fatty acids, total monounsaturated (g)
Fatty_acids_total_polyunsaturated	Fatty acids, total polyunsaturated (g)
Cholesterol	Cholesterol (mg)
Retinol	Retinol (mcg)
Vitamin_A_RAE	Vitamin A, RAE (mcg_RAE)
Carotene_alpha	Carotene, alpha (mcg)
Carotene_beta	Carotene, beta (mcg)
Cryptoxanthin_beta	Cryptoxanthin, beta (mcg)
Lycopene	Lycopene (mcg)
Lutein_zeaxanthin	Lutein + zeaxanthin (mcg)
Thiamin	Thiamin (mg)
Riboflavin	Riboflavin (mg)
Niacin	Niacin (mg)
Vitamin_B6	Vitamin B-6 (mg)
Folic_acid	Folic acid (mcg)
Folate_food	Folate, food (mcg)
Folate_DFE	Folate, DFE (mcg_DFE)
Folate_total	Folate, total (mcg)
Choline_total	Choline, total (mg)
Vitamin_B12	Vitamin B-12 (mcg)
Vitamin_B12_added	Vitamin B-12, added (mcg)
Vitamin_C	Vitamin C (mg)
Vitamin_D	Vitamin D (D2 + D3) (mcg)
Vitamin_E_alphatocopherol	Vitamin E (alpha-tocopherol) (mg)
Vitamin_E_added	Vitamin E, added (mg)
Vitamin_K_phylloquinone	Vitamin K (phylloquinone) (mcg)
Calcium	Calcium (mg)

Column	Description
Phosphorus	Phosphorus (mg)
Magnesium	Magnesium (mg)
Iron	Iron (mg)
Zinc	Zinc (mg)
Copper	Copper (mg)
Selenium	Selenium (mcg)
Potassium	Potassium (mg)
Sodium	Sodium (mg)
Caffeine	Caffeine (mg)
Theobromine	Theobromine (mg)
Alcohol	Alcohol (g)
Fatty_acid_4_0	4:0 (g)
Fatty_acid_6_0	6:0 (g)
Fatty_acid_8_0	8:0 (g)
Fatty_acid_10_0	10:0 (g)
Fatty_acid_12_0	12:0 (g)
Fatty_acid_14_0	14:0 (g)
Fatty_acid_16_0	16:0 (g)
Fatty_acid_18_0	18:0 (g)
Fatty_acid_16_1	16:1 (g)
Fatty_acid_18_1	18:1 (g)
Fatty_acid_20_1	20:1 (g)
Fatty_acid_22_1	22:1 (g)
Fatty_acid_18_2	18:2 (g)
Fatty_acid_18_3	18:3 (g)
Fatty_acid_18_4	18:4 (g)
Fatty_acid_20_4	20:4 (g)
Fatty_acid_20_5_n3	20:5 n-3 (g)
Fatty_acid_22_5_n3	22:5 n-3 (g)
Fatty_acid_22_6_n3	22:6 n-3 (g)
Water	Water (g)

F.9 Clustering Benchmarks

In the Exercises section of Chapter XXX we study a few clustering benchmark datasets. They are two-dimensional, so that it's possible to visualise the obtained results and quality of the obtained clusterings.

The `wut_isolation`, `wut_mk2` and `wut_z3` datasets were created by the author's fantastic

students at Warsaw University of Technology and are part of the *Benchmark Suite for Clustering Algorithms - Version 1*.

wut_*

Last updated	23 April 2020
Authors	Anna Gierlak, Mateusz Kobylka, Aleksander Truszczyński
License	Creative Commons Attribution 4.0 International License
Source	https://github.com/gagolews/clustering_benchmarks_v1

The `sipu_aggregation` (Gionis et al. 2007), `sipu_pathbased` (Chang & Yeung 2008) and `sipu_unbalance` (Rezaei & Fränti 2016) datasets are available at the SIPU (Speech and Image Processing Unit, School of Computing, University of Eastern Finland) website.

sipu_*

Last updated	2018
Provider	Pasi Fränti and Sami Sieranoja (Fränti & Sieranoja 2018)
License	Public Domain
Source	http://cs.joensuu.fi/sipu/datasets/

F.10 Movie Lens (TODO)

`movies`, `ratings`

F.11 Other (TODO)

the famous Fisher's Iris flower dataset, see `?iris` in R and https://en.wikipedia.org/wiki/Iris_flower_data_set.

`titanic::`

The famous `Boston` dataset from the `MASS` package records the historical (in the 1970s) median house value (`medv` column, in 1,000s USD) for 506 suburbs around Boston, MA, USA.

You can access the dataset by calling:

```
# call install.packages("MASS") first (only once)
```

```
library("MASS")
head(Boston, 3)

##      crim zn indus chas   nox    rm   age    dis rad tax ptratio black
## 1 0.00632 18 2.31     0 0.538 6.575 65.2 4.0900    1 296   15.3 396.90
## 2 0.02731  0 7.07     0 0.469 6.421 78.9 4.9671    2 242   17.8 396.90
## 3 0.02729  0 7.07     0 0.469 7.185 61.1 4.9671    2 242   17.8 392.83
##    lstat medv
## 1  4.98 24.0
## 2  9.14 21.6
## 3  4.03 34.7
```

Read the description of each of the 14 columns in the dataset's manual, see `?Boston`.

Here is a famous illustrative example proposed by the statistician Francis Anscombe in the early 1970s.

```
print(anscombe) # `anscombe` is a built-in object
```

```
##    x1 x2 x3 x4    y1    y2    y3    y4
## 1 10 10 10  8 8.04 9.14 7.46 6.58
## 2  8   8   8  8 6.95 8.14 6.77 5.76
## 3 13 13 13  8 7.58 8.74 12.74 7.71
## 4  9   9   9  8 8.81 8.77 7.11 8.84
## 5 11 11 11  8 8.33 9.26 7.81 8.47
## 6 14 14 14  8 9.96 8.10 8.84 7.04
## 7  6   6   6  8 7.24 6.13 6.08 5.25
## 8  4   4   4 19 4.26 3.10 5.39 12.50
## 9 12 12 12  8 10.84 9.13 8.15 5.56
## 10 7   7   7  8 4.82 7.26 6.42 7.91
## 11 5   5   5  8 5.68 4.74 5.73 6.89
```

What we see above is a single data frame that encodes four separate datasets: `anscombe$x1` and `anscombe$y1` define the first pair of variables, `anscombe$x2` and `anscombe$y2` define the second pair and so forth.

MNIST digits

MNIST fashion

Wisconsin Diagnostic Breast Cancer (WDBC)

```
wdbc <- read.csv("datasets/wdbc.csv", comment.char="#")
head(wdbc, 2)
```

```
##      id diagnosis mean_radius mean_texture mean_perimeter mean_area
## 1 842302          M       17.99        10.38       122.8      1001
## 2 842517          M       20.57        17.77       132.9      1326
##    mean_smoothness mean_compactness mean_concavity mean_concave_points
```

```
## 1      0.11840      0.27760      0.3001      0.14710
## 2      0.08474      0.07864      0.0869      0.07017
##   mean_symmetry mean_fractal_dimension stderr_radius stderr_texture
## 1      0.2419       0.07871      1.0950      0.9053
## 2      0.1812       0.05667      0.5435      0.7339
##   stderr_perimeter stderr_area stderr_smoothness stderr_compactness
## 1      8.589        153.40      0.006399     0.04904
## 2      3.398        74.08      0.005225     0.01308
##   stderr_concavity stderr_concave_points stderr_symmetry
## 1      0.05373      0.01587      0.03003
## 2      0.01860      0.01340      0.01389
##   stderr_fractal_dimension largest_radius largest_texture
## 1      0.006193     25.38       17.33
## 2      0.003532     24.99       23.41
##   largest_perimeter largest_area largest_smoothness largest_compactness
## 1      184.6         2019       0.1622      0.6656
## 2      158.8         1956       0.1238      0.1866
##   largest_concavity largest_concave_points largest_symmetry
## 1      0.7119        0.2654      0.4601
## 2      0.2416        0.1860      0.2750
##   largest_fractal_dimension
## 1      0.11890
## 2      0.08902
```

References

- Bezdek JC, Ehrlich R, Full W (1984) FCM: The fuzzy c-means clustering algorithm. *Computer and Geosciences* 10, 191–203.
- Bishop C (2006) *Pattern recognition and machine learning*. Springer-Verlag <https://www.microsoft.com/en-us/research/people/cmbishop/>.
- Blum A, Hopcroft J, Kannan R (2020) *Foundations of data science*. Cambridge University Press.
- Boyd S, Vandenberghe L (2004) *Convex optimization*. Cambridge University Press https://web.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf.
- Breiman L, Friedman J, Stone CJ, Olshen RA (1984) *Classification and regression trees*. Chapman; Hall/CRC.
- Caliński T, Harabasz J (1974) A dendrite method for cluster analysis. *Communications in Statistics* 3, 1–27.
- Campello RJGB, Moulavi D, Zimek A, Sander J (2015) Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data* 10, 5:1–5:51.
- Chandola V, Banerjee A, Kumar V (2009) Anomaly detection: A survey. *ACM Computing Surveys* 41.
- Chang H, Yeung D (2008) Robust path-based spectral clustering. *Pattern Recognition* 41, 191–203.
- Cortez P, Cerdeira A, Almeida F, Matos T, Reis J (2009) Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems* 47, 547–553.
- Deisenroth MP, Faisal AA, Ong CS (2020) *Mathematics for machine learning*. Cambridge University Press <https://mml-book.com/>.
- Devroye L, Györfi L, Lugosi G (1996) *A probabilistic theory of pattern recognition*. Springer.
- Deza MM, Deza E (2014) *Encyclopedia of distances*. Springer.
- Dua D, Graff C, others (eds) (2020) UCI machine learning repository. <http://archive.ics.uci.edu/ml>.
- Eddelbuettel D (2013) *Seamless R and C++ integration with Rcpp*. Springer, New York.

- Edwards AWF, Cavalli-Sforza LL (1965) A method for cluster analysis. *Biometrics* 21, 362–375.
- Efron B, Hastie T (2016) *Computer age statistical inference: Algorithms, evidence, and data science*. Cambridge University Press.
- Ester M, Kriegel H-P, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise *Proc. KDD'96*, pp. 226–231.
- Estivill-Castro V (2002) Why so many clustering algorithms: A position paper. *ACM SIGKDD Explorations Newsletter* 4, 65–75.
- Fix E, Hodges J (1951) *Discriminatory analysis. Nonparametric discrimination; consistency properties*. Randolph Field, Texas.
- Fix E, Hodges J (1952) *Discriminatory analysis. Nonparametric discrimination; small sample performance*. Randolph Field, Texas.
- Fletcher R (2008) *Practical methods of optimization*. Wiley.
- Florek K, Łukaszewicz J, Perkal J, Steinhaus H, Zubrzycki S (1951) Sur la liaison et la division des points d'un ensemble fini. 2, 282–285. <http://matwbn.icm.edu.pl/ksiazki/cm/cm2/cm2145.pdf>.
- Fränti P, Sieranoja S (2018) K-means properties on six clustering benchmark datasets. *Applied Intelligence* 48, 4743–4759. <http://cs.uef.fi/sipu/datasets/>.
- Gagolewski M, Bartoszuk M, Cena A (2016) Genie: A new, fast, and outlier-resistant hierarchical clustering algorithm. *Information Sciences* 363, 8–23.
- Gionis A, Mannila H, Tsaparas P (2007) Clustering aggregation. *ACM Transactions on Knowledge Discovery from Data* 1, art. no. 4.
- Goodfellow I, Bengio Y, Courville A (2016) *Deep learning*. MIT Press <https://www.deeplearningbook.org/>.
- Harper FM, Konstan JA (2015) The MovieLens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems* 5, 19:1–19:19.
- Hastie T, Tibshirani R, Friedman J (2017) *The elements of statistical learning*. Springer-Verlag <https://web.stanford.edu/~hastie/ElemStatLearn/>.
- Herlocker JL, Konstan JA, Terveen LG, Riedl JT (2004) Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems* 22, 5–53. https://web.archive.org/web/20070306161407/http://web.engr.oregonstate.edu/~herlock/papers/eval_tois.pdf.
- Hubert L, Arabie P (1985) Comparing partitions. *Journal of Classification* 2, 193–218.
- Jain AK (2010) Data clustering: 50 years beyond K-means. *Pattern Recognition Letters* 31, 651–666.

- James G, Witten D, Hastie T, Tibshirani R (2017) *An introduction to statistical learning with applications in R*. Springer-Verlag <http://faculty.marshall.usc.edu/gareth-james/ISL/>.
- Koren Y (2009) *The BellKor solution to the Netflix grand prize*. https://netflixprize.com/assets/GrandPrize2009_BPC_BellKor.pdf.
- Lance GN, Williams WT (1967) A general theory of classificatory sorting strategies: 1. Hierarchical systems. *The Computer Journal* 9, 373–380.
- Ling RF (1973) A probability theory of cluster analysis. *Journal of the American Statistical Association* 68, 159–164.
- Lü L, others (2012) Recommender systems. *Physics Reports* 519, 1–49. <https://arxiv.org/pdf/1202.1112.pdf>.
- Luxburg U von, Williamson RC, Guyon I (2012) Clustering: Science or art? *Proc. ICML workshop on unsupervised and transfer learning*, pp. 65–79. <http://proceedings.mlr.press/v27/luxburg12a.html>.
- MacQueen J (1967) Some methods for classification and analysis of multivariate observations *Proc. Fifth berkeley symposium on mathematical statistics and probability*, pp. 281–297. University of California Press, Berkeley.
- Müller AC, Nowozin S, Lampert CH (2012) Information theoretic clustering using minimum spanning trees *Proc. German conference on pattern recognition*, pp. 205–215. <https://github.com/amueller/information-theoretic-mst>.
- Müllner D (2011) Modern hierarchical, agglomerative clustering algorithms. <https://arxiv.org/abs/1109.2378v1>.
- Müllner D (2013) Fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python. *Journal of Statistical Software* 53, 1–18.
- Ng AY, Jordan MI, Weiss Y (2001) On spectral clustering: Analysis and an algorithm *Proc. Advances in neural information processing systems 14 (NIPS'01)*, pp. 849–856. <https://papers.nips.cc/paper/2092-on-spectral-clustering-analysis-and-an-algorithm.pdf>.
- Nocedal J, Wright SJ (2006) *Numerical optimization*. Springer.
- Peng RD (2019) *R programming for data science*. <https://bookdown.org/rdpeng/rprogdatascience/>.
- Piotte M, Chabbert M (2009) *The Pragmatic Theory solution to the Netflix grand prize*. https://netflixprize.com/assets/GrandPrize2009_BPC_PragmaticTheory.pdf.
- Quinlan R (1986) Induction of decision trees. *Machine Learning* 1, 81–106.
- Quinlan R (1993) *C4.5: Programs for machine learning*. Morgan Kaufmann Publishers.
- Rao C (1964) The use and interpretation of principal component analysis in applied research. *Sankhyā A* 26, 329–358.

- R Development Core Team (2020) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria <http://www.R-project.org>.
- Rezaei M, Fränti P (2016) Set-matching measures for external cluster validity. *IEEE Transactions on Knowledge and Data Engineering* 28, 2173–2186.
- Ricci F, Rokach L, Shapira B, Kantor P (eds) (2011) *Recommender systems handbook*. Springer <http://www.inf.unibz.it/~ricci/papers/intro-rec-sys-handbook.pdf>.
- Sarle WS, others (eds) (2002) The comp.ai.neural-nets FAQ. <http://www.faqs.org/faqs/ai-faq/neural-nets/part1/>.
- Sokal RR, Michener CD (1958) A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin* 38, 1409–1438. https://archive.org/details/cbarchive_33927_a statistical method for evaluation in 1902.
- Sørensen TJ (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons. *Biologiske Skrifter* 5, 1–34. http://www.royalacademy.dk/Publications/High/295_S%C3%B8rensen,%20Thorvald.pdf.
- Therneau TM, Atkinson EJ (2019) *An introduction to recursive partitioning using the RPART routines*. <https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf>.
- Töscher A, Jahrer M, Bell RM (2009) *The BigChaos solution to the Netflix grand prize*. https://netflixprize.com/assets/GrandPrize2009_BPC_BigChaos.pdf.
- Van de Kerk G, Manuel AR (2008) A comprehensive index for a sustainable society: The SSI – The Sustainable Society Index. *Ecological Economics* 66, 228–242.
- Venables WN, Smith DM, R Core Team (2020) *An introduction to R*. <https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf>.
- Ward Jr. JH (1963) Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* 58, 236–244.
- Wickham H, Grolemund G (2017) *R for data science*. O'Reilly <https://r4ds.had.co.nz/>.
- Wierzchoń ST, Kłopotek MA (2018) *Modern algorithms for cluster analysis*. Springer.
- World Commission on Environment and Development (1987) *Our common future*. Oxford University Press.
- Zhang T, Ramakrishnan R, Livny M (1996) BIRCH: An efficient data clustering method for large databases *Proc. ACM SIGMOD international conference on management of data – SIGMOD'96*, pp. 103–114.