

Spis treści

Przedmowa	15
Wprowadzenie	17
O autorze	23
Część I	
Pierwsze kroki	25
Rozdział 1	
Podyskutujmy o uczeniu się	27
1.1. Witaj	27
1.2. Zakres, terminologia, predykcja i dane	28
1.2.1. Cechy	28
1.2.2. Wartości docelowe i predykcje	31
1.3. Rola maszyny w uczeniu maszynowym	31
1.4. Przykład systemów uczących się	33
1.4.1. Predykcja kategorii: przykłady klasyfikacji	33
1.4.2. Predykcja wartości — przykłady regresorów	35
1.5. Ocena systemów uczących się	35
1.5.1. Poprawność	36
1.5.2. Wykorzystanie zasobów	37
1.6. Proces budowania systemów uczących się	38
1.7. Założenia i realia uczenia się	40
1.8. Zakończenie rozdziału	42
1.8.1. Droga przed nami	42
1.8.2. Uwagi	43
Rozdział 2	
Kontekst techniczny	45
2.1. O naszej konfiguracji	45
2.2. Potrzeba posiadania języka matematycznego	45
2.3. Nasze oprogramowanie do zmierzenia się z uczeniem maszynowym	46
2.4. Prawdopodobieństwo	47
2.4.1. Zdarzenia elementarne	48
2.4.2. Niezależność zdarzeń	50

2.4.3.	Prawdopodobieństwo warunkowe	50
2.4.4.	Rozkłady	52
2.5.	Kombinacje liniowe, sumy ważone i iloczyny skalarne	54
2.5.1.	Średnia ważona	57
2.5.2.	Suma kwadratów	59
2.5.3.	Suma kwadratów błędów	59
2.6.	Perspektywa geometryczna: punkty w przestrzeni	60
2.6.1.	Linie	61
2.6.2.	Coś więcej niż linie	65
2.7.	Notacja sztuczki plus jeden	69
2.8.	Odjazd, zrywanie kaftana bezpieczeństwa i nieliniowość ..	71
2.9.	NumPy kontra „cała matematyka”	73
2.9.1.	Wracamy do 1D i 2D	75
2.10.	Problemy z wartościami zmiennoprzecinkowymi	78
2.11.	Zakończenie rozdziału	79
2.11.1.	Podsumowanie	79
2.11.2.	Uwagi	79
Rozdział 3	Predykcja kategorii — początki klasyfikacji	81
3.1.	Zadania klasyfikacji	81
3.2.	Prosty zestaw danych do klasyfikacji	82
3.3.	Trenowanie i testowanie: nie ucz się do testu	84
3.4.	Ocena — wystawienie stopni	87
3.5.	Prosty klasyfikator nr 1: najbliżsi sąsiedzi, związki na odległość i założenia	88
3.5.1.	Definiowanie podobieństwa	88
3.5.2.	k w k-NN	90
3.5.3.	Kombinacja odpowiedzi	90
3.5.4.	k-NN, parametry i metody bezparametrowe	90
3.5.5.	Budowa modelu klasyfikacji k-NN	91
3.6.	Prosty klasyfikator nr 2: naiwny klasyfikator bayesowski, prawdopodobieństwo i złamane obietnice	93
3.7.	Uproszczona ocena klasyfikatorów	96
3.7.1.	Wydajność uczenia się	96
3.7.2.	Wykorzystanie zasobów w klasyfikacji	97
3.7.3.	Szacowanie zasobów w aplikacjach samodzielnych	103
3.8.	Koniec rozdziału	106
3.8.1.	Ostrzeżenie: ograniczenia i otwarte kwestie	106
3.8.2.	Podsumowanie	107
3.8.3.	Uwagi	107
3.8.4.	Ćwiczenia	109

Rozdział 4 Predykcja wartości numerycznych: początki regresji 111

4.1.	Prosty zbiór danych dla regresji	111
4.2.	Regresja z najbliższymi sąsiadami i statystyki sumaryczne ..	113
4.2.1.	Miary środka: mediana i średnia	114
4.2.2.	Budowa modelu regresji k-NN	116
4.3.	Błędy regresji liniowej	117
4.3.1.	Ziemia nie jest płaska, czyli dlaczego potrzebujemy pochyłości	118
4.3.2.	Przekrzywienie pola	120
4.3.3.	Wykonanie regresji liniowej	122
4.4.	Optymalizacja — wybór najlepszej odpowiedzi	123
4.4.1.	Zgadywanie losowe	124
4.4.2.	Losowe kroki	124
4.4.3.	Sprytne kroki	125
4.4.4.	Obliczony skrót	125
4.4.5.	Wykorzystanie w regresji liniowej	126
4.5.	Prosta ocena i porównanie regresorów	126
4.5.1.	Pierwiastek średniego błędu kwadratowego	126
4.5.2.	Wydajność uczenia się	127
4.5.3.	Wykorzystanie zasobów w regresji	127
4.6.	Zakończenie rozdziału	129
4.6.1.	Ograniczenia i kwestie otwarte	129
4.6.2.	Podsumowanie	130
4.6.3.	Uwagi	130
4.6.4.	Ćwiczenia	130

Część II Ocena 131**Rozdział 5 Ocena i porównywanie metod uczenia się 133**

5.1.	Ocena i dlaczego mniej znaczy więcej	133
5.2.	Terminologia dla faz uczenia się	134
5.2.1.	Powrót do maszyn	135
5.2.2.	Mówiąc bardziej technicznie... ..	137
5.3.	Majorze Tom, coś jest nie tak — nadmierne dopasowanie i niedopasowanie	141
5.3.1.	Dane syntetyczne i regresja liniowa	141
5.3.2.	Ręczna modyfikacja złożoności modelu	143
5.3.3.	Złotowłosa — wizualizacja nadmiernego dopasowania, niedopasowania oraz „w sam raz”	145
5.3.4.	Prostota	148
5.3.5.	Uwagi na temat nadmiernego dopasowania	148

5.4.	Od błędów do kosztów	149
5.4.1.	Strata	149
5.4.2.	Koszt	150
5.4.3.	Punktacja	151
5.5.	(Powtórne) próbkowanie — zamienić mniej w więcej	152
5.5.1.	Walidacja krzyżowa	152
5.5.2.	Rozwarstwienie	156
5.5.3.	Powtarzany podział na dane treningowe i testowe ..	158
5.5.4.	Lepszy sposób i tasowanie	161
5.5.5.	Walidacja krzyżowa z odłożeniem jednego	164
5.6.	Rozbicie: dekonstrukcja błędu na błąd systematyczny i wariancję	166
5.6.1.	Wariancja danych	167
5.6.2.	Wariancja modelu	167
5.6.3.	Błąd systematyczny modelu	168
5.6.4.	A teraz wszystko razem	168
5.6.5.	Przykłady kompromisów związanych z błędem systematycznym i wariancją	169
5.7.	Ocena graficzna i porównanie	173
5.7.1.	Krzywe uczenia — jak dużo danych potrzebujemy?	173
5.7.2.	Krzywe złożoności	177
5.8.	Porównywanie metod uczących się za pomocą walidacji krzyżowej	178
5.9.	Koniec rozdziału	179
5.9.1.	Podsumowanie	179
5.9.2.	Uwagi	179
5.9.3.	Ćwiczenia	181

Rozdział 6 Ocena klasyfikatorów183

6.1.	Klasyfikatory bazowe	183
6.2.	Więcej niż dokładność — wskaźniki dla klasyfikacji	186
6.2.1.	Eliminacja zamieszania za pomocą macierzy błędu	187
6.2.2.	W jaki sposób można się mylić	188
6.2.3.	Wskaźniki z macierzy błędu	189
6.2.4.	Kodowanie macierzy błędu	190
6.2.5.	Radzenie sobie z wieloma klasami — uśrednianie wieloklasowe	192
6.2.6.	F1	194
6.3.	Krzywe ROC	194
6.3.1.	Wzorce w ROC	197
6.3.2.	Binarny ROC	199

6.3.3. AUC — obszar pod krzywą ROC	201
6.3.4. Wieloklasowe mechanizmy uczące się, jeden kontra reszta i ROC	203
6.4. Inne podejście dla wielu klas: jeden-kontra-jeden	205
6.4.1. Wieloklasowe AUC, część druga — w poszukiwaniu pojedynczej wartości	206
6.5. Krzywe precyzji i skuteczności wyszukiwania	209
6.5.1. Uwaga o kompromisie precyzji i skuteczności wyszukiwania	209
6.5.2. Budowanie krzywej precyzji i skuteczności wyszukiwania	210
6.6. Krzywe kumulacyjnej odpowiedzi i wzniesienia	211
6.7. Bardziej wyrafinowana ocena klasyfikatorów — podejście drugie	213
6.7.1. Binarne	213
6.7.2. Nowy problem wieloklasowy	217
6.8. Koniec rozdziału	222
6.8.1. Podsumowanie	222
6.8.2. Uwagi	222
6.8.3. Ćwiczenia	224

Rozdział 7 Ocena metod regresji 225

7.1. Metody regresji będące punktem odniesienia	225
7.2. Dodatkowe miary w metodach regresji	227
7.2.1. Tworzenie własnych miar oceny	227
7.2.2. Inne wbudowane miary regresji	228
7.2.3. R^2	229
7.3. Wykresy składników resztowych	235
7.3.1. Wykresy błędów	235
7.3.2. Wykresy składników resztowych	237
7.4. Pierwsze podejście do standaryzacji	241
7.5. Ocena mechanizmów regresji w bardziej zaawansowany sposób: podejście drugie	245
7.5.1. Wyniki po sprawdzanie krzyżowym z użyciem różnych miar	246
7.5.2. Omówienie wyników ze sprawdzianu krzyżowego	249
7.5.3. Składniki resztowe	250
7.6. Koniec rozdziału	251
7.6.1. Podsumowanie	251
7.6.2. Uwagi	251
7.6.3. Ćwiczenia	254

Część III	Jeszcze o metodach i podstawach	255
Rozdział 8	Inne metody klasyfikacji	257
8.1.	Jeszcze o klasyfikacji	257
8.2.	Drzewa decyzyjne	259
8.2.1.	Algorytmy budowania drzewa	262
8.2.2.	Do pracy. Pora na drzewa decyzyjne	265
8.2.3.	Obciążenie i wariancja w drzewach decyzyjnych	268
8.3.	Klasyfikatory oparte na wektorach nośnych	269
8.3.1.	Stosowanie klasyfikatorów SVC	272
8.3.2.	Obciążenie i wariancja w klasyfikatorach SVC	275
8.4.	Regresja logistyczna	277
8.4.1.	Szanse w zakładach	278
8.4.2.	Prawdopodobieństwo, szanse i logarytm szans ...	280
8.4.3.	Po prostu to zrób: regresja logistyczna	285
8.4.4.	Regresja logistyczna: osobliwość przestrzenna ...	286
8.5.	Analiza dyskryminacyjna	287
8.5.1.	Kowariancja	289
8.5.2.	Metody	299
8.5.3.	Przeprowadzanie analizy dyskryminacyjnej	301
8.6.	Założenia, obciążenie i klasyfikatory	302
8.7.	Porównanie klasyfikatorów: podejście trzecie	304
8.7.1.	Cyfry	305
8.8.	Koniec rozdziału	307
8.8.1.	Podsumowanie	307
8.8.2.	Uwagi	307
8.8.3.	Ćwiczenia	310
Rozdział 9	Inne metody regresji	313
9.1.	Regresja liniowa na ławce kar — regularyzacja	313
9.1.1.	Przeprowadzanie regresji z regularyzacją	318
9.2.	Regresja z użyciem wektorów nośnych	319
9.2.1.	Zawiasowa funkcja straty	319
9.2.2.	Od regresji liniowej przez regresję z regularyzacją do regresji SVR	323
9.2.3.	Po prostu to zrób — w stylu SVR	324
9.3.	Regresja segmentowa ze stałymi	325
9.3.1.	Implementowanie regresji segmentowej ze stałymi	327
9.3.2.	Ogólne uwagi na temat implementowania modeli ...	328
9.4.	Drzewa regresyjne	331
9.4.1.	Przeprowadzanie regresji z użyciem drzew	331

9.5. Porównanie metod regresji: podejście trzecie	332
9.6. Koniec rozdziału	334
9.6.1. Podsumowanie	334
9.6.2. Uwagi	334
9.6.3. Ćwiczenia	335

Rozdział 10 Ręczna inżynieria cech

— manipulowanie danymi dla zabawy i dla zysku	337
10.1. Terminologia i przyczyny stosowania inżynierii cech	337
10.1.1. Po co stosować inżynierię cech?	338
10.1.2. Kiedy stosuje się inżynierię cech?	339
10.1.3. Jak przebiega inżynieria cech?	340
10.2. Wybieranie cech i redukcja danych — pozbywanie się śmieci	341
10.3. Skalowanie cech	342
10.4. Dyskretyzacja	346
10.5. Kodowanie kategorii	348
10.5.1. Inna metoda kodowania i niezwykły przypadek braku punktu przecięcia z osią	351
10.6. Relacje i interakcje	358
10.6.1. Ręczne tworzenie cech	358
10.6.2. Interakcje	360
10.6.3. Dodawanie cech na podstawie transformacji	364
10.7. Manipulowanie wartościami docelowymi	366
10.7.1. Manipulowanie przestrzenią danych wejściowych	367
10.7.2. Manipulowanie wartościami docelowymi	369
10.8. Koniec rozdziału	371
10.8.1. Podsumowanie	371
10.8.2. Uwagi	371
10.8.3. Ćwiczenia	372

Rozdział 11 Dopracowywanie hiperparametrów i potoki **375**

11.1. Modele, parametry i hiperparametry	376
11.2. Dostrajanie hiperparametrów	378
11.2.1. Uwaga na temat słownictwa informatycznego i z dziedziny uczenia maszynowego	378
11.2.2. Przykład przeszukiwania kompletnego	378
11.2.3. Używanie losowości do szukania igły w stogu siana	384
11.3. Wyprawa w rekurencyjną króliczą norę — zagnieżdżony sprawdzian krzyżowy	385
11.3.1. Opakowanie w sprawdzian krzyżowy	386
11.3.2. Przeszukiwanie siatki jako model	387

11.3.3.	Sprawdzian krzyżowy zagnieżdżony w sprawdzanie krzyżowym	388
11.3.4.	Uwagi na temat zagnieżdżonych SK	391
11.4.	Potoki	393
11.4.1.	Prosty potok	393
11.4.2.	Bardziej skomplikowany potok	394
11.5.	Potoki i dostrajanie całego procesu	395
11.6.	Koniec rozdziału	397
11.6.1.	Podsumowanie	397
11.6.2.	Uwagi	397
11.6.3.	Ćwiczenia	398
Część IV	Zwiększanie złożoności	399
Rozdział 12	Łączenie mechanizmów uczących się	401
12.1.	Zespoły	401
12.2.	Zespoły głosujące	404
12.3.	Bagging i lasy losowe	404
12.3.1.	Technika bootstrap	404
12.3.2.	Od techniki bootstrap do metody bagging	408
12.3.3.	Przez losowy las	410
12.4.	Boosting	412
12.4.1.	Szczegółowe omówienie boostingu	413
12.5.	Porównywanie metod opartych na zespołach drzew	415
12.6.	Koniec rozdziału	418
12.6.1.	Podsumowanie	418
12.6.2.	Uwagi	419
12.6.3.	Ćwiczenia	420
Rozdział 13	Modele z automatyczną inżynierią cech	423
13.1.	Wybieranie cech	425
13.1.1.	Filtrowanie jednoetapowe z wybieraniem cech na podstawie miar	426
13.1.2.	Wybieranie cech na podstawie modelu	437
13.1.3.	Integrowanie wybierania cech z potokiem procesu uczenia	440
13.2.	Tworzenie cech za pomocą jąder	441
13.2.1.	Powód używania jąder	441
13.2.2.	Ręczne metody wykorzystujące jądra	446
13.2.3.	Metody wykorzystujące jądro i opcje jądra	450
13.2.4.	Klasyfikatory SVC dostosowane do jądra — maszyny SVM	454
13.2.5.	Uwagi do zapamiętania na temat maszyn SVM i przykładów	456

13.3. Analiza głównych składowych	
— technika nienadzorowana	457
13.3.1. Rozgrzewka — centrowanie	458
13.3.2. Znajdowanie innej najlepszej linii	459
13.3.3. Pierwsza analiza głównych składowych	461
13.3.4. Analiza głównych składowych od kuchni	463
13.3.5. Wielki finał — uwagi na temat analizy głównych składowych	469
13.3.6. Analiza głównych składowych dla jądra i metody oparte na rozmaitościach	470
13.4. Koniec rozdziału	473
13.4.1. Podsumowanie	473
13.4.2. Uwagi	474
13.4.3. Ćwiczenia	478

Rozdział 14 Inżynieria cech dla dziedzin — uczenie specyficzne dla dziedziny 481

14.1. Praca z tekstem	482
14.1.1. Kodowanie tekstu	484
14.1.2. Przykład maszynowego klasyfikowania tekstu	488
14.2. Klastrowanie	490
14.2.1. Klastrowanie metodą k-średnich	491
14.3. Praca z obrazami	492
14.3.1. Worek słów graficznych	492
14.3.2. Dane graficzne	493
14.3.3. Kompletny system	494
14.3.4. Kompletny kod transformacji obrazów na postać WGSZ	501
14.4. Koniec rozdziału	503
14.4.1. Podsumowanie	503
14.4.2. Uwagi	503
14.4.3. Ćwiczenia	505

Rozdział 15 Powiązania, rozwinięcia i kierunki dalszego rozwoju 507

15.1. Optymalizacja	507
15.2. Regresja liniowa z prostych składników	510
15.2.1. Graficzne ujęcie regresji liniowej	513
15.3. Regresja logistyczna z prostych składników	514
15.3.1. Regresja logistyczna i kodowanie zerojedynekowe	515
15.3.2. Regresja logistyczna z kodowaniem plus jeden – minus jeden	517
15.3.3. Graficzne ujęcie regresji logistycznej	518
15.4. Maszyna SVM z prostych składników	518

15.5. Sieci neuronowe	520
15.5.1. Regresja liniowa za pomocą sieci neuronowych	521
15.5.2. Regresja logistyczna za pomocą sieci neuronowych	523
15.5.3. Poza podstawowe sieci neuronowe	524
15.6. Probabilistyczne modele grafowe	525
15.6.1. Próbkowanie	527
15.6.2. Regresja liniowa za pomocą modelu PGM	528
15.6.3. Regresja logistyczna za pomocą modelu PGM	531
15.7. Koniec rozdziału	534
15.7.1. Podsumowanie	534
15.7.2. Uwagi	534
15.7.3. Ćwiczenia	535
Dodatek A Kod z pliku mlwpy.py	537