
Python w uczeniu maszynowym

Podejście sterowane testami

Matthew Kirk

przekład: Jakub Niedźwiedź

Spis treści

Wstęp	ix
Podziękowania.....	xi
1. W przybliżeniu prawdopodobnie poprawne oprogramowanie	1
Prawidłowe pisanie oprogramowania.....	2
SOLID	2
Testowanie albo TDD.....	4
Refaktoring	6
Pisanie prawidłowego oprogramowania	7
Pisanie odpowiedniego oprogramowania przy zastosowaniu uczenia maszynowego	7
Czym dokładnie jest uczenie maszynowe?	8
Wysoko oprocentowany dług uczenia maszynowego	8
Zastosowanie zasad SOLID w uczeniu maszynowym.....	9
Kod uczenia maszynowego jest skomplikowany	13
TDD: metoda naukowa 2.0.....	13
Refaktoring wiedzy.....	13
Plan tej książki	14
2. Szybkie wprowadzenie do uczenia maszynowego	15
Czym jest uczenie maszynowe?.....	15
Uczenie nadzorowane	16
Uczenie nienadzorowane	17
Uczenie wzmacniane.....	17
Co może osiągnąć uczenie maszynowe?	17
Notacja matematyczna używana w tej książce.....	19
Podsumowanie.....	20
3. K najbliższych sąsiadów	21
Jak ustalić, czy chcemy kupić dom?	21
Ile wart jest dany dom?.....	22
Regresja hedonistyczna.....	22
Czym jest sąsiedztwo?.....	23
K najbliższych sąsiadów	24
Najbliższe sąsiedztwo	24

Odległości	25
Nierówność trójkąta	25
Odległość geometryczna	26
Odległości obliczeniowe	27
Odległości statystyczne	30
Przekleństwo wymiarowości	31
Jak wybrać K?	32
Zgadywanie K	33
Heurystyka wyboru K	33
Wycenianie domów w Seattle	36
Informacje o danych	36
Ogólna strategia	37
Projekt kodowania i testowania	37
Konstrukcja regresora dla algorytmu K najbliższych sąsiadów	38
Testowanie algorytmu K najbliższych sąsiadów	40
Podsumowanie	43
4. Naiwna klasyfikacja bayesowska	45
Wykorzystanie twierdzenia Bayesa do znajdowania oszukańczych zamówień	45
Prawdopodobieństwa warunkowe	46
Symbole prawdopodobieństwa	46
Odwrócone prawdopodobieństwo warunkowe (czyli twierdzenie Bayesa)	48
Naiwny klasyfikator bayesowski	49
Reguła łańcuchowa	49
Naiwność w rozumowaniu bayesowskim	49
Pseudozliczanie	51
Filtr spamu	52
Uwagi przygotowawcze	52
Projekt kodowania i testowania	52
Źródło danych	53
EmailObject	53
Analiza leksykalna i kontekst	58
SpamTrainer	60
Minimalizacja błędów przez sprawdzanie krzyżowe	67
Podsumowanie	70
5. Drzewa decyzyjne i losowe lasy decyzyjne	71
Niuanse dotyczące grzybów	72
Klasyfikowanie grzybów przy wykorzystaniu wiedzy ludowej	73
Znajdowanie optymalnego punktu zwrotnego	74
Zysk informacyjny	75
Niejednorodność Gini	76

Redukcja wariancji	77
Przycinanie drzew	77
Uczenie zespołowe	77
Pisanie klasyfikatora grzybów.	79
Podsumowanie.	87
6. Ukryte modele Markowa	89
Śledzenie zachowania użytkownika przy użyciu automatów skończonych. . .	89
Emisje/obserwacje stanów	91
Uproszczenie poprzez założenie Markowa	93
Wykorzystanie łańcuchów Markowa zamiast automatu skończonego	93
Ukryty model Markowa	94
Ocena: algorytm Naprzód-Wstecz	94
Matematyczne przedstawienie algorytmu Naprzód-Wstecz	94
Wykorzystanie zachowania użytkownika	96
Problem dekodowania poprzez algorytm Viterbiego.	98
Problem uczenia	99
Oznaczanie części mowy z wykorzystaniem korpusu Browna	100
Uwagi przygotowawcze	100
Projekt kodowania i testowania	100
Podstawa naszego narzędzia do oznaczania części mowy: CorpusParser .	101
Pisanie narzędzia do oznaczania części mowy	103
Sprawdzanie krzyżowe w celu potwierdzenia poprawności modelu.	110
Jak ulepszyć ten model.	111
Podsumowanie.	112
7. Maszyny wektorów nośnych	113
Zadowolenie klientów jako funkcja tego, co mówią.	113
Klasyfikacja nastrojów przy użyciu maszyn wektorów nośnych	114
Teoria stojąca za maszynami wektorów nośnych	115
Granica decyzyjna	117
Maksymalizowanie granic.	117
Sztuczka jądrowa: transformacja cech.	118
Optymalizacja przez poluzowanie	120
Analizator nastrojów.	121
Uwagi przygotowawcze	121
Projekt kodowania i testowania	121
Strategie testowania maszyny wektorów nośnych	122
Klasa Corpus	122
Klasa CorpusSet	125
Sprawdzanie poprawności modelu i klasyfikator nastrojów	128
Agregowanie nastrojów	132

Wykładnicza ważona średnia ruchoma	133
Mapowanie nastroju do wyniku finansowego	134
Podsumowanie	134
8. Sieci neuronowe	135
Czym jest sieć neuronowa?	135
Historia sieci neuronowych	136
Logika boolowska	136
Perceptrony	137
Jak konstruować sieci neuronowe ze sprzężeniem w przód	137
Warstwa wejściowa	138
Warstwy ukryte	140
Neurony	141
Funkcje aktywacyjne	142
Warstwa wyjściowa	147
Algorytmy uczące	147
Zasada delty	148
Propagacja wsteczna	149
QuickProp	149
RProp	150
Budowanie sieci neuronowych	151
Ile ukrytych warstw?	151
Ile neuronów dla każdej warstwy?	152
Tolerancja błędów i maksymalna liczba epok	152
Wykorzystanie sieci neuronowej do klasyfikowania języków	153
Uwagi przygotowawcze	153
Projekt kodowania i testowania	154
Dane	154
Pisanie testu podstawowego dla języka	154
Przejście do klasy Network	157
Dostrajanie sieci neuronowej	161
Precyzja i czułość w sieciach neuronowych	161
Podsumowanie przykładu	161
Podsumowanie	161
9. Grupowanie	163
Badanie danych bez żadnego błędu systematycznego	163
Kohorty użytkowników	164
Testowanie mapowań do grup	166
Zdatność grupy	166
Współczynnik zarysu	166
Porównywanie wyników z prawdą bazową	167

Grupowanie K-średnich	167
Algorytm K-średnich	168
Słabe strony grupowania K-średnich	169
Grupowanie przez maksymalizację wartości oczekiwanej	169
Algorytm	170
Twierdzenie o niemożności	171
Przykład: kategoryzowanie muzyki	172
Uwagi przygotowawcze	172
Zbieranie danych	173
Projekt kodowania	173
Analizowanie danych przez algorytm K-średnich	174
Grupowanie naszych danych	175
Wyniki z grupowania danych dotyczących muzyki jazzowej z użyciem maksymalizowania wartości oczekiwanej	180
Podsumowanie	182
10. Poprawianie modeli i wydobywania danych	183
Klub dyskusyjny	183
Wybieranie lepszych danych	184
Wybieranie cech	184
Wyczerpujące wyszukiwanie	187
Losowe wybieranie cech	188
Lepszy algorytm wybierania cech	189
Wybieranie cech przez minimalizowanie redundancji i maksymalizowanie istotności	190
Transformacja cech i rozkład macierzy	191
Analiza głównych składowych	191
Analiza niezależnych składowych	193
Uczenie zespołowe	195
Grupowanie typu bootstrap	195
Boosting	195
Podsumowanie	197
11. Łączenie wszystkiego razem: Podsumowanie	199
Przypomnienie algorytmów uczenia maszynowego	199
Jak wykorzystywać te informacje do rozwiązywania problemów	201
Co dalej?	202
O autorze	203
Kolofon	203