Περιεχόμενα

Πρόλογος xv

1. Εισαγωγή 1

1.1. Τα δεδομένα στη σύγχρονη πραγματικότητα 2
1.2. Η Επιστήμη των Δεδομένων και οι Συγγενείς Κλάδοι 3
1.3. Πλάνο του βιβλίου 6
1.4. Βιβλιογραφικές Πηγές 7

2. Η Εξόρυξη Δεδομένων ως Διαδικασία 9

2.1. Μοντέλο Ωρίμανσης της Αναλυτικής Δεδομένων 10
2.2. Διαδικασία Ανακάλυψης Γνώσης από Βάσεις Δεδομένων 11
2.3. Το CRISP-DM Μοντέλο 12
2.4. Βιβλιογραφικές Πηγές 15

3. Περιβάλλοντα και Εργαλεία 17

3.1. Η γλώσσα προγραμματισμού R 21
3.1.1. Λήψη και εγκατάσταση της R 22
3.1.2. Το γραφικό περιβάλλον της R 24
3.1.3. Ο χώρος εργασίας 25
3.1.4. Λήψη βοήθειας στην R 27
3.2. Αντικείμενα και κλάσεις αντικειμένων στην R 28
3.2.1. Ατομικές κλάσεις αντικειμένων 29
3.2.2. Πράξεις μεταξύ αντικειμένων 31
3.2.3. Διάνυσμα 33
3.2.4. Λίστα 39
3.2.5. Μητρώο 43
3.2.6. Παράγοντας 47
3.2.7. Πλαίσιο δεδομένων 49
3.2.8. Μετατροπή κλάσης αντικειμένου 56
3.2.9. Ειδικές τιμές 58
3.3. Δομημένος προγραμματισμός 60

3.3.1. Δομές ελέγχου....60
3.3.2. Δομές επανάληψης....61
3.3.3. Συναρτήσεις....66
3.3.4. Αναδρομικές συναρτήσεις....70
3.3.5. Επαναληπτικές συναρτήσεις....70
3.4. Βασικές εργασίες....76
3.4.1. Ανάγνωση δεδομένων από αρχείο....76
3.4.2. Εγγραφή δεδομένων σε αρχείο....78
3.4.3. Δουλεύοντας με υποσύνολα δεδομένων....79
3.4.4. Παραγωγή ακολουθιών....80
3.4.5. Δειγματοληψία & τυχαιοποιημένα δεδομένα....81
3.5. Τα πακέτα της R....83
3.5.1. Το πακέτο dplyr....85
3.5.2. Το πακέτο mlr....92
3.5.3. Το πακέτο OpenML....93
3.5.4. Το πακέτο ggplot2....95
3.6. Γραφικά περιβάλλοντα της R....96
3.6.1. RStudio....96
3.6.2. Rattle....97
3.6.3. Jupyter Notebook....98
3.7. Βιβλιογραφικές Πηγές....101
3.8. Κατάλογος Πακέτων της R....102
3.9. Απαντήσεις Ασκήσεων Αυτοαξιολόγησης....102
3.10. Ενδεικτικές Απαντήσεις Δραστηριοτήτων....106

4. Δεδομένα και Προεπεξεργασία....111

4.1. Δεδομένα και τύποι δεδομένων....115
4.1.1. Ποιοτικά (quantitative) ή κατηγορικά (categorical) δεδομένα....117
4.1.2. Ποσοτικά (quantitative) ή αριθμητικά (numerical) δεδομένα....118
4.1.3. Συνεχή (continuous) και διακριτά (discrete) δεδομένα....118
4.1.4. Ιδιότητες τύπων δεδομένων....120
4.2. Προεπεξεργασία δεδομένων....122
4.3. Ορθότητα ανάγνωσης αρχείων εισόδου....123
4.4. Ελλιπείς τιμές....124

4.4.1. Η συνάρτηση impute του πακέτου mlr....128
4.4.2. Οπτικοποίηση ελλιπών τιμών.. 130
4.5. Καθαρισμός δεδομένων με θόρυβο 132
4.5.1. Το πακέτο NoiseFiltersR....136
4.6. Μετασχηματισμός και κανονικοποίηση δεδομένων....140
4.6.1. Κανονικοποίηση ελαχίστου-μεγίστου 141
4.6.2. Κανονικοποίηση Z-score.. 141
4.6.3. Κανονικοποίηση δεκαδικής κλίμακας.. 141
4.7. Ακραίες τιμές....145
4.7.1. Γραφικές μέθοδοι εντοπισμού ακραίων τιμών....145
4.7.2. Αριθμητικές μέθοδοι εντοπισμού ακραίων τιμών.. 146
4.8. Διακριτοποίηση δεδομένων....147
4.9. Το πακέτο tidyr....150
4.9.1. Αναδιαμόρφωση των δεδομένων 151
4.9.2. Διάσπαση και ενοποίηση στηλών....153
4.9.3. Χειρισμός ελλιπών τιμών.. 155
4.10. Βιβλιογραφικές Πηγές....157
4.11. Κατάλογος Πακέτων της R....157
4.12. Απαντήσεις Ασκήσεων Αυτοαξιολόγησης 158
4.13. Ενδεικτικές Απαντήσεις Δραστηριοτήτων 164

5. Περιγραφική Στατιστική και Οπτικοποίηση 169

5.1. Μέτρα Θέσης....172
5.1.1. Μέση τιμή 172
5.1.2. Διάμεσος 173
5.1.3. Επικρατούσα τιμή 175
5.2. Μέτρα Διασποράς....175
5.2.1. Εύρος....176
5.2.2. Διακύμανση, τυπική απόκλιση & συντελεστής μεταβλητότητας...177
5.2.3. Άλλα αριθμητικά μέτρα 179
5.2.4. Συνδιασπορά και συντελεστής συσχέτισης δύο αριθμητικών μεταβλητών 180
5.3. Δείκτες σχετικής θέσης 181
5.3.1. Ποσοστιαία σημεία....181

5.3.2. Τεταρτημόρια 182
5.3.3. Ενδοτεταρτημορικό εύρος 182
5.3.4. Οι συναρτήσεις summary και aggregate 183
5.3.5. Το πακέτο psych και η συνάρτηση describe() 184
5.4. Οπτικοποίηση αριθμητικών δεδομένων 184
5.4.1. Πίνακες συχνοτήτων και σχετικών συχνοτήτων 185
5.4.2. Ιστογράμματα 187
5.4.3. Διάγραμμα στελέχους-φύλλου 192
5.4.4. Το θηκόγραμμα 194
5.4.5. Διάγραμμα θεωρητικής ποσοστιαίας-ποσοστιαίας 196
5.4.6. Διαγράμματα διασποράς 197
5.5. Οπτικοποίηση κατηγορικών δεδομένων 200
5.5.1. Πίνακες συχνοτήτων και σχετικών συχνοτήτων 202
5.5.2. Πίνακες συνάφειας 203
5.5.3. Ραβδογράμματα 206
5.5.4. Κυκλικά διαγράμματα 207
5.5.5. Ομαδοποιημένα ραβδογράμματα 208
5.5.6. Στοιβαγμένα ραβδογράμματα 208
5.5.7. Το πακέτο DescTools και η συνάρτηση Desc() 210
5.6. Βιβλιογραφικές Πηγές 211
5.7. Κατάλογος Πακέτων της R 212
5.8. Απαντήσεις Ασκήσεων Αυτοαξιολόγησης 212
5.9. Απαντήσεις Δραστηριοτήτων 221

6. Παλινδρόμηση 223

6.1. Βασικές έννοιες 225
6.2. Διερεύνηση της σχέσης μεταξύ μεταβλητών 229
6.3. Εξερεύνηση σχέσεων μεταβλητών: Η μήτρα διαγραμμάτων διασποράς 236
6.4. Στόχοι ενός μοντέλου παλινδρόμησης 240
6.5. Γραμμική και μη-γραμμική παλινδρόμηση 248
6.6. Μοντέλα γραμμικής παλινδρόμησης 251
6.6.1. Εκτίμηση συντελεστών γραμμικών μοντέλων παλινδρόμησης 251
6.6.2. Εκτίμηση συντελεστών γραμμικών μοντέλων παλινδρόμησης: Η μέθοδος των ελαχίστων τετραγώνων. 253

6.6.2.1. Εκτίμηση συντελεστών γραμμικού μοντέλου παλινδρόμησης: με τη μέθοδο των ελαχίστων τετραγώνων στο περιβάλλον της R.....259
6.6.3. Εκτίμηση συντελεστών γραμμικών μοντέλων παλινδρόμησης: Η μέθοδος της Σταδιακής Καθόδου (Gradient Descent).....267
6.6.3.1. Ο αλγόριθμος της Σταδιακής Καθόδου.....270
6.6.3.2. Τύπος υπολογισμού συντελεστών για πολλαπλό γραμμικό μοντέλο παλινδρόμησης με τη μέθοδο της Σταδιακής Καθόδου.....276
6.6.3.3. Ο ρόλος της παραμέτρου μάθησης α. 277
6.6.3.4. Η μέθοδος της Σταδιακής Καθόδου στο περιβάλλον της R.....281
6.6.3.5. Χρήση της μεθόδου Σταδιακής Καθόδου για την εκτίμηση συντελεστών. 285
6.6.3.6. Εκδοχές της μεθόδου Σταδιακής Καθόδου: Χρήση σε Περιβάλλοντα Μεγάλων Δεδομένων.....291
6.6.4. Σύγκριση μεθόδων Ελαχίστων Τετραγώνων και Σταδιακής Καθόδου..295
6.7. Αξιολόγηση και ερμηνεία μοντέλων γραμμικής παλινδρόμησης 298
6.7.1. Μοντέλα γραμμικής παλινδρόμησης με στόχο την εξήγηση της διακύμανσης 299
6.7.1.1. Τεκμηρίωση του γραμμικού μοντέλου παλινδρόμησης. 299
6.7.1.2. Αξιολόγηση και ερμηνεία γραμμικού μοντέλου παλινδρόμησης. .313
6.7.2. Γραμμικά μοντέλα γραμμικής παλινδρόμησης με στόχο την πρόβλεψη 322
6.7.2.1. Αξιολόγηση και ερμηνεία γραμμικού μοντέλου παλινδρόμησης με στόχο τη πρόβλεψη. 323
6.7.2.2. Υποπροσαρμογή, Υπερπροσαρμογή και Κανονικοποίηση μοντέλου παλινδρόμησης. 335
6.8. Κατάλογος Πακέτων της R.....344
6.9. Βιβλιογραφικές Πηγές.....344
6.10. Απαντήσεις Ασκήσεων Αυτοαξιολόγησης 345
6.11. Ενδεικτικές Απαντήσεις Δραστηριοτήτων 355

7. Κατηγοριοποίηση.....365

7.1. Αξιολόγηση απόδοσης κατηγοριοποιητών και στατιστική σύγκριση 367
7.2. Δέντρα απόφασης.....370
7.2.1. Κατασκευή Δένδρου Απόφασης 371
7.2.2. Δημιουργία Δέντρων Αποφάσεων με Χρήση R 381

7.3. Μπεϋζιανοί κατηγοριοποιητές. 392
7.3.1. Αφελής κατηγοριοποιητής....393
7.3.2. Δημιούργια αφελή κατηγοριοποιητή Μπέυζ μέσω R 397
7.3.3. Μπεϋζιανά Δίκτυα 403
7.4. Λογιστική Παλινδρόμηση....406
7.4.1. Λογιστική Παλινδρόμηση με χρήση R....408
7.5. Κανόνες κατηγοριοποίησης 411
7.5.1. Εξαγωγή Κανόνων Κατηγοριοποίησης με χρήση R....415
7.6. Μάθηση βασισμένη στα στιγμιότυπα 423
7.6.1. Παράδειγμα χρήσης κοντινότερου γείτονα με χρήση R....431
7.7. Μηχανές διανυσμάτων υποστήριξης....441
7.7.1. Διαχωρίσιμη περίπτωση 441
7.7.2. Μη γραμμικές μηχανές υποστήριξης διανυσμάτων 446
7.7.3. Εκπαίδευση Μηχανών Διανυσματων Υποστήριξης με R....451
7.8. Πακέτα της R με Αλγορίθμους Κατηγοριοποίησης....461
7.9. Επιλογή μεταβλητών με τη μέθοδο του περιτυλίγματος 463
7.9.1. Επιλογή μεταβλητών με χρήση R....466
7.10. Μάθηση βαθμωτής συνάρτησης στόχου....470
7.11. Μάθηση σε ανομοιογενή (imbalanced) δεδομένα....472
7.11.1. Χειρισμός ανομοιογενών δεδομένων με χρήση R....475
7.12. Επαυξητική μάθηση 479
7.13. Βιβλιογραφικές Πηγές....480
7.14. Ενδεικτικές Απαντήσεις Δραστηριοτήτων 482

8. Κανόνες Συσχέτισης 485

8.1. Εισαγωγή 486
8.2. Περιγραφή προβλήματος....487
8.3. Εξόρυξη Συχνών Στοιχειοσυνόλων....490
8.4. Αλγόριθμοι Εξόρυξης Συχνών Στοιχειοσυνόλων....495
8.4.1. Ο κατά επίπεδα αλγόριθμος Apriori....496
8.4.2. Ο αλγόριθμος Eclat για την τομή των tidsets 499
8.4.3. Ο αλγόριθμος FPGrowth....500
8.5. Εξόρυξη Συχνών Στοιχειοσυνόλων με χρήση της R....506
8.6. Βιβλιογραφικές Πηγές....513

8.7. Απαντήσεις Ασκήσεων Αυτοαξιολόγησης 514
8.8. Ενδεικτικές Απαντήσεις Δραστηριοτήτων 515

9. Συσταδοποίηση 527

9.1. Ιεραρχικοί αλγόριθμοι 528
9.1.1. Ιεραρχική συσταδοποίηση με χρήση R.....531
9.2. Εκτίμηση του αποτελέσματος συσταδοποίησης 534
9.3. Διαμεριστικοί αλγόριθμοι.....536
9.3.1. Τυχαία Αρχικοποίηση Κεντροειδών.....537
9.3.2. Επιλογή του Αριθμού Συστάδων 542
9.3.3. Διαμεριστικοί αλγόριθμοι με χρήση R.....544
9.4. Αυτό-Οργανωμένοι Χάρτες 553
9.5. Αλγόριθμοι βασισμένοι στη πυκνότητα (Density based) 556
9.5.1. Περιγραφή Αλγορίθμου 558
9.5.2. DBSCAN με χρήση R 563
9.6. Πακέτα αλγορίθμων συσταδοποίησης 564
9.7. Αναπαράσταση των συστάδων.....565
9.8. Βιβλιογραφικές Πηγές.....566
9.9. Απαντήσεις Ασκήσεων Αυτοαξιολόγησης 567
9.10. Ενδεικτικές Απαντήσεις Δραστηριοτήτων 569

10. Ομάδες μοντέλων μάθησης.....575

10.1. Εισαγωγή 576
10.2. Τεχνικές που βασίζονται σε έναν αλγόριθμο μάθησης και την δειγματο-
ληψία των δεδομένων εκπαίδευσης.....578
10.3. Συνδυασμός αποφάσεων διαφορετικών κατηγοριοποιητών 586
10.4. Αλλαγή της συνάρτησης-στόχου.....591
10.5. Εισαγωγή τυχαιότητας στον αλγόριθμο μάθησης.....593
10.6. Συνδυασμός Μοντέλων με χρήση R σε προβλήματα κατηγοριοποίησης.....593
10.7. Πακέτα στη R για δημιουργία ομάδων μοντέλων κατηγοριοποίησης...602
10.8. Συνδυασμός Μοντέλων με χρήση R σε προβλήματα παλινδρόμησης...603
10.9. Πακέτα της R για δημιουργία ομάδων μοντέλων παλινδρόμησης 609
10.10. Βιβλιογραφικές Πηγές.....610
10.11. Ενδεικτικές Απαντήσεις Δραστηριοτήτων 611

11. Εξόρυξη γνώσης από κείμενα.....613

11.1. Εισαγωγή.....614
11.2. Εξαγωγή γνώσης από κείμενο με χρήση R.....618
11.3. Ανάλυση Συναισθήματος.....626
11.3.1. Εξόρυξη άποψης με χρήση R.....629
11.4. Προβλήματα Πολλαπλών κλάσεων.....632
11.5. Αυτοταξινομούμενες Μέθοδοι.....637
11.6. Ενεργή μάθηση.....640
11.7. Βιβλιογραφικές πηγές.....641

12 Ανάλυση Κοινωνικών Δικτύων.....643

12.1. Οπτικοποίηση των κοινωνικών δικτύων.....643
12.2. Βιβλιογραφικές Πηγές.....654

13. Χρονοσειρές με αλγόριθμους παλινδρόμησης.....655

13.1. Παλινδρόμηση.....656
13.1.1. Γραμμική Παλινδρόμηση με χρήση mlr.....657
13.1.2. Αλγόριθμος κοντινότερων γειτόνων με χρήση R.....662
13.1.3. Δέντρα παλινδρόμησης με χρήση R.....667
13.1.4. Παλινδρόμηση Διανυσμάτων Υποστήριξης (Support Vector Regression).....674
13.2. Επιλογή μεταβλητών με τη μέθοδο του περιτυλίγματος.....680
13.2.1. Χρήση R.....681
13.3. Πακέτα για αλγόριθμους παλινδρόμησης.....683
13.4. Χειρισμός Χρονοσειρών.....684
13.5. Βιβλιογραφικές Πηγές.....701
13.6. Ενδεικτικές Απαντήσεις Δραστηριοτήτων.....702

14. Νευρωνικά Δίκτυα και Βαθιά Μάθηση.....711

14.1. Τεχνητά Νευρωνικά Δίκτυα.....712
14.1.1. Το μοντέλο του τεχνητού νευρώνα.....713
14.1.2. Νευρωνικά δίκτυα.....722
14.2. Εκπαίδευση Νευρωνικών Δικτύων με χρήση R σε προβλήματα κατηγοριοποίησης.....730

14.3. Πακέτα της R εκπαίδευσης νευρωνικών δικτύων για κατηγοριοποίηση 740
14.4. Εκπαίδευση Νευρωνικών Δικτύων με χρήση R σε προβλήματα παλινδρόμησης 740
14.5. Βαθιά μάθηση και εξόρυξη γνώσης από εικόνες 747
14.5.1. Μάθηση από Εικόνες με χρήση R 750
14.6. Εξόρυξη γνώσης από ήχο και βίντεο 760
14.6.1. Εξόρυξη γνώσης από ήχο με χρήση R 761
14.7. Βιβλιογραφικές Πηγές 765

15. Μεγάλα Δεδομένα 767

15.1. Εισαγωγικές έννοιες 768
15.2. Χαρακτηριστικά μεγάλων δεδομένων 769
15.3. Τεχνολογικές προκλήσεις των μεγάλων δεδομένων 771
15.4. Apache Hadoop 772
15.4.1. Hadoop Distributed File System (HDFS) 774
15.4.1.1. Βασικές έννοιες στο σύστημα HDFS. 778
15.4.1.2. Η έννοια του block. 778
15.4.1.3. Ρόλοι υπολογιστών στο σύστημα HDFS. 779
15.4.1.4. Τοπολογία συστάδας HDFS. 781
15.4.1.5. Σύστημα διαχείρισης αρχείων HDFS 782
15.4.1.6. Ομοσπονδιοποίηση στο HDFS (HDFS Federation) 788
15.4.2. Επεξεργασία μεγάλων δεδομένων με το σύστημα Hadoop: Το προγραμματιστικό μοντέλο MapReduce 791
15.4.2.1. Παράδειγμα χρήσης του προγραμματιστικού μοντέλου MapReduce: Εύρεση κοινών φίλων μεταξύ χρηστών σε ιστοτόπους κοινωνικής δικτύωσης.. 793
15.4.2.2. Εκτέλεση προγραμμάτων MapReduce σε συστάδα κόμβων... 797
15.5. Βιβλιογραφικές Πηγές *801*

16. Επίλογος 803