Nowa Biblioteka. Usługi, Technologie Informacyjne i Media 2018, nr 3 (30), s. 189–192
ISSN 2451-2575 (wersja elektroniczna)
http://www.nowabiblioteka.us.edu.pl/
GICID 71.0000.1500.0677

ARTYKUŁY RECENZYJNE, RECENZJE I OMÓWIENIA

Anna Małgorzata Kamińska
Zakład Bibliotekoznawstwa
Instytut Bibliotekoznawstwa i Informacji Naukowej
Uniwersytet Śląski w Katowicach
e-mail: anna.kaminska@us.edu.pl

Data science i uczenie maszynowe / Marcin Szeliga. – Warszawa : Wydawnictwo Naukowe PWN SA, 2017. – XXVI, [2], 371, [1] s. : il. ; 24 cm. – ISBN 978-83-01-19232-7

Termin *data science* został po raz pierwszy użyty przez Petera Naura w 1960 r. na określenie metod automatycznego przetwarzania danych przy użyciu komputerów. Podstawowym założeniem *data science* jest uczenie się na podstawie danych. Dotyczy to zarówno ludzi, jak i maszyn. Dzięki rozwojowi technologii informatycznych liczba cyfrowych danych rośnie. Dotyczą one otaczającego nas świata, a ukryta w nich informacja ma ogromną wartość. Ponieważ jest ich dużo i są gromadzone szybko, do ich analizy wykorzystuje się komputery. Takie rozwiązanie stosują nie tylko ośrodki naukowe, lecz także firmy działające w najróżniejszych sektorach, od finansów, przez sektor publiczny i produkcyjny, po handel. Wynikiem wspomnianych analiz są modele eksploracji danych, które poprzez zastosowanie takich algorytmów, jak sieci neuronowe, lasy drzew decyzyjnych czy maszyny wektorów nośnych, analizują ukryte w danych wzorce i na tej podstawie tworzą ogólne reguły.

Ciekawe refleksje na temat tej dziedziny przedstawił Marcin Szeliga w opracowaniu pt. *Data science i uczenie maszynowe*. Można je podzielić na trzy części. W pierwszej autor wyjaśnia, czym tak naprawdę jest uczenie maszynowe i jakiego typu problemy można przy jego uży-

ciu rozwiązać. W drugiej omawia jeden z najważniejszych elementów procesu *data science*, czyli przygotowanie danych. W trzeciej prezentuje klasyczne algorytmy używane do rozwiązywania określonych klas problemów, takich jak: klasyfikacja, analiza regresji, skupień czy prognozowanie.

W kolejnych rozdziałach książki M. Szeliga opisuje typowy proces automatycznej analizy danych (w terminologii *data science* nazywany eksperymentem), prezentując etapy od zdefiniowania problemu w kontekście posiadanych danych, aż po wdrożenie wybranego modelu w celu jego rozwiązania i przedstawienie wyników użytkownikom. Eksperymenty *data science* można podzielić na te, których celem jest uzupełnienie brakujących danych (predykcja), i te pomagające odkrywać ogólne wzorce ukryte w danych (deskrypcja). Analiza deskrypcyjna obejmuje znalezienie i przedstawienie zależności (wzorców) ukrytych w danych oraz wyjaśnienie przyczyn ich występowania. Jakość wyników tego typu analizy oceniana jest przede wszystkim na podstawie ich przydatności dla użytkowników. Cechami charakterystycznymi są: konieczność posługiwania się terminologią z modelowanej dziedziny oraz wymóg uwzględniania wszystkich, również nietypowych, przypadków i zjawisk. W analizie predykcyjnej utworzony model danych służy do uzupełniania brakujących danych. W takich analizach wykorzystuje się możliwości statystycznej oceny dokładności oraz wiarygodności wyników i przekształceń obejmujących celowe zniekształcanie danych treningowych.

Celem książki jest przedstawienie naukowej metody tworzenia modeli eksploracji danych. Autor opisuje zasady działania poszczególnych algorytmów, ale też prowadzi czytelnika przez wszystkie etapy eksperymentu *data science*, skupiając się przede wszystkim na analizie predykcyjnej. Książka jest adresowana do tych, którzy chcieliby: zapoznać się z zagadnieniami związanymi z *data science* lub poszerzyć swoją wiedzę w tym zakresie; zdobyć specjalistyczną wiedzę z dziedziny, w ramach której realizowany jest eksperyment uczenia maszynowego; pozyskać informacje o odpowiedniej ocenie danych i wstępnym przygotowaniu ich do analizy przy użyciu specjalistycznego języka (np. SQL czy R), wyborze i parametryzacji właściwych algorytmów uczenia maszynowego, ocenie jakości utworzonych modeli i prezentacji uzyskanych wyników użytkownikom. Opisane w książce zagadnienia autor zobrazował praktycznymi przykładami, wybierając usługę Azure Machine Learning Studio i język R. Za pomocą tych narzędzi, dostępnych za darmo i bogato udokumentowanych w internecie, można w prosty sposób wykonywać skomplikowane analizy. Język R jest używany do statystycznego

analizowania danych, zaś skrypty R w usłudze Azure ML służą do oceny i przygotowania danych, tworzenia modeli predykcyjnych i wizualizacji wyników. Do utworzenia modeli klasyfikacyjnych użyto dwóch algorytmów: drzew decyzyjnych oraz maszyn wektorów nośnych.

W rozdziale pierwszym M. Szeliga opisuje uczenie maszynowe jako element eksperymentów *data science* i eksploracji danych – techniki wspomagania decyzji, ponadto definiuje pojęcia związane z modelowaniem i założeniami eksperymentu. W rozdziale drugim wyjaśnia, na czym polega ocena przydatności danych. Przedstawia metody zbierania i oceniania danych źródłowych (zarówno pojedynczych zmiennych, jak i zależności między zmiennymi) oraz sposoby sprawdzania integralności danych, ustalania wymaganej liczby przypadków i upraszczania modeli. Rozdział trzeci zawiera informacje na temat wstępnego przetwarzania danych, które obejmuje uzupełnienie brakujących wartości, poprawianie błędnych danych, przekształcenie zmiennych, wydzielenie danych testowych i kontrolnych oraz specyficzne operacje potrzebne do przekształcenia danych na potrzeby eksperymentu. W kolejnym rozdziale autor opracowania opisuje etap wzbogacania danych, od którego wyników w dużym stopniu zależą wyniki całego eksperymentu, podkreślając, że etap ten wymaga ścisłej współpracy analityka z ekspertem z dziedziny eksperymentu, ponieważ efektywne wzbogacanie danych wymaga znajomości modelowanego zagadnienia. Ponadto analizuje zagadnienie zrównoważenia danych źródłowych, metody wzbogacania danych przez tworzenie zmiennych pochodnych i uniwersalną metodę wzbogacania danych przez zastąpienie zmiennych wejściowych wspólnym rozkładem prawdopodobieństwa, z jakim wpływają one na zmienną wejściową. Kolejne rozdziały są poświęcone poszczególnym algorytmom eksploracji danych. Rozdział piąty dotyczy jednej z najstarszych i najczęściej stosowanych metod eksploracji danych – klasyfikacji. Jej celem jest znalezienie modelu klasyfikacyjnego – klasyfikatora, który, nauczony na podstawie danych historycznych, treningowych, będzie przypisywał nowe przypadki do jednej z możliwych klas. W rozdziale szóstym autor omawia kolejną metodę uczenia nadzorowanego eksploracji danych – regresję, służącą do znalezienia modelu, który na podstawie znanych danych wystarczająco dokładnie obliczy brakujące wartości. Następną metodą, przedstawioną w rozdziale siódmym, jest grupowanie. W przeciwieństwie do klasyfikacji i regresji jest to technika uczenia nienadzorowanego, a jej celem jest podział obserwacji na grupy obiektów o podobnych cechach, czyli na klastry lub skupienia, bez wcześniejszej wiedzy na temat tego, jak docelowe grupy powinny wyglądać. W rozdziale ósmym autor przybliża zagadnienie systemów rekomendu-

jących, których zadaniem jest wybranie z wielu dostępnych opcji tych odpowiadających potrzebom użytkownika. Systemy te są powszechnie stosowane do szukania asocjacji zgodnych z kryteriami wcześniejszego wyszukiwania, np. do wybierania wyświetlanych w serwisach społecznościowych reklam, sugerowania produktów w e-sklepach, polecania książek, filmów, muzyki itp. Rozdział dziewiąty zawiera wiadomości na temat modeli prognozujących, służących do sporządzania prognoz, czyli przewidywania przyszłych wartości zmiennej numerycznej na podstawie poprzednich wartości zmiennych (wartości historycznych), ze szczególnym uwzględnieniem prognozowania na podstawie szeregów czasowych.

W rozdziale dziesiątym autor skupia się na ocenie i poprawie jakości modeli, możliwym dzięki zastosowaniu odpowiednich kryteriów. Modele aktualizuje się co pewien czas, biorąc pod uwagę opinie użytkowników na co dzień z nimi pracujących, a także techniczną ocenę ich jakości, przeprowadzoną w momencie zakończenia eksperymentu. Modele oceniane są pod kątem łatwości ich interpretacji, trafności, wiarygodności, wydajności i skalowalności, a także przydatności.

W rozdziale jedenastym, będącym podsumowaniem książki, autor zamieszcza informacje z wszystkich poprzednich rozdziałów w celu utworzenia wzorcowego eksperymentu *data science*, który następnie udostępniono poprzez usługę WWW. Eksperyment bazował na popularnym zbiorze danych zawierającym wymiary płatków i kielichów kwiatu irysa, na podstawie których dokonano klasyfikacji gatunków tych roślin. Publikacja została wzbogacona o bibliografię oraz polsko-angielski i angielsko-polski słownik terminów *data science*.

Tekst wpłynął do redakcji 23 maja 2018 r.