

Spis treści

Przedmowa	XI
0.1. Co to jest eksploracja danych?	XI
0.2. Dlaczego ta książka jest potrzebna?	XI
0.3. Podejście typu „biała skrzynka” — zrozumienie podstawowych struktur algorytmów i modeli	XII
0.3.1. Omówienie działania algorytmów	XII
0.3.2. Zastosowanie algorytmów do dużego zbioru danych	XII
0.3.3. Ćwiczenia do rozdziałów — upewnij się, że rozumiesz	XII
0.3.4. Ćwiczenia praktyczne — nauka eksploracji danych poprzez wykonywanie eksploracji danych	XIII
0.3.5. Studium przypadku — podejście całościowe	XIII
0.4. Eksploracja danych jako proces	XIII
0.5. Oprogramowanie	XIV
0.5.1. WEKA — alternatywa typu <i>open source</i>	XIV
0.6. Strona internetowa: www.dataminingconsultant.com	XV
0.7. <i>Metody i modele eksploracji danych</i> jako podręcznik	XV
0.8. Podziękowania	XVI
1. Metody redukcji wymiaru	1
1.1. Konieczność redukcji wymiaru w eksploracji danych	1
1.2. Analiza składowych głównych	2
1.2.1. Zastosowanie analizy składowych głównych do zbioru danych <i>houses</i> (domy)	5
1.2.2. Ile składowych należy wyodrębnić?	10
1.2.3. Tworzenie profili składowych głównych	13
1.2.4. Zasób zmienności wspólnej	16
1.2.5. Walidacja składowych głównych	18
1.3. Analiza czynnikowa	19
1.3.1. Zastosowanie analizy czynnikowej do zbioru danych <i>adult</i>	20
1.3.2. Rotacja czynników	22
1.4. Komponenty zdefiniowane przez użytkownika	25
1.4.1. Przykład komponentu zdefiniowanego przez użytkownika	26
1.5. Podsumowanie	27
1.6. Literatura	30
1.7. Ćwiczenia	31

2. Modelowanie regresji	36
2.1. Przykład prostej regresji liniowej	36
2.2. Oszacowanie najmniejszych kwadratów	39
2.3. Współczynnik determinacji	42
2.4. Błąd standardowy estymatora	47
2.5. Współczynnik korelacji	48
2.6. Tablica ANOVA	50
2.7. Punkty oddalone, punkty wysokiej dźwigni i obserwacje wpływowe	51
2.8. Model regresji	59
2.9. Wnioskowanie w modelu regresji liniowej	61
2.9.1. Test t dla hipotezy o zachodzeniu liniowego związku między x a y	62
2.9.2. Przedział ufności dla nachylenia linii regresji	64
2.9.3. Przedział ufności dla wartości średniej y przy danej wartości x	64
2.9.4. Przedział predykcji dla wartości y przy danej wartości x	65
2.9.5. Sprawdzanie założeń regresji	68
2.10. Przykład: zbiór danych <i>baseball</i>	73
2.11. Przykład: zbiór danych <i>California</i>	79
2.12. Transformacje w celu osiągnięcia liniowości	83
2.12.1. Transformacja Boxa–Coxa	88
2.13. Podsumowanie	88
2.14. Literatura	91
2.15. Ćwiczenia	91
3. Regresja wielokrotna i budowa modelu	98
3.1. Przykład regresji wielokrotnej	98
3.2. Model regresji wielokrotnej	104
3.3. Wnioskowanie w modelu regresji wielokrotnej	105
3.3.1. Test t dla zależności pomiędzy y a x_i	105
3.3.2. Test t dla zależności pomiędzy zmienną <i>wartość odżywcza</i> a zmienną <i>cukry</i>	106
3.3.3. Test t dla zależności pomiędzy zmienną <i>wartość odżywcza</i> a zmienną <i> błonnik</i>	106
3.3.4. Test F istotności całego modelu regresji	107
3.3.5. Test F dla zależności pomiędzy zmienną <i>wartość odżywcza</i> a zmiennymi <i>cukry</i> i <i> błonnik</i>	108
3.3.6. Przedział ufności dla danego współczynnika	109
3.3.7. Przedział ufności dla wartości średniej y przy danych $x_1, x_2, \dots, x_m$	110
3.3.8. Przedział predykcji dla y przy danych $x_1, x_2, \dots, x_m$	110
3.4. Regresja z jakościowymi zmiennymi objaśniającymi	110
3.4.1. Skorygowany R^2 — sposób na wyeliminowanie z modelu zmiennych objaśniających, które nie są przydatne	119
3.4.2. Sekwencyjne sumy kwadratów	120
3.5. Współliniowość	122
3.6. Metody wyboru zmiennych objaśniających	128
3.6.1. Częściowy test F	128
3.6.2. Metoda dołączania	130
3.6.3. Metoda eliminacji	131
3.6.4. Regresja krokowa	131
3.6.5. Metoda najlepszych podzbiorów	131
3.6.6. Metoda wszystkich możliwych regresji	132
3.7. Zastosowanie metod wyboru zmiennych objaśniających	132
3.7.1. Metoda dołączania zastosowana do zbioru danych <i>cereals</i>	133

3.7.2. Metoda eliminacji zastosowana do zbioru danych <i>cereals</i>	135
3.7.3. Metoda regresji krokowej zastosowana do zbioru danych <i>cereals</i>	137
3.7.4. Metoda najlepszych podzbiorów zastosowana do zbioru danych <i>cereals</i>	138
3.8. Statystyka C_p Mallowsa	138
3.9. Kryteria wyboru zmiennych	140
3.10. Zastosowanie składowych głównych jako zmiennych objaśniających	148
3.11. Podsumowanie	154
3.12. Literatura	156
3.13. Ćwiczenia	156
4. Regresja logistyczna	163
4.1. Przykład prostej regresji logistycznej	163
4.2. Estymacja największej wiarygodności	166
4.3. Interpretacja wyników regresji logistycznej	166
4.4. Wnioskowanie — czy zmienne objaśniające są istotne?	168
4.5. Interpretacja modelu regresji logistycznej	170
4.5.1. Interpretacja modelu dla zmiennych jakościowych o dwóch wartościach	171
4.5.2. Interpretacja modelu dla zmiennych objaśniających z wieloma wartościami	175
4.5.3. Interpretacja modelu z ciągłą zmienną objaśniającą	179
4.6. Założenie liniowości	184
4.7. Problem komórek z wartością zero	187
4.8. Wielokrotna regresja logistyczna	189
4.9. Wprowadzenie funkcji wyższego rzędu, aby uwzględnić nieliniowość	194
4.10. Sprawdzenie modelu regresji logistycznej	201
4.11. WEKA — praktyczna analiza za pomocą regresji logistycznej	205
4.12. Podsumowanie	209
4.13. Literatura	210
4.14. Ćwiczenia	211
5. Naiwna estymacja bayesowska i sieci bayesowskie	217
5.1. Podejście bayesowskie	217
5.2. Klasyfikacja maksymalnego <i>a posteriori</i>	219
5.2.1. Iloraz szans <i>a posteriori</i>	223
5.2.2. Równoważenie danych	225
5.3. Naiwny klasyfikator bayesowski	228
5.3.1. Numeryczne zmienne objaśniające	233
5.4. WEKA — praktyczna analiza zastosowania naiwnego Bayesa	237
5.5. Sieci przekonań Bayesa	241
5.5.1. Przykład zakupów odzieży	242
5.5.2. Zastosowanie sieci Bayesa do znalezienia prawdopodobieństw	244
5.6. WEKA — praktyczna analiza z użyciem klasyfikatora sieci Bayesa	247
5.7. Podsumowanie	250
5.8. Literatura	251
5.9. Ćwiczenia	252
6. Algorytmy genetyczne	255
6.1. Wprowadzenie do algorytmów genetycznych	255
6.2. Podstawowy szkielet algorytmu genetycznego	256
6.3. Prosty przykład działania algorytmu genetycznego	258
6.4. Modyfikacje i rozszerzenia — selekcja	260

6.5.	Modyfikacje i rozszerzenia — krzyżowanie	262
6.5.1.	Krzyżowanie wielopunktowe	262
6.5.2.	Krzyżowanie równomierne	263
6.6.	Algorytm genetyczny dla zmiennych rzeczywistych	263
6.7.	Zastosowanie algorytmów genetycznych do uczenia sieci neuronowych	265
6.8.	WEKA — praktyczna analiza użycia algorytmów genetycznych	270
6.9.	Podsumowanie	277
6.10.	Literatura	278
6.11.	Ćwiczenia	280
7.	Studium przypadku — modelowanie odpowiedzi na kampanię marketingową	282
7.1.	<i>Cross-Industry Standard Process for Data Mining</i>	282
7.2.	Zrozumienie uwarunkowań biznesowych	284
7.2.1.	Problem odpowiedzi na kampanię marketingową	284
7.2.2.	Budowanie tabeli koszt/zysk	284
7.3.	Zrozumienie danych i przygotowanie danych	286
7.3.1.	Przekształcenia w celu uzyskania normalności albo symetrii	289
7.3.2.	Standaryzacja i zmienne flagowe	293
7.3.3.	Otrzymywanie nowych zmiennych	294
7.3.4.	Badanie relacji pomiędzy zmiennymi objaśniającymi a zmienną celu	296
7.3.5.	Badanie struktury korelacji pomiędzy zmiennymi objaśniającymi	305
7.4.	Etap modelowania i ewaluacji	307
7.4.1.	Analiza składowych głównych	310
7.4.2.	Analiza skupień — algorytm grupowania BIRCH	312
7.4.3.	Równoważenie zbioru uczącego	316
7.4.4.	Określenie punktu odniesienia dla działania modelu	317
7.4.5.	Zbiór modeli A — wykorzystanie analizy składowych głównych	319
7.4.6.	Przeważanie jako substytut kosztów błędnej klasyfikacji	321
7.4.7.	Łączenie modeli — głosowanie	323
7.4.8.	Zbiór modeli B — modele bez składowych głównych	325
7.4.9.	Łączenie modeli za pomocą średniego prawdopodobieństwa odpowiedzi	327
7.5.	Podsumowanie	332
7.6.	Literatura	335
	Literatura uzupełniająca	336
	Indeks	337