

Uczenie maszynowe w języku R : tworzenie i doskonalenie model - od przygotowania danych po dostrajanie, ewaluację i pracę z big data / Brett Lantz. – Gliwice, copyright © 2024

Spis treści

O autorze	11
O recenzencie	12
Przedmowa	13
Rozdział 1. Wprowadzenie do uczenia maszynowego	17
Początki uczenia maszynowego	18
Użycia i nadużycia uczenia maszynowego	21
Sukcesy uczenia maszynowego	23
Ograniczenia uczenia maszynowego	24
Etyka uczenia maszynowego	25
Jak uczą się maszyny?	30
Zachowywanie danych	32
Abstrakcja	32
Generalizacja	35
Ewaluacja	37
Uczenie maszynowe w praktyce	39
Typy danych wejściowych	40
Typy algorytmów uczenia maszynowego	42
Dopasowywanie danych wejściowych do algorytmów	47
Uczenie maszynowe w języku R	48
Instalowanie pakietów R	49
Wczytywanie pakietów R i usuwanie ich z pamięci	50
Instalowanie RStudio	51
Dlaczego R i dlaczego teraz?	52
Podsumowanie	54
Rozdział 2. Zarządzanie danymi	55
Struktury danych języka R	56
Wektory	56
Czynniki	58
Listy	60
Ramki danych	62
Macierze i tablice	65
Zarządzanie danymi w języku R	67
Wczytywanie, zapisywanie i usuwanie struktur danych R	67

Importowanie i zapisywanie zbiorów danych z plików CSV	69
Importowanie typowych formatów zbiorów danych do RStudio	71
Badanie i rozumienie danych	73
Badanie struktury danych	73
Badanie cech liczbowych	75
Badanie cech kategoriowych	86
Eksplorowanie relacji między cechami	88
Podsumowanie	94

Rozdział 3. Uczenie leniwe — klasyfikacja metodą najbliższych sąsiadów

96

Klasyfikacja metodą najbliższych sąsiadów	97
Algorytm k-NN	97
Dlaczego algorytm k-NN jest „leniwy”?	106
Przykład — diagnozowanie raka piersi za pomocą algorytmu k-NN	107
Etap 1. Zbieranie danych	107
Etap 2. Badanie i przygotowywanie danych	108
Etap 3. Trenowanie modelu na danych	113
Etap 4. Ewaluacja modelu	114
Etap 5. Poprawianie działania modelu	116
Podsumowanie	119

Rozdział 4. Uczenie probabilistyczne — naiwny klasyfikator bayesowski

120

Naiwny klasyfikator bayesowski	121
Podstawowe założenia metod bayesowskich	122
Naiwny klasyfikator bayesowski	128
Przykład — filtrowanie spamu w telefonach komórkowych za pomocą naiwnego klasyfikatora bayesowskiego	135
Etap 1. Zbieranie danych	136
Etap 2. Badanie i przygotowywanie danych	137
Etap 3. Trenowanie modelu na danych	152
Etap 4. Ocena działania modelu	154
Etap 5. Ulepszanie modelu	155
Podsumowanie	156

Rozdział 5. Dziel i zwyciężaj — klasyfikacja z wykorzystaniem drzew decyzyjnych i reguł

158

Drzewa decyzyjne	159
Dziel i zwyciężaj	160
Algorytm drzewa decyzyjnego C5.0	163
Przykład — identyfikowanie ryzykownych pożyczek za pomocą drzew decyzyjnych C5.0	169
Etap 1. Zbieranie danych	169
Etap 2. Badanie i przygotowywanie danych	170

Etap 3. Trenowanie modelu na danych	173
Etap 4. Ocena działania modelu	178
Etap 5. Poprawianie działania modelu	179
Reguły klasyfikacji	183
Wydzielaj i zwyciężaj	184
Algorytm 1R	187
Algorytm RIPPER	189
Reguły z drzew decyzyjnych	190
Dlaczego drzewa i reguły są „zachłanne”?	191
Przykład — identyfikowanie trujących grzybów za pomocą algorytmu uczącego się reguł	194
Etap 1. Zbieranie danych	194
Etap 2. Badanie i przygotowywanie danych	195
Etap 3. Trenowanie modelu na danych	196
Etap 4. Ewaluacja modelu	198
Etap 5. Poprawianie działania modelu	198
Podsumowanie	201

Rozdział 6. Prognozowanie danych liczbowych — metody regresji

203

Regresja	204
Prosta regresja liniowa	206
Metoda zwykłych najmniejszych kwadratów	209
Korelacje	211
Wieloraka regresja liniowa	213
Uogólnione modele liniowe i regresja logistyczna	217
Przykład — przewidywanie kosztów likwidacji szkód z wykorzystaniem regresji liniowej	223
Etap 1. Zbieranie danych	224
Etap 2. Badanie i przygotowywanie danych	226
Etap 3. Trenowanie modelu na danych	231
Etap 4. Ewaluacja modelu	234
Etap 5. Poprawianie działania modelu	236
Krok dalej — przewidywanie odpływu ubezpieczonych z wykorzystaniem regresji logistycznej	242
Drzewa regresji i drzewa modeli	249
Dodawanie regresji do drzew	250
Przykład — ocenianie jakości win za pomocą drzew regresji i drzew modeli	252
Etap 1. Zbieranie danych	253
Etap 2. Badanie i przygotowywanie danych	254
Etap 3. Trenowanie modelu na danych	255
Etap 4. Ewaluacja modelu	259
Etap 5. Poprawianie działania modelu	261
Podsumowanie	264

Rozdział 7. Czarne skrzynki — sieci neuronowe i maszyny wektorów nośnych	266
Sieci neuronowe	267
Od neuronów biologicznych do sztucznych	268
Funkcje aktywacji	270
Topologia sieci	273
Trenowanie sieci neuronowej za pomocą propagacji wstecznej	278
Przykład — modelowanie wytrzymałości betonu za pomocą sieci ANN	281
Etap 1. Zbieranie danych	281
Etap 2. Badanie i przygotowywanie danych	282
Etap 3. Trenowanie modelu na danych	283
Etap 4. Ewaluacja modelu	286
Etap 5. Poprawianie działania modelu	287
Maszyny wektorów nośnych	293
Klasyfikacja za pomocą hiperpłaszczyzn	294
Używanie funkcji jądrowych w przestrzeniach nieliniowych	299
Przykład — optyczne rozpoznawanie znaków za pomocą modelu SVM	301
Etap 1. Zbieranie danych	302
Etap 2. Badanie i przygotowywanie danych	303
Etap 3. Trenowanie modelu na danych	304
Etap 4. Ewaluacja modelu	307
Etap 5. Poprawianie działania modelu	308
Podsumowanie	311
 Rozdział 8. Znajdowanie wzorców — analiza koszyka z wykorzystaniem reguł asocjacyjnych	 312
Reguły asocjacyjne	313
Algorytm Apriori do nauki reguł asocjacyjnych	314
Mierzenie istotności reguł — wsparcie i ufność	316
Budowanie zbioru reguł z wykorzystaniem zasady Apriori	317
Przykład — identyfikowanie często kupowanych artykułów spożywczych za pomocą reguł asocjacyjnych	319
Etap 1. Gromadzenie danych	319
Etap 2. Badanie i przygotowywanie danych	320
Etap 3. Trenowanie modelu na danych	328
Etap 4. Ewaluacja modelu	332
Etap 5. Poprawianie działania modelu	336
Podsumowanie	341
 Rozdział 9. Znajdowanie grup danych — klasteryzacja metodą k-średnich	 342
Klasteryzacja	343
Klasteryzacja jako zadanie uczenia maszynowego	343
Klastry algorytmów klasteryzacji	346

Klasteryzacja metodą k-średnich	350
Znajdowanie segmentów rynkowych wśród nastolatków poprzez klasteryzację metodą k-średnich	358
Etap 1. Zbieranie danych	359
Etap 2. Badanie i przygotowywanie danych	360
Etap 3. Trenowanie modelu na danych	364
Etap 4. Ewaluacja modelu	367
Etap 5. Poprawianie działania modelu	370
Podsumowanie	372
Rozdział 10. Ewaluacja działania modelu	373
Mierzenie trafności klasyfikacji	374
Rozumienie prognoz klasyfikatora	374
Bliższe spojrzenie na macierze błędów	378
Używanie macierzy błędów do mierzenia trafności	380
Nie tylko dokładność — inne miary trafności	382
Wizualizacja kompromisów za pomocą krzywych ROC	394
Szacowanie przyszłej trafności	406
Metoda wstrzymywania	407
Walidacja krzyżowa	411
Próbkowanie bootstrapowe	415
Podsumowanie	416
Rozdział 11. Jak odnieść sukces w uczeniu maszynowym?	418
Co decyduje o sukcesie praktyka uczenia maszynowego?	419
Co decyduje o sukcesie modelu uczenia maszynowego?	422
Unikanie oczywistych prognoz	425
Przeprowadzanie uczciwych ewaluacji	427
Uwzględnianie realiów	431
Budowanie zaufania do modelu	436
Więcej „nauki” w „nauce o danych”	440
Notatniki R i znakowanie R	443
Zaawansowane badanie danych	447
Podsumowanie	466
Rozdział 12. Zaawansowane przygotowywanie danych	467
Inżynieria cech	468
Rola człowieka i maszyny	469
Wpływ big data i uczenia głębokiego	473
Praktyczna inżynieria cech	479
Podpowiedź 1. Znajdź nowe cechy podczas burzy mózgów	481
Podpowiedź 2. Znajdź spostrzeżenia ukryte w tekście	482
Podpowiedź 3. Przekształcaj zakresy liczbowe	484
Podpowiedź 4. Obserwuj zachowanie sąsiadów	485
Podpowiedź 5. Wykorzystaj powiązane wiersze	487

Podpowiedź 6. Dekomponuj szeregi czasowe	488
Podpowiedź 7. Dołącz dane zewnętrzne	492
tidyverse	495
„Schludne” struktury tabelaryczne — obiekty tibble	496
Szybsze odczytywanie plików prostokątnych za pomocą pakietów readr i readxl	497
Przygotowywanie i potokowe przetwarzanie danych za pomocą pakietu dplyr	499
Przekształcanie tekstu za pomocą pakietu stringr	504
Czyszczenie danych za pomocą pakietu lubridate	509
Podsumowanie	513
Rozdział 13. Trudne dane — za duże, za małe, zbyt złożone	514
Dane wysokowymiarowe	515
Stosowanie selekcji cech	516
Ekstrakcja cech	527
Używanie danych rozrzedzonych	539
Identyfikowanie danych rozrzedzonych	540
Przykład — zmiana odwzorowania rozrzedzonych danych kategorycznych	541
Przykład — dzielenie rozrzedzonych danych liczbowych na przedziały	545
Obsługa brakujących danych	549
Typy brakujących danych	550
Imputacja brakujących wartości	552
Problem niezrównoważonych danych	556
Proste strategie przywracania równowagi danych	557
Generowanie syntetycznego zrównoważonego zbioru danych z wykorzystaniem algorytmu SMOTE	559
Czy zrównoważone zawsze znaczy lepsze?	563
Podsumowanie	565
Rozdział 14. Budowanie lepiej uczących się modeli	567
Dostrajanie standardowych modeli	568
Określanie zakresu dostrajania hiperparametrów	569
Przykład — automatyczne dostrajanie za pomocą pakietu caret	573
Zwiększanie trafności modeli za pomocą zespołów	584
Uczenie zespołowe	584
Popularne algorytmy zespołowe	588
Spiętrzanie modeli do celów metanauki	612
Spiętrzanie i mieszanie modeli	614
Praktyczne metody mieszania i spiętrzania w języku R	616
Podsumowanie	619
Rozdział 15. Praca z big data	621
Praktyczne zastosowania uczenia głębokiego	622
Pierwsze kroki w uczeniu głębokim	623

Konwolucyjne sieci neuronowe	630
Uczenie nienadzorowane a big data	640
Reprezentowanie koncepcji wysokowymiarowych jako osadzeń	641
Wizualizacja danych wysokowymiarowych	653
Adaptowanie języka R do obsługi dużych zbiorów danych	664
Odpytywanie baz danych SQL	664
Szybsza praca dzięki przetwarzaniu równoległemu	670
Używanie wyspecjalizowanego sprzętu i algorytmów	677
Podsumowanie	684

oprac. BPK