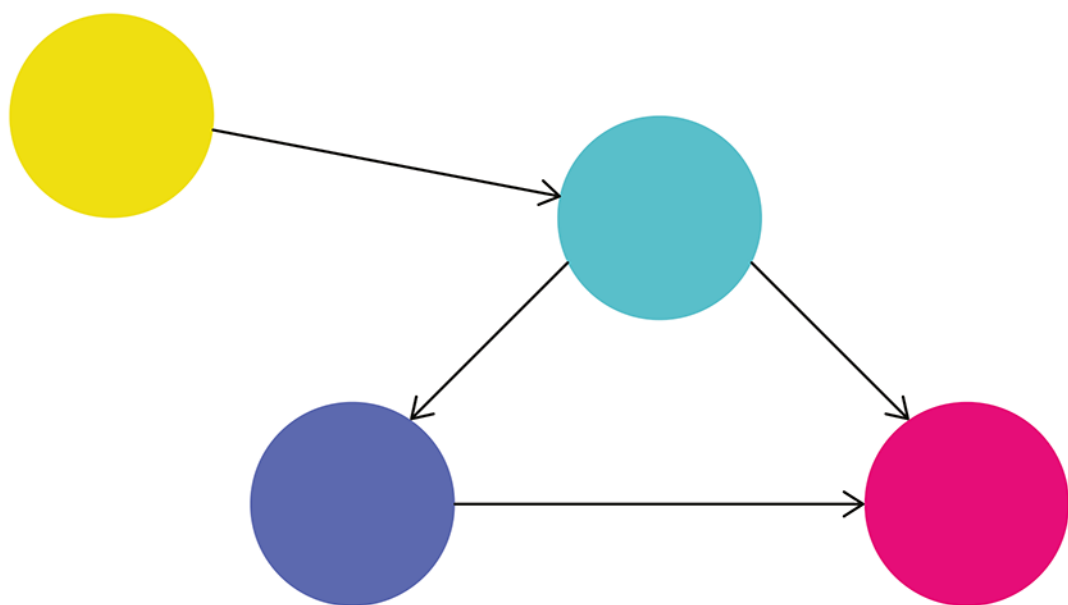


⟨packt⟩

Wnioskowanie i związki przyczynowe w Pythonie

Nowoczesne uczenie maszynowe
z wykorzystaniem bibliotek DoWhy, EconML,
PyTorch i nie tylko



ALEKSANDER MOLAK

Tytuł oryginału: Causal Inference and Discovery in Python – Machine Learning and Pearlman Perspective: Unlock the secrets of modern causal machine learning with DoWhy, EconML, PyTorch and more

Tłumaczenie: Radosław Meryk

ISBN: 978-83-289-0832-1

Copyright © Packt Publishing 2023. First published in the English language under the title 'Causal Inference and Discovery in Python' – (9781804612989).

Polish edition copyright © 2024 by Helion S.A.

All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage retrieval system, without permission from the Publisher.

Wszelkie prawa zastrzeżone. Nieautoryzowane rozpowszechnianie całości lub fragmentu niniejszej publikacji w jakiegokolwiek postaci jest zabronione. Wykonywanie kopii metodą kserograficzną, fotograficzną, a także kopiowanie książki na nośniku filmowym, magnetycznym lub innym powoduje naruszenie praw autorskich niniejszej publikacji.

Wszystkie znaki występujące w tekście są zastrzeżonymi znakami firmowymi bądź towarowymi ich właścicieli.

Autor oraz wydawca dołożyli wszelkich starań, by zawarte w tej książce informacje były kompletne i rzetelne. Nie biorą jednak żadnej odpowiedzialności ani za ich wykorzystanie, ani za związane z tym ewentualne naruszenie praw patentowych lub autorskich. Autor oraz wydawca nie ponoszą również żadnej odpowiedzialności za ewentualne szkody wynikłe z wykorzystania informacji zawartych w książce.

Drogi Czytelniku!

Jeżeli chcesz ocenić tę książkę, zajrzyj pod adres

<https://helion.pl/user/opinie/wnizwi>

Możesz tam wpisać swoje uwagi, spostrzeżenia, recenzję.

Helion S.A.

ul. Kościuszki 1c, 44-100 Gliwice

tel. 32 230 98 63

e-mail: helion@helion.pl

WWW: <https://helion.pl> (księgarnia internetowa, katalog książek)

Printed in Poland.

- Kup książkę
- Poleć książkę
- Oceń książkę

- Księgarnia internetowa
- Lubię to! » Nasza społeczność

Spis treści |

O autorze	15
O recenzentach	16
Podziękowania	17
Słowo wstępne	19
Przedmowa	21

CZĘŚĆ 1. Przyczynowość — wprowadzenie

ROZDZIAŁ 1

Związki przyczynowe? Przecież jest uczenie maszynowe, więc po co zawracać sobie tym głowę?	29
Krótka historia przyczynowości	30
Dlaczego przyczynowość? Zapytaj dzieci!	31
Interakcje ze światem	31
Zakłócenia — związki, które nie są prawdziwe	32
Jak nie stracić pieniędzy... i ludzkich istnień	35
Dylemat marketera	35
Pobawmy się w doktora!	36
Asocjacje w realnym świecie	37
Podsumowanie	37
Bibliografia	38

ROZDZIAŁ 2

Judea Pearl i drabina przyczynowości	39
Od asocjacji do logiki i wyobraźni. Drabina przyczynowości	39
Asocjacje	42
Ćwiczenie	44
Czym są interwencje?	47
Zmienianie świata	48
Korelacja i przyczynowość	50

Czym są kontrfakty?	52
Zanurzymy się w dziwactwa (zapis formalny)	53
Podstawowy problem wnioskowania przyczynowego	54
Obliczanie kontrfaktów	54
Czas na kodowanie!	56
Dodatek. Czym jest uczenie maszynowe z perspektywy przyczynowości? ...	57
Przyczynowość a uczenie ze wzmocnieniem	57
Przyczynowość a uczenie półnadzorowane i nienadzorowane	58
Podsumowanie	58
Bibliografia	59

ROZDZIAŁ 3

Regresja, obserwacje i interwencje	61
Wprowadzenie. Dane obserwacyjne a regresja liniowa	61
Regresja liniowa	62
Wartości p i istotność statystyczna	65
Interpretacja geometryczna regresji liniowej	66
Odwrocenie kolejności	67
Czy zawsze należy kontrolować wszystkie dostępne współzmiennne?	69
Poruszanie się po labiryncie	69
Jeśli nie wiesz, dokąd zmierzasz, możesz trafić gdzie indziej	70
Pójdźmy dalej!	73
Kontrolować czy nie kontrolować?	73
Modele regresyjne a modele strukturalne	73
Modele SCM	73
Regresja liniowa a modele SCM	74
Szukanie powiązania	74
Regresja a skutki przyczynowe	76
Podsumowanie	78
Bibliografia	78

ROZDZIAŁ 4

Modele grafów	79
Grafy, grafy, grafy	79
Rodzaje grafów	80
Reprezentacje grafów	83
Grafy w Pythonie	85
Czym jest model grafów?	87
Skierowane grafy acykliczne w świecie związków przyczynowych	88
Definicje przyczynowości	88
Grafy DAG a przyczynowość	89
Formalna definicja grafów DAG	90
Ograniczenia grafów DAG	90

Źródła grafów przyczynowych w świecie rzeczywistym	91
Odkrywanie związków przyczynowych	91
Wiedza ekspercka	91
Połączenie technik odkrywania związków przyczynowych i wiedzy eksperckiej	92
Dodatek. Czy można opisywać związki przyczynowe bez grafów DAG?	92
Układy dynamiczne	92
Cykliczne modele SCM	92
Podsumowanie	93
Bibliografia	93

ROZDZIAŁ 5

Rozwidlenia, łańcuchy i kolidery	95
Grafy i rozkłady oraz sposoby mapowania między nimi	95
Jak opisywać niezależność?	96
Wybór właściwego kierunku	97
Warunki i założenia	98
Łańcuchy, rozwidlenia i kolidery	102
łańcuch zdarzeń	102
łańcuchy	103
Rozwidlenia	104
Kolidery lub struktury v	106
Przypadki niejednoznaczne	108
Rozwidlenia, łańcuchy, kolidery i regresja	109
Tworzenie zbioru danych dla łańcucha	110
Tworzenie zestawu danych dla rozwidlenia	112
Tworzenie zbioru danych dla kolidera	112
Dopasowanie modeli regresji	113
Podsumowanie	116
Bibliografia	116

CZĘŚĆ 2. Wnioskowanie związków przyczynowych

ROZDZIAŁ 6

Węzły, krawędzie i statystyczna (nie)zależność	119
Zadbaj o separację d!	120
Trening czyni mistrza — separacja d	121
Najpierw estymandy!	124
Żyjemy w świecie estymatorów	124
Czym są estymandy?	124

Kryterium back-door	126
Czym jest kryterium back-door?	127
Kryterium back-door a estymandy równoważne	127
Kryterium front-door	130
Czy GPS może nas wyprowadzić na manowce?	130
Londyńskie taksówki i magiczny kamień	131
Otwarcie frontowych drzwi	132
Trzy proste kroki w kierunku kryterium front-door	134
Kryterium front-door w praktyce	134
Czy są jakieś inne kryteria? Zastosujmy rachunek do!	140
Trzy zasady rachunku do	141
Zmienne instrumentalne	142
Podsumowanie	144
Odpowiedź	144
Bibliografia	145

ROZDZIAŁ 7

Czteroetapowy proces wnioskowania przyczynowego	146
Wprowadzenie do bibliotek DoWhy i EconML	147
Ekosystem analizy przyczynowej Pythona	147
Dlaczego DoWhy?	149
Czym jest pakiet DoWhy?	150
A co z biblioteką EconML?	150
Krok 1. Modelowanie problemu	151
Utworzenie grafu	151
Tworzenie obiektu CausalModel	153
Krok 2. Identyfikacja estymand	154
Krok 3. Wyznaczanie oszacowań	156
Krok 4. Zestaw walidacyjny. Testy obalające	156
Jak walidować modele przyczynowe?	157
Wprowadzenie do testów obalających	158
Pełny przykład	160
Krok 1. Zakodowanie założeń	161
Krok 2. Wyznaczenie estymandy	163
Krok 3. Wyznaczenie oszacowania	163
Krok 4. Obalenie oszacowania	165
Podsumowanie	169
Bibliografia	169

ROZDZIAŁ 8

Modele przyczynowe. Założenia i wyzwania	170
Jestem królem świata! Czy rzeczywiście tak jest?	170
Gdzieś pośrodku	171
Identyfikowalność	172
Brak grafów przyczynowych	172
Za mało danych	173
Nieweryfikowalne założenia	175
Słów w pokoju — nadzieja czy beznadzieja?	175
Zjedźmy słonia	175
Dodatniość	177
Wymiennność	179
Podmioty wymienne	180
Wymiennność a zakłócenia	180
...i inne	181
Modułowość	181
SUTVA	183
Spójność	183
Nazywaj mnie po imieniu — relacje pozorne	184
Nazwy, nazwy, nazwy	184
Czy powinienem zapytać Ciebie, czy kogoś, kogo tu nie ma?	185
Stwórzcie graf DAG!	185
Dodatkowe informacje o stronniczości wyboru	187
Podsumowanie	189
Bibliografia	189

ROZDZIAŁ 9

Wnioskowanie związków przyczynowych i uczenie maszynowe — od dopasowywania do metalearnerów	191
Podstawy I. Dopasowywanie	192
Rodzaje dopasowywania	192
Efekty interwencji — ATE w porównaniu z ATT i ATC	193
Estymatory dopasowywania	194
Implementacja dopasowywania	196
Podstawy II. Współczynniki skłonności	201
Dopasowywanie w praktyce	201
Zmniejszenie wymiarowości za pomocą współczynników skłonności ...	202
Dopasowywanie współczynników skłonności (PSM)	203
Odwrotne ważenie prawdopodobieństwa (IPW)	204
Wiele twarzy współczynników skłonności	204
Formalizacja techniki IPW	205
Implementacja IPW	205
IPW — względy praktyczne	206

S-Learner — samotny stróż	207
Diabeł tkwi w szczegółach	207
Mamo, tato, poznajcie CATE	208
Żarty na bok. Pozdrowienia dla heterogenicznego tłumu	209
Machanie flagą założeń	210
Jesteś jedyny. Modelowanie z wykorzystaniem techniki S-Learner	211
Dane o niewielkiej objętości	217
Słabe punkty modelu S-Learner	218
T-Learner. Razem możemy więcej	218
Wymuszenie właściwego podziału zmiennych	219
T-Learner w czterech krokach i wzory	219
Implementacja modelu T-Learner	220
X-Learner. Krok dalej	222
Wyciskanie cytryny	222
Rekonstrukcja modelu X-Learner	223
X-Learner. Formuła alternatywna	225
Implementacja X-Learner	226
Podsumowanie	230
Bibliografia	231

ROZDZIAŁ 10

Wnioskowanie związków przyczynowych i uczenie maszynowe — zaawansowane estymatory, eksperymenty, oceny i nie tylko 233

Metody DR. Spróbujmy uzyskać więcej!	234
Czy potrzebujemy czegoś więcej?	234
Podwójnie wzmocniony nie oznacza niezniszczalny...	236
...ale pozwala wiele zyskać	236
Sekretny, podwójnie mocny sos	236
Estymator DR a założenia	238
DR-Learner. Przechodzenie nad przepaścią	238
Modele DR-Learner — opcje dodatkowe	242
Ukierunkowany estymator maksymalnego prawdopodobieństwa	242
Jeśli uczenie maszynowe jest fajne, to co powiesz na podwójne uczenie maszynowe?	246
Dlaczego DML i co jest w nim podwójnego?	246
Implementacja DML za pomocą bibliotek DoWhy i EconML	249
Dostrajanie hiperparametrów za pomocą bibliotek DoWhy i EconML ...	252
Czy DML jest srebrną kulą?	256
Techniki DR a DML	258
Co z tego będę miał?	259

Lasy przyczynowe i nie tylko	260
Drzewa przyczynowe	260
Przepelnienia lasów	260
Zalety lasów przyczynowych	260
Lasy przyczynowe z wykorzystaniem bibliotek DoWhy i EconML	261
Niejednorodne efekty interwencji z danymi eksperymentalnymi — odyseja upliftingu	263
Dane	263
Wybór frameworka	268
Nie znamy połowy tej historii	268
Wyzwanie Kevina	269
Otwarcie skrzynki z narzędziami	270
Modele uplift a wydajność	274
Inne wskaźniki dla wyników ciągłych z wieloma interwencjami	279
Przedziały ufności	279
Zwycięski wynik w wyzwaniu Kevina	280
Kiedy należy stosować estymatory CATE dla danych eksperymentalnych?	280
Wybór modelu. Uproszczony przewodnik	281
Dodatek. Objaśnienia kontrfaktyczne	283
Zła wola czy nieodpowiednia technologia?	283
Podsumowanie	284
Bibliografia	285

ROZDZIAŁ 11

Wnioskowanie związków przyczynowych i uczenie maszynowe — uczenie głębokie, przetwarzanie języka naturalnego i inne techniki 288

Wykorzystanie technik uczenia głębokiego do wyznaczania heterogenicznych efektów interwencji	289
Wskaźniki CATE sięgają głębiej	289
SNet	291
Transformatory i wnioskowanie związków przyczynowych	299
Teoria znaczenia w pięciu akapitach	299
Co zrobić, by komputery rozumiały język naturalny?	300
Od filozofii do kodu Pythona	301
Modele LLM a przyczynowość	301
Trzy scenariusze	303
CausalBert	306
Przyczynowość i szeregi czasowe, czyli kiedy ekonometryk przechodzi na Bayesa	312
Metody quasi-eksperymentalne	312
Przejęcie Twittera i wzorce googlowania	313

Logika syntetycznych kontroli	313
Wizualne wprowadzenie do logiki kontroli syntetycznej	314
Na początek dane	317
Kontrola syntetyczna w kodzie	317
Wyzwania	322
Podsumowanie	323
Bibliografia	324

CZĘŚĆ 3. Odkrywanie związków przyczynowych

ROZDZIAŁ 12

Czy można prosić o graf przyczynowy?	329
Źródła wiedzy przyczynowej	329
Zalew informacji	330
Siła zaskoczenia	330
Spostrzeżenia naukowe	331
Logika nauki	331
Hipotezy są gatunkiem	332
Jedna logika, wiele dróg	333
Eksperymenty kontrolowane	333
Randomizowane badania kontrolowane	334
Od eksperymentów do grafów	334
Symulacje	335
Osobiste doświadczenia i wiedza dziedzinowa	335
Osobiste doświadczenia	336
Wiedza dziedzinowa	337
Uczenie się struktury przyczynowej	337
Podsumowanie	338
Bibliografia	339

ROZDZIAŁ 13

Odkrywanie związków przyczynowych i uczenie maszynowe — od założeń do zastosowań	340
Odkrywanie związków przyczynowych — przypomnienie informacji o założeniach	341
Przygotowania	341
Należy zawsze dążyć do zapewnienia wierności... ..	341
...ale czasami to jest trudne	341
Minimalizm jest cnotą	342
Cztery (i pół) rodziny	342
Cztery strumienie	343

Wprowadzenie do pakietu gCastle	344
Witaj, gCastle!	344
Dane syntetyczne w gCastle	345
Dopasowywanie pierwszego modelu odkrywania związków przyczynowych	348
Wizualizacja modelu	349
Wskaźniki oceny modelu	350
Odkrywanie związków przyczynowych oparte na ograniczeniach	353
Ograniczenia i niezależność	353
Wykorzystanie struktury niezależności w celu odtworzenia grafu	354
Algorytm PC — ukryte wyzwania	357
Algorytm PC dla danych kategoryalnych	358
Odkrywanie związków przyczynowych na podstawie punktacji	359
Tabula rasa — zaczynamy od nowa	359
GES — punktacja	359
Algorytm GES w bibliotece gCastle	360
Funkcyjne odkrywanie związków przyczynowych	361
Błogosławieństwa asymetrii	361
Model ANM	362
Ocena niezależności	365
Czas na LiNGAM	366
Odkrywanie związków przyczynowych oparte na gradientach	372
Czym jest ten gradient?	372
Proszę nie ronić łez!	374
Nie płacz, GOLEM!	374
Porównanie	375
Kodowanie wiedzy eksperckiej	377
Czym jest wiedza ekspercka?	378
Wiedza ekspercka w bibliotece gCastle	378
Podsumowanie	379
Bibliografia	379

ROZDZIAŁ 14

Odkrywanie związków przyczynowych i uczenie maszynowe — zaawansowane uczenie głębokie i nie tylko	382
Zaawansowane odkrywanie związków przyczynowych za pomocą uczenia głębokiego	383
Od modeli generatywnych do przyczynowości	384
Spójrz wstecz, aby dowiedzieć się, kim jesteś	384
Elementy składowe frameworka DECI	385
Implementacja DECI	386
DECI to rozwiązanie kompleksowe	398

Odkrywanie związków przyczynowych w wypadku występowania ukrytych zakłóceń	398
Algorytm FCI	398
Inne podejścia do danych z zakłóceniami	403
Dodatek. Nie tylko obserwacje	404
ENCO	404
ABCI	404
Odkrywanie związków przyczynowych — praktyczne zastosowania, wyzwania i otwarte problemy	405
Podsumowanie	406
Bibliografia	407

ROZDZIAŁ 15

Epilog	409
Czego nauczyłeś się z tej książki?	409
Pięć kroków do jak najlepszego wykorzystania projektów przyczynowych	410
Zadaj pytanie	410
Zdobądź wiedzę ekspercką	411
Wygeneruj hipotetyczny graf (grafy)	412
Sprawdź identyfikowalność	412
Dokonaj falsyfikacji hipotez	413
Przyczynowość w biznesie	414
Jak eksperci analizy przyczynowej przechodzą od wizji do implementacji?	414
Przyszłość przyczynowego uczenia maszynowego	416
Gdzie jesteśmy dziś i dokąd zmierzamy?	417
Wskaźniki przyczynowe	418
Fuzja danych przyczynowych	418
Agenty interwencji	419
Uczenie się struktury przyczynowej	419
Uczenie się przez naśladowanie	420
Studiowanie przyczynowości	421
Pozostańmy w kontakcie	422
Podsumowanie	422
Bibliografia	423
Skorowidz	425

Związki przyczynowe? Przecież jest uczenie maszynowe, więc po co zawracać sobie tym głowę?

Rozdział

1

Tutaj zaczyna się nasza podróż.

W tym rozdziale zadam kilka pytań dotyczących analizy związków przyczynowych.

Czym ona jest? Czy wnioskowanie przyczynowe różni się od wnioskowania statystycznego? Jeśli tak, to czym?

Czy analiza przyczynowa w ogóle jest potrzebna? Uczenie maszynowe wydaje się wystarczająco dobre.

Jeśli śledziłeś szybko zmieniający się krajobraz uczenia maszynowego przez ostatnie 5 – 10 lat, prawdopodobnie zauważyłeś wiele przykładów — jak lubimy to nazywać w społeczności uczenia maszynowego — *irracjonalnej skuteczności* nowoczesnych algorytmów uczenia maszynowego w dziedzinach rozpoznawania obrazów, przetwarzania języka naturalnego i innych.

Takie algorytmy jak DALL-E 2 czy GPT-3/4 trafiły do świadomości nie tylko środowiska naukowców, ale także ogółu społeczeństwa.

Mógłbyś zadać sobie pytanie — jeśli te algorytmy tak dobrze się sprawdzają, to po co w ogóle zawracać sobie głowę czymś innym?

Ten rozdział rozpocznę krótkim omówieniem historii przyczynowości. Następnie podam kilka powodów stosowania w modelowaniu podejścia przyczynowego zamiast czysto statystycznego i wprowadzę pojęcie zakłóceń (ang. *confounding*).

Na koniec zaprezentuję przykłady zastosowania podejścia przyczynowego do rozwiązywania problemów związanych z marketingiem i medycyną. Po zakończeniu lektury tego rozdziału czytelnik powinien mieć jasny obraz obszarów, w których może być przydatne wnioskowanie przyczynowe, oraz powodów tego stanu. Powinien umieć wyjaśnić, czym jest zakłócanie i dlaczego jest ono ważne.

W tym rozdziale omówię następujące zagadnienia:

- Krótka historia przyczynowości.
- Motywacje do stosowania przyczynowego podejścia do modelowania.
- Jak nie stracić pieniędzy... i ludzkich istnień.

Krótka historia przyczynowości

Przyczynowość (ang. *causality*) ma długą historię. Była przedmiotem zainteresowania większości znanych kultur. Jeden z najpłodniejszych filozofów starożytnej Grecji, Arystoteles, twierdził, że zrozumienie przyczynowej struktury procesu jest niezbędnym elementem wiedzy o tym procesie. Ponadto argumentował, że istotą wyjaśniania naukowego jest umiejętność odpowiadania na pytania w rodzaju *dlaczego* (Falcon, 2006, 2022). Arystoteles wyróżnił cztery typy przyczyn (materialne, formalne, sprawcze i ostateczne). O ile taka klasyfikacja pozwala uwzględnić w równym stopniu pewne interesujące aspekty rzeczywistości, o tyle dla współczesnego czytelnika może sprawiać wrażenie sprzecznej z intuicją.

Słynny szkocki filozof z XVIII wieku, David Hume, zaproponował bardziej ujednolicone ramy związków przyczynowo-skutkowych. Hume wyszedł z założenia, że w rzeczywistym świecie nigdy nie obserwujemy związków przyczynowo-skutkowych. Jedyne, czego doświadczamy, to obserwacja, że niektóre zdarzenia są ze sobą powiązane:

„Dostrzegamy jedynie fakt, że jedno zdarzenie wynika z drugiego. Uderzenie jednej kuli bilardowej powoduje ruch drugiej. To wszystko, co ukazuje się zewnętrznym zmysłom. Umysł nie odczuwa ani żadnego uczucia, ani wewnętrznego wrażenia wynikających z tego następstwa zdarzeń. W związku z tym nie ma, w żadnym pojedynczym, konkretnym przypadku przyczyny i skutku, niczego co mogłoby sugerować ideę mocy lub koniecznego związku” (Hume i Millican, 2007; pierwotnie opublikowany w 1739 r.).

Jedna z interpretacji teorii przyczynowości Hume’a (tu uproszczona dla zachowania przejrzystości) brzmi następująco:

- Dostrzegamy jedynie to, że ruch lub pojawienie się obiektu *A* poprzedza ruch lub pojawienie się obiektu *B*.
- Jeśli zaobserwujemy takie następstwo zdarzeń wystarczającą liczbę razy, rozwija się w nas poczucie oczekiwania.
- To poczucie oczekiwania jest sednem pojmowania przyczynowości (nie dotyczy świata, lecz uczucia, które rozwijamy).

Teoria przyczynowości Hume’a

Interpretacja Hume’owskiej teorii przyczynowości, którą podałem powyżej, nie jest jedyna. Warto zauważyć, że nawet Hume w swoim późniejszym dziele *An Inquiry Concerning the Human Understanding* (1758) podał inną definicję przyczynowości. Co więcej, nie wszyscy naukowcy podzielają moją wizję (na przykład Archie, 2005).

Teoria Hume’a jest bardzo interesująca z co najmniej dwóch punktów widzenia.

Po pierwsze elementy jego teorii wykazują duże podobieństwo do bardzo ważnego terminu w psychologii zwanego warunkowaniem (ang. *conditioning*). **Warunkowanie** jest formą uczenia się. Choć istnieje wiele rodzajów warunkowania, wszystkie opierają się na wspólnym fundamencie — mianowicie na skojarzeniach, nazywanych również **asocjacjami** (stąd nazwa tego typu uczenia się — **uczenie asocjacyjne**). W każdym typie warunkowania bierzemy jakieś zdarzenie lub obiekt (zwykle nazywane bodźcem) i kojarzymy je z pewnym zachowaniem lub reakcją. Uczenie asocjacyjne obserwuje się w przyrodzie u różnych gatunków. Jest właściwe ludziom, małpom człekokształtnym, psom i kotom, ale także występuje u znacznie prostszych organizmów, takich jak ślimaki (Alexander, Audesirk i Audesirk, 1985).

Warunkowanie

Więcej informacji na temat różnych rodzajów warunkowania znajdziesz na stronie <https://bit.ly/MoreOnConditioning>. Możesz również poszukać na przykład frazy *warunkowanie klasyczne* kontra *warunkowanie instrumentalne* i nazwisk, np. Ivan Pavlov i Burrhus Skinner.

Po drugie asocjacje są również podstawą większości klasycznych algorytmów uczenia maszynowego. Kiedy szkolimy sieć neuronową z wykorzystaniem uczenia nadzorowanego, staramy się znaleźć funkcję, która przekształca określone wejście na określone wyjście. Aby zrobić to skutecznie, trzeba ustalić, które elementy danych wejściowych są przydatne do prognozowania wyników. W większości przypadków asocjacje są wystarczające do osiągnięcia tego celu.

Dlaczego przyczynowość? Zapytaj dzieci!

Czy Hume'owskiej teorii przyczynowości czegoś brakuje? Chociaż na to pytanie próbowało odpowiedzieć wielu innych filozofów, w tej książce skupię się na jednej szczególnie interesującej odpowiedzi, która pochodzi od... dzieci.

Interakcje ze światem

Alison Gopnik to amerykańska psycholog dziecięca zajmująca się badaniami nad rozwojem modeli świata przez dzieci. Alison Gopnik pomaga również informatykom w zrozumieniu sposobu budowania zdroworozsądkowego rozumienia przez dzieci świata zewnętrznego. Dzieci korzystają z uczenia asocjacyjnego w jeszcze większym stopniu niż dorośli, ale są także nienasyconymi eksperymentatorami.

Czy widziałeś kiedyś rodzica próbującego przekonać swoje dziecko, aby przestało rzucać zabawką? Niektórzy rodzice interpretują tego typu zachowanie jako *niegrzeczne*, *destrukcyjne* lub *agresywne*, ale dzieci często kierują się innymi motywacjami. Prowadzą systematyczne eksperymenty, które pozwalają im poznać prawa fizyki i zasady społecznych interakcji (Gopnik, 2009). Już 11-miesięczne niemowlęta wolą przeprowadzać eksperymenty z przedmiotami, które wykazują nieprzewidywalne właściwości (na przykład sprawiają wrażenie przechodzenia przez ścianę), niż z obiektami, które zachowują się przewidywalnie (Stahl i Feigenson, 2015). Ta preferencja pozwala im budować skuteczne modele świata.

Od dzieci możemy się nauczyć, że nie powinniśmy ograniczać się, jak sugerował Hume, do obserwacji świata. Możemy także wchodzić z nim w interakcje. W kontekście wnioskowania przyczynowego te interakcje są nazywane **interwencjami**. Więcej informacji na ich temat można znaleźć w rozdziale 2. Interwencje stanowią sedno tego, co wielu uważa za Świętego Graala metody naukowej: **randomizowanych badań kontrolowanych** (ang. *randomized controlled trial* — RCT).

Zakłócenia — związki, które nie są prawdziwe

Zdolność przeprowadzania eksperymentów poszerza paletę możliwości w stosunku do tego, o czym myślał Hume. Eksperymenty stwarzają wielkie możliwości! Chociaż nie pozwalają rozwiązać wszystkich problemów filozoficznych związanych ze zdobywaniem nowej wiedzy, mogą rozwiązać część z nich. Bardzo ważną cechą odpowiednio zaprojektowanego, randomizowanego eksperymentu jest możliwość unikania **zakłóceń** (ang. *confounding*). Dlaczego to jest ważne?

Zmienna zakłócająca wpływa na dwie lub więcej innych zmiennych i generuje pomiędzy nimi *fałszywe* powiązanie. Z czysto statystycznego punktu widzenia takie powiązania są nie do odróżnienia od tych wynikających z mechanizmu przyczynowego. Dlaczego to stwarza problemy? Przyjrzyjmy się przykładowi.

Wyobraź sobie, że pracujesz w instytucie badawczym i próbujesz zrozumieć przyczyny tonięcia ludzi. Twoja organizacja udostępniła Ci ogromną bazę danych zmiennych społeczno-ekonomicznych. Aby przewidzieć liczbę dziennych utonień w obszarze zainteresowania, zdecydowałeś się na skorzystanie z modelu regresji na obszernym zbiorze tych zmiennych. Po sprawdzeniu wyników okazało się, że największy uzyskany współczynnik dotyczy dziennej sprzedaży lodów. To bardzo interesujące! Lody zwykle zawierają duże ilości cukru, więc być może cukier wpływa na koncentrację lub kondycję fizyczną ludzi przebywających w wodzie.

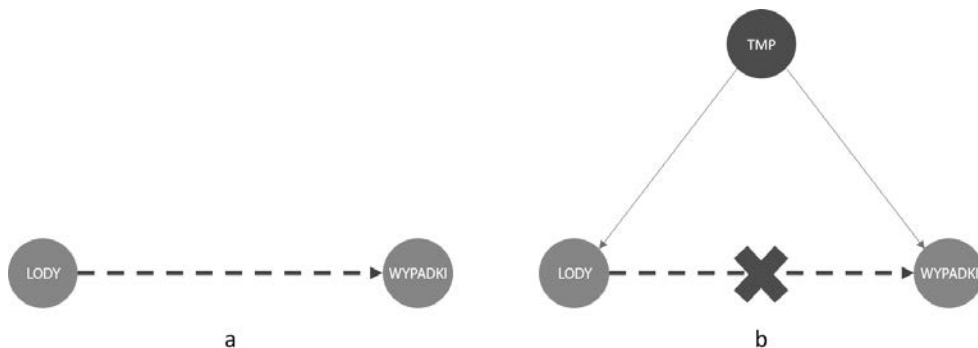
Ta hipoteza może mieć sens, ale zanim przejdziemy dalej, zadajmy kilka pytań. Co z innymi zmiennymi, których nie uwzględniliśmy w modelu? Czy w celu opisanie wszystkich istotnych aspektów problemu dodaliśmy do modelu wystarczającą liczbę predyktorów? A co, jeśli dodaliśmy ich za dużo? Czy dodanie do modelu tylko jednej zmiennej może całkowicie zmienić wynik?

Dodanie zbyt wielu predyktorów

Dodanie do modelu *zbyt wielu* predyktorów może być szkodliwe zarówno z punktu widzenia statystycznego, jak i przyczynowego. Więcej na ten temat dowiemy się w rozdziale 3.

Okazuje się, że jest to możliwe.

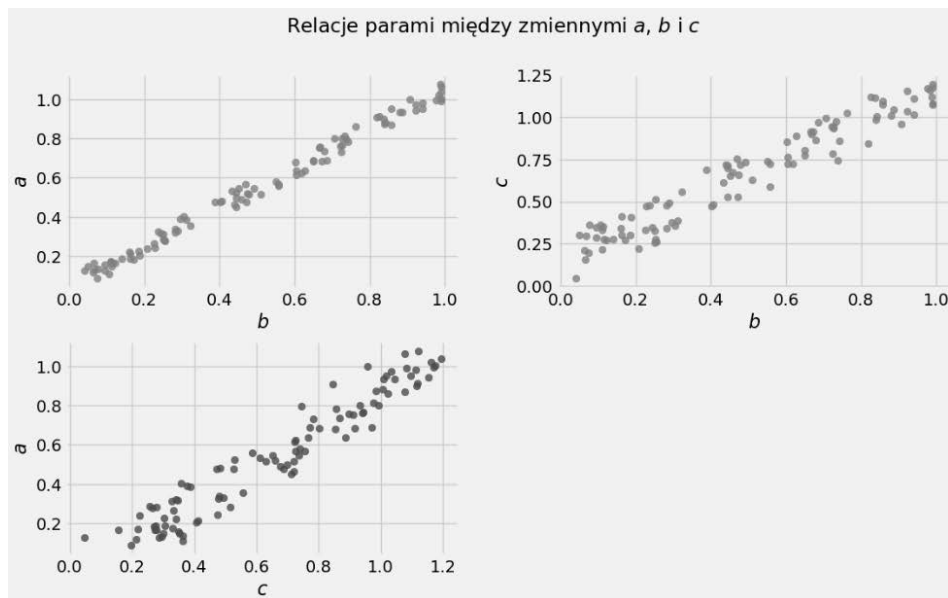
Pozwólcie, że przedstawię *czynnik zakłócający* — średnią dobową temperaturę. Wyższa dzienna temperatura sprawia, że ludzie chętniej kupują lody i chętniej pływają. Im więcej osób pływa, tym częściej dochodzi do wypadków. Spróbujmy zobrazować tę zależność (rysunek 1.1):



Rysunek 1.1. Graficzna reprezentacja modeli z dwiema (a) i trzema zmiennymi (b). Linie przerywane przedstawiają powiązanie, linie ciągłe przedstawiają związek przyczynowy. LODY = sprzedaż lodów, WYPADKI = liczba wypadków, TMP = temperatura

Na rysunku 1.1 widać, że dodanie do modelu średniej dziennej temperatury usuwa związek między sprzedażą lodów a dzienną liczbą utonięć. Dla niektórych czytelników może to być zaskakujące, dla innych nie. Więcej o mechanizmie tego efektu dowiemy się w rozdziale 3.

Zanim przejdę dalej, zwrócę szczególną uwagę na ważną rzecz: zakłócanie jest *pojęciem ściśle przyczynowym*. Co to znaczy? Chodzi o to, że nie da się wiele powiedzieć o zakłóceniach, posługując się językiem czysto statystycznym (zauważmy, że to oznacza, iż definicja Hume’a w takiej formie, w jakiej ją przedstawiłem, *nie jest w stanie tego uchwycić*). Aby to wyraźnie zobaczyć, spójrzmy na rysunek 1.2:

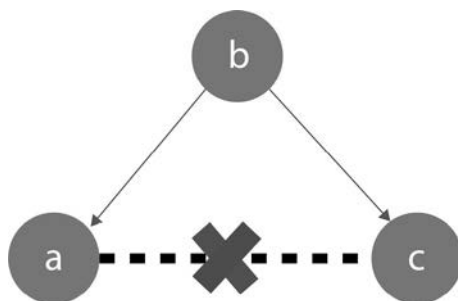


Rysunek 1.2. Wykresy punktowe relacji parami pomiędzy zmiennymi a, b i c. Kod umożliwiający odtworzenie poprzedniego wykresu można znaleźć w notatniku Chapter_01.ipynb (https://github.com/PacktPublishing/Causal-Inferenceand-Discovery-in-Python/blob/main/Chapter_01.ipynb)

Niebieskie punkty na rysunku 1.2 oznaczają związek przyczynowy, czerwone oznaczają związek pozorny, a zmienne a , b i c są powiązane w następujący sposób:

- b powoduje a i c ,
- a i c są przyczynowo niezależne.

Graficzną reprezentację tych zależności przedstawia rysunek 1.3:



Rysunek 1.3. Zależności między zmiennymi a , b i c

Czarna linia przerywana z czerwonym krzyżykiem oznacza, że pomiędzy zmiennymi a i c nie ma związku przyczynowego w żadnym kierunku

Ale przecież na rysunku 1.2 widać pewną zależność! Spróbujmy się jej przyjrzeć!

Przedstawione na rysunku 1.2 zależności niefałszywe (niebieskie) i fałszywe (czerwone) wyglądają dość podobnie, a ich współczynniki korelacji są podobnie duże. W praktyce w większości przypadków po prostu nie można ich rozróżnić na podstawie wyłącznie kryteriów statystycznych. Do takiego rozróżnienia potrzebna jest wiedza przyczynowa.

Asymetrie i odkrywanie przyczyn

W gruncie rzeczy, aby dowiedzieć się, który kierunek jest przyczynowy, w niektórych przypadkach można wykorzystać rozkład szumów lub funkcjonalne asymetrie. Informacje te można wykorzystać do odtworzenia struktury przyczynowej na podstawie danych obserwacyjnych, ale wymagają one również pewnych założeń dotyczących procesu generowania tych danych. Więcej informacji na ten temat podam w części III „Odkrywanie związków przyczynowych” (rozdział 13.).

Napisałem, że w danych naszego eksperymentu niektóre relacje były fałszywe. Następnie dodałem do modelu kolejną zmienną, co zmieniło generowany przez niego wynik. Mimo tego byłem w stanie wykonać użyteczne prognozy bez tej zmiennej. Jeśli to prawda, dlaczego miałoby mnie obchodzić, czy związek jest fałszywy, czy nie? Dlaczego miałoby mnie interesować, czy związek jest przyczynowy, czy nie?

Jak nie stracić pieniędzy... i ludzkich istnień

Dowiedziałeś się, że problemów wynikających z zakłóceń można uniknąć dzięki eksperymentom randomizowanym. Niestety, nie zawsze są one dostępne. Czasami wykonanie takich eksperymentów może być zbyt kosztowne, nieetyczne lub praktycznie niemożliwe (na przykład przeprowadzenie eksperymentu polegającego na migracji dużej grupy jakiejś populacji). W tym podrozdziale omówię kilka scenariuszy, w których pożądane jest wyciąganie wniosków przyczynowych mimo dysponowania wyłącznie danymi obserwacyjnymi. Te przykłady stworzą solidną podstawę do kolejnych rozdziałów.

Dylemat marketera

Wyobraź sobie, że jesteś marketerem znającym się na technologii i chcesz efektywnie alokować swój budżet na marketing bezpośredni. Jak podszedłbyś do tego zadania? Gdy alokujesz budżet na kampanię marketingu bezpośredniego, chciałbyś wiedzieć, jakiego zwrotu możesz oczekiwać, jeśli wydasz na daną osobę określoną kwotę. Mówiąc inaczej, chcesz oszacować wpływ swoich działań na wyniki niektórych klientów (Gutierrez, Gérardy, 2017). Być może, w celu rozwiązania problemu można by skorzystać z technik uczenia nadzorowanego? Aby odpowiedzieć na to pytanie, przyjrzyjmy się bliżej temu, co chcemy prognozować.

Interesuje nas reakcja konkretnej osoby na prezentowane jej treści. Spróbujmy zakodować to w następującym wzorze:

$$\tau_i = Y_i(1) - Y_i(0)$$

Oto opis poszczególnych elementów powyższego wzoru:

- τ_i to efekt eksperymentu na osobę i ;
- $Y_i(1)$ to wynik dla osoby i , kiedy została poddana eksperymentowi T (w tym przykładzie otrzymała od Ciebie treści marketingowe);
- $Y_i(0)$ to wynik dla tej samej osoby przy założeniu, że nie została poddana eksperymentowi T .

Zgodnie z powyższym wzorem od wyniku Y_i osoby i , gdy ta osoba nie została poddana eksperymentowi T , chcesz odjąć wynik uzyskany w przypadku, gdyby osobę tę poddano eksperymentowi T .

Interesujące jest to, że aby rozwiązać to równanie, trzeba wiedzieć, która odpowiedź osoby i została uzyskana po przeprowadzeniu eksperymentu, a która bez jego przeprowadzenia. W rzeczywistości nigdy nie można obserwować tej samej osoby w dwóch wzajemnie wykluczających się warunkach jednocześnie. Aby rozwiązać równanie z poprzedniego wzoru, potrzebne są kontryfakty (ang. *counterfactuals*).

Kontryfakty to szacunki wyników po zmianie wartości jednej lub większej liczby zmiennych, gdyby wszystkie inne pozostały niezmiennione. Ponieważ kontryfaktów nie można zaobserwować, prawdziwy skutek przyczynowy τ jest nieznany. Jest to jeden z powodów, dla których wyżej sformułowanego problemu nie da się rozwiązać z wykorzystaniem

klasycznych technik uczenia maszynowego. Rodzina technik przyczynowych zwykle stosowanych do takich problemów nazywa się **modelowaniem różnicowym** (ang. *uplift modelling*), które omówię dokładnie w rozdziałach 9. i 10.

Pobawmy się w doktora!

Spróbuję posłużyć się innym przykładem. Wyobraź sobie, że jesteś lekarzem. Jedna z Twoich pacjentek, Joanna, cierpi na rzadką chorobę *D*. Dodatkowo zdiagnozowano u niej wysokie ryzyko wystąpienia zakrzepów krwi. Przestudiowałeś informacje dotyczące dwóch najpopularniejszych leków na chorobę *D*. Obydwa leki mają praktycznie identyczną skuteczność na *D*, ale na podstawie diagnozy postawionej Joannie nie masz pewności, który lek będzie dla niej bezpieczniejszy. Przeglądasz dane badawcze przedstawione w tabeli 1.1:

Tabela 1.1. Dane dla leków A i B

Lek	A		B	
Zakrzepy	Tak	Nie	Tak	Nie
Razem	27	95	23	99
Procent	22%	78%	19%	81%

Liczbę w tabeli 1.1 przedstawiają liczbę pacjentów, u których zdiagnozowano chorobę *D*, poddanych leczeniu za pomocą specyfików *A* lub *B*. Wiersz 2 (**zakrzepy**) zawiera informacje o tym, czy u pacjentów wykryto zakrzepy krwi, czy nie. Warto zwrócić uwagę, że wyniki procentowe są zaokrąglone. Który lek byś wybrał na podstawie tych danych? Odpowiedź wydaje się dość oczywista. U 81% pacjentów, którzy otrzymali lek *B*, nie wystąpiły zakrzepy krwi. To samo dotyczyło jedynie 78% pacjentów, którzy otrzymali lek *A*. Ryzyko powstania zakrzepów krwi jest o około 3% niższe u pacjentów otrzymujących lek *B* w porównaniu z pacjentami otrzymującymi lek *A*.

To wygląda na dobry wniosek, ale jesteś sceptyczny. Wiesz, że zakrzepy krwi mogą być bardzo ryzykowne i chcesz dowiedzieć się więcej. Znalazłeś bardziej szczegółowe dane, które uwzględniają płeć pacjenta. Spójrzmy na tabelę 1.2:

Tabela 1.2. Dane dla leku A i leku B z dodanymi wynikami dla płci. K = kobieta, M = mężczyzna. Dla ułatwienia interpretacji dodano kodowanie kolorami, lepsze wyniki zaznaczono na zielono, a gorsze na pomarańczowo

Lek	A		B	
Zakrzepy	Tak	Nie	Tak	Nie
Kobiety	24	56	17	25
Mężczyźni	3	39	6	74
Razem	27	95	23	99
Procent	22%	78%	18%	82%
Procent (K)	30%	70%	40%	60%
Procent (M)	7%	93%	7,5%	92,5%

Stało się tu coś dziwnego. Uzyskane liczby są takie same jak poprzednio: lek *B* nadal jest preferowany dla wszystkich pacjentów. Wydaje się jednak, że lek *A* działa lepiej w przypadku i kobiet, i mężczyzn! Czy właśnie znaleźliśmy medycznego kota Schrödingera (https://en.wikipedia.org/wiki/Schr%C3%B6dinger%27s_cat), który odwraca działanie leku, gdy zaobserwuje płęć pacjenta?

Jeśli sądzisz, że mogłem pomylić się w obliczeniach, nie wierz mi na słowo, po prostu samodzielnie sprawdź dane. Można je znaleźć w pliku `data/ch_01_drug_data.csv` (https://github.com/PacktPublishing/Causal-Inference-and-Discovery-in-Python/blob/main/data/ch_01_drug_data.csv).

To, czego właśnie doświadczyliśmy, nazywa się **paradoksem Simpsona** (znanym również jako **efekt Yule-Simpsona**). Paradoks Simpsona występuje wtedy, gdy podział danych uwzględniający dodatkowe zmienne w ustawieniach regresji znacząco zmienia wynik analizy. W rzeczywistym świecie zwykle istnieje wiele sposobów partycjonowania danych. Mógłbyś zapytać: skąd więc mam wiedzieć, który podział jest właściwy?

Można by spróbować odpowiedzieć na to pytanie ściśle z punktu widzenia uczenia maszynowego: przeprowadzić selekcję cech z walidacją krzyżową i wybrać zmienne, które w znaczący sposób wpływają na wynik. To rozwiązanie w niektórych sytuacjach jest wystarczająco dobre. Sprawdzi się np. wtedy, gdy zależy nam jedynie na prognozowaniu (a nie na podejmowaniu decyzji) i gdy wiadomo, że dane produkcyjne będą niezależne i równomiernie rozłożone. Innymi słowy, dane produkcyjne muszą mieć rozkład praktycznie identyczny (lub przynajmniej wystarczająco podobny) do rozkładu danych szkoleniowych i walidacyjnych. Jeśli chcesz czegoś więcej, będziesz potrzebować jakiegoś (przyczynowego) modelu świata.

Asocjacje w realnym świecie

Niektórzy uważają, że relacje czysto asocjacyjne (skojarzeniowe) rzadko występują w prawdziwym świecie lub że zazwyczaj są słabe, więc nie mogą zbyt mocno wpływać na uzyskiwane wyniki. Aby przekonać się, jak zaskakująco silne i spójne mogą być fałszywe relacje w prawdziwym świecie, odwiedź stronę Tylera Vigena: <https://www.tylervigen.com/spurious-correlations>. Warto zauważyć, że zależności pomiędzy wieloma zmiennymi są czasami bardzo silne i utrzymują się przez długi czas! Osobiście podoba mi się przykład ze startami kosmicznymi i doktoratami z socjologii. Często wykorzystuję go podczas moich wykładów i prezentacji. Który jest Twoim ulubionym? Udostępnij go innym i oznacz mnie na LinkedIn, Twitterze (aby nawiązać kontakt, zapoznaj się z punktem „Pozostajmy w kontakcie” w rozdziale 15.). Możemy o tym porozmawiać!

Podsumowanie

„Niech przemówią dane” to chwytliwe i mocne hasło, ale jak można się było przekonać podczas lektury tego rozdziału, same dane nie zawsze wystarczą. Warto pamiętać, że w wielu przypadkach „dane nie mówią same za siebie” (Hernán, Robins, 2020) i że do odpowiedzi na niektóre pytania mogą być potrzebne dodatkowe informacje poza obserwacjami.

W tym rozdziale dowiedziałeś się, że gdy myślisz o przyczynowości, nie powinieneś, wbrew temu, co uważał David Hume, ograniczać się do obserwacji. Możesz też eksperymentować, zupełnie jak dzieci.

Niestety, eksperymenty nie zawsze są dostępne. W takiej sytuacji możesz spróbować wykorzystać dane obserwacyjne do wyciągnięcia wniosku przyczynowego, jednak same dane zwykle nie wystarczą do osiągnięcia tego celu. Potrzebujesz także modelu przyczynowego. W następnym rozdziale przedstawię *drabinę przyczynowości* — zaproponowaną przez Judeę Pearla zgrabną metaforę pozwalającą zrozumieć trzy poziomy przyczynowości.

Bibliografia

- Alexander J.E., Audesirk T.E., Audesirk G.J. *Classical Conditioning in the Pond Snail Lymnaea stagnalis*. „The American Biology Teacher”, 47(5), 1985, s. 295 – 298.
<https://doi.org/10.2307/4448054>.
- Archie L., *Hume’s Considered View on Causality*. [Preprint]: <http://philsci-archive.pitt.edu/id/eprint/2247>, 2005 (dostęp 23.04.2022).
- Falcon A. „Aristotle on Causality”, *The Stanford Encyclopedia of Philosophy* (wydanie wiosna 2022), pod red. Edwarda N. Zalta,
<https://plato.stanford.edu/archives/spr2022/entries/aristotle-causality/>. Dostęp 23.04.2022.
- Gopnik A., *The philosophical baby: What children’s minds tell us about truth, love, and the meaning of life*, Nowy Jork, Farrar, Straus and Giroux, 2009.
- Gutierrez P., Gérardy, J., *Causal Inference and Uplift Modelling: A Review of the Literature*, materiały z 3rd International Conference on Predictive Applications and APIs, opublikowane w „Proceedings of Machine Learning Research”, 67, 2017, s. 1 – 13.
- Hernán M.A., Robins J.M., *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020.
- Hume D., Millican P.F., *An enquiry concerning human understanding*. Oxford: Oxford University Press, 2007.
- Kahneman D., *Thinking, Fast and Slow*, Farrar, Straus and Giroux, 2011.
- Lorkowski C.M., <https://iep.utm.edu/hume-causation/>. Dostęp 23.04.2022.
- Stahl A.E., Feigenson L., *Cognitive development. Observing the unexpected enhances infants’ learning and exploration*. „Science”, 348(6230), 2015, s. 91 – 94.
<https://doi.org/10.1126/science.aaa3799>.

Skorowidz |

A

ABCI, Active Bayesian Casual Inference, 404
abdukcja, 54
addytywne modele szumów, ANM, 168
agenty interwencji, 419
algorytm
 CORTH, 403
 DECI, 390
 DR-Learner, 270
 FCI, 398
 GES, 360
 GOLEM, 374
 liniowy DML, 270
 NOTEARS, 372
 PC, 348, 355, 357
 dla danych kategoryalnych, 358
 wyniki, 379
 S-Learner, 270
 T-Learner, 270
 word2vec, 301
 X-Learner, 270
algorytmy identyfikacji, 412
analiza
 czułości, sensitivity analysis, 176
 przyczynowa, 414
 przyczynowa w Pythonie, 147
 regresji, 75
 składowych niezależnych, ICA, 367
ANM, Additive Noise Model, 168, 361, 362
API GCM, 166
asocjacje, 31, 37, 42
 nieliniowe, 63
ATC, average treatment effect on the control, 191, 193
ATE, average treatment effect, 191, 193

B

badania kontrolowane randomizowane, RCT, 41, 312, 334
bazowe mechanizmy uczenia, 211
biblioteka
 CATENets, 292, 293
 DoWhy, 149, 249
 EconML, 150, 249
 gCastle, 344
 statsmodels, 63
błąd
 nadmiernego dopasowania, 247
 regularyzacji, 248

C

CATE, conditional average treatment effects, 191, 207, 208
CATENets, 292
CausalBert, 306
 architektura, 307
 implementacja, 308
 używanie modelu, 308
CausalForestDML, 282
ChatGPT, 301
 rozumowanie kontrfaktyczne, 302
CV, cross-validation, 157
cykliczne modele SCM, 92

D

DAG, directed acyclic graphs, 84
dane
 eksperymentalne, 263
 gaussowskie i niegaussowskie, 367
 kategoryalne, 358
 syntetyczne, 345
 z zakłóceniami, 403

DECI

- elementy frameworka, 385
- implementacja, 386
- konfiguracja, 389
- moduły modelu, 392
- przygotowanie danych, 390
- szkolenie modelu, 395
- wiedza ekspercka, 390
- wyniki, 396

dodatniość, 177

dopasowywanie, matching, 192

- dokładne, 203
- estymatory, 194
- implementacja, 196
- modeli regresji, 113
- nadmierne, 247
- krzyżowe, cross-fitting, 238
- prawie dokładne, 203
- przybliżone, 203
- wielowymiarowe, 201
- współczynników skłonności, PSM, 203

dostrajanie hiperparametrów, 252, 254

DR, double robust, 233

drabina przyczynowości, ladder of causation, 39

DR-Learner, 238, 282

- opcje, 242

- wyniki, 240, 241

drzewa przyczynowe, 260

dylemat marketera, 35

E

efekt Yule-Simpsona, 37

eksperyment z wielkością próby, 174

eksperymenty kontrolowane, 333

ENCO, Efficient Neural Causal Discovery, 404

estymandy, estimands, 124

- identyfikowanie, 154

- równoważne, 127

estymator, 124

- CATE, 280

- dopasowywania, 194

- DR, 237, 238

- kontroli syntetycznej, 314, 317

- TMLE, 242

- ukierunkowany prawdopodobieństwa, 242

F

falsyfikacja hipotez, 413

FCI, 398

- implementacja, 399

- wiedza ekspercka, 402

framework

- DECI, 385

- potencjalnych wyników, 180

- TCDF, 405

funkcja

- uciążliwości, 249

- wpływu, influence function, 244

- wskaźnikowa, 45

fuzja danych przyczynowych, 418

G

gCastle, 344

- algorytm GES, 360

- algorytm PC, 348

- dane syntetyczne, 345

- wiedza ekspercka, 378

- wizualizacja modelu, 349

- wskaźniki oceny modelu, 350

GCM, graphical causal models, 87

generowanie hipotetycznego grafu, 412

GES, 359

globalna własność Markowa, 100

GML, graph modeling language, 85, 146

GOLEM, 374

gradient, 372

graf

- CPDAG, 109

- modelu SCM, 43

- przyczynowy, 212

- z czynnikiem zakłócającym, 400

grafy

- acykliczne, 81

- CPDAG, 80

- cykliczne, 81

- hipotetyczne, 412

- język GML, 85

- macierze przyległości, 83

- modele GCM, 87

- nieskierowane, 80

- niespójne, 82

- nieważone, 82

- niezależność, 98

- okaleczanie, 89, 182

- przyczynowe, 91, 172

- skierowane, 80

- skierowane acykliczne, DAG, 8, 84, 88, 90, 234

- a przyczynowość, 89

- ograniczenia, 90

- spójne, 82

struktura
kolidera, 106
łańcucha, 103
rozwidlenia, 104
ważone, 82

H

hipoteza
falsyfikowalna, 332
zerowa, 65
HSIC, Hilbert-Schmidt independence
criterion, 51
HTE, heterogeneous treatment effects, 207

I

ICA, independent component analysis, 367
identyfikowalność, identifiability, 171, 172, 412
implementacja
algorytmu FCI, 399
DECI, 386
DML, 249
dopasowywania, 196
estymatora TMLE, 243
IPW, 205
modelu T-Learner, 220
obliczeń kontrfaktów, 56
X-Learner, 226
indywidualny efekt interwencji, ITE, 207
interakcja, 63
interfejs API GCM, 166
interwencje, 40, 47
IPW, inverse probability weighing, 204,
206, 236
implementacja, 205
istotność statystyczna, 65
ITE, individualized treatment effects, 207
IV, instrumental variables, 142

J

jednostka
ELU, 296
kontrolna syntetyczna, 316
SELU, 296
język
DOT, 86
GML, 85, 146
naturalny, 300

K

klasy równoważności Markowa, MEC, 108,
109, 158
koliderzy, 106, 109
tworzenie zbioru danych, 112
kontrfakty, 35, 41, 52
deterministyczne, 55
obliczanie, 54
probabilistyczne, 55
kontrola
statystyczna, 69
syntetyczna, 314
implementacja, 317
korelacja, 50
kryterium
back-door, 126, 127
front-door, 134
niezależności Hilberta-Schmidta, HSIC,
51, 365
tylnych drzwi, 57
kwantyfikacja związków przyczynowych, 191

L

lasy przyczynowe, 260, 261
wyniki, 262
learnery bazowe drugiej fazy, 224
LinearDML, 282
LiNGAM, 361, 366
wyniki, 369, 370
LLM, large language models, 301

Ł

łańcuchy, 103, 109
tworzenie zbioru danych, 110
zdarzeń, 102

M

macierz
modelu interwencji, 209
przyległości, 83
sąsiedztwa, 83
maksymalny współczynnik informacji, MIC, 51
metauczenie, 211

metody

- oparte na gradiencie, 343
- oparte na ograniczeniach, 343, 353
- oparte na punktacji, 343
- podwójnie niezawodne, DR, 233
- porównanie wydajności, 375
- RCT, 312

MIC, maximal information coefficient, 51

minimalizm, minimality, 341, 342

model

- ANM, 362
- CausalBert, 306
- DirectLINGAM, 376
- DR-Learner, 242
- GCM, 87
- GES, 376
- GOLEM, 376
- LiNGAM, 366
- LLM
 - a przyczynowość, 301
- NotearsNonlinear, 376
- PC, 376
- SCM, 44, 70, 73, 92, 134, 181, 186
 - graficzna reprezentacja, 75
- szumów addytywnych CCANM, 403

modele

- interwencji, treatment models, 234
- porównawcze
 - wyniki, 297
- przyczynowe, 170
- regresji
 - dopasowanie, 113
- regresyjne, 73
- strukturalne, 73
- uczenia głębokiego
 - wyniki, 297
- uplift, 263, 274
- różnicowe, uplift modelling, 36
- wyników, outcome models, 234

modelowanie

- użycie techniki S-Learner, 211
- heterogenicznych efektów interwencji, 263, 289
- problemu
 - tworzenie grafu, 151
 - tworzenie obiektu CausalModel, 153

modułowość, 181

modyfikacja, 54

N

- niezależność, 353
 - kryteria, 365
 - przyczynowa, 71
 - w grafie, 98
 - warunkowa, 97, 120
 - zmiennych, 96
- NLP, natural language processing, 288
 - scenariusze, 303
- norma Frobeniusa, 385
- NOTEARS, 372

O

- obliczanie kontrfaktów, 54
- odkrywanie związków przyczynowych, 91, 98, 100, 329
 - ABCI, 404
 - DECI, 398
 - ENCO, 404
 - funkcyjne, 343, 361
 - na podstawie punktacji, 359
 - oparte na gradientach, 372
 - oparte na ograniczeniach, 100, 353
 - osobiste doświadczenia, 335
 - pakiet gCastle, 344
 - problemy, 405
 - sposzczenia naukowe, 331
 - uczenie głębokie, 383
 - ukryte zakłócenia, 398
 - założenia, 341
 - zastosowania, 405
 - źródła wiedzy przyczynowej, 329
- odległość
 - euklidesowa, 192
 - Mahalanobisa, 192
 - Minkowskiego, 192
- odwrotne ważenie prawdopodobieństwa, IPW, 204, 236
- odwrócony model regresji, 67
- ograniczenia, 353
- operator
 - do, 129
 - morsa, 44
 - przypisania, 44
 - wartości oczekiwanej, 62
- optymalizator Lagrangiana, 386
- ortogonalizacja, 238
- oszacowania, estimates, 124
 - obliczanie, 156

P

pakiet gCastle, 344
 paradoks Simpsona, 37
 paradygmat RLFH, 301
 parametr
 fluktuacji, 244
 uciążliwości, nuisance parameter, 249
 polisemia, 301
 porównanie wydajności metod, 375
 prawdopodobieństwo warunkowe, 42
 prawo Twymana, 255
 problem
 cocktail party, 367
 Walda, 186
 proces wnioskowania przyczynowego, 146
 prognoza, 54
 przedziały ufności, 279
 przetwarzanie języka naturalnego, NLP, 288
 przyczynowe uczenie maszynowe, 416
 przyczynowość, causality, 30, 50, 88, 312, 421
 a modele LLM, 301
 a uczenie nienadzorowane, 58
 a uczenie półnadzorowane, 58
 a uczenie ze wzmocnieniem, 57
 w biznesie, 414
 PSM, propensity score matching, 203
 punktacja, 359
 Python
 ekosystem analizy przyczynowej, 147

R

rachunek do, 140
 zasady, 141
 RCT, Randomized Controlled Trials, 41, 312, 334
 regresja, 67, 109
 a skutki przyczynowe, 76
 liniowa, 62, 74
 interpretacja geometryczna, 66
 wielozmienna, 113
 ważonych najmniejszych kwadratów, WLS, 205
 wielozmienna, multiple regression, 6, 692
 regularyzacja, 248
 reguła
 front-door, 133
 potęgi, 77
 wpływu przyczynowego, 126
 relacje
 parami, 33
 pozorne, 184

RL, reinforcement learning, 57
 rozwidlenia, 104, 109
 tworzenie zestawu danych, 112

S

SCM, structural causal model, 42, 70, 342
 separacja d, d-separation, 120
 skojarzenia, 40
 S-Learner, 207, 211, 282
 słabe punkty, 218
 szkolenie modelu, 213, 217
 wyniki, 217
 SNet, 291
 architektura, 291
 spójność, 183
 sprytna zmienna kowariantowa, clever covariate, 244
 strategia „dziel i zwyciężaj”, 132
 stronniczość
 ocalałych, survivorship bias, 185
 wyboru, selection bias, 185, 187
 strukturalne modele przyczynowe, SCM, 42, 70, 342
 struktury v, 106
 SUTVA, 183
 symulacje, 335
 szeregi czasowe, 312
 szkielet grafu, 109

T

TARNet, 290
 architektura, 290
 TCDF, Temporal Causal Discovery Framework, 405
 tekst
 jako czynnik zakłócający, 305
 jako interwencja, 304
 jako wynik, 304
 teoria znaczenia, 299
 teorie Poppera, 158
 testy
 dwóch próbek klasyfikatora, 268
 obalające, refutation tests, 156, 158
 porównawcze CRASS, 303
 T-Learner, 218, 282
 implementacja modelu, 220
 wyniki, 221
 TMLE
 implementacja, 243
 transformatory, 299

twierdzenie

- Frischa-Waugh-Lovella, 248
- o współczynnikach skłonności, 235

U

uczenie

- asocjacyjne, 31
- głębokie, 289, 383
- maszynowe podwójne, DML, 173, 246, 256
 - a estymator DR-Learner, 258
 - implementacja, 249
 - wyniki, 251–254
- się przez naśladowanie, 420
- się struktury przyczynowej, 337, 419
- ze wzmocnieniem, RL, 57

uplift, 263

- według decyla, 274

W

walidacja

- krzyżowa, CV, 157
- modeli przyczynowych, 157

wartość

- oczekiwana, 62
- p, 65

warunek przyczynowości Markowa, 98, 99

warunkowanie, conditioning, 31

wiedza

- dziedzinowa, 337
- eksperycka, 378, 390, 402, 411
- przyczynowa
 - osobiste doświadczenia, 335
 - źródła, 329

wielkość próby, 173

wierność, faithfulness, 341

WLS, weighted least squares, 205

wnioskowanie

- kontrafaktyczne, 41
- przyczynowe, 54, 97, 98, 191
 - modelowanie problemu, 151
 - obalenie oszacowania, 165
 - testy obalające, 156
 - wyznaczanie estymand, 154, 163
 - wyznaczanie oszacowań, 156, 163
 - zakodowanie założeń, 161

wskaźnik

- CATE, 211, 289
- MAPE, 227, 239
- oczekiwanej odpowiedzi, 277
- rzadkości, sparsity score, 385

wskaźniki przyczynowe, 418

współczynnik

- AUUC, 274
- korelacji, 315
- MAPE, 216
- Qini, 274
- skłonności, propensity scores, 201, 202, 204, 234

wybór modelu, 281

wydajność, 274

wykorzystanie projektów przyczynowych, 410

wykres

- linii regresji, 66, 138
- modelu odwróconego, 68
- punktowy, 111, 112
 - danych z próbkowaniem selektywnym, 51
 - nieliniowego zbioru danych, 363
 - relacji parami, 33
 - rozzrzu tu danych nieliniowych, 364
 - wartości reszt, 365
- upliftu według decyli, 275, 276

wymienność, 179

- a zakłócenia, 180

wynik kontrafaktyczny, 193

wystarczalność przyczynowa, causal sufficiency, 341

X

X-Learner, 222, 225, 282

- implementacja, 226
- rekonstrukcja modelu, 223
- wyniki, 228, 229

Z

zakłócenia, confounding, 32, 184

- ukryte, 398, 403

założenie

- braku ukrytych zakłóceń, 101
- dodatniości, 177
- minimalności przyczynowej, 101, 342
- modułowości, 181
- spójności, 183
- SUTVA, 181
- wierności, faithfulness assumption, 100, 341
- wymienności, 179

zasada wspólnej przyczyny Reichenbacha, 314

zmienne

- egzogeniczne, 135
- heterogeniczny wpływ na wynik, 77
- instrumentalne, IV, 142, 175
- kontrolne, 70
- niezależne, 96
- szumów, 43
- zależne, 67
- zakłócające, 32

związki

- pozorne, spurious relationships, 115
- przyczynowe, 91, 92

PROGRAM PARTNERSKI

— GRUPY HELION —



1. ZAREJESTRUJ SIĘ
2. PREZENTUJ KSIĄŻKI
3. ZBIERAJ PROWIZJĘ

Zmień swoją stronę WWW w działający bankomat!

Dowiedz się więcej i dołącz już dzisiaj!

<http://program-partnerski.helion.pl>

GRUPA
Helion 

Przyczyna i skutek, nic więcej. Pomyłki jako takie nie istnieją...

José Antonio Cotrina, hiszpański pisarz science fiction

W uczeniu maszynowym odkrywanie związków przyczynowych daje możliwości, jakich nie można uzyskać tradycyjnymi technikami statystycznymi. Najnowsze trendy w programowaniu pokazują, że przyczynowość staje się kluczowym zagadnieniem dla generatywnej sztucznej inteligencji. Niezbędna okazuje się więc znajomość grafów przyczynowych i zapytań konfrontacyjnych.

Dzięki tej książce łatwo przyswoisz teoretyczne podstawy i zaczniesz je płynnie wdrażać w rzeczywistych scenariuszach. Dowiesz się, w jaki sposób myślenie przyczynowe ułatwia rozwiązywanie problemów, i poznasz pojęcia Pearla, takie jak strukturalny model przyczynowy, interwencje, kontryfakty itp. Każde zagadnienie zostało dokładnie wyjaśnione i opatrzone zbiorem praktycznych ćwiczeń z kodem w Pythonie. Nauczysz się także implementować poszczególne modele i zrozumiesz, czym się kierować przy wyborze technik i algorytmów do rozwiązywania konkretnych scenariuszy przyczynowych. To przewodnik, który docenią szczególnie inżynierowie uczenia maszynowego i analitycy danych.

W książce:

- wnioskowanie związków przyczynowych
- budowa i działanie strukturalnych modeli przyczynowych
- czteroetapowy proces wnioskowania związków przyczynowych w Pythonie
- techniki modelowania efektu interwencji
- nowoczesne metody odkrywania związków przyczynowych za pomocą Pythona
- korzystanie z wnioskowania związków przyczynowych

ALEKSANDER MOLAK jest niezależnym badaczem i konsultantem w dziedzinie uczenia maszynowego. Współpracował z licznymi firmami w Europie, USA i Izraelu, gdzie uczestniczył w tworzeniu wielkoskalowych systemów uczenia maszynowego. Jest też współzałożycielem firmy Lespire.io, dostawcy szkoleń z zakresu sztucznej inteligencji dla zespołów korporacyjnych.

Helion	KOD KORZYŚCI Sięgnij po więcej!	
 helion.pl	ISBN 978-83-289-0832-1	
 HELION SA ul. Kościuszki 1c 44-100 Gliwice tel.: 32 230 98 63 helion@helion.pl		
Cena: 109,00 zł		