
Spis treści

Przedmowa	13
1. Badania eksploracyjne	17
Elementy danych uporządkowanych	18
Dla pogłębienia wiedzy	20
Dane stabelaryzowane	20
Ramki danych i indeksy	22
Niestabelaryzowane struktury danych	22
Dla pogłębienia wiedzy	23
Miary położenia	23
Średnia	24
Mediana i estymatory odporne	26
Przykład: miara położenia dla wielkości populacji i wskaźnika morderstw	27
Dla pogłębienia wiedzy	28
Miary rozproszenia	28
Odchylenie standardowe i powiązane estymatory	30
Estymatory oparte na percentylach	32
Przykład: szacowanie zmienności dla populacji Stanów Zjednoczonych	33
Dla pogłębienia wiedzy	34
Badanie rozkładu danych	34
Percentyle i boxploty	34
Tablica częstości i histogramy	36
Szacowanie i wykresy gęstości rozkładu	38
Dla pogłębienia wiedzy	40
Badanie danych binarnych i skategoryzowanych	40
Moda	42
Wartość oczekiwana	42
Prawdopodobieństwo	43
Dla pogłębienia wiedzy	43

Korelacja	43
Wykres punktowy	46
Dla pogłębienia wiedzy	48
Badanie dwóch lub więcej zmiennych	48
Wykres przedziałów heksagonalnych i wykres konturowy (przedstawianie danych numerycznych względem danych numerycznych)	48
Dwie zmienne skategoryzowane	51
Dane kategoryzowane i numeryczne	52
Wizualizacja wielu zmiennych	53
Dla pogłębienia wiedzy	56
Podsumowanie	56
2. Rozkłady danych i prób	57
Losowy dobór i obciążenie próby	58
Obciążenie	60
Dobór losowy	61
Rozmiar a jakość: kiedy rozmiar ma znaczenie?	61
Średnia z próby a średnia z populacji	62
Dla pogłębienia wiedzy	63
Błąd doboru	63
Regresja do średniej	64
Dla pogłębienia wiedzy	66
Rozkład próbkowania dla statystyki	66
Centralne twierdzenie graniczne	69
Błąd standardowy	69
Dla pogłębienia wiedzy	70
Próby bootstrapowe	70
Ponowne próbkowanie a próby bootstrapowe	73
Dla pogłębienia wiedzy	73
Przedziały ufności	74
Dla pogłębienia wiedzy	76
Rozkład normalny	76
Standaryzowany rozkład normalny i wykres K-K	78
Rozkłady z długimi ogonami	79
Dla pogłębienia wiedzy	81
Rozkład t-Studenta	81
Dla pogłębienia wiedzy	83
Rozkład binarny	83
Dla pogłębienia wiedzy	85

Rozkład chi-kwadrat	85
Dla pogłębienia wiedzy	86
Rozkład F	86
Dla pogłębienia wiedzy	87
Rozkład Poissona i jego pochodne	87
Rozkład Poissona	88
Rozkład wykładniczy	88
Szacowanie współczynnika porażki	89
Rozkład Weibulla	89
Dla pogłębienia wiedzy	90
Podsumowanie	90
3. Eksperymenty statystyczne i testowanie istotności	91
Test A/B	91
Po co Ci grupa kontrolna?	94
Dlaczego tylko A/B? Dlaczego nie C, D itd.?	94
Dla pogłębienia wiedzy	95
Testowanie hipotezy	96
Hipoteza zerowa	97
Hipoteza alternatywna	97
Test jednostronny i test dwustronny	97
Dla pogłębienia wiedzy	98
Testy randomizacyjne	98
Test permutacyjny	99
Przykład: licznik odwiedzin strony	100
Zupełny test permutacyjny i bootstrap	103
Test permutacyjny: podstawa w data science	104
Dla pogłębienia wiedzy	104
Istotność statystyczna i p-wartość	104
p-wartość	106
Alfa	108
Błędy pierwszego i drugiego rodzaju	109
Data science i p-wartość	109
Dla pogłębienia wiedzy	110
Test t	110
Dla pogłębienia wiedzy	112
Testowanie wielokrotne	112
Dla pogłębienia wiedzy	115
Stopnie swobody	115
Dla pogłębienia wiedzy	116

ANOVA	116
Statystyka F	119
Dwustronna ANOVA	121
Dla pogłębienia wiedzy	121
Test chi-kwadrat	121
Test chi-kwadrat: podejście randomizacyjne	122
Test chi-kwadrat: teoria	124
Dokładny test Fishera	125
Znaczenie testu chi-kwadrat w data science	127
Dla pogłębienia wiedzy	128
Algorytm Wielorękiego Bandyty	128
Dla pogłębienia wiedzy	131
Moc i rozmiar próby	131
Rozmiar próby	132
Dla pogłębienia wiedzy	134
Podsumowanie	134
4. Regresja i predykcja	135
Prosta regresja liniowa	135
Równanie regresji	135
Dopasowanie wartości i rezydua	139
Metoda najmniejszych kwadratów	139
Predykcja a objaśnienie (profilowanie)	141
Dla pogłębienia wiedzy	141
Regresja wieloraka	142
Przykład: wartość domów w King County	143
Ocena modelu	144
Kroswalidacja	146
Dobór modelu i regresja krokowa	147
Regresja ważona	150
Dla pogłębienia wiedzy	151
Predykcja z wykorzystaniem regresji	151
Niebezpieczeństwa związane z ekstrapolacją	151
Przedziały ufności i predykcji	152
Zmienne skategoryzowane w regresji	153
Zmienne fikcyjne	154
Zmienne skategoryzowane na wielu poziomach	156
Uporządkowane zmienne skategoryzowane	158
Interpretacja równania regresji	159
Predyktory skorelowane	160
Współliniowość	161

Zmienne zakłócające	161
Interakcje i efekty główne	163
Diagnostyka regresji	164
Wartości odstające	165
Obserwacje wpływowe	167
Heteroskedastyczność, anormalność i błędy skorelowane	170
Wykresy częściowych rezyduów i nieliniowość	172
Regresja wielomianowa i regresja sklejana	174
Wielomian	175
Funkcja sklejana	176
Uogólnione modele addytywne	178
Dla pogłębienia wiedzy	180
Podsumowanie	180
5. Klasyfikacja	181
Naiwny klasyfikator bayesowski	182
Dlaczego klasyfikator bayesowski jest niepraktyczny?	183
Naiwne rozwiązanie	183
Numeryczne zmienne objaśniające	186
Dla pogłębienia wiedzy	186
Analiza dyskryminacyjna	186
Macierz kowariancji	187
Liniowy dyskryminator Fishera	188
Prosty przykład	188
Dla pogłębienia wiedzy	191
Regresja logistyczna	192
Funkcja odpowiedzi logistycznej i logit	192
Regresja logistyczna i GLM	194
Uogólnione modele liniowe	195
Wartości prognozowane na podstawie regresji logistycznej	195
Interpretacja współczynników i iloraz szans	196
Regresja liniowa i regresja logistyczna: podobieństwa i różnice	197
Ocena modelu	198
Dla pogłębienia wiedzy	201
Ewaluacja modeli klasyfikacji	202
Macierz błędów	203
Problem mało licznych klas	204
Precyzja, czułość i swoistość	205
Krzywa ROC	206
Pole pod wykresem krzywej ROC	208

Lift	209
Dla pogłębienia wiedzy	210
Strategie dla niezbilansowanych danych	210
Undersampling	211
Oversampling i zwiększenie/obniżenie wag	212
Generowanie danych	213
Klasyfikacja oparta na kosztach	214
Badanie prognozy	214
Dla pogłębienia wiedzy	216
Podsumowanie	216
6. Statystyczne uczenie maszynowe	217
K-najbliższych sąsiadów	218
Przykład: przewidywanie opóźnienia w spłacie pożyczki	219
Metryki odległości	221
Kodowanie 1 z n	222
Standaryzacja (normalizacja, z-wartość)	223
Dobór K	225
KNN w doborze cech	226
Drzewa decyzyjne	228
Prosty przykład	229
Algorytm rekursywnego podziału	231
Pomiar homogeniczności lub zanieczyszczenia	232
Zatrzymanie wzrostu drzewa	234
Prognoza ciągłych wartości	235
Jak są wykorzystywane drzewa	235
Dla pogłębienia wiedzy	236
Bagging i lasy losowe	236
Bagging	238
Las losowy	238
Istotność zmiennej	242
Hiperparametry	244
Boosting	245
Algorytm wzmacniania	247
XGBoost	247
Regularyzacja: unikanie nadmiernego dopasowania	249
Hiperparametry i krosvalidacja	253
Podsumowanie	256

7. Uczenie nienadzorowane	257
Analiza głównych składowych	258
Prosty przykład	259
Obliczanie głównych składowych	261
Interpretacja głównych składowych	261
Analiza odpowiedniości	264
Dla pogłębienia wiedzy	266
Metoda K-średnich (centroidów)	266
Prosty przykład	267
Algorytm K-średnich	269
Interpretacja klastrów	270
Dobór liczby klastrów	272
Klasteryzacja hierarchiczna	274
Prosty przykład	274
Dendrogram	275
Algorytm aglomeracyjny	276
Miary podobieństwa	277
Klasteryzacja oparta na modelu	278
Wielowymiarowy rozkład normalny	278
Mieszaniny rozkładów normalnych	280
Dobór liczby klastrów	282
Dla pogłębienia wiedzy	284
Skalowanie i zmienne skategoryzowane	284
Skalowanie zmiennych	285
Zmienne dominujące	287
Zmienne skategoryzowane i odległość Gowera	288
Problem z klasteryzacją danych mieszanych	290
Podsumowanie	291
Bibliografia	293