
**SOLUTIONS MANUAL FOR FUNDAMENTALS OF
MACHINE LEARNING FOR PREDICTIVE DATA
ANALYTICS**

SOLUTIONS MANUAL FOR FUNDAMENTALS OF MACHINE LEARNING FOR PREDICTIVE DATA ANALYTICS

Algorithms, Worked Examples, and Case Studies

John D. Kelleher
Brian Mac Namee
Aoife D'Arcy

The MIT Press
Cambridge, Massachusetts
London, England

Contents

Notation		vii
	Notational Conventions	vii
	Notational Conventions for Probabilities	ix
1	Machine Learning for Predictive Data Analytics: Exercise Solutions	1
2	Data to Insights to Decisions: Exercise Solutions	5
3	Data Exploration: Exercise Solutions	11
4	Information-based Learning: Exercise Solutions	29
5	Similarity-based Learning: Exercise Solutions	45
6	Probability-based Learning: Exercise Solutions	55
7	Error-based Learning: Exercise Solutions	65
8	Evaluation: Exercise Solutions	77
	Bibliography	91
	Index	93

5. The following table⁴ lists a dataset containing the details of five participants in a heart disease study, and a target feature RISK which describes their risk of heart disease. Each patient is described in terms of four binary descriptive features
- EXERCISE, how regularly do they exercise
 - SMOKER, do they smoke
 - OBESE, are they overweight
 - FAMILY, did any of their parents or siblings suffer from heart disease

ID	EXERCISE	SMOKER	OBESE	FAMILY	RISK
1	daily	false	false	yes	low
2	weekly	true	false	yes	high
3	daily	false	false	no	low
4	rarely	true	true	yes	high
5	rarely	true	true	no	high

- a. As part of the study researchers have decided to create a predictive model to screen participants based on their risk of heart disease. You have been asked to implement this screening model using a **random forest**. The three tables below list three bootstrap samples that have been generated from the above dataset. Using these bootstrap samples create the decision trees that will be in the random forest model (use entropy based information gain as the feature selection criterion).

ID	EXERCISE	FAMILY	RISK	ID	SMOKER	OBESE	RISK	ID	OBESE	FAMILY	RISK
1	daily	yes	low	1	false	false	low	1	false	yes	low
2	weekly	yes	high	2	true	false	high	1	false	yes	low
2	weekly	yes	high	2	true	false	high	2	false	yes	high
5	rarely	no	high	4	true	true	high	4	true	yes	high
5	rarely	no	high	5	true	true	high	5	true	no	high
Bootstrap Sample A				Bootstrap Sample B				Bootstrap Sample C			

⁴ The data in this table has been artificially generated for this question, but is inspired by the results from the Framingham Heart Study: www.framinghamheartstudy.org.

ID	Target	Model 1	Error	Error ²	Error	SST	SST ²
		Pred.					
1	2,623.4	2,664.3	40.9	1,674.2	40.9	173.5	30,089.5
2	2,423.0	2,435.9	12.9	167.4	12.9	-54.9	3,017.9
3	2,423.3	2,398.5	-24.8	615.0	24.8	-92.3	8,528.0
4	2,448.1	2,447.1	-1.1	1.2	1.1	-43.8	1,918.8
5	2,761.7	2,847.3	85.7	7,335.9	85.7	356.4	127,043.9
6	2,434.9	2,411.2	-23.7	560.9	23.7	-79.6	6,341.4
7	2,519.0	2,516.4	-2.6	6.7	2.6	25.5	652.8
8	2,771.6	2,870.2	98.6	9,721.7	98.6	379.4	143,913.2
9	2,601.4	2,585.9	-15.6	242.0	15.6	95.0	9,028.8
10	2,422.3	2,414.2	-8.1	65.0	8.1	-76.7	5,875.6
11	2,348.8	2,406.7	57.9	3,352.0	57.9	-84.1	7,079.6
12	2,514.7	2,505.2	-9.4	89.3	9.4	14.4	206.2
13	2,548.4	2,581.2	32.8	1,075.2	32.8	90.3	8,157.2
14	2,281.4	2,276.9	-4.5	20.4	4.5	-214.0	45,776.8
15	2,295.1	2,279.7	-15.4	238.5	15.4	-211.2	44,597.1
16	2,570.5	2,576.6	6.1	37.2	6.1	85.7	7,346.2
17	2,528.1	2,510.2	-17.9	320.8	17.9	19.4	375.1
18	2,342.2	2,380.9	38.7	1,496.9	38.7	-110.0	12,093.6
19	2,456.0	2,452.1	-3.9	15.1	3.9	-38.8	1,501.8
20	2,451.1	2,436.7	-14.4	208.5	14.4	-54.2	2,934.9
21	2,295.8	2,307.2	11.4	129.8	11.4	-183.7	33,730.7
22	2,405.0	2,354.9	-50.1	2,514.9	50.1	-136.0	18,492.1
23	2,388.9	2,418.1	29.2	853.2	29.2	-72.8	5,297.2
24	2,629.5	2,582.4	-47.1	2,215.7	47.1	91.5	8,380.0
25	2,583.8	2,563.5	-20.3	411.7	20.3	72.7	5,281.6
26	2,658.2	2,662.0	3.9	15.1	3.9	171.2	29,298.7
27	2,482.3	2,491.8	9.4	88.6	9.4	0.9	0.8
28	2,470.8	2,477.7	6.9	47.7	6.9	-13.1	172.6
29	2,604.9	2,619.8	14.9	221.7	14.9	128.9	16,624.4
30	2,441.6	2,444.9	3.3	10.9	3.3	-46.0	2,117.8
Mean	2,490.9	2,497.3	6.5	1,125.1	23.7	6.5	19,529.1
Std Dev	127.0	142.0	33.5	2,204.9	24.1	142.0	34,096.6

$$\begin{aligned} \text{sum of squared errors} &= \frac{1}{2} \sum_{i=1}^n (t_i - \mathbb{M}(\mathbf{d}_i))^2 \\ &= 16,876.6 \end{aligned}$$

The sum of squared errors for Model 2 can be calculated as follows.