

Contents

1	Introduction to Data Science	1
1.1	Linear Algebra for data science	2
1.2	Linear Equations	3
1.3	Distance	4
1.4	Hyperplane	5
1.5	Eigen Values	6
1.6	Eigen Vectors	6
1.7	Exercises	11
1.7.1	Short and Long Answer Questions	11
1.7.2	Fill in the Blanks	12
1.7.3	Multiple Choice Questions	13
2	Foundations of Statistical Analysis	17
2.1	Statistical Modelling	17
2.2	Random Variables	17
2.3	Probability Mass/Density Functions	18
2.4	Sample Statistics	19
2.5	Hypothesis Testing	20
2.6	Exercises	20
2.6.1	Short and Long Answer Questions	20
2.6.2	Fill in the Blanks	21
2.6.3	Multiple Choice Questions	22
3	Predictive Modelling	25
3.1	Predictive Modelling	25
3.2	Linear Regression	26
3.3	Simple Linear Regression Model Building	26
3.3.1	The Regression Equation in Matrix Form	27
3.4	Multiple Linear Regression	29
3.5	Logistic Regression	30
3.6	Building SLR Example	31
3.7	Building MLR Example	33
3.8	Multinomial regression	35

3.9	Exercises	36
3.9.1	Short and Long Answer Questions	36
3.9.2	Fill in the Blanks	37
3.9.3	Multiple Choice Questions	37
4	Introduction to R Programming	41
4.1	Introduction to R Programming	41
4.2	Installation of R S/W using the interface	42
4.3	Variables and Data Types	43
4.4	Operators	45
4.5	Control Structures	47
4.6	R Objects	48
4.6.1	Creating Arrays	48
4.6.2	Matrices	50
4.6.3	Vectors	52
4.6.4	Lists	53
4.6.5	Data Frames	54
4.7	Factors in R	55
4.8	Functions	56
4.9	Debugging and Simulation in R	58
4.10	Introduction to Data Preprocessing in R	60
4.11	Experiments	63
4.12	Exercises	70
4.12.1	Short and Long Answer Questions	70
4.12.2	Fill in the Blanks	71
4.12.3	Multiple Choice Questions	72
5	Classification and Clustering	75
5.1	Classification	75
5.2	Classifier Performance Metrics	76
5.2.1	Confusion Matrix	76
5.2.2	Numerical Example	77
5.2.3	Receiver Operating Characteristic (ROC) Curve	78
5.3	Ensemble Methods	79
5.3.1	Overview of Ensemble Methods	79
5.3.2	Common Ensemble Methods	79
5.3.3	Advantages of Ensemble Methods	80
5.4	Logistic Regression in R Programming	80
5.4.1	Theory of Logistic Regression	80
5.4.2	Implementation in R	81
5.5	K-Nearest Neighbors (KNN) Algorithm	81
5.5.1	K-Nearest Neighbors Implementation in R	82
5.6	Introduction to Clustering	83
5.6.1	Introduction to Clustering	83
5.6.2	K-Means Algorithm	84

5.6.3	Numerical Example	84
5.6.4	K-Means implementation in R	85
5.7	Advanced Data Analysis and Integration with R	85
5.7.1	Time Series Analysis	87
	Example of Time Series Analysis in R	87
5.8	Social Network Analysis in Data Science	88
5.9	Reading data from Relational databases	89
5.10	Reading Data from MongoDB	91
5.11	Exercises	91
	5.11.1 Short and Long Answer Questions	91
	5.11.2 Fill in the Blanks	92
	5.11.3 Multiple Choice Questions	93
6	Association Rule Mining	97
6.1	Introduction	97
6.2	Algorithms for Association Rule Mining	98
6.3	Measures of Finding Interesting Patterns	99
6.4	R Code for Finding frequent itemset and Generating Rules	100
6.5	Finding Frequent Itemsets using Apriori	101
6.6	Rules Generation	102
6.7	Exercises	102
	6.7.1 Short and Long Answer Questions	102
	6.7.2 Fill in the Blanks	102
	6.7.3 Multiple choice Questions	103
7	Additional Experiments	107
8	Real-world Data Science Case Studies	125
8.1	Insightful House Pricing	125
8.2	Flower Power: Iris Study	132
8.3	Automobile Clustering Analysis	139
A	Answers	143