

Spis treści

Informacje o autorach 11

Informacje o recenzentach 13

Wstęp 15

Rozdział 1. Umożliwianie komputerom uczenia się z danych 23

Tworzenie inteligentnych maszyn służących do przekształcania danych w wiedzę 24

Trzy różne rodzaje uczenia maszynowego 24

Prognostowanie przyszłości za pomocą uczenia nadzorowanego 25

Rozwiązywanie problemów interaktywnych za pomocą uczenia przez wzmacnianie 28

Odkrywanie ukrytych struktur za pomocą uczenia nienadzorowanego 29

Wprowadzenie do podstawowej terminologii i notacji 30

Strategia tworzenia systemów uczenia maszynowego 32

Wstępne przetwarzanie - nadawanie danym formy 32

Trenowanie i dobór modelu predykcyjnego 34

Ewaluacja modeli i przewidywanie wystąpienia nieznanych danych 34

Wykorzystywanie środowiska Python do uczenia maszynowego 35

Instalacja środowiska Python i pakietów z repozytorium Python Package Index 35

Korzystanie z platformy Anaconda i menedżera pakietów 36

Pakiety przeznaczone do obliczeń naukowych, analizy danych i uczenia maszynowego 36

Podsumowanie 37

Rozdział 2. Trenowanie prostych algorytmów uczenia maszynowego w celach klasyfikacji 39

Sztuczne neurony - rys historyczny początków uczenia maszynowego 40

Formalna definicja sztucznego neuronu 41

Reguła uczenia perceptronu 43

Implementacja algorytmu uczenia perceptronu w Pythonie 45

Obiektowy interfejs API perceptronu 45

Trenowanie modelu perceptronu na zestawie danych Iris 48

Adaptacyjne neurony liniowe i zbieżność uczenia 53

Minimalizacja funkcji kosztu za pomocą metody gradientu prostego 55

Implementacja algorytmu Adaline w Pythonie 56

Usprawnianie gradientu prostego poprzez skalowanie cech 60

Wielkoskalowe uczenie maszynowe i metoda stochastycznego spadku wzdłuż gradientu 62

Podsumowanie 66

Rozdział 3. Stosowanie klasyfikatorów uczenia maszynowego za pomocą biblioteki scikit-learn 67

Wybór algorytmu klasyfikującego 68

Pierwsze kroki z biblioteką scikit-learn - uczenie perceptronu 68

Modelowanie prawdopodobieństwa przynależności do klasy za pomocą regresji logistycznej 74

Teoretyczne podłoże regresji logistycznej i prawdopodobieństwa warunkowego 74

Wyznaczanie wag logistycznej funkcji kosztu 78

Przekształcanie implementacji Adaline do postaci algorytmu regresji logistycznej 80

Uczenie modelu regresji logistycznej za pomocą biblioteki scikit-learn 84

Zapobieganie przetrenowaniu za pomocą regularyzacji 86

Wyznaczanie maksymalnego marginesu za pomocą maszyn wektorów nośnych 88

Teoretyczne podłoże maksymalnego marginesu 89

Rozwiązywanie przypadków nieliniowo rozdzielnych za pomocą zmiennych uzupełniających 90

Alternatywne implementacje w interfejsie scikit-learn 92

Rozwiązywanie nieliniowych problemów za pomocą jądra SVM 93

Metody jądrowe dla danych nierozdzielnych liniowo 93

Stosowanie sztuczki z funkcją jądra do znajdowania przestrzeni rozdzielających w przestrzeni wielowymiarowej 95

Uczenie drzew decyzyjnych 99

Maksymalizowanie przyrostu informacji - osiągnięcie jak największych korzyści 100

Budowanie drzewa decyzyjnego 103

Łączenie wielu drzew decyzyjnych za pomocą modelu losowego lasu 107

Algorytm k-najbliższych sąsiadów - model leniwego uczenia 109

Podsumowanie 113

Rozdział 4. Tworzenie dobrych zbiorów uczących - wstępne przetwarzanie danych 115

Kwestia brakujących danych 115

Wykrywanie brakujących wartości w danych tabelarycznych 116

Usuwanie próbek lub cech niezawierających wartości 117

Wstawianie brakujących danych 118

Estymatory interfejsu scikit-learn 119

Przetwarzanie danych kategoryzujących 119

Cechy nominalne i porządkowe 120

Tworzenie przykładowego zestawu danych 120

Mapowanie cech porządkowych 121

Kodowanie etykiet klas 121

Kodowanie "gorącojedynkowe" cech nominalnych (z użyciem wektorów własnych) 122

Rozdzielanie zestawu danych na oddzielne podzbiory uczące i testowe 124

Skalowanie cech 127

Dobór odpowiednich cech 129

Regularyzacje L1 i L2 jako kary ograniczające złożoność modelu 129

Interpretacja geometryczna regularyzacji L2 130

Rozwiązania rzadkie za pomocą regularyzacji L1 131

Algorytmy sekwencyjnego wyboru cech 135

Ocenianie istotności cech za pomocą algorytmu losowego lasu 140

Podsumowanie 142

Rozdział 5. Kompresja danych poprzez redukcję wymiarowości 143

Nienadzorowana redukcja wymiarowości za pomocą analizy głównych składowych 144

Podstawowe etapy analizy głównych składowych 144

Wydobywanie głównych składowych krok po kroku 146

Wyjaśniona wariancja całkowita 148

Transformacja cech 149

Analiza głównych składowych w interfejsie scikit-learn 152

Nadzorowana kompresja danych za pomocą liniowej analizy dyskryminacyjnej 154

Porównanie analizy głównych składowych z liniową analizą dyskryminacyjną 155

Wewnętrzne mechanizmy działania liniowej analizy dyskryminacyjnej 156

Obliczanie macierzy rozproszenia 157

Dobór dyskryminant liniowych dla nowej podprzestrzeni cech 159

Rzutowanie próbek na nową przestrzeń cech 161

Implementacja analizy LDA w bibliotece scikit-learn 161

Jądrowa analiza głównych składowych jako metoda odwzorowywania nierozdzielnych liniowo klas 163

Funkcje jądra oraz sztuczka z funkcją jądra 164

Implementacja jądrowej analizy głównych składowych w Pythonie 168

Rzutowanie nowych punktów danych 175

Algorytm jądrowej analizy głównych składowych w bibliotece scikit-learn 178

Podsumowanie 179

Rozdział 6. Najlepsze metody oceny modelu i strojenie parametryczne 181

Usprawnianie cyklu pracy za pomocą kolejkowania 181

Wczytanie zestawu danych Breast Cancer Wisconsin 182

Łączenie funkcji transformujących i estymatorów w kolejce czynności 183

Stosowanie k-krotnego sprawdzianu krzyżowego w ocenie skuteczności modelu 184

Metoda wydzielenia	185
K-krotny sprawdzian krzyżowy	186
Sprawdzanie algorytmów za pomocą krzywych uczenia i krzywych walidacji	190
Diagnozowanie problemów z obciążeniem i wariancją za pomocą krzywych uczenia	190
Rozwiązywanie problemów przetrenowania i niedotrenowania za pomocą krzywych walidacji	193
Dostrajanie modeli uczenia maszynowego za pomocą metody przeszukiwania siatki	195
Strojenie hiperparametrów przy użyciu metody przeszukiwania siatki	195
Dobór algorytmu poprzez zagnieżdżony sprawdzian krzyżowy	196
Przegląd metryk oceny skuteczności	198
Odczytywanie macierzy pomyłek	198
Optymalizacja precyzji i pełności modelu klasyfikującego	200
Wykres krzywej ROC	202
Metryki zliczające dla klasyfikacji wieloklasowej	204
Kwestia dysproporcji klas	205
Podsumowanie	208
 Rozdział 7. Łączenie różnych modeli w celu uczenia zespołowego	 209
Uczenie zespołów	209
Łączenie klasyfikatorów za pomocą algorytmu głosowania większościowego	213
Implementacja prostego klasyfikatora głosowania większościowego	214
Stosowanie reguły głosowania większościowego do uzyskiwania prognoz	219
Ewaluacja i strojenie klasyfikatora zespołowego	221
Agregacja - tworzenie zespołu klasyfikatorów za pomocą próbek początkowych	226
Agregacja w pigułce	227
Stosowanie agregacji do klasyfikowania przykładów z zestawu Wine	228
Usprawnianie słabych klasyfikatorów za pomocą wzmocnienia adaptacyjnego	231
Wzmacnianie - mechanizm działania	232

Stosowanie algorytmu AdaBoost za pomocą biblioteki scikit-learn 236

Podsumowanie 239

Rozdział 8. Wykorzystywanie uczenia maszynowego w analizie sentymentów 241

Przygotowywanie zestawu danych IMDb movie review do przetwarzania tekstu 242

Uzyskiwanie zestawu danych IMDb 242

Przetwarzanie wstępne zestawu danych IMDb do wygodniejszego formatu 243

Wprowadzenie do modelu worka słów 244

Przekształcanie słów w wektory cech 245

Ocena istotności wyrazów za pomocą ważenia częstości termów - odwrotnej częstości w tekście 246

Oczyszczanie danych tekstowych 248

Przetwarzanie tekstu na znaczniki 249

Uczenie modelu regresji logistycznej w celu klasyfikowania tekstu 251

Praca z większą ilością danych - algorytmy sieciowe i uczenie pozardzeniowe 253

Modelowanie tematyczne za pomocą alokacji ukrytej zmiennej Dirichleta 256

Rozkładanie dokumentów tekstowych za pomocą analizy LDA 257

Analiza LDA w bibliotece scikit-learn 258

Podsumowanie 261

Rozdział 9. Wdrażanie modelu uczenia maszynowego do aplikacji sieciowej 263

Serializacja wyuczonych estymatorów biblioteki scikit-learn 264

Konfigurowanie bazy danych SQLite 266

Tworzenie aplikacji sieciowej za pomocą środowiska Flask 269

Nasza pierwsza aplikacja sieciowa 269

Sprawdzanie i wyświetlanie formularza 271

Przekształcanie klasyfikatora recenzji w aplikację sieciową 275

Pliki i katalogi - wygląd drzewa katalogów 277

Implementacja głównej części programu w pliku app.py	277
Konfigurowanie formularza recenzji	280
Tworzenie szablonu strony wynikowej	281
Umieszczanie aplikacji sieciowej na publicznym serwerze	282
Tworzenie konta w serwisie PythonAnywhere	283
Przesyłanie aplikacji klasyfikatora filmowego	283
Aktualizowanie klasyfikatora recenzji filmowych	284
Podsumowanie	286
Rozdział 10. Przewidywanie ciągłych zmiennych docelowych za pomocą analizy regresywnej	287
Wprowadzenie do regresji liniowej	288
Prosta regresja liniowa	288
Wielowymiarowa regresja liniowa	288
Zestaw danych Housing	290
Wczytywanie zestawu danych Housing do obiektu DataFrame	290
Wizualizowanie ważnych elementów zestawu danych	292
Analiza związków za pomocą macierzy korelacji	293
Implementacja modelu regresji liniowej wykorzystującego zwykłą metodę najmniejszych kwadratów	296
Określanie parametrów regresyjnych za pomocą metody gradientu prostego	296
Szacowanie współczynnika modelu regresji za pomocą biblioteki scikit-learn	300
Uczenie odpornego modelu regresyjnego za pomocą algorytmu RANSAC	301
Ocenianie skuteczności modeli regresji liniowej	304
Stosowanie regularyzowanych metod regresji	307
Przekształcanie modelu regresji liniowej w krzywą - regresja wielomianowa	308
Dodawanie członów wielomianowych za pomocą biblioteki scikit-learn	309
Modelowanie nieliniowych zależności w zestawie danych Housing	310
Analiza nieliniowych relacji za pomocą algorytmu losowego lasu	314

Podsumowanie 318

Rozdział 11. Praca z nieoznakowanymi danymi - analiza skupień 319

Grupowanie obiektów na podstawie podobieństwa przy użyciu algorytmu centroidów 320

Algorytm centroidów w bibliotece scikit-learn 320

324 Inteligentniejszy sposób dobierania pierwotnych centroidów za pomocą algorytmu k-means++

Klasteryzacja twarda i miękka 325

Stosowanie metody łokcia do wyszukiwania optymalnej liczby skupień 327

Ujęcie ilościowe jakości klasteryzacji za pomocą wykresu profilu 328

Organizowanie skupień do postaci drzewa klastrow 333

Oddolne grupowanie skupień 333

Przeprowadzanie hierarchicznej analizy skupień na macierzy odległości 335

Dołączanie dendrogramów do mapy cieplnej 338

Aglomeracyjna analiza skupień w bibliotece scikit-learn 339

Wyznaczanie rejonów o dużej gęstości za pomocą algorytmu DBSCAN 340

Podsumowanie 345

Rozdział 12. Implementowanie wielowarstwowej sieci neuronowej od podstaw 347

Modelowanie złożonych funkcji przy użyciu sztucznych sieci neuronowych 348

Jednowarstwowa sieć neuronowa - powtórzenie 349

Wstęp do wielowarstwowej architektury sieci neuronowych 351

Aktywacja sieci neuronowej za pomocą propagacji w przód 354

Klasyfikowanie pisma odręcznego 356

Zestaw danych MNIST 357

Implementacja perceptronu wielowarstwowego 362

Trenowanie sztucznej sieci neuronowej 371

Obliczanie logistycznej funkcji kosztu 371

Ujęcie intuicyjne algorytmu wstecznej propagacji 374

Uczenie sieci neuronowych za pomocą algorytmu propagacji wstecznej 375

Zbieżność w sieciach neuronowych 378

Jeszcze słowo o implementacji sieci neuronowej 380

Podsumowanie 380

Rozdział 13. Równoległe przetwarzanie sieci neuronowych za pomocą biblioteki TensorFlow 381

Biblioteka TensorFlow a skuteczność uczenia 382

Czym jest biblioteka TensorFlow? 383

W jaki sposób będziemy poznawać bibliotekę TensorFlow? 384

Pierwsze kroki z biblioteką TensorFlow 384

Praca ze strukturami tablicowymi 386

Tworzenie prostego modelu za pomocą podstawowego interfejsu TensorFlow 387

Skuteczne uczenie sieci neuronowych za pomocą wyspecjalizowanych interfejsów biblioteki TensorFlow 391

Tworzenie wielowarstwowych sieci neuronowych za pomocą interfejsu Layers 392

Projektowanie wielowarstwowej sieci neuronowej za pomocą interfejsu Keras 395

Dobór funkcji aktywacji dla wielowarstwowych sieci neuronowych 400

Funkcja logistyczna - powtórzenie 400

Szacowanie prawdopodobieństw przynależności do klas w klasyfikacji wieloklasowej za pomocą funkcji softmax 402

Rozszerzanie zakresu wartości wyjściowych za pomocą funkcji tangensa hiperbolicznego 403

Aktywacja za pomocą prostowanej jednostki liniowej (ReLU) 405

Podsumowanie 407

Rozdział 14. Czas na szczegóły - mechanizm działania biblioteki TensorFlow 409

Główne funkcje biblioteki TensorFlow 410

Rzędy i tensory 410

Sposób uzyskania rzędu i wymiarów tensora	411
Grafy obliczeniowe	412
Węzły zastępcze	414
Definiowanie węzłów zastępczych	414
Wypełnianie węzłów zastępczych danymi	415
Definiowanie węzłów zastępczych dla tablic danych o różnych rozmiarach pakietów danych	416
Zmienne	417
Definiowanie zmiennych	417
Inicjowanie zmiennych	419
Zakres zmiennych	420
Wielokrotne wykorzystywanie zmiennych	421
Tworzenie modelu regresyjnego	423
Realizowanie obiektów w grafie TensorFlow przy użyciu ich nazw	426
Zapisywanie i wczytywanie modelu	428
Przekształcanie tensorów jako wielowymiarowych tablic danych	430
Wykorzystywanie mechanizmów przebiegu sterowania do tworzenia grafów	433
Wizualizowanie grafów za pomocą modułu TensorBoard	436
Zdobywanie doświadczenia w używaniu modułu TensorBoard	439
Podsumowanie	440
 Rozdział 15. Klasyfikowanie obrazów za pomocą splotowych sieci neuronowych	441
Podstawowe elementy splotowej sieci neuronowej	442
Splotowe sieci neuronowe i hierarchie cech	442
Splot dyskretny	444
Podpróbkiowanie	452
Konstruowanie sieci CNN	454
Praca z wieloma kanałami wejściowymi/barw	454

Regularyzowanie sieci neuronowej metodą porzucania 457

Implementacja głębokiej sieci splotowej za pomocą biblioteki TensorFlow 459

Architektura wielowarstwowej sieci CNN 459

Wczytywanie i wstępne przetwarzanie danych 460

Implementowanie sieci CNN za pomocą podstawowego interfejsu TensorFlow 461

Implementowanie sieci CNN za pomocą interfejsu Layers 471

Podsumowanie 476

Rozdział 16. Modelowanie danych sekwencyjnych za pomocą rekurencyjnych sieci neuronowych 477

Wprowadzenie do danych sekwencyjnych 478

Modelowanie danych sekwencyjnych - kolejność ma znaczenie 478

Przedstawianie sekwencji 478

Różne kategorie modelowania sekwencji 479

Sieci rekurencyjne służące do modelowania sekwencji 480

Struktura sieci RNN i przepływ danych 480

Obliczanie aktywacji w sieciach rekurencyjnych 482

Problemy z uczeniem długofalowych oddziaływań 485

Jednostki LSTM 486

Implementowanie wielowarstwowej sieci rekurencyjnej przy użyciu biblioteki TensorFlow do modelowania sekwencji 488

Pierwszy projekt - analiza sentymentów na zestawie danych IMDb za pomocą wielowarstwowej sieci rekurencyjnej 489

Przygotowanie danych 489

Wektor właściwościowy 492

Budowanie modelu sieci rekurencyjnej 494

Konstruktor klasy SentimentRNN 495

Metoda build 495

Metoda train 499

Metoda predict	500
Tworzenie wystąpienia klasy SentimentRNN	500
Uczenie i optymalizowanie modelu sieci rekurencyjnej przeznaczonej do analizy sentymentów	501
Drugi projekt - implementowanie sieci rekurencyjnej modelującej język na poziomie znaków	502
Przygotowanie danych	503
Tworzenie sieci RNN przetwarzającej znaki	506
Konstruktor	506
Metoda build	507
Metoda train	509
Metoda sample	510
Tworzenie i uczenie modelu CharRNN	512
Model CharRNN w trybie próbkowania	512
Podsumowanie rozdziału i książki	513
Skorowidz	515