
Spis treści

Przedmowa	7
1. Wprowadzenie	13
Dlaczego uczenie maszynowe?	13
Problemy, które może rozwiązać uczenie maszynowe	14
Znajomość zadania i znajomość danych	16
Dlaczego Python?	17
scikit-learn	17
Instalacja scikit-learn	17
Podstawowe biblioteki i narzędzia	18
Jupyter Notebook	18
NumPy	19
SciPy	19
matplotlib	20
pandas	21
mglearn	22
Python 2 a Python 3	22
Wersje użyte w tej książce	23
Pierwsza aplikacja: klasyfikacja gatunków irysa	24
Zapoznaj się z danymi	25
Sprawdzanie osiągnięcia sukcesu: dane treningowe i testowe	27
Najpierw najważniejsze: zapoznaj się z danymi	28
Budowa pierwszego modelu: k-najbliżsi sąsiedzi	30
Przewidywania	31
Ocena modelu	32
Podsumowanie i przegląd	32

2. Nadzorowane uczenie maszynowe	35
Klasyfikacja i regresja	35
Uogólnianie, nadmierne dopasowanie i niedopasowanie	36
Relacja złożoności modelu do rozmiaru zestawu danych	38
Nadzorowane algorytmy uczenia maszynowego	39
Przykładowe zestawy danych	39
k-najbliższych sąsiadów	43
Modele liniowe	50
Naiwne klasyfikatory Bayesa	69
Drzewa decyzyjne	70
Zespoły drzew decyzyjnych	81
Maszyny wektorów nośnych	88
Sieci neuronowe (głębokie uczenie)	98
Szacunki niepewności na podstawie klasyfikatorów	109
Funkcja decyzyjna	110
Prognozy prawdopodobieństw	112
Niepewność w klasyfikacji wieloklasowej	114
Podsumowanie i przegląd	116
 3. Uczenie nienadzorowane i przetwarzanie wstępne	 119
Rodzaje nienadzorowanego uczenia maszynowego	119
Wyzwania związane z uczeniem nienadzorowanym	120
Przetwarzanie wstępne i skalowanie	120
Różne rodzaje przetwarzania wstępnego	121
Zastosowanie transformacji danych	122
Skalowanie danych treningowych i testowych w ten sam sposób	124
Wpływ przetwarzania wstępnego na uczenie nadzorowane	126
Redukcja wymiarowości, wyodrębnianie cech i wielorakie uczenie	127
Analiza głównych komponentów (PCA)	127
Nieujemna faktoryzacja macierzy (NMF)	140
Manifold learning z t-SNE	146
Grupowanie	150
Grupowanie k-średnich	150
Grupowanie aglomeracyjne	160
DBSCAN	164
Porównanie i ocena algorytmów grupowania	168
Podsumowanie metod grupowania	181
Podsumowanie i przegląd	181

4. Reprezentacja danych i cechy inżynierskie	183
Zmienne kategoryjne	184
Kodowanie jeden-z-N (zmienne fikcyjne)	185
Liczby mogą kodować zmienne kategoryjne	189
Dzielenie, dyskretyzacja, modele liniowe i drzewa	190
Interakcje i wielomiany	194
Jednowymiarowe transformacje nieliniowe	200
Automatyczny wybór cechy	203
Statystyki jednoczynnikowe	203
Wybór cechy na podstawie modelu	205
Iteracyjny wybór cech	206
Wykorzystanie wiedzy eksperckiej	208
Podsumowanie i przegląd	215
5. Ocena i doskonalenie modelu	217
Walidacja krzyżowa	218
Walidacja krzyżowa w scikit-learn	218
Korzyści z walidacji krzyżowej	219
Stratyfikowana k-krotna walidacja krzyżowa i inne strategie	220
Przeszukiwanie siatki	225
Proste przeszukiwanie siatki	226
Nadmierne dopasowanie parametrów i zestaw walidacyjny	226
Przeszukiwanie siatki z walidacją krzyżową	228
Wskaźniki oceny i punktacja	238
Pamiętaj o celu	239
Metryki klasyfikacji binarnej	239
Metryki klasyfikacji wieloklasowej	257
Metryki regresji	260
Używanie metryk oceny w wyborze modelu	260
Podsumowanie i przegląd	262
6. Łańcuchy algorytmów i potoki	263
Wybór parametrów z przetwarzaniem wstępnym	264
Tworzenie potoków	265
Używanie potoków w przeszukiwaniu siatki	266
Ogólny interfejs potoku	269
Wygodne tworzenie potoków za pomocą funkcji <code>make_pipeline</code>	270
Dostęp do atrybutów kroku	271
Dostęp do atrybutów klasy <code>GridSearchCV</code>	271
Kroki przetwarzania wstępnego przeszukiwania siatki i parametry modelu	273
Przeszukiwanie siatki modeli	275
Podsumowanie i przegląd	276

7. Praca z danymi tekstowymi	277
Typy danych przedstawione jako ciągi znaków	277
Przykładowe zastosowanie: analiza recenzji filmowych	279
Przedstawianie danych tekstowych w postaci worka słów	281
Stosowanie worka słów do przykładowego zestawu danych	282
Zastosowanie worka słów do recenzji filmowych	283
Słowa pomijalne	287
Skalowanie danych z tf-idf	288
Badanie współczynników modelu	290
Worek słów z więcej niż jednym słowem (n-gram)	291
Zaawansowana tokenizacja, stemming i lematyzacja	295
Modelowanie tematów i grupowanie dokumentów	298
Utajniona alokacja Dirichleta	299
Podsumowanie i przegląd	305
8. Podsumowanie	307
Podejście do problemu uczenia maszynowego	307
Informowanie ludzi	308
Od prototypu do produkcji	308
Testowanie systemów na produkcji	309
Tworzenie własnego estymatora	310
Co dalej	311
Teoria	311
Inne narzędzia i pakiety do uczenia maszynowego	311
Ranking, systemy rekomendujące i inne rodzaje uczenia	312
Modelowanie probabilistyczne, wnioskowanie i programowanie probabilistyczne	312
Sieci neuronowe	313
Skalowanie do większych zestawów danych	313
Doskonalenie umiejętności	314
Podsumowanie	315