

Spis treści

Przedmowa do drugiego wydania	11
Przedmowa do pierwszego wydania	14
1. Wprowadzenie	17
Znaczenie danych	17
Czym jest analiza danych?	17
Hipotetyczna motywacja	18
Określanie najważniejszych węzłów	19
Analitycy, których możesz znać	21
Wynagrodzenie i doświadczenie	23
Płatne konta	25
Tematy interesujące użytkowników	26
Co dalej?	27
2. Błyskawiczny kurs Pythona	29
Zasady tworzenia kodu Pythona	29
Skąd wziąć interpreter Pythona?	30
Środowiska wirtualne	30
Formatowanie za pomocą białych znaków	31
Moduły	32
Polskie znaki diakrytyczne	33
Funkcje	33
Łańcuchy	34
Wyjątki	35
Listy	35
Krotki	36
Słowniki	37
defaultdict	38
Counter	39
Zbiory	39
Przepływ sterowania	40
Wartości logiczne	41
Sortowanie	42
Składanie list	42
Testy automatyczne i instrukcja assert	43
Programowanie obiektowe	43
Obiekty iterowalne i generatory	45
Losowość	46
Wyrażenia regularne	47

Narzędzia funkcyjne	48
Funkcja zip i rozpakowywanie argumentów	48
Argumenty nazwane i nienazwane	49
Adnotacje typów	50
Jak pisać adnotacje typów	52
Witaj w firmie DataSciencester!	53
Dalsza eksploracja	53
3. Wizualizacja danych	55
Pakiet matplotlib	55
Wykres słupkowy	57
Wykresy liniowe	60
Wykresy punktowe	60
Dalsza eksploracja	63
4. Algebra liniowa	65
Wektory	65
Macierze	69
Dalsza eksploracja	71
5. Statystyka	73
Opis pojedynczego zbioru danych	73
Tendencje centralne	74
Dyspersja	76
Korelacja	78
Paradoks Simpsona	80
Inne pułapki związane z korelacją	81
Korelacja i przyczynowość	81
Dalsza eksploracja	82
6. Prawdopodobieństwo	83
Zależność i niezależność	83
Prawdopodobieństwo warunkowe	84
Twierdzenie Bayesa	85
Zmienne losowe	87
Ciągły rozkład prawdopodobieństwa	87
Rozkład normalny	89
Centralne twierdzenie graniczne	91
Dalsza eksploracja	93
7. Hipotezy i wnioski	95
Sprawdzanie hipotez	95
Przykład: rzut monetą	95
Wartości p	98
Przedziały ufności	99
Hakowanie wartości p	100
Przykład: przeprowadzanie testu A-B	101

Wnioskowanie bayesowskie	102
Dalsza eksploracja	105
8. Metoda gradientu prostego	107
Podstawy metody gradientu prostego	107
Szacowanie gradientu	108
Korzystanie z gradientu	111
Dobór właściwego rozmiaru kroku	111
Używanie metody gradientu do dopasowywania modeli	112
Metody gradientu prostego: stochastyczna i minibatch	113
Dalsza eksploracja	114
9. Uzyskiwanie danych	117
Strumienie stdin i stdout	117
Wczytywanie plików	119
Podstawowe zagadnienia dotyczące plików tekstowych	119
Pliki zawierające dane rozdzielone separatorem	120
Pobieranie danych ze stron internetowych	122
HTML i parsowanie	122
Przykład: wypowiedzi kongresmenów	124
Korzystanie z interfejsów programistycznych	126
Format JSON (i XML)	126
Korzystanie z interfejsu programistycznego bez uwierzytelniania	127
Poszukiwanie interfejsów programistycznych	128
Przykład: korzystanie z interfejsów programistycznych serwisu Twitter	128
Uzyskiwanie danych uwierzytelniających	129
Dalsza eksploracja	132
10. Praca z danymi	133
Eksploracja danych	133
Eksploracja danych jednowymiarowych	133
Dwa wymiary	135
Wiele wymiarów	136
Wykorzystanie klasy NamedTuple	137
Dekorator dataclass	139
Oczyszczanie i wstępne przetwarzanie danych	140
Przetwarzanie danych	141
Przeskalowanie	144
Dygresja: tqdm	145
Redukcja liczby wymiarów	146
Dalsza eksploracja	151
11. Uczenie maszynowe	153
Modelowanie	153
Czym jest uczenie maszynowe?	154
Nadmierne i zbyt małe dopasowanie	154
Poprawność	157

Kompromis pomiędzy wartością progową a wariancją	159
Ekstrakcja i selekcja cech	161
Dalsza eksploracja	162
12. Algorytm k najbliższych sąsiadów	163
Model	163
Przykład: dane dotyczące irysów	165
Przekleństwo wymiarowości	168
Dalsza eksploracja	171
13. Naiwny klasyfikator bayesowski	173
Bardzo prosty filtr antyspamowy	173
Bardziej zaawansowany filtr antyspamowy	174
Implementacja	175
Testowanie modelu	177
Używanie modelu	178
Dalsza eksploracja	180
14. Prosta regresja liniowa	181
Model	181
Korzystanie z algorytmu spadku gradientowego	184
Szacowanie maksymalnego prawdopodobieństwa	185
Dalsza eksploracja	185
15. Regresja wieloraka	187
Model	187
Dalsze założenia dotyczące modelu najmniejszych kwadratów	188
Dopasowywanie modelu	189
Interpretacja modelu	190
Poprawność dopasowania	191
Dygresja: ładowanie wstępne	192
Błędy standardowe współczynników regresji	193
Regularyzacja	194
Dalsza eksploracja	196
16. Regresja logistyczna	197
Problem	197
Funkcja logistyczna	199
Stosowanie modelu	201
Poprawność dopasowania	202
Maszyny wektorów nośnych	203
Dalsza eksploracja	206
17. Drzewa decyzyjne	207
Czym jest drzewo decyzyjne?	207
Entropia	209
Entropia podziału	211

Tworzenie drzewa decyzyjnego	211
Łączenie wszystkiego w całość	214
Lasy losowe	216
Dalsza eksploracja	217
18. Sztuczne sieci neuronowe	219
Perceptrony	219
Jednokierunkowe sieci neuronowe	221
Propagacja wsteczna	224
Przykład: Fizz Buzz	226
Dalsza eksploracja	228
19. Uczenie głębokie	229
Tensor	229
Abstrakcja Layer	231
Warstwa Linear	233
Sieci neuronowe jako sekwencje warstw	235
Abstrakcja Loss i optymalizacja	235
Przykład: kolejne podejście do bramki XOR	237
Inne funkcje aktywacji	238
Przykład: kolejne podejście do gry Fizz Buzz	239
Funkcja softmax i entropia krzyżowa	240
Dropout	242
Przykład: MNIST	243
Zapisywanie i wczytywanie modeli	246
Dalsza eksploracja	247
20. Grupowanie	249
Idea	249
Model	250
Przykład: spotkania	252
Wybór wartości parametru k	253
Przykład: grupowanie kolorów	255
Grupowanie hierarchiczne z podejściem aglomeracyjnym	257
Dalsza eksploracja	261
21. Przetwarzanie języka naturalnego	263
Chmury wyrazowe	263
Modele n-gram	264
Gramatyka	267
Na marginesie: próbkowanie Gibbsa	269
Modelowanie tematu	271
Wektory słów	275
Rekurencyjne sieci neuronowe	283
Przykład: używanie rekurencyjnej sieci neuronowej na poziomie pojedynczych znaków	285
Dalsza eksploracja	288

22. Analiza sieci społecznościowych	289
Pośrednictwo	289
Centralność wektorów własnych	294
Mnożenie macierzy	294
Centralność	295
Grafy skierowane i metoda PageRank	297
Dalsza eksploracja	299
23. Systemy rekomendujące	301
Ręczne rozwiązywanie problemu	301
Rekomendowanie tego, co jest popularne	302
Filtrowanie kolaboratywne oparte na użytkownikach	303
Filtrowanie kolaboratywne oparte na zainteresowaniach	305
Faktoryzacja macierzy	307
Dalsza eksploracja	311
24. Bazy danych i SQL	313
Polecenia CREATE TABLE i INSERT	313
Polecenie UPDATE	316
Polecenie DELETE	316
Polecenie SELECT	317
Polecenie GROUP BY	319
Polecenie ORDER BY	321
Polecenie JOIN	322
Zapytania składowe	324
Indeksy	324
Optymalizacja zapytań	325
Bazy danych NoSQL	326
Dalsza eksploracja	326
25. Algorytm MapReduce	327
Przykład: liczenie słów	327
Dlaczego warto korzystać z algorytmu MapReduce?	329
Algorytm MapReduce w ujęciu bardziej ogólnym	330
Przykład: analiza treści statusów	331
Przykład: mnożenie macierzy	332
Dodatkowe informacje: zespaland	334
Dalsza eksploracja	334
26. Etyka przetwarzania danych	335
Czym jest etyka danych?	335
Ale tak naprawdę to czym jest etyka danych?	336
Czy powinienem przejmować się etyką danych?	336
Tworzenie złych produktów wykorzystujących dane	337
Kompromis między dokładnością a uczciwością	337
Współpraca	339

Interpretowalność	339
Rekomendacje	340
Tendencyjne dane	340
Ochrona danych	341
Podsumowanie	342
Dalsza eksploracja	342
27. Praktyka czyni mistrza	343
IPython	343
Matematyka	343
Korzystanie z gotowych rozwiązań	344
NumPy	344
Pandas	344
scikit-learn	344
Wizualizacja	345
R	345
Uczenie głębokie	346
Szukanie danych	346
Zabierz się za analizę	346
Hacker News	347
Wozy straży pożarnej	347
Koszulki	347
Tweety na kuli ziemskiej	348
A Ty?	348

oprac. BPK