

Spis treści

O autorze 11

O korektorach merytorycznych 13

Przedmowa 15

Rozdział 1. Pierwsze kroki z uczeniem maszynowym w Pythonie 19

Wprowadzenie do uczenia maszynowego 20

Dlaczego uczenie maszynowe jest potrzebne? 20

Różnice między uczeniem maszynowym a automatyką 22

Zastosowania uczenia maszynowego 23

Wstępne wymagania 24

Trzy rodzaje uczenia maszynowego 25

Krótką historia rozwoju algorytmów uczenia maszynowego 27

Istota uczenia maszynowego 28

Uogólnianie danych 29

Nadmierne i niedostateczne dopasowanie modelu

oraz kompromis między obciążeniem a wariancją 30

Zapobieganie nadmiernemu dopasowaniu poprzez weryfikację krzyżową 34

Zapobieganie nadmiernemu dopasowaniu za pomocą regularyzacji 37

Zapobieganie nadmiernemu dopasowaniu poprzez selekcję cech i redukcję
wymiarowości 39

Wstępne przetwarzanie danych i inżynieria cech 40

Wstępne przetwarzanie i eksploracja danych 41

Inżynieria cech 44

Łączenie modeli 45

Głosowanie i uśrednianie 46

Agregacja bootstrap 46

Wzmacnianie 46

Składowanie 49

Instalacja i konfiguracja oprogramowania 49

Przygotowanie Pythona i środowiska pracy	50
Instalacja najważniejszych pakietów Pythona	51
Wprowadzenie do pakietu TensorFlow 2	53
Podsumowanie	54
Ćwiczenia	54
Rozdział 2. Tworzenie systemu rekomendacji filmów na bazie naiwnego klasyfikatora Bayesa	55
Pierwsze kroki z klasyfikacją	56
Klasyfikacja binarna	56
Klasyfikacja wieloklasowa	57
Klasyfikacja wieloetykietowa	59
Naiwny klasyfikator Bayesa	59
Twierdzenie Bayesa w przykładach	60
Mechanizm naiwnego klasyfikatora Bayesa	62
Implementacja naiwnego klasyfikatora Bayesa	65
Implementacja od podstaw	66
Implementacja z wykorzystaniem pakietu scikit-learn	69
Budowanie systemu rekomendacyjnego na bazie klasyfikatora Bayesa	69
Ocena jakości klasyfikacji	74
Strojenie modeli poprzez weryfikację krzyżową	78
Podsumowanie	80
Ćwiczenia	81
Bibliografia	81
Rozdział 3. Rozpoznawanie twarzy przy użyciu maszyny wektorów nośnych	82
Określanie granic klas za pomocą maszyny wektorów nośnych	83
Scenariusz 1. Określenie hiperpłaszczyzny rozdzielającej	84
Scenariusz 2. Określenie optymalnej hiperpłaszczyzny rozdzielającej	84
Scenariusz 3. Przetwarzanie punktów odstających	88
Implementacja maszyny wektorów nośnych	88
Scenariusz 4. Więcej niż dwie klasy	90
Scenariusz 5. Rozwiązywanie nierozdzielnego liniowo problemu	

za pomocą jądra 94

Wybór między jądrem liniowym a radialną funkcją bazową 98

Klasyfikowanie zdjęć twarzy za pomocą maszyny wektorów nośnych 101

Badanie zbioru zdjęć twarzy 101

Tworzenie klasyfikatora obrazów opartego na maszynie wektorów nośnych 103

Zwiększanie skuteczności klasyfikatora obrazów za pomocą analizy głównych składowych 105

Klasyfikacja stanu płodu w kardiokardigrafii 106

Podsumowanie 108

Ćwiczenia 108

Rozdział 4. Prognozowanie kliknięć reklam internetowych

przy użyciu algorytmów drzewiastych 109

Wprowadzenie do prognozowania kliknięć reklam 110

Wprowadzenie do dwóch typów danych: liczbowych i kategoryalnych 111

Badanie drzewa decyzyjnego od korzeni do liści 112

Budowanie drzewa decyzyjnego 113

Wskaźniki jakości podziału zbioru 116

Implementacja drzewa decyzyjnego od podstaw 122

Implementacja drzewa decyzyjnego za pomocą biblioteki scikit-learn 129

Prognozowanie kliknięć reklam za pomocą drzewa decyzyjnego 129

Gromadzenie drzew decyzyjnych: las losowy 135

Gromadzenie drzew decyzyjnych: drzewa ze wzmocnieniem gradientowym 137

Podsumowanie 139

Ćwiczenia 140

Rozdział 5. Prognozowanie kliknięć reklam internetowych

przy użyciu regresji logistycznej 141

Przekształcanie cech kategoryalnych w liczbowe: kodowanie porządkowe i „1 z n” 142

Klasyfikowanie danych z wykorzystaniem regresji logistycznej 144

Wprowadzenie do funkcji logistycznej 145

Przejście od funkcji logistycznej do regresji logistycznej 146

Trening modelu opartego na regresji logistycznej 149

Trening modelu opartego na regresji logistycznej z gradientem prostym	150
Prognozowanie kliknięć reklam z wykorzystaniem regresji logistycznej z gradientem prostym	154
Trening modelu opartego na regresji logistycznej ze stochastycznym gradientem prostym	156
Trening modelu opartego na regresji logistycznej z regularyzacją	158
Selekcja cech w regularyzacji L1	159
Trening modelu na dużym zbiorze danych z uczeniem online	160
Klasyfikacja wieloklasowa	163
Implementacja regresji logistycznej za pomocą pakietu TensorFlow	165
Selekcja cech z wykorzystaniem lasu losowego	167
Podsumowanie	168
Ćwiczenia	168
Rozdział 6. Skalowanie modelu prognozującego do terabajtowych dzienników kliknięć	169
Podstawy Apache Spark	170
Komponenty	170
Instalacja	172
Uruchamianie i wdrażanie programów	173
Programowanie z wykorzystywaniem modułu PySpark	174
Trenowanie modelu na bardzo dużych zbiorach danych za pomocą narzędzia Apache Spark	177
Ładowanie danych o kliknięciach reklam	177
Podzielenie danych i umieszczenie ich w pamięci	180
Zakodowanie „1 z n” cech kategoryalnych	181
Trening i testy modelu regresji logistycznej	184
Inżynieria cech i wartości kategoryalnych przy użyciu narzędzia Apache Spark	186
Mieszanie cech kategoryalnych	186
Interakcja cech, czyli łączenie zmiennych	189
Podsumowanie	192
Ćwiczenia	192

Rozdział 7. Prognozowanie cen akcji za pomocą algorytmów regresji 194

Krótkie wprowadzenie do giełdy i cen akcji 195

Co to jest regresja? 196

Pozyskiwanie cen akcji 197

Pierwsze kroki z inżynierią cech 199

Pozyskiwanie danych i generowanie cech 202

Szacowanie za pomocą regresji liniowej 205

Jak działa regresja liniowa? 206

Implementacja regresji liniowej od podstaw 207

Implementacja regresji liniowej z wykorzystaniem pakietu scikit-learn 210

Implementacja regresji liniowej z wykorzystaniem pakietu TensorFlow 211

Prognozowanie za pomocą regresyjnego drzewa decyzyjnego 212

Przejście od drzewa klasyfikacyjnego do regresyjnego 212

Implementacja regresyjnego drzewa decyzyjnego 214

Implementacja lasu regresyjnego 218

Prognozowanie za pomocą regresji wektorów nośnych 218

Implementacja regresji wektorów nośnych 219

Ocena jakości regresji 220

Prognozowanie cen akcji za pomocą trzech algorytmów regresji 222

Podsumowanie 225

Ćwiczenia 226

Rozdział 8. Prognozowanie cen akcji za pomocą sieci neuronowych 227

Demistyfikacja sieci neuronowych 228

Pierwsze kroki z jednowarstwową siecią neuronową 228

Funkcje aktywacji 229

Propagacja wstecz 231

Wprowadzanie kolejnych warstw do sieci neuronowej i uczenie głębokie 232

Tworzenie sieci neuronowej 234

Implementacja sieci neuronowej od podstaw 234

Implementacja sieci neuronowej przy użyciu pakietu scikit-learn 237

Implementacja sieci neuronowej przy użyciu pakietu TensorFlow 237

Dobór właściwej funkcji aktywacji 239

Zapobieganie nadmiernemu dopasowaniu sieci 240

Dropout 240

Wczesne zakończenie treningu 241

Prognozowanie cen akcji za pomocą sieci neuronowej 242

Trening prostej sieci neuronowej 242

Dostrojenie parametrów sieci neuronowej 243

Podsumowanie 248

Ćwiczenie 249

Rozdział 9. Badanie 20 grup dyskusyjnych przy użyciu technik analizy tekstu 250

Jak komputery rozumieją ludzi, czyli przetwarzanie języka naturalnego 251

Czym jest przetwarzanie języka naturalnego? 251

Historia przetwarzania języka naturalnego 252

Zastosowania przetwarzania języka naturalnego 253

Przegląd bibliotek Pythona i podstawy przetwarzania języka naturalnego 254

Instalacja najważniejszych bibliotek 254

Korpusy 255

Tokenizacja 257

Oznaczanie części mowy 258

Rozpoznawanie jednostek nazwanych 259

Stemming i lematyzacja 260

Modelowanie semantyczne i tematyczne 261

Pozyskiwanie danych z grup dyskusyjnych 261

Badanie danych z grup dyskusyjnych 264

Przetwarzanie cech danych tekstowych 267

Zliczanie wystąpień wszystkich tokenów 267

Wstępne przetwarzanie tekstu 270

Usuwanie stop-słów 270

Upraszczenie odmian 271

Wizualizacja danych tekstowych z wykorzystaniem techniki t-SNE 272

Co to jest redukcja wymiarowości? 272

Redukcja wymiarowości przy użyciu techniki t-SNE 273

Podsumowanie 276

Ćwiczenia 276

Rozdział 10. Wyszukiwanie ukrytych tematów w grupach dyskusyjnych
poprzez ich klastrowanie i modelowanie tematyczne 277

Nauka bez wskazówek, czyli uczenie nienadzorowane 278

Klastrowanie grup dyskusyjnych metodą k-średnich 280

Jak działa klastrowanie metodą k-średnich? 280

Implementacja klastrowania metodą k-średnich od podstaw 281

Implementacja klastrowania metodą k-średnich

z wykorzystaniem pakietu scikit-learn 289

Dobór wartości k 291

Klastrowanie danych z grup dyskusyjnych metodą k-średnich 293

Odkrywanie ukrytych tematów grup dyskusyjnych 296

Modelowanie tematyczne z wykorzystaniem nieujemnej faktoryzacji macierzy 297

Modelowanie tematyczne z wykorzystaniem ukrytej alokacji Dirichleta 300

Podsumowanie 303

Ćwiczenia 304

Rozdział 11. Dobre praktyki uczenia maszynowego 305

Proces rozwiązywania problemów uczenia maszynowego 306

Dobre praktyki przygotowywania danych 307

Dobra praktyka nr 1. Dokładne poznanie celu projektu 307

Dobra praktyka nr 2. Zbieranie wszystkich istotnych pól 307

Dobra praktyka nr 3. Ujednolicenie danych 308

Dobra praktyka nr 4. Opracowanie niekompletnych danych 308

Dobra praktyka nr 5. Przechowywanie dużych ilości danych 311

Dobre praktyki tworzenia zbioru treningowego 312

Dobra praktyka nr 6. Oznaczanie cech kategoryalnych liczbami 312

Dobra praktyka nr 7. Rozważenie kodowania cech kategoryalnych 313

Dobra praktyka nr 8. Rozważenie selekcji cech i wybór odpowiedniej metody 313

Dobra praktyka nr 9. Rozważenie redukcji wymiarowości

i wybór odpowiedniej metody 314

Dobra praktyka nr 10. Rozważenie normalizacji cech 315

Dobra praktyka nr 11. Inżynieria cech na bazie wiedzy eksperckiej 316

Dobra praktyka nr 12. Inżynieria cech bez wiedzy eksperckiej 316

Dobra praktyka nr 13. Dokumentowanie procesu tworzenia cech 318

Dobra praktyka nr 14. Wyodrębnianie cech z danych tekstowych 318

Dobre praktyki trenowania, oceniania i wybierania modelu 322

Dobra praktyka nr 15. Wybór odpowiedniego algorytmu początkowego 323

Dobra praktyka nr 16. Zapobieganie nadmiernemu dopasowaniu 325

Dobra praktyka nr 17. Diagnozowanie nadmiernego

i niedostatecznego dopasowania 325

Dobra praktyka nr 18. Modelowanie dużych zbiorów danych 327

Dobre praktyki wdrażania i monitorowania modelu 328

Dobra praktyka nr 19. Zapisywanie, ładowanie

i wielokrotne stosowanie modelu 328

Dobra praktyka nr 20. Monitorowanie skuteczności modelu 330

Dobra praktyka nr 21. Regularne aktualizowanie modelu 331

Podsumowanie 331

Ćwiczenia 331

Rozdział 12. Kategoryzacja zdjęć odzieży

przy użyciu konwolucyjnej sieci neuronowej 332

Bloki konstrukcyjne konwolucyjnej sieci neuronowej 333

Warstwa konwolucyjna 333

Warstwa nieliniowa 335

Warstwa redukująca 335

Budowanie konwolucyjnej sieci neuronowej na potrzeby klasyfikacji 337

Badanie zbioru zdjęć odzieży 338

Klasyfikowanie zdjęć odzieży za pomocą konwolucyjnej sieci neuronowej 342

Tworzenie sieci 342

Trening sieci 345

Wizualizacja filtrów konwolucyjnych 347

Wzmacnianie konwolucyjnej sieci neuronowej poprzez uzupełnianie danych	349
Odwracanie obrazów w poziomie i pionie	349
Obracanie obrazów	351
Przesuwanie obrazów	352
Usprawnianie klasyfikatora obrazów poprzez uzupełnianie danych	354
Podsumowanie	356
Ćwiczenia	356
Rozdział 13. Prognozowanie sekwencji danych przy użyciu rekurencyjnej sieci neuronowej	357
Wprowadzenie do uczenia sekwencyjnego	358
Architektura rekurencyjnej sieci neuronowej na przykładzie	358
Mechanizm rekurencyjny	359
Sieć typu „wiele do jednego”	361
Sieć typu „jedno do wielu”	362
Sieć synchroniczna typu „wiele do wielu”	362
Sieć niesynchroniczna typu „wiele do wielu”	363
Trening rekurencyjnej sieci neuronowej	364
Długoterminowe zależności i sieć LSTM	365
Analiza recenzji filmowych za pomocą sieci neuronowej	367
Analiza i wstępne przetworzenie recenzji	368
Zbudowanie prostej sieci LSTM	370
Poprawa skuteczności poprzez wprowadzenie dodatkowych warstw	372
Pisanie nowej powieści „Wojna i pokój” za pomocą rekurencyjnej sieci neuronowej	374
Pozyskanie i analiza danych treningowych	374
Utworzenie zbioru treningowego dla generatora tekstu	376
Utworzenie generatora tekstu	378
Trening generatora tekstu	380
Zaawansowana analiza języka przy użyciu modelu Transformer	383
Architektura modelu	383
Samouwaga	383
Podsumowanie	386

Ćwiczenia 386

Rozdział 14. Podejmowanie decyzji w skomplikowanych warunkach
z wykorzystaniem uczenia przez wzmacnianie 387

Przygotowanie środowiska do uczenia przez wzmacnianie 388

Instalacja biblioteki PyTorch 388

Instalacja narzędzi OpenAI Gym 390

Wprowadzenie do uczenia przez wzmacnianie z przykładami 391

Komponenty uczenia przez wzmacnianie 391

Sumaryczna nagroda 392

Algorytmy uczenia przez wzmacnianie 393

Problem FrozenLake i programowanie dynamiczne 394

Utworzenie środowiska FrozenLake 394

Rozwiązanie problemu przy użyciu algorytmu iteracji wartości 397

Rozwiązanie problemu przy użyciu algorytmu iteracji polityki 400

Metoda Monte Carlo uczenia przez wzmacnianie 403

Utworzenie środowiska Blackjack 403

Ocenianie polityki w metodzie Monte Carlo 405

Sterowanie Monte Carlo z polityką 407

Problem taksówkarza i algorytm Q-uczenia 411

Utworzenie środowiska Taxi 411

Implementacja algorytmu Q-uczenia 414

Podsumowanie 418

Ćwiczenia 418

Skorowidz 419