

Spis treści

Przedmowa	9
1. Wprowadzenie	13
Znaczenie danych	13
Czym jest analiza danych?	13
Hipotetyczna motywacja	14
2. Błyskawiczny kurs Pythona	25
Podstawy	25
Bardziej skomplikowane zagadnienia	36
Dalsza eksploracja	43
3. Wizualizacja danych	45
Pakiet matplotlib	45
Wykres słupkowy	47
Wykresy liniowe	50
Wykresy punktowe	51
Dalsza eksploracja	52
4. Algebra liniowa	55
Wektory	55
Macierze	59
Dalsza eksploracja	61
5. Statystyka	63
Opis pojedynczego zbioru danych	63
Korelacja	67
Paradoks Simpsona	70
Inne pułapki związane z korelacją	71
Korelacja i przyczynowość	71
Dalsza eksploracja	72
6. Prawdopodobieństwo	73
Zależność i niezależność	73
Prawdopodobieństwo warunkowe	74
Twierdzenie Bayesa	75
Zmienne losowe	77
Ciągły rozkład prawdopodobieństwa	77
Rozkład normalny	79
Centralne twierdzenie graniczne	81
Dalsza eksploracja	83

7. Hipotezy i wnioski	85
Sprawdzanie hipotez	85
Przykład: rzut monetą	85
Przedziały ufności	89
Hakowanie wartości p	90
Przykład: przeprowadzanie testu A-B	91
Wnioskowanie bayesowskie	92
Dalsza eksploracja	95
8. Metoda gradientu prostego	97
Podstawy metody gradientu prostego	97
Szacowanie gradientu	98
Korzystanie z gradientu	101
Dobór właściwego rozmiaru kroku	101
Łączenie wszystkich elementów	102
Stochastyczna metoda gradientu prostego	103
Dalsza eksploracja	104
9. Uzyskiwanie danych	105
Strumienie stdin i stdout	105
Wczytywanie plików	107
Pobieranie danych ze stron internetowych	109
Korzystanie z interfejsów programistycznych	115
Przykład: korzystanie z interfejsów programistycznych serwisu Twitter	117
Dalsza eksploracja	120
10. Praca z danymi	121
Eksploracja danych	121
Oczyszczanie i wstępne przetwarzanie danych	126
Przetwarzanie danych	127
Przeskalowanie	130
Redukcja liczby wymiarów	132
Dalsza eksploracja	137
11. Uczenie maszynowe	139
Modelowanie	139
Czym jest uczenie maszynowe?	140
Nadmierne i zbyt małe dopasowanie	140
Poprawność	143
Kompromis pomiędzy wartością progową a wariancją	145
Ekstrakcja i selekcja cech	146
Dalsza eksploracja	147
12. Algorytm k najbliższych sąsiadów	149
Model	149
Przykład: ulubione języki	151
Przekleństwo wymiarowości	155

Dalsza eksploracja	159
13. Naiwny klasyfikator bayesowski	161
Bardzo prosty filtr antyspamowy	161
Bardziej zaawansowany filtr antyspamowy	162
Implementacja	163
Testowanie modelu	165
Dalsza eksploracja	167
14. Prosta regresja liniowa	169
Model	169
Korzystanie z algorytmu spadku gradientowego	172
Szacowanie maksymalnego prawdopodobieństwa	172
Dalsza eksploracja	173
15. Regresja wieloraka	175
Model	175
Dalsze założenia dotyczące modelu najmniejszych kwadratów	176
Dopasowywanie modelu	177
Interpretacja modelu	178
Poprawność dopasowania	178
Dygresja: ładowanie wstępne	179
Błędy standardowe współczynników regresji	180
Regularyzacja	181
Dalsza eksploracja	183
16. Regresja logistyczna	185
Problem	185
Funkcja logistyczna	187
Stosowanie modelu	189
Poprawność dopasowania	190
Maszyny wektorów nośnych	190
Dalsza eksploracja	194
17. Drzewa decyzyjne	195
Czym jest drzewo decyzyjne?	195
Entropia	197
Entropia podziału	198
Tworzenie drzewa decyzyjnego	199
Łączenie wszystkiego w całość	202
Lasy losowe	204
Dalsza eksploracja	205
18. Sztuczne sieci neuronowe	207
Perceptrony	207
Jednokierunkowe sieci neuronowe	209
Propagacja wsteczna	211
Przykład: pokonywanie zabezpieczenia CAPTCHA	213

Dalsza eksploracja	216
19. Grupowanie	217
Idea	217
Model	218
Przykład: spotkania	219
Wybór wartości parametru k	221
Przykład: grupowanie kolorów	222
Grupowanie hierarchiczne z podejściem aglomeracyjnym	224
Dalsza eksploracja	228
20. Przetwarzanie języka naturalnego	229
Chmury wyrazowe	229
Modele n-gram	231
Gramatyka	234
Na marginesie: próbkowanie Gibbsa	236
Modelowanie tematu	237
Dalsza eksploracja	241
21. Analiza sieci społecznościowych	243
Pośrednictwo	243
Centralność wektorów własnych	248
Grafy skierowane i metoda PageRank	251
Dalsza eksploracja	253
22. Systemy rekomendujące	255
Ręczne rozwiązywanie problemu	255
Rekomendowanie tego, co jest popularne	256
Filtrowanie kolaboratywne oparte na użytkownikach	257
Filtrowanie kolaboratywne oparte na zainteresowaniach	260
Dalsza eksploracja	261
23. Bazy danych i SQL	263
Polecenia CREATE TABLE i INSERT	263
Polecenie UPDATE	265
Polecenie DELETE	265
Polecenie SELECT	266
Polecenie GROUP BY	267
Polecenie ORDER BY	269
Polecenie JOIN	270
Zapytania składowe	272
Indeksy	272
Optymalizacja zapytań	273
Bazy danych NoSQL	274
Dalsza eksploracja	274
24. Algorytm MapReduce	275
Przykład: liczenie słów	275

Dlaczego warto korzystać z algorytmu MapReduce?	277
Algorytm MapReduce w ujęciu bardziej ogólnym	277
Przykład: analiza treści statusów	278
Przykład: mnożenie macierzy	280
Dodatkowe informacje: zespalanie	281
Dalsza eksploracja	281
25. Po prostu zabierz się za praktykę	283
IPython	283
Matematyka	283
Korzystanie z gotowych rozwiązań	284
Szukanie danych	286
Zabierz się za analizę	287
Skorowidz	289

oprac. BPK