Data Mining în mediul R

Teorie și aplicații

Stelian STANCU

Data Mining în mediul R

Teorie și aplicații

Colecția
Cibernetică

Editura ASE
București
2022

ACADEMIA DE STUDII ECONOMICE DIN BUCUREŞTI

Editura ASE
Piața Romană nr. 6, sector 1, București, România
cod 010374
www.ase.ro
www.editura.ase.ro
editura@ase.ro

Descrierea CIP a Bibliotecii Naţionale a României
STANCU, STELIAN
Data Mining în mediul R : teorie și aplicații / Stelian Stancu. – București : Editura ASE, 2022
Conține bibliografie
ISBN 978-606-34-0433-7

005

Editura ASE
Redactor, tehnoredactor și copertă: Claudia-Marinela Dumitru

Datele sunt precum vremea: le vezi, le simți,
dar nu știi dacă te vor ajuta în timpul următor!

Totul este sub falsul control
doar atunci când mergi suficient de încet.

Cheia succesului este să faci în fiecare zi
cel puțin ceva ce este pentru prima dată!

Critica dispare doar atunci când te aventurezi
în marele neant al acțiunii de a nu face nimic!

Gândirea trebuie să fie mereu în antiteză cu durata vieții:
gândește mult, căci de trăit oricum vei trăi puțin.

Cuprins

Listă termeni **11**

Despre autor **13**

Cuvânt-înainte **15**

Capitolul 1
Introducere în Text Mining (TM) şi implementare R **17**
1.1 Introducere în Text Mining 17
1.2 Pachete disponibile în R, pentru Text Mining 18
1.3 Termeni importanți în analiza Text Mining 20
1.4 Implementarea Text Mining în mediul R 20

Capitolul 2
Text Mining. Preprocesarea datelor de tip text
şi constituirea matricei DTM în mediul R **26**
2.1 Preprocesarea datelor 26
2.1.1 Prezentarea datelor de intrare 26
2.1.2 Prelucrarea inițială a datelor 27
2.1.3 Tratarea erorilor de ortografie și a abrevierilor 27
2.1.4 Steem-ul cuvântului 31
2.2 Matricea termenilor documentului (DTM) 32

Capitolul 3
Text Mining în R. Analiza cluster **35**
3.1 Introducere 35
3.2 Analiza ierarhică a clusterelor 36
3.3.1 k-means clustering în R 39
3.3.2 Determinarea numărului optim de clustere 39
3.4 k-medoids clustering 40
3.4.1 Aspecte generale 40
3.4.2 Vizualizarea medoizilor 41

Capitolul 4
Text Mining în R. Analiza sentimentelor **42**
4.1 Introducere 42
Cuvântul sentiment 42
4.2 Încărcarea pachetelor necesare şi a datelor folosite 44
4.3 Analiza de bază a teoriei sentimentelor 45
4.4 Analiza de tip 2-Gram, în cazul teoriei sentimentelor 49

4.5 Scorul qdap pentru alegerea cuvintelor/sentimentelor pozitive
și, respectiv, negative 50
4.6 Revizuirea norilor de cuvinte ce exprimă sentimente 53
4.7 Obținerea datelor 55
4.8 Definirea TFID 57

Capitolul 5
Arbori de decizie și randomForest în R 58
5.1 Introducere 58
5.2 Tipuri clasice de arbori de decizie 59
5.2.1 Arbori de regresie 59
5.2.2 Arbori de clasificare 61
5.2.3 Arbori de regresie și clasificare (CART) 63
5.2.4 Avantaje și dezavantaje ale arborilor de decizie 63
5.3 Tipuri specifice de arbori de decizie 63
5.3.1 Bagging-ul (agregarea bootstrap – colecții de date) 63
5.3.2 Random Forest 65
5.3.3 Boosting-ul (agregarea cu ponderi). AdaBoost 67
5.4 Încărcarea pachetelor necesare în R 70
5.5 Arbori de decizie în R 71
5.5.1 Arbori de clasificare în R 71
5.5.2 Arbori de regresie în R 83
5.5.3 Random Forest și boosting în R 86

Capitolul 6
Metode de tip ansamblu: bagging, boosting, stacking
și Random Forest în R 95
6.1 Prezentare generală 95
Un singur „elev slab” 95
Combinarea „cursanților slabi” 96
6.2.2 Bagging-ul 100
6.3 Random Forest 103
6.4 Boosting 104
6.4.1 Boosting – caracterizare 104
6.4.2 Adaptive Boosting (AdaBoost) 106
6.4.3 Gradient Boosting 109
6.5 Stacking (stivuirea) 110
6.5.1 Stacking (stivuirea) 112
6.5.2 Multi-levels Stacking
(stivuire pe mai multe niveluri/straturi) 113
6.6 Concluzii 114

Capitolul 7
Metode de tip ansamblu: bagging versus boosting în R 116
7.1 Introducere în învățarea de tip ansamblu, în context bagging - boosting 116
7.2 Bootstrapping 117
7.3 Bagging 118
7.4 Boosting 119
7.5 Obținerea de *N* cursanți pentru bagging și boosting 122
7.6 Elemente de ponderare a datelor 122
7.7 Etapa de clasificare a datelor 123
7.8 Selectarea celei mai bune tehnici: bagging sau boosting 125
7.9 Asemănări și deosebiri dintre bagging și boosting 126

Capitolul 8
Utilizarea metodelor de tip ansamblu în analiza de clasificare: bagging, boosting, stacking și randomForest în R 127
8.1 Bagging-ul, folosit în clasificare 127
8.1.1 Încărcarea pachetelor necesare 127
8.1.2 Aplicarea bagging-ului în mediul R, pentru clasificare 127
8.2 AdaBoost în R, folosit în clasificare 130
8.2.1 Încărcarea pachetelor necesare 130
8.2.2 Aplicarea AdaBoost în R, pentru clasificare 130
8.3 Random Forest în R, folosit în clasificare 135
8.3.1 Încărcarea pachetelor necesare 135
8.3.2 Aplicarea Random Forest în R, pentru clasificare 135

Capitolul 9
Utilizarea metodelor de tip ansamblu în analiza de regresie: bagging, boosting și randomForest în R 139
9.1 Bagging-ul, folosit în regresie 139
9.1.1 Încărcarea pachetelor necesare 139
9.1.2 Aplicarea bagging-ului în mediul R, pentru regresie 139
9.2 AdaBoost în R, folosit în regresie 141
9.2.1 Încărcarea pachetelor necesare 141
9.2.2 Aplicarea AdaBoost în R, pentru regresie 141
9.3 Random Forest în R, folosit în regresie 145
9.3.1 Încărcarea pachetelor necesare 145
9.3.2 Aplicarea Random Forest în R, pentru regresie 146

Capitolul 10
Construirea modelelor de tip bagging, boosting, stacking și Random Forest în R, prin combinarea mai multor modele de bază. Exemplificări și comparații 149
10.1 Încărcarea pachetelor necesare 149
10.2 Citirea și formatarea datelor în R 149
10.3 Construirea de modele de tip Boosting în R, prin combinarea mai multor modele de bază 149
10.3.1 Construirea modelului C5.0 în R 150
10.3.2 Construirea modelului Stochastic Gradient Boosting în R 150
10.3.3 Sumarizarea rezultatelor de la ambele modele 150
10.4 Construirea de modele de tip Bagging în R, prin combinarea mai multor modele de bază 151
10.4.1 Construirea modelului Bagged CART în R 151
10.4.2 Construirea modelului Random Forest în R 153
10.5 Construirea unui algoritm/model de stivuire, prin combinarea mai multor modele de bază în R 156
10.5.1 Construirea unui model de stivuire, prin combinarea a patru modele de bază în R 156
10.5.2 Stivuirea modelelor folosind Random Forest în R 159

Capitolul 11
Utilizarea metodelor de tip ansamblu în alte analize: ACP, rețele neuronale, bagging, boosting, stacking și Random Forest în R 160
11.1 Încărcarea pachetelor necesare 160
11.2 Citirea și formatarea/pre-procesarea datelor în R 160
11.3 Analiza în componente principale (ACP) 162
11.4 Splitarea (împărțirea) setului de date în set de instruire și set de testare 168
11.5 Rețele neuronale artificiale (RNA) 169
11.6 Arbori de decizie 172
11.7 Random Forest 173
11.8 Suport Vector Mașină (SVM) 174
11.9 Modele de tip ansamblu. Bagging 176
11.9.1 Împărțirea setului de Instruire/antrenare în două părți în funcție de rezultat: 80% și 20% 176
11.9.2 Definirea controalelor de instruire/antrenare pentru mai multe modele 176
11.9.3 Definirea predictorilor și a rezultatului 178
11.10 Model de tip stivă 182

Bibliografie 187