
Praktyczne uczenie nienadzorowane przy użyciu języka Python

*Jak budować użytkowe rozwiązania uczenia
maszynowego na podstawie
nieoznakowanych danych.*

Ankur A. Patel

przekład: Jakub Niedźwiedź

Spis treści

Wstęp	xi
-------------	----

Część I. Podstawy uczenia nienadzorowanego

1. Uczenie nienadzorowane w ekosystemie uczenia maszynowego	3
Podstawowa terminologia związana z uczeniem maszynowym	3
System oparty na zasadach a uczenie maszynowe	4
Uczenie nadzorowane a nienadzorowane	5
Mocne i słabe strony uczenia nadzorowanego	6
Mocne i słabe strony uczenia nienadzorowanego	7
Używanie uczenia nienadzorowanego do poprawy rozwiązań wykorzystujących uczenie maszynowe ..	8
Bliższe spojrzenie na algorytmy nadzorowane	10
Metody liniowe	12
Metody oparte na sąsiedztwie	13
Metody oparte na drzewach	14
Maszyny wektorów nośnych	16
Sieci neuronowe	16
Bliższe spojrzenie na algorytmy nienadzorowane	16
Redukcja wymiarowości	17
Analiza skupień	19
Wyodrębnianie cech	21
Nienadzorowane uczenie głębokie	22
Problemy z danymi sekwencyjnymi przy użyciu uczenia nienadzorowanego	24
Uczenie wzmacniane przy użyciu uczenia nienadzorowanego	25
Uczenie pół-nadzorowane	26
Udane zastosowania uczenia nienadzorowanego	26
Wykrywanie anomalii	26
Podsumowanie	28

2. Kompleksowy projekt uczenia maszynowego	29
Konfiguracja środowiska	29
Kontrola wersji: Git.....	29
Klonowanie repozytorium Git dla tej książki	30
Biblioteki naukowe: dystrybucja Anaconda dla języka Python	30
Sieci neuronowe: TensorFlow i Keras.....	30
Wzmacnianie gradientowe, wersja pierwsza: XGBoost.....	31
Wzmacnianie gradientowe, wersja druga: LightGBM.....	31
Algorytmy analizy skupień (grupowania).....	32
Interaktywne środowisko obliczeniowe: Jupyter Notebook	32
Przegląd danych	32
Przygotowanie danych.....	33
Pozyskiwanie danych.....	33
Badanie danych	35
Generowanie macierzy cech tablicy oznakowań	38
Konstruowanie cech i wybieranie cech	39
Wizualizacja danych.....	40
Przygotowanie modelu	41
Podział na zestaw szkoleniowy i testowy	41
Wybranie funkcji kosztu	41
Tworzenie zestawów k-krotnego sprawdzania krzyżowego.....	42
Modele uczenia maszynowego (część I)	43
Model #1: Regresja logistyczna	43
Metryki oceny	46
Macierz pomyłek	46
Krzywa precyzji-czułości	47
Krzywa ROC.....	49
Modele uczenia maszynowego (część II).....	51
Model #2: Losowe lasy.....	51
Model #3: Automat wzmacniania gradientowego (XGBoost)	54
Model #4: Automat wzmacniania gradientowego (LightGBM)	57
Ocena czterech modeli przy użyciu zestawu testowego	60
Zespoły modeli	64
Układanie warstwowe.....	65
Ostateczny wybór modelu.....	68
Potok produkcyjny	69
Podsumowanie.....	70

Część II. Uczenie nienadzorowane przy użyciu Scikit-Learn

3. Redukcja wymiarowości	73
Motywacja do redukcji wymiarowości	73
Baza danych MNIST	74
Algorytmy redukcji wymiarowości	78
Rzutowanie liniowe a uczenie rozmaitościowe	78
Analiza głównych składowych	78
Pojęcie analizy PCA	78
Analiza PCA w praktyce	79
Przyrostowa analiza PCA	84
Rzadka analiza PCA	84
Rdzeniowa analiza PCA	86
Rozkład według wartości osobliwych	87
Losowe rzutowanie	88
Losowe rzutowanie Gaussa	88
Rzadkie losowe rzutowanie	89
Isomap	90
Skalowanie wielowymiarowe	91
Lokalnie liniowe osadzanie	92
Stochastyczne osadzanie sąsiadów z t-rozkładem	93
Inne metody redukcji wymiarowości	95
Uczenie słownikowe	95
Analiza niezależnych składowych	96
Podsumowanie	97
4. Wykrywanie anomalii	99
Wykrywanie oszustw na kartach kredytowych	100
Przygotowanie danych	100
Definiowanie funkcji oceniającej anomalie	100
Definiowanie metryk oceny	101
Definiowanie funkcji wykreślającej	103
Wykrywanie anomalii przy użyciu normalnej analizy PCA	103
Składowe PCA równe liczbie oryginalnych wymiarów	104
Szukanie optymalnej liczby głównych składowych	106
Wykrywanie anomalii przy użyciu rzadkiej analizy PCA	108
Wykrywanie anomalii przy użyciu rdzeniowej analizy PCA	111
Wykrywanie anomalii przy użyciu losowego rzutowania Gaussa	113
Wykrywanie anomalii przy użyciu rzadkiego losowego rzutowania	115
Nieliniowe wykrywanie anomalii	116
Wykrywanie anomalii przy użyciu uczenia słownikowego	117

Wykrywanie anomalii przy użyciu ICA	119
Wykrywanie oszustw na zestawie testowym	120
Wykrywanie anomalii na zestawie testowym przy użyciu normalnej analizy PCA	120
Wykrywanie anomalii na zestawie testowym przy użyciu analizy ICA	122
Wykrywanie anomalii na zestawie testowym przy użyciu uczenia słownikowego	124
Podsumowanie	125
5. Analiza skupień	127
Zestaw danych MNIST	128
Przygotowanie danych	128
Algorytmy analizy skupień (grupowania)	129
k-średnich	130
Bezładność k-średnich	130
Ocena wyników grupowania	131
Dokładność k-średnich	133
k-średnich a liczba głównych składowych	134
k-średnich na oryginalnym zestawie danych	136
Grupowanie hierarchiczne	137
Aglomeracyjne grupowanie hierarchiczne	138
Dendrogram	139
Ocena wyników grupowania	141
DBSCAN	143
Algorytm DBSCAN	143
Zastosowanie DBSCAN wobec naszego zestawu danych	144
HDBSCAN	145
Podsumowanie	147
6. Segmentacja grup	149
Dane Lending Club	149
Przygotowanie danych	150
Przekształcenie formatu tekstowego w format liczbowy	151
Przypisywanie brakujących wartości	152
Konstruowanie cech	154
Wybieranie ostatecznego zestawu cech i przeprowadzanie skalowania	154
Wyznaczanie oznakowań do oceny	155
Ocena grup	156
Aplikacja k-średnich	158
Aplikacja grupowania hierarchicznego	160
Aplikacja HDBSCAN	164
Podsumowanie	166

Część III. Uczenie nienadzorowane przy użyciu TensorFlow i Keras

7. Autokodery	169
Sieci neuronowe	170
TensorFlow	171
Keras	172
Autokoder: koder i dekodek	173
Autokodery niepełne	173
Autokodery nadmiarowe	174
Gęste i rzadkie autokodery	175
Autokoder odszumiający	175
Autokoder wariacyjny	176
Podsumowanie	176
8. Praktyczny autokoder	179
Przygotowanie danych	179
Elementy składowe autokodera	182
Funkcje aktywacji	182
Nasz pierwszy autokoder	183
Funkcja straty	184
Optymalizator	184
Szkolenie modelu	185
Ocenianie na zestawie testowym	187
Dwuwarstwowy, niepełny autokoder z liniową funkcją aktywacji	190
Zwiększanie liczby węzłów	193
Dodawanie więcej ukrytych warstw	195
Autokoder nieliniowy	196
Nadmiarowy autokoder z aktywacją liniową	198
Nadmiarowy autokoder z aktywacją liniową i wykluczeniem	201
Rzadki, nadmiarowy autokoder z aktywacją liniową	203
Rzadki, nadmiarowy autokoder z aktywacją liniową i wykluczeniem	205
Praca z zasumionymi zestawami danych	207
Autokoder odszumiający	208
Dwuwarstwowy, odszumiający, niepełny autokoder z aktywacją liniową	208
Dwuwarstwowy, odszumiający, nadmiarowy autokoder z aktywacją liniową	211
Dwuwarstwowy, odszumiający, nadmiarowy autokoder z aktywacją ReLU	213
Podsumowanie	215
9. Uczenie pół-nadzorowane	217
Przygotowanie danych	217
Model nadzorowany	220

Model nienadzorowany	222
Model pół-nadzorowany	224
Siła uczenia nadzorowanego i nienadzorowanego	226
Podsumowanie	227

Część IV. Głębokie uczenie nienadzorowane przy użyciu TensorFlow i Keras

10. Systemy rekomendacyjne przy użyciu ograniczonych automatów Boltzmanna	231
Automaty Boltzmanna	231
Ograniczone automaty Boltzmanna	232
Systemy rekomendacyjne	233
Filtrowanie kolektywne	233
The Netflix Prize	234
Zestaw danych MovieLens	234
Przygotowanie danych	234
Definiowanie funkcji kosztu: błąd średniokwadratowy	238
Przeprowadzenie podstawowych eksperymentów	239
Rozkład macierzy	240
Jeden utajony czynnik	241
Trzy utajone czynniki	242
Pięć utajonych czynników	243
Filtrowanie kolektywne przy użyciu automatów RBM	243
Architektura sieci neuronowej automatu RBM	244
Budowanie składników klasy RBM	245
Szkolenie systemu rekomendacyjnego opartego na automacie RBM	248
Podsumowanie	249
11. Wykrywanie cech przy użyciu sieci głębokiego przekonania	251
Sieci głębokiego przekonania w szczegółach	251
Klasyfikacja obrazów MNIST	252
Ograniczone automaty Boltzmanna	254
Budowanie składników klasy RBM	254
Generowanie obrazów przy użyciu modelu RBM	257
Przeglądanie pośrednich detektorów cech	257
Szkolenie trzech automatów RBM dla sieci DBN	258
Badanie detektorów cech	260
Przeglądanie wygenerowanych obrazów	261
Pełna sieć DBN	265
Jak działa szkolenie sieci DBN	269
Szkolenie sieci DBN	269

Jak uczenie nienadzorowane pomaga uczeniu nadzorowanemu	270
Generowanie obrazów do zbudowania lepszego klasyfikatora obrazów	271
Klasyfikator obrazów wykorzystujący LightGBM	278
Tylko nadzorowany	278
Rozwiązanie nienadzorowane i nadzorowane	279
Podsumowanie	280
12. Generujące sieci antagonistyczne	283
Pojęcie sieci GAN	283
Siła sieci GAN	284
Głębokie splotowe sieci GAN	284
Splotowe sieci neuronowe	285
Powrót do sieci DCGAN	289
Generator sieci DCGAN	290
Dyskryminator sieci DCGAN	291
Modele dyskryminatora i antagonistyczny	292
Sieć DCGAN dla zestawu danych MNIST	293
Sieć DCGAN dla MNIST w działaniu	295
Generowanie syntetycznych obrazów	296
Podsumowanie	297
13. Grupowanie szeregów czasowych	299
Dane z EKG	300
Podejście do grupowania szeregów czasowych	300
k-kształtów	300
Grupowanie szeregów czasowych przy użyciu k-kształtów na danych ECGFiveDays	301
Przygotowanie danych	301
Szkolenie i ocena	306
Grupowanie szeregów czasowych przy użyciu k-kształtów na danych ECG5000	307
Przygotowanie danych	307
Szkolenie i ocena	311
Grupowanie szeregów czasowych przy użyciu k-średnich na danych ECG5000	313
Grupowanie szeregów czasowych przy użyciu hierarchicznego DBSCAN na danych ECG5000	314
Porównanie algorytmów grupowania szeregów czasowych	314
Pełny przebieg dla algorytmu k-kształtów	315
Pełny przebieg dla algorytmu k-średnich	317
Pełny przebieg dla algorytmu HDBSCAN	318
Porównanie wszystkich trzech podejść do grupowania szeregów czasowych	319
Podsumowanie	321
14. Podsumowanie	323
Uczenie nadzorowane	324

Uczenie nienadzorowane	324
Scikit-Learn	325
TensorFlow i Keras	325
Uczenie wzmacniane	326
Najbardziej obiecujące obecnie obszary uczenia nienadzorowanego	327
Przyszłość uczenia nienadzorowanego	328
Ostatnia uwaga	329
<i>Indeks</i>	331
<i>O autorze</i>	339
<i>Kolofon</i>	339