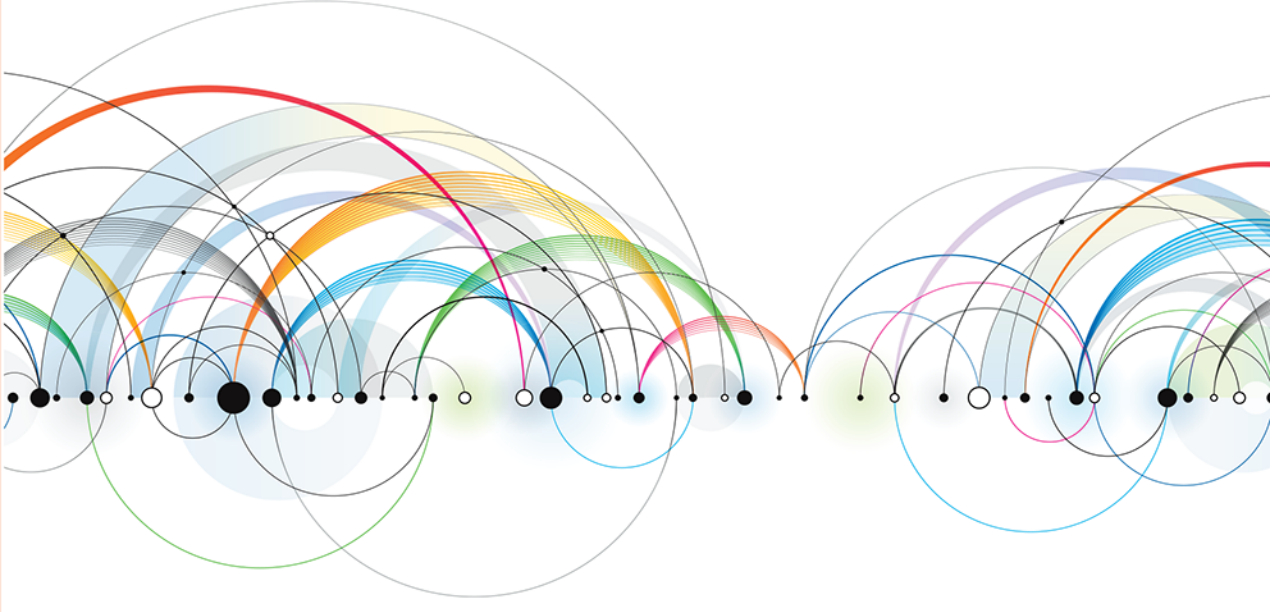
Foster Provost, Tom Fawcett

Analiza danych *w* biznesie

Sztuka podejmowania skutecznych decyzji



onepress

Helion

Tytuł oryginału: Data Science for Business

Tłumaczenie: Leszek Sielicki

ISBN: 978-83-8322-580-7

Drogi Czytelniku!
Jeżeli chcesz ocenić tę książkę, zajrzyj pod adres
https://helion.pl/user/opinie/andavv
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.

Spis treści

ROZDZIAŁ 6.

Podobieństwo, sąsiedzi i klastry

Podstawowe pojęcia: *Obliczanie podobieństwa obiektów opisanych przez dane; Wykorzystywanie podobieństwa do celów predykcji; Klastrowanie jako segmentacja oparta na podobieństwie.*

Przykładowe techniki: *Poszukiwanie podobnych jednostek; Metody najbliższych sąsiadów; Metody klastrowania; Miary odległości do obliczania podobieństwa.*

Podobieństwo leży u podstaw wielu metod nauki o danych i rozwiązań problemów biznesowych. Jeżeli dwa obiekty (osoby, firmy, produkty) są pod jakimiś względami podobne, to często dzielą także inne cechy. Procedury eksploracji danych bywają często oparte na grupowaniu obiektów według podobieństwa lub na poszukiwaniu „właściwego" rodzaju podobieństwa. W sposób dorozumiany zapoznaliśmy się z tym w poprzednich rozdziałach, w których procedury modelowania tworzyły granice grupowania wystąpień mających podobne wartości zmiennych docelowych. W tym rozdziale przyjrzymy się podobieństwu bezpośrednio i pokażemy, w jaki sposób odnosi się ono do wielu różnych zadań. Zawarliśmy w nim także podrozdziały dotyczące pewnych szczegółów technicznych, umożliwiające lepsze zrozumienie koncepcji podobieństwa czytelnikom mającym przygotowanie matematyczne; te podrozdziały można pominąć.

Wnioskowanie na podstawie podobnych przykładów jest elementem wielu różnorodnych zadań biznesowych.

- Moglibyśmy chcieć *wyszukać* podobne obiekty bezpośrednio. Na przykład firmie IBM mogłoby zależeć na znalezieniu przedsiębiorstw, które są podobne do jej najlepszych klientów biznesowych, aby personel działu sprzedaży mógł je potraktować jako potencjalnych klientów. Hewlett-Packard utrzymuje dla swoich klientów wiele serwerów wysokiej wydajności; ich utrzymywanie jest wspomagane za pomocą narzędzia, które przy danej konfiguracji serwera pobiera informacje o innych podobnie skonfigurowanych serwerach. Reklamodawcom często zależy na kierowaniu reklam internetowych do podmiotów, które są podobne do ich aktualnych dobrych klientów.
- Podobieństwo może być wykorzystywane do przeprowadzania *klasyfikacji* i *regresji*. Ponieważ wiemy już sporo na temat klasyfikacji, zilustrujemy wykorzystanie podobieństwa za pomocą przykładu klasyfikacyjnego poniżej.
- Może zależeć nam na grupowaniu podobnych elementów w *klastrach*, na przykład w celu sprawdzenia, czy nasza baza klientów zawiera grupy klientów podobnych i co te grupy mają ze sobą wspólnego. Wcześniej omawialiśmy nadzorowaną segmentację; tutaj mamy do czynienia z segmentacją nienadzorowaną. Po omówieniu podobieństwa do celów klasyfikacji omówimy jego zastosowanie do celów klastrowania.

- Współcześni detaliści, tacy jak Amazon i Netflix, wykorzystują podobieństwo do tworzenia *rekomendacji* podobnych produktów lub pochodzących od podobnych osób. Ilekroć widzimy stwierdzenia typu: „Osoby, które lubią X, lubią także Y" albo „Klienci o podobnej do Twojej historii przeglądania poszukiwali również...", mamy do czynienia z wykorzystywaniem podobieństwa. W rozdziale 12. opiszemy, w jaki sposób klient może być podobny do filmu, jeśli zarówno on, jak i film opisywani są przez takie same „wymiary związane z gustem". W takim przypadku, aby dokonać rekomendacji, możemy znaleźć filmy, które są najbardziej podobne do klienta (i których klient jeszcze nie obejrzał).
- Wnioskowanie z podobnych przypadków oczywiście wykracza poza aplikacje biznesowe; jest naturalne w takich dziedzinach jak medycyna czy prawo. Lekarz może rozpatrywać nowy trudny przypadek, przypominając sobie podobny (znany mu z własnego doświadczenia lub opisany w piśmie fachowym) i jego diagnozę. Prawnicy często prowadzą sprawy, powołując się na precedensy prawne, czyli przypadki historyczne, w których zapadły orzeczenia i które umieszczono w zbiorach orzecznictwa. W ramach obszaru sztucznej inteligencji od dawna tworzone są systemy wspomagające lekarzy i prawników w tego rodzaju wnioskowaniu opartym na konkretnych przypadkach. Orzeczenia na podstawie podobieństwa są tutaj elementem kluczowym.

W celu bardziej szczegółowego omówienia tych zastosowań musimy poświęcić chwilę na uporządkowanie wiedzy dotyczącej podobieństwa i ściśle z nim powiązanej odległości.

Podobieństwo i odległość

Gdy obiekt może zostać przedstawiony w formie danych, to możemy zacząć bardziej szczegółowo mówić o podobieństwie obiektów lub o odległości między nimi. Zastanówmy się na przykład nad reprezentacją danych, którą posługiwaliśmy się w naszej książce do tej pory. W tym ujęciu każdy obiekt reprezentowany jest przez wektor cech. Im bliżej siebie znajdują się dwa obiekty w przestrzeni definiowanej przez cechy, tym bardziej są podobne.

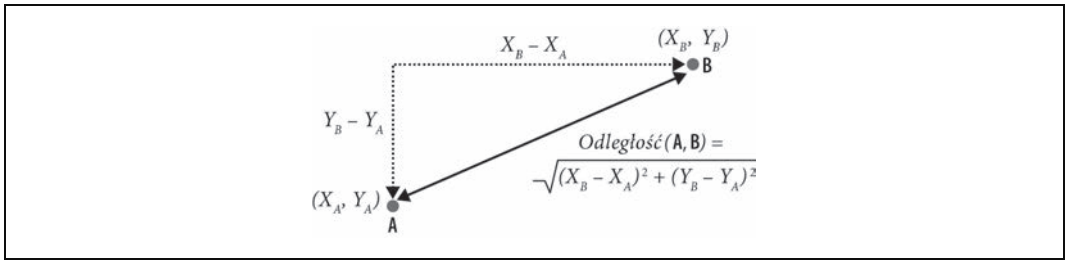
Przypomnijmy, że kiedy budujemy i stosujemy modele predykcyjne, to naszym celem jest określenie wartości cechy docelowej. Robiąc to, wykorzystywaliśmy już dorozumiane podobieństwo obiektów. W rozdziale 3., w podrozdziale „Wizualizacja segmentacji", omawialiśmy geometryczną interpretację pewnych modeli klasyfikacyjnych, a w rozdziale 4., w podrozdziale „Klasyfikacja za pomocą funkcji matematycznych", przyglądaliśmy się, w jaki sposób różne typy modeli dzielą przestrzeń wystąpień na obszary na podstawie bliskości wystąpień o podobnych etykietach klasy. Wiele metod wykorzystywanych w nauce o danych można postrzegać w tym świetle: jako metody organizowania przestrzeni wystąpień zawierającej dane (odwzorowania istotnych obiektów) w taki sposób, że wystąpienia znajdujące się blisko siebie są traktowane podobnie w określonym celu. Zarówno drzewa klasyfikacyjne, jak i klasyfikatory liniowe określają granice między obszarami o różniących się klasyfikacjach. Ich wspólną cechą jest uznawanie wystąpień znajdujących się na tym samym obszarze za takie, które powinny być podobne; obie metody odróżnia sposób odwzorowywania i odkrywania obszarów.

Dlaczego więc nie traktować podobieństwa i odległości między obiektami wprost? Aby to zrobić, musimy dysponować podstawową metodą pomiaru podobieństwa lub odległości. Co to znaczy, że dwie firmy lub dwaj klienci są podobni? Przyjrzyjmy się temu zagadnieniu uważniej. Rozważmy dwa wystąpienia z naszej uproszczonej domeny zastosowań kredytowych:

Atrybut	osoba A	osoba B
Wiek	23	40
Liczba lat pod aktualnym adresem	2	10
Status rezydenta (1 = Właściciel, 2 = Najemca, 3 = Inny)	2	1

Te elementy danych mają wiele atrybutów i nie istnieje jedna, najlepsza metoda umożliwiająca zredukowanie ich do jednego pomiaru podobieństwa czy odległości. Istnieje wiele różnych sposobów pomiaru podobieństwa lub odległości pomiędzy Osobą A i Osobą B. Dobrym początkiem będą miary odległości znane z podstaw geometrii.

Przypomnijmy z naszych wcześniejszych rozważań dotyczących interpretacji geometrycznej, że jeśli dysponujemy dwiema cechami (liczbowymi), to każdy obiekt jest punktem w przestrzeni dwuwymiarowej. Na rysunku 6.1 widać dwa elementy danych, A i B, które znajdują się na płaszczyźnie dwuwymiarowej. Obiekt A ma współrzędne (x_A, y_A), a obiekt B (x_B, y_B). Ryzykując, że zbyt często się powtarzamy, przypomnijmy jeszcze, że te współrzędne to po prostu wartości dwóch cech obiektów. Wykorzystując nasze dwa obiekty, możemy, jak pokazano na rysunku, narysować trójkąt prostokątny, którego podstawą będzie różnica odciętych x: $x_A - x_B$, a wysokością różnica rzędnych y: $y_A - y_B$. Z twierdzenia Pitagorasa wiemy, że odległość między A i B jest określona długością przeciwprostokątnej i jest równa pierwiastkowi kwadratowemu z sumy kwadratów długości dwóch pozostałych boków trójkąta, czyli w naszym przypadku $\sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}$. Możemy zasadniczo obliczyć odległość całkowitą poprzez obliczenie odległości poszczególnych wymiarów — w naszym ujęciu konkretnych cech. Wielkość tę nazywamy **odległością euklidesową**[1] pomiędzy dwoma punktami i jest to prawdopodobnie najpowszechniej stosowana geometryczna miara odległości.



Rysunek 6.1. Odległość euklidesowa

Odległość euklidesowa nie ogranicza się do dwóch wymiarów. Gdyby A i B były obiektami opisanymi za pomocą trzech cech, to mogłyby zostać odwzorowane przez punkty w przestrzeni trójwymiarowej, a ich umiejscowienie byłoby wtedy określone jako (x_A, y_A, z_A) i (x_B, y_B, z_B). Odległość między A i B zawierałaby wtedy warunek $(z_A - z_B)^2$. Możemy dodać dowolną liczbę cech, a każda z nich będzie nowym wymiarem. Gdy obiekt będzie opisany przez n cech, będziemy dysponowali n wymiarami (d_1, d_2, ..., d_n), a ogólną postacią równania odległości euklidesowej w n wymiarach będzie sformułowanie przedstawione w równaniu 6.1:

Równanie 6.1. Ogólna odległość euklidesowa

$$\sqrt{(d_{1,A} - d_{1,B})^2 + (d_{2,A} - d_{2,B})^2 + \ldots + (d_{n,A} - d_{n,B})^2}$$

[1] Od imienia Euklidesa, greckiego matematyka żyjącego w IV w. p.n.e., znanego jako ojciec geometrii.

Dysponujemy więc teraz miarą odległości pomiędzy dwoma dowolnymi obiektami opisanymi przez wektory cech — prostym wzorem opartym na odległościach poszczególnych cech obiektów. Dla osób A i B opisanych powyżej odległość euklidesowa wynosi:

$$d(A,B) = \sqrt{(23-40)^2 + (2-10)^2 + (2-1)^2}$$
$$\approx 18{,}8$$

Odległość między tymi przykładami wynosi więc około 19. Ta odległość to tylko liczba — nie posiada jednostek ani jednoznacznej interpretacji. Naprawdę użyteczna jest tylko do porównywania podobieństwa jednej pary wystąpień do innej pary. Okazuje się, że porównywanie podobieństw jest niezwykle użyteczne.

Wnioskowanie metodą najbliższych sąsiadów

Skoro dysponujemy sposobem pomiaru odległości, to możemy wykorzystać go do wielu różnych zadań z zakresu analizy danych. Przywołując przykłady z początku tego rozdziału, moglibyśmy wykorzystać tę miarę do znalezienia firm najbardziej podobnych do naszych najlepszych klientów korporacyjnych czy klientów internetowych najbardziej podobnych do naszych najlepszych klientów detalicznych. Gdy uda nam się znaleźć tych podobnych klientów, będziemy mogli podjąć najbardziej odpowiednie w kontekście biznesowym działania. Tak postępuje IBM w odniesieniu do klientów korporacyjnych, wspomagając kierunkowanie działań swojego personelu handlowego. Reklamodawcy internetowi robią tak, targetując reklamy. Takie najbardziej podobne wystąpienia nazywane są **najbliższymi sąsiadami**.

Przykład: analityka whisky

Porozmawiajmy o nowym przykładzie. Jeden z nas (Foster) lubi szkocką whisky single malt. Jeśli próbowałeś więcej niż jednej czy dwu, to zdajesz sobie sprawę, że istnieją setki różnorodnych słodowych whisky. Kiedy Foster znajduje taką, która naprawdę mu smakuje, to chce znaleźć podobne — dlatego że lubi eksplorować „przestrzeń" single maltów, ale także dlatego, że w sklepach z alkoholami i restauracjach zwykle nie można dostać wszystkich gatunków. Foster chce mieć możliwość wyboru whisky, która naprawdę będzie mu smakowała. Pewnego wieczoru podczas kolacji polecono mu na przykład whisky „Bunnahabhain"[2]. Była to niezwykła i bardzo dobra whisky. W jaki sposób pośród wielu single maltów Foster mógłby znaleźć podobne do niej?

Zastosujmy podejście z zakresu nauki o danych. Przypomnijmy z rozdziału 2., że przede wszystkim powinniśmy pomyśleć, na jakie dokładnie pytanie chcielibyśmy odpowiedzieć i jakie dane są właściwe, aby znaleźć odpowiedź. Jak możemy opisać różne gatunki whisky single malt jako wektory cech, aby móc wyodrębnić te, które mają podobny smak? Tym właśnie zagadnieniem zajęli się François-Joseph Lapointe i Pierre Legendre z Uniwersytetu w Montrealu (Lapointe i Legendre, 1994). Zainteresowały ich pewne klasyfikacyjne i organizacyjne pytania dotyczące szkockich whisky. Posłużmy się elementami ich metody[3].

[2] Nie, on także nie potrafi prawidłowo wymówić tej nazwy.

[3] Aby zapoznać się z analityką whisky w rzeczywistym zastosowaniu biznesowym, patrz: *http://www.whisky ↪classified.com/*.

Okazuje się, że dla wielu gatunków whisky publikowane są notki smakowe. Na przykład Michael Jackson, znany koneser whisky i piwa, napisał książkę *Michael Jackson's Malt Whisky Companion: a Connoisseur's Guide to the Malt Whiskies of Scotland* (Jackson, 1989), w której opisuje 109 różnych gatunków szkockiej whisky single malt. Opisy poszczególnych gatunków whisky mają formę notek smakowych, w rodzaju: „apetyczny zapach dymu torfowego, bardzo przypominający kadzidło, miód wrzosowy z owocową delikatnością".

Jako badacze danych robimy postępy. Znaleźliśmy potencjalnie użyteczne źródło danych. Nie dysponujemy jednak jeszcze opisami whisky w postaci wektorów cech. Mamy notki smakowe. Musimy kontynuować opracowywanie danych. Biorąc przykład z Lapointe'a i Legendre'a (1994), stwórzmy szereg cech liczbowych podsumowujących informacje zawarte w notkach smakowych. Zdefiniujmy pięć ogólnych atrybutów whisky, z których każdy może przyjmować wiele możliwych wartości:

1. **Kolor:** *żółty, bardzo jasny, jasny, jasnozłoty, złoty, stare złoto, pełne złoto, bursztyn itp.* (14 wartości)
2. **Nos:** *aromatyczny, torfowy, słodki, lekki, świeży, suchy, trawiasty itp.* (12 wartości)
3. **Usta:** *miękkie, średnie, pełne, krągłe, gładkie, lekkie, jędrne, oleiste.* (8 wartości)
4. **Podniebienie:** *pełne, suche, sherry, duże, owocowe, trawiaste, dymne, słone itp.* (15 wartości)
5. **Finisz:** *pełny, suchy, ciepły, lekki, gładki, czysty, owocowy, trawiasty, dymny itp.* (19 wartości)

Należy zauważyć, że wartości poszczególnych kategorii *nie* wykluczają się wzajemnie (np. podniebienie w przypadku Aberlour jest opisane jako średnie, pełne, miękkie, krągłe i gładkie). Ogólnie rzecz biorąc, wszystkie wartości mogą współwystępować (choć niektóre, jak na przykład kolor lekki i dymny, nie występują jednocześnie nigdy), ale ponieważ mogą współwystępować, to każda wartość każdej zmiennej została przez Lapointe'a i Legendre'a zakodowana jako odrębna cecha. W związku z tym istnieje 68 binarnych cech każdej whisky.

Foster lubi Bunnahabhain, więc możemy wykorzystać zestawienie whisky Lapointe'a i Legendre'a oraz odległość euklidesową, aby znaleźć dla niego podobne gatunki. Jako punkt odniesienia posłuży nam sporządzony przez nich opis Bunnahabhain:

- *Kolor:* złoty
- *Nos:* świeży i morski
- *Usta:* mocne, średnie i lekkie
- *Podniebienie:* słodkie, owocowe i czyste
- *Finisz:* pełny

Oto opis Bunnahabhain i pięciu szkockich whisky single malt najbardziej podobnych do Bunnahabhain, uszeregowanych zgodnie ze wzrostem odległości:

Whisky	Odległość	Deskryptory
Bunnahabhain	—	*złoty; pewny, średni, jasny; słodkie, owocowe, czyste; świeży, morski; pełny*
Glenglassaugh	0,643	złoty; twardy, lekki, gładki; słodkie, trawiaste; świeży, trawiasty
Tullibardine	0,647	złoty; pewny, średni, gładki; słodkie, owocowe, pełne, trawiaste, czyste; słodki; duży, aromatyczny, słodki
Ardberg	0,667	sherry; pewny, średni, pełny, jasny; słodkie; suchy, torfowy, morski, słony
Bruichladdich	0,667	jasny; twardy, lekki, gładki; suche, słodkie, dymne, czyste; jasny; pełny
Glenmorangie	0,667	jasnozłoty; średni, tłusty, jasny; słodkie, trawiaste, korzenne; słodki, pikantny, trawiasty, morski, świeży; pełny, długi

Korzystając z tej listy, moglibyśmy znaleźć whisky podobną do Bunnahabhain. Być może w konkretnym sklepie musielibyśmy chwilę podążać w dół listy, aby znaleźć whisky będącą akurat na składzie, ale ponieważ whisky na liście uszeregowane są według podobieństwa, nie byłoby to nic trudnego (a oprócz tego moglibyśmy mniej więcej stwierdzić, na ile whisky dostępna w sklepie podobna jest do innych, których w nim nie ma).

Jeżeli interesuje Cię zbiór danych dotyczący szkockich whisky, Lapointe i Legendre udostępniają te dane oraz swój artykuł pod adresem: *http://adn.biol.umontreal.ca/~numericalecology/data/scotch.html*.

Oto przykład bezpośredniego zastosowania podobieństwa do rozwiązania problemu. Rozumiejąc to podstawowe pojęcie, zyskujemy skuteczne narzędzie koncepcyjne, umożliwiające analizę różnych problemów, takich jak opisane wcześniej (znajdowanie podobnych firm, podobnych klientów itp.). Jak widzimy na przykładzie z whisky, badacze danych często muszą przede wszystkim prawidłowo zdefiniować dane, aby podobieństwo odnosiło się do użytecznego zbioru cech. Dalej zaprezentujemy kilka innych koncepcji podobieństwa i odległości. Teraz przejdźmy do innego bardzo powszechnego sposobu wykorzystywania podobieństwa w nauce o danych.

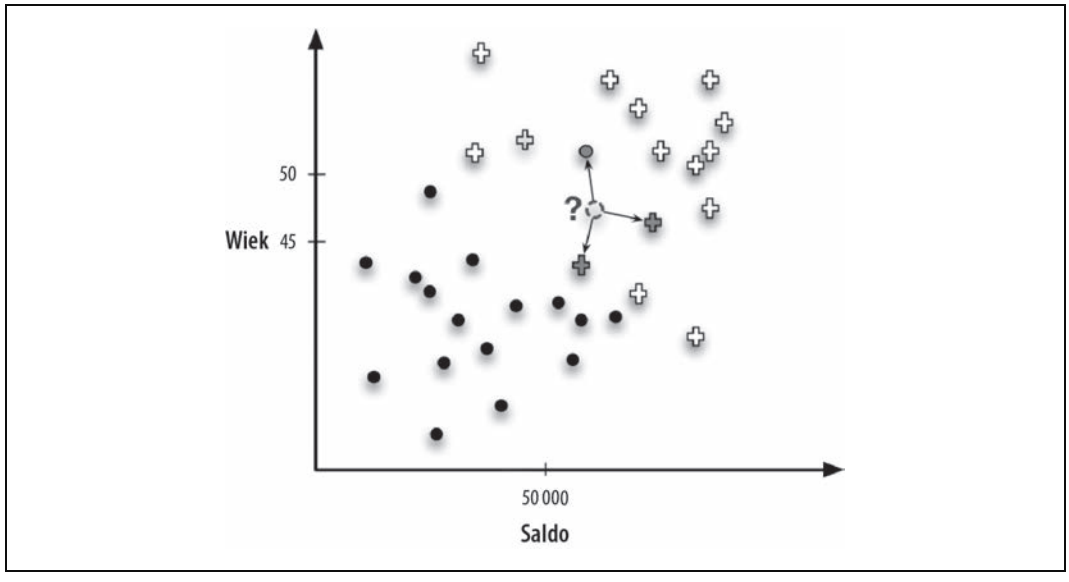
Najbliżsi sąsiedzi w modelowaniu predykcyjnym

Możemy również wykorzystać koncepcję najbliższych sąsiadów do modelowania predykcyjnego w inny sposób. Przypomnijmy sobie wszystko, co już wiemy o modelowaniu predykcyjnym z poprzednich rozdziałów. Podstawowa procedura umożliwiająca wykorzystanie podobieństwa do modelowania predykcyjnego jest wyjątkowo prosta: mając nowy przykład, którego zmienną docelową chcemy przewidzieć, możemy przejrzeć wszystkie przykłady uczące i wybrać kilka najbardziej podobnych do nowego przykładu. Następnie dokonujemy predykcji wartości zmiennej docelowej nowego przykładu na podstawie (znanych) wartości tej zmiennej dla najbliższych sąsiadów. Sposób przeprowadzenia tego ostatniego etapu musi zostać zdefiniowany; na razie powiedzmy po prostu, że dysponujemy jakąś **funkcją łączącą** (jak głosowanie lub uśrednianie) operującą na znanych wartościach docelowych sąsiadów. Dzięki funkcji łączącej otrzymamy naszą predykcję.

Klasyfikacja

Ponieważ poświęciliśmy do tej pory w naszej książce wiele miejsca zadaniom klasyfikacyjnym, zacznijmy od przyjrzenia się, jak sąsiedzi mogą zostać wykorzystani do sklasyfikowania nowego wystąpienia w bardzo uproszczonych warunkach. Rysunek 6.2 przedstawia nowy przykład, oznaczony jako „?”, którego etykietę chcemy przewidzieć. Zgodnie z podstawową procedurą przedstawioną powyżej wyszukujemy najbliższych sąsiadów (w tym przykładzie jest ich troje) i sprawdzamy ich znane zmienne docelowe (klasy). W tym przykładzie dwa przykłady są dodatnie, a jeden ujemny. Jaka powinna być nasza funkcja łącząca? Prostą funkcją łączącą byłaby w tym przypadku większość głosów, więc przewidywana klasa byłaby dodatnia.

Rozważmy nieco bardziej złożony problem, związany z marketingiem w sferze kart kredytowych. Naszym celem jest przewidzenie, czy nowy klient zareaguje na ofertę przyjęcia karty kredytowej na podstawie tego, jak zareagowali inni podobni klienci. Dane (oczywiście nadal uproszczone) przedstawiono w tabeli 6.1.



Rysunek 6.2. Klasyfikacja metodą najbliższych sąsiadów. Punkt, który ma zostać sklasyfikowany, oznaczony znakiem zapytania, zostanie sklasyfikowany jako +, ponieważ większość z jego najbliższych trzech sąsiadów także posiada wartość +

Tabela 6.1. Przykład najbliższych sąsiadów: Czy David odpowie, czy nie?

Klient	Wiek	Dochód (w tysiącach)	Karty	Odpowiedź (wielkość docelowa)	Odległość od Davida
David	*37*	*20*	*2*	*?*	*0*
John	35	35	3	Tak	$\sqrt{(35-37)^2+(35-50)^2+(3-2)^2}=15{,}16$
Rachael	22	50	2	Nie	$\sqrt{(22-37)^2+(50-50)^2+(2-2)^2}=15$
Ruth	63	200	1	Nie	$\sqrt{(63-37)^2+(200-50)^2+(1-2)^2}=152{,}23$
Jefferson	59	170	1	Nie	$\sqrt{(59-37)^2+(170-50)^2+(1-2)^2}=122$
Norah	25	40	4	Tak	$\sqrt{(25-37)^2+(40-50)^2+(4-2)^2}=15{,}74$

W tym przykładzie dysponujemy danymi pięciorga dotychczasowych klientów, którym wcześniej oferowaliśmy karty kredytowe. W przypadku każdego z nich znamy imię, wiek, dochód i liczbę już posiadanych kart; wiemy też, czy odpowiedzieli na ofertę. W przypadku nowej osoby, Davida, chcemy przewidzieć, czy odpowie on na ofertę czy nie (abyśmy nie musieli ponosić kosztów wysłania oferty, na którą i tak nie odpowie).

Ostatnia kolumna w tabeli 6.1 przedstawia kalkulację odległości każdego ze znanych klientów od Davida, przeprowadzoną za pomocą równania 6.1. Troje klientów (John, Rachael i Norah) jest dosyć podobnych do Davida, znajdując się w odległości około 15. Pozostałych dwoje klientów (Ruth i Jefferson) jest znacznie dalej. Wobec tego troje najbliższych sąsiadów Davida to Rachael, następnie John i dalej Norah. Ich odpowiedzi to odpowiednio: *Nie*, *Tak* i *Tak*. Jeśli zastosujemy do tych wartości większość głosów, to przewidzimy odpowiedź „Tak" (David

zareaguje). W tym miejscu pojawia się jednak szereg istotnych problemów związanych z metodami najbliższych sąsiadów: ilu powinno być sąsiadów? Czy powinni oni mieć jednakowe wagi w funkcji łączącej? Te kwestie omówimy w dalszej części rozdziału.

Szacowanie prawdopodobieństwa

Zauważyliśmy, że zwykle istotne jest nie tylko sklasyfikowanie nowego przykładu, ale oszacowanie związanego z nim prawdopodobieństwa — przypisanie mu określonego wskaźnika, bo wskaźnik zawiera w sobie więcej informacji niż decyzja typu tak/nie. Klasyfikację najbliższych sąsiadów można stosować do tego celu w całkiem prosty sposób. Rozważmy ponownie zadanie klasyfikacyjne związane z ustaleniem, czy David odpowie na ofertę czy nie. Jego najbliżsi sąsiedzi (Rachael, John i Norah) mają odpowiednio klasę `Nie`, `Tak` i `Tak`. Jeśli nadamy wartość klasie `Tak` w ten sposób, że `Tak = 1`, a `Nie = 0`, to będziemy mogli uśrednić poszczególne wyniki dla Dawida do wartości ⅔. Gdybyśmy mieli zrobić to w praktyce, to być może zechcielibyśmy posłużyć się do obliczenia szacunków prawdopodobieństwa większą liczbą najbliższych sąsiadów, a nie tylko trojgiem (i przypomnielibyśmy sobie omówienie szacowania prawdopodobieństw z małych próbek w podrozdziale „Szacowanie prawdopodobieństwa”).

Regresja

Skoro możemy wyszukać najbliższych sąsiadów, to możemy także wykorzystać ich do dowolnego zadania predykcyjnego, łącząc ich na różne sposoby. Właśnie dowiedzieliśmy się, jak przeprowadzić klasyfikację na podstawie większości głosów za wielkością docelową lub przeciw niej. Regresję możemy wykonać w podobny sposób.

Załóżmy, że dysponujemy zbiorem danych z tabeli 6.1, ale tym razem chcemy przewidzieć dochód Davida. Nie będziemy powtarzali obliczeń odległości; założymy tylko, że troje najbliższych sąsiadów Davida to znowu Rachael, John i Norah. Ich dochody wynoszą odpowiednio 50, 35 i 40 (w tysiącach). Następnie za pomocą tych wartości stworzymy predykcję dochodu Davida. Możemy użyć średniej (około 42) lub mediany (40).

Należy pamiętać, że wyszukując sąsiadów, nie używamy zmiennej docelowej, ponieważ to ją staramy się przewidzieć. W ten sposób dochód nie będzie brany pod uwagę przy obliczaniu odległości, jak to ma miejsce w tabeli 6.1. Do obliczenia odległości możemy jednak wykorzystywać inne zmienne, których wartości znamy.

Ilu sąsiadów i jak duży wpływ?

Wyjaśniając sposób przeprowadzania klasyfikacji, regresji i scoringu, wykorzystywaliśmy przykład z tylko trojgiem sąsiadów. Może to rodzić kilka pytań. Po pierwsze, dlaczego *troje* sąsiadów, a nie tylko jeden albo pięciu czy stu? Po drugie, czy powinniśmy traktować wszystkich sąsiadów jednakowo? Choć wszyscy nazywani są „najbliższymi” sąsiadami, to niektórzy są bliżsi niż inni, czy więc nie powinno to mieć wpływu na sposób, w jaki ich wykorzystujemy?

Nie istnieje prosta odpowiedź na pytanie, ilu sąsiadów powinniśmy brać pod uwagę. Wygodne bywają liczby nieparzyste, bo pozwalają unikać remisów w przypadku klasyfikacji metodą większościową w ramach problemów dwuklasowych. Algorytmy najbliższych sąsiadów określa się często skrótem *k*-NN, gdzie *k* oznacza liczbę użytych sąsiadów, na przykład 3-NN.

Na ogół im większa wartość k, tym dokładniejsze są szacunki spomiędzy sąsiadów. Jeśli jak dotąd wszystko zrozumiałeś, to po chwili zastanowienia powinieneś zdać sobie sprawę, że jeżeli maksymalnie zwiększymy wartość k (tak że $k = n$), to do każdej predykcji będzie wykorzystywany cały zbiór danych. Ujmując to w elegancki sposób, dla każdego przykładu będzie to po prostu predykcja średniej w całym zbiorze danych. Dla klasyfikacji byłaby to predykcja klasy większościowej w całym zbiorze danych, dla regresji średnia wszystkich wartości wielkości docelowej, dla szacowania prawdopodobieństwa klasy prawdopodobieństwo o „stopie bazowej" (patrz „Uwaga: stopa bazowa" w podrozdziale „Dane wydzielone i wykresy dopasowania" w rozdziale 5.).

Nawet jeżeli mamy pewność co do tego, ilu sąsiadów użyć, to możemy zauważyć, że poziom podobieństwa sąsiadów do przykładu, który próbujemy przewidzieć, jest różny. Czy nie powinno to wpływać na sposób, w jaki się nimi posługujemy?

Zaczęliśmy od prostej strategii głosowania *większościowego*, pobierając nieparzystą liczbę sąsiadów, aby uniknąć remisu. Pomijamy tutaj jednak istotną informację: jak blisko wystąpienia znajduje się każdy sąsiad. Zastanówmy się na przykład, co by się stało, gdybyśmy użyli $k = 4$ sąsiadów do sklasyfikowania Davida. Otrzymalibyśmy odpowiedzi (`Tak`, `Nie`, `Tak`, `Nie`), które dawałyby remis. Pierwsi trzej sąsiedzi znajdują się jednak bardzo blisko Davida (odległość ≈ 15), a czwarty znacznie dalej (odległość ≈ 122). Intuicja podpowiada nam, że to czwarte wystąpienie nie powinno mieć takiej samej wagi w głosowaniu jak pierwsza trójka. Aby uwzględnić tę wątpliwość, metody najbliższych sąsiadów często wykorzystują **ważenie głosów** lub **głosowanie moderowane podobieństwem**, w przypadku których udział każdego sąsiada jest proporcjonalny do wartości podobieństwa.

Przyjrzyjmy się ponownie danym z tabeli 6.1, dotyczącej prognozowania, czy David odpowie na ofertę przyznania karty kredytowej. Wykazaliśmy, że jeśli dokonamy predykcji klasy Davida metodą większości głosów, to będzie ona zależała w znaczący sposób od liczby sąsiadów, których wybierzemy. Przeprowadźmy naszą kalkulację ponownie, tym razem uwzględniając *wszystkich* sąsiadów, ale proporcjonalnie do stopnia ich podobieństwa do Davida, wykorzystując jako wagę skalowania odwrotność kwadratu odległości. Oto sąsiedzi uszeregowani na podstawie odległości od Davida:

Imię	Odległość	Waga podobieństwa	Udział	Klasa
Rachael	15,0	0,004444	0,344	Nie
John	15,2	0,004348	0,336	Tak
Norah	15,7	0,004032	0,312	Tak
Jefferson	122,0	0,000067	0,005	Nie
Ruth	152,2	0,000043	0,003	Nie

W kolumnie *Udział* ujęto wartości wkładu każdego sąsiada do ostatecznej kalkulacji predykcji prawdopodobieństwa wielkości docelowej (udziały są proporcjonalne do wag, ale dają w sumie 1). Widzimy, że odległości mają znaczący wpływ na wartość udziałów: Rachael, John i Norah są najbardziej podobni do Davida i skutecznie określają naszą predykcję jego reakcji, a Jefferson i Ruth znajdują się tak daleko, że ich udział jest w zasadzie nieistotny. Podsumowując udziały dla przypadków dodatnich i ujemnych, ostateczne szacunki prawdopodobieństwa dla Davida wynoszą 0,65 dla `Tak` i 0,35 dla `Nie`.

Ta koncepcja odnosi się także do innych rodzajów zadań predykcyjnych, na przykład do regresji i szacowania prawdopodobieństwa klas. Ogólnie możemy traktować tę procedurę jako **scoring** ważony. Miłą cechą scoringu ważonego jest to, że zmniejsza on znaczenie określania, ilu powinniśmy użyć sąsiadów. Ponieważ udział każdego sąsiada jest moderowany przez jego odległość, wpływ sąsiadów w naturalny sposób zmniejsza się, im dalej znajdują się oni od wystąpienia. W związku z tym w przypadku stosowania scoringu ważonego dokładna wartość k jest znacznie mniej istotna niż w przypadku metody większościowej. Istnieją metody, które, aby uniezależnić się od k, pobierają po prostu bardzo dużą liczbę wystąpień (na przykład wszystkie, $k = n$) i opierają się na ważeniu odległości w celu moderowania wpływów.

Ramka: wiele nazw dla wnioskowania metodą najbliższego sąsiedztwa

Podobnie jak w wielu innych przypadkach z dziedziny eksploracji danych, istnieją różne nazwy klasyfikatorów najbliższego sąsiedztwa, po części dlatego, że podobne koncepcje były rozwijane niezależnie od siebie. Klasyfikatory najbliższego sąsiedztwa powstały dawno temu, w ramach statystyki i rozpoznawania wzorców (Cover i Hart, 1967). Koncepcja klasyfikowania nowych wystąpień bezpośrednio poprzez konsultacje z bazą danych („pamięcią") wystąpień została nazwana **uczeniem się z przykładów** (Aha, Kibler i Albert, 1991) i **uczeniem się opartym na pamięci** (Lin i Vitter, 1994). Ze względu na to, że po napotkaniu wystąpień nie jest przeprowadzana indukcja i większość działań odkłada się do momentu pobrania wystąpień, ta ogólna koncepcja znana jest jako **„ograniczone" uczenie się** (Aha, 1997).

Pokrewną techniką z obszaru sztucznej inteligencji jest **wnioskowanie na podstawie przypadków** (*Case-Based Reasoning* — CBR; Kolodner, 1993; Aamodt i Plaza, 1994). Jak już wspomnieliśmy, przypadki historyczne są powszechnie wykorzystywane przez lekarzy i prawników poszukujących rozwiązań nowych problemów, więc wnioskowanie na podstawie przypadków ma w tych dziedzinach ugruntowaną tradycję.

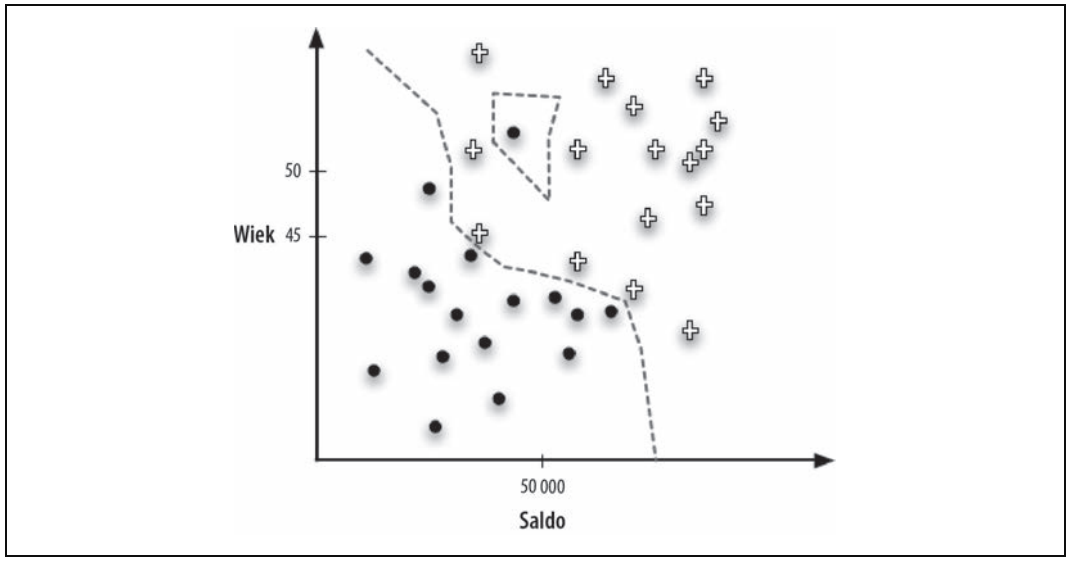
Istnieją jednak także znaczące różnice pomiędzy wnioskowaniem na podstawie przypadków i metodami najbliższego sąsiedztwa. Przypadki w ramach CBR zazwyczaj nie są prostymi wystąpieniami w postaci wektorów cech, ale bardzo szczegółowymi opisami epizodów chorobowych, zawierającymi takie elementy jak objawy, historia choroby, diagnoza, przebieg leczenia i jego wyniki; albo szczegóły dotyczące sprawy sądowej, w tym argumentacje powoda i pozwanego, przywoływane precedensy i orzeczenie. Ponieważ przypadki są tak szczegółowe, CBR wykorzystuje je nie tylko do etykietowania klas, ale jako źródła informacji dotyczącej diagnoz i planowania, które mogą być wykorzystywane do opracowywania przypadku, gdy już zostanie on wyszukany. Dostosowywanie historycznych przypadków, aby można je było wykorzystać w nowej sytuacji, jest zwykle skomplikowanym procesem, który wymaga znacznego wysiłku.

Interpretacja geometryczna, nadmierne dopasowanie i kontrola złożoności

Tak jak w przypadku innych modeli, które omawialiśmy, duże znaczenie poznawcze ma wizualizacja obszarów klasyfikacyjnych utworzonych metodą najbliższych sąsiadów. Chociaż nie występuje tutaj wyraźna granica, to istnieją dorozumiane obszary sąsiedztwa wystąpień. Obszary te można wyznaczyć poprzez systematyczne badanie punktów w przestrzeni wystą-

pień, określanie klasyfikacji każdego punktu i budowanie granicy tam, gdzie następują zmiany klasyfikacji

Rysunek 6.3 przedstawia taki obszar, utworzony przez klasyfikator 1-NN wokół wystąpień naszej domeny „Spisań w straty". Porównajmy je z obszarami drzewa klasyfikacji z rysunku 3.15 i obszarami utworzonymi przez granicę liniową na rysunku 4.3.



Rysunek 6.3. Granice wyznaczone przez klasyfikator 1-NN

Możemy zauważyć, że granice nie są liniami, nie przyjmują także żadnego rozpoznawalnego kształtu geometrycznego; są nieregularne i obrazują rozgraniczenia między wystąpieniami uczącymi o różnych klasach. Klasyfikator najbliższego sąsiada tworzy bardzo specyficzne granice wokół wystąpień uczących. Zauważmy także, że pojedyncze ujemne wystąpienie, odizolowane w obszarze wystąpień dodatnich, tworzy „ujemną wyspę" wokół samego siebie. Ten punkt mógłby zostać uznany za szum lub element odstający i inny typ modelu mógłby go „wygładzić".

Ta wrażliwość na elementy odstające wynika w pewnym stopniu z użycia klasyfikatora 1-NN, który wyszukuje tylko pojedyncze wystąpienia, a więc ma bardziej nieregularne granice niż taki, który uśrednia wielu sąsiadów. Wrócimy do tego za chwilę. Ujmując rzecz bardziej ogólnie, nieregularne granice koncepcyjne są charakterystyczne dla wszystkich klasyfikatorów najbliższych sąsiadów, ponieważ nie nakładają one na klasyfikator żadnej konkretnej formy geometrycznej. Zamiast tego tworzą w przestrzeni wystąpień granice dostosowane do konkretnych danych wykorzystanych do uczenia.

W tym miejscu należy przypomnieć naszą dyskusję na temat nadmiernego dopasowania i kontroli złożoności z rozdziału 5. Jeśli uważasz, że w przypadku 1-NN nadmierne dopasowanie musi być bardzo silne, to masz rację. Wystarczy zastanowić się, co by się stało, gdybyśmy ewaluowali klasyfikator 1-NN na danych uczących. Przy klasyfikacji każdego punktu danych uczących każda rozsądna miara odległości prowadziłaby do pobrania danego punktu jako jego własnego najbliższego sąsiada! Następnie wartość jego własnej zmiennej docelowej zostałaby wykorzystana do prognozowania jej samej i proszę bardzo, otrzymalibyśmy doskonałą

klasyfikację. To samo dotyczy regresji. Klasyfikator 1-NN zapamiętuje dane uczące. Radzi sobie jednak nieco lepiej niż nasza tabela wyszukiwania z początkowej części rozdziału 5. Ponieważ tabela wyszukiwania nie posiadała żadnego ujęcia podobieństwa, po prostu doskonale prognozowała w odniesieniu do przykładów uczących i prezentowała pewne domyślne prognozy dla wszystkich innych. Klasyfikator 1-NN prognozuje doskonale w przypadku przykładów uczących, ale może również dokonywać często rozsądnych predykcji na podstawie innych przykładów: wykorzystując najbardziej podobny przykład uczący.

W odniesieniu do przeuczenia i jego unikania, k w klasyfikatorze k-NN jest więc parametrem złożoności. Z jednej strony, możemy wyznaczyć $k = n$ i nie pozwolić na jakąkolwiek większą złożoność naszego modelu. Jak opisaliśmy to wcześniej, model n-NN (pomijając ważenie podobieństwa) po prostu prognozuje dla każdego przypadku średnią wartość w zbiorze danych. Z drugiej strony, możemy wyznaczyć $k = 1$ i otrzymamy wyjątkowo złożony model, który wyznacza skomplikowane granice tak, że każdy przykład uczący znajdzie się w obszarze etykietowanym przez jego własną klasę.

Wróćmy teraz do wcześniejszego pytania: w jaki sposób powinniśmy wybrać k? Możemy użyć tej samej procedury, którą opisaliśmy w rozdziale 5., w podrozdziale „Ogólna metoda unikania nadmiernego dopasowania", do wyznaczenia innych parametrów złożoności: możemy przeprowadzić sprawdzian krzyżowy lub inne zagnieżdżone testowanie na zbiorze uczącym dla różnych wartości k, w poszukiwania takiej, która przyniesie najlepszą skuteczność na bazie danych uczących. Gdy już wybierzemy wartość k, budujemy model k-NN z całego zbioru uczącego. Jak omówiliśmy to szczegółowo w rozdziale 5., skoro ta procedura wykorzystuje tylko dane uczące, to wciąż możemy dokonać jej ewaluacji na bazie danych testowych i uzyskać obiektywne oszacowanie skuteczności jej generalizacji. Narzędzia eksploracji danych są zwykle w stanie przeprowadzać taki zagnieżdżony sprawdzian krzyżowy, aby wyznaczyć k automatycznie.

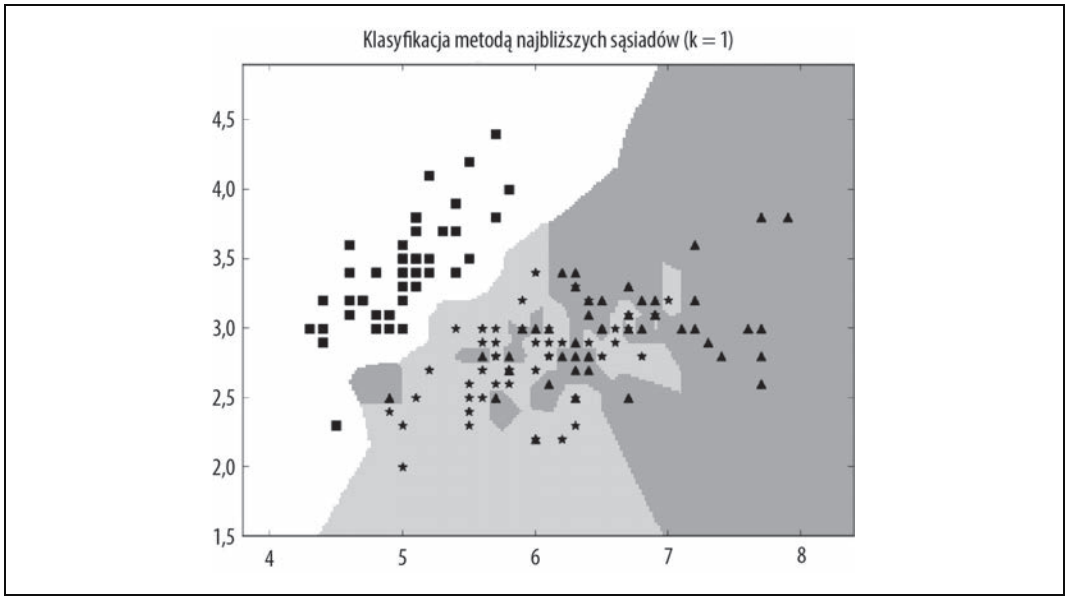
Rysunki 6.4 i 6.5 przedstawiają różne granice utworzone przez klasyfikatory najbliższych sąsiadów. Klasyfikujemy prosty problem trzyklasowy, wykorzystując różną liczbę sąsiadów. Na rysunku 6.4 posługujemy się tylko jednym sąsiadem. Granice są nieregularne i bardzo specyficzne dla przykładów uczących w zbiorze danych. Na rysunku 6.5 w celu dokonania klasyfikacji uśredniono 30 najbliższych sąsiadów. Granice są oczywiście inne niż na rysunku 6.4 i znacznie mniej postrzępione. Należy jednak zauważyć, że w żadnym z przypadków granice nie są gładkimi krzywymi czy odcinkowo regularnymi obszarami, których spodziewalibyśmy się dla modelu liniowego lub modelu o strukturze drzewa. Granice dla k-NN są silniej zdefiniowane przez dane.

Problemy z metodami najbliższych sąsiadów

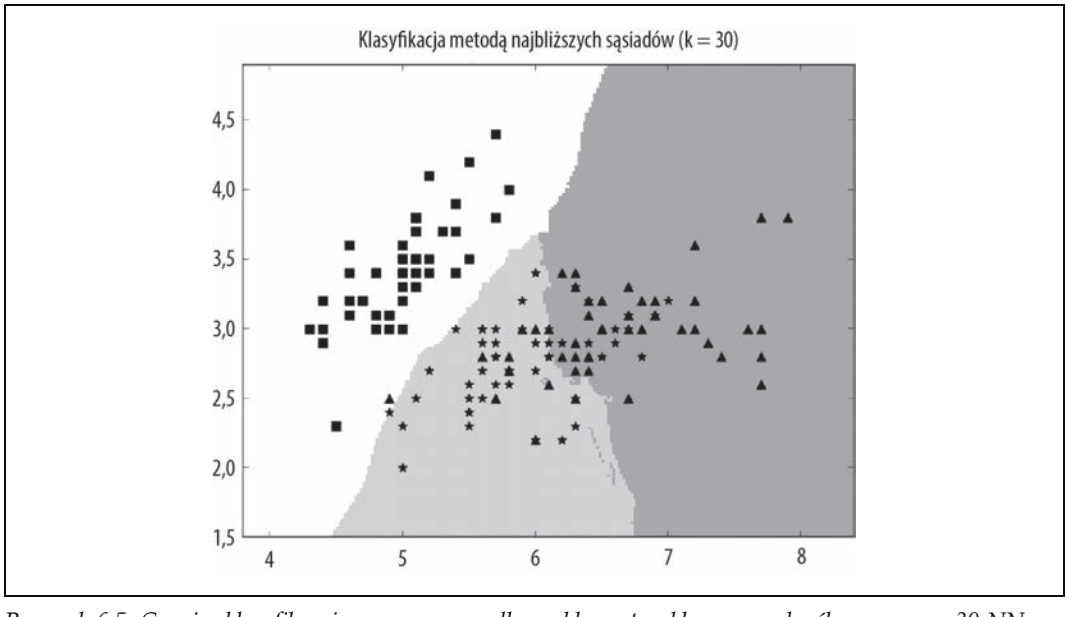
Przed podsumowaniem dyskusji o metodach najbliższych sąsiadów jako modelach predykcyjnych powinniśmy wspomnieć o szeregu problemów związanych z ich wykorzystywaniem, które często pojawiają się podczas ich stosowania w praktyce.

Zrozumiałość

Kwestia zrozumiałości klasyfikatorów najbliższych sąsiadów jest zagadnieniem złożonym. Jak wspomnieliśmy, w niektórych dziedzinach, takich jak medycyna czy prawo, wnioskowanie na podstawie podobnych historycznych przypadków jest naturalnym sposobem podejmowania



Rysunek 6.4. Granice klasyfikacyjne wyznaczone dla problemu trzyklasowego określonego przez 1-NN (pojedynczy najbliższy sąsiad)



Rysunek 6.5. Granice klasyfikacyjne wyznaczone dla problemu trzyklasowego określonego przez 30-NN (uśrednienie 30 najbliższych sąsiadów)

decyzji o nowym przypadku. W takich dziedzinach metoda najbliższego sąsiedztwa może być dobrym rozwiązaniem. W innych obszarach brak jednoznacznego, możliwego do interpretowania modelu może stanowić problem.

Tak naprawdę istnieją dwa aspekty problemu zrozumiałości: uzasadnianie konkretnych *decyzji* i zrozumiałość całego *modelu*.

W przypadku *k*-NN zwykle łatwo opisać, w jaki sposób podejmowana jest decyzja w pojedynczym przypadku: można przedstawić zbiór sąsiadów uczestniczących w procesie podejmowania decyzji, wraz z wartością ich wkładu w jej podjęcie. Zrobiliśmy tak w przykładzie prognozy reakcji Davida, przedstawionym wcześniej w tabeli 6.1. Przydatne jest tutaj staranne formułowanie i wyważona prezentacja najbliższych sąsiadów. Netflix wykorzystuje na przykład pewnego rodzaju klasyfikację metodą najbliższych sąsiadów w swoich rekomendacjach i wyjaśnia rekomendacje dotyczące filmów za pomocą zdań w rodzaju:

> „Film *Billy Elliot* zarekomendowaliśmy na podstawie Twojego zainteresowania filmami *Amadeusz*, *Wierny ogrodnik* i *Mała Miss*".

Amazon przedstawia rekomendacje, stosując takie wyrażenia, jak: „Klienci o podobnej historii wyszukiwania nabywali..." i „Powiązane wyszukiwania".

To, czy te uzasadnienia są odpowiednie, zależy od konkretnego zastosowania. Klient Amazona mógłby być zadowolony z takiego wyjaśnienia otrzymania rekomendacji. Z drugiej jednak strony, osoba starająca się o kredyt hipoteczny mogłaby nie być zadowolona z wyjaśnienia: „Odrzuciliśmy Pański wniosek o kredyt hipoteczny, ponieważ przypomina nam Pan Kowalskich i Nowaków, którzy przestali spłacać kredyty". I rzeczywiście, istnieją regulacje prawne ograniczające rodzaje modeli, które mogą być stosowane do oceny zdolności kredytowej. Mogą to być tylko modele podające bardzo proste wyjaśnienia na podstawie konkretnych, istotnych zmiennych. Na przykład na bazie modelu liniowego można byłoby stwierdzić: „Przy założeniu identycznych pozostałych danych, gdyby Pana dochód był wyższy o 20 000 dolarów, to ten konkretny kredyt hipoteczny zostałby Panu przyznany".

Łatwo również wyjaśnić, w jaki sposób cały model działający na bazie metody najbliższych sąsiadów podejmuje decyzje w nowych przypadkach. Koncepcja wyszukiwania najbardziej podobnych przypadków i badania, jak zostały sklasyfikowane lub jaką miały wartość, jest dla wielu osób intuicyjna.

Trudniejsze jest bardziej szczegółowe wyjaśnienie, jaka „wiedza" została wydobyta z danych. Odpowiedź na pytanie decydenta: „Czego wasz system dowiedział się z danych o moich klientach? Na jakiej podstawie podejmuje on swoje decyzje?" może być niełatwa, ponieważ nie ma tutaj jasno sprecyzowanego modelu. Ściśle mówiąc, „model" najbliższych sąsiadów składa się z całego zbioru przypadków (bazy danych), funkcji odległości i funkcji łączącej. W dwóch wymiarach możemy zwizualizować to bezpośrednio, jak miało to miejsce na wcześniejszych rysunkach. Nie jest to jednak możliwe, gdy istnieje wiele wymiarów. Wiedza zawarta w tym modelu zazwyczaj nie jest łatwa do zrozumienia, więc jeśli zrozumiałość i uzasadnienie stworzenia modelu mają podstawowe znaczenie, to metod najbliższych sąsiadów należy unikać.

Wymiarowość i wiedza fachowa

Przy obliczaniu odległości między dwoma wystąpieniami metody najbliższego sąsiedztwa zazwyczaj uwzględniają wszystkie cechy. W podrozdziale „Atrybuty heterogeniczne" omawiamy jedną z komplikacji związanych z atrybutami: atrybuty liczbowe mogą znacząco różnić się zakresami i o ile nie zostaną odpowiednio wyskalowane, to efekt wywoływany przez jeden atrybut o szerokim zakresie może wchłonąć efekt innego, o zakresie znacznie mniejszym.

Poza tym istnieje także problem związany z nadmierną liczbą atrybutów lub wieloma takimi atrybutami, które są nieistotne z punktu widzenia oceny podobieństwa.

W obszarze kart kredytowych baza danych klientów mogłaby na przykład zawierać wiele przypadkowych informacji, takich jak liczba dzieci, czas zatrudnienia, kubatura domu, średni dochód, marka i model samochodu, średni poziom wykształcenia i tak dalej. Niewykluczone, że część z nich mogłaby być istotna w odniesieniu do ewentualnego zaakceptowania przez klienta oferty karty kredytowej, ale prawdopodobnie większość byłaby bez znaczenia. Takie problemy określane są jako wysoko wymiarowe — ciąży na nich tak zwana **klątwa wymiarowości** — a to stwarza problemy dla metod najbliższego sąsiedztwa. Większość z nich ma zasadniczo charakter techniczny[4], ale mówiąc ogólnie, skoro wszystkie atrybuty (wymiary) biorą udział w obliczeniach odległości, to ze względu na obecność zbyt wielu nieistotnych atrybutów podobieństwo wystąpień może zostać zachwiane i błędnie określone.

Istnieje kilka sposobów rozwiązania problemu z wieloma prawdopodobnie nieistotnymi atrybutami. Jednym z nich jest **wybór cech**, czyli uważne określenie tych, które powinny zostać uwzględnione w modelu eksploracji danych. Wybór cech może zostać przeprowadzony ręcznie przez badacza danych, z wykorzystaniem wiedzy fachowej dotyczącej tego, które atrybuty są istotne. Jest to jeden z głównych sposobów wprowadzania przez zespół badaczy danych wiedzy fachowej do procesu eksploracji danych. Jak wspomnieliśmy w rozdziale 3. i w rozdziale 5., istnieją także zautomatyzowane metody selekcji cech, które potrafią przetwarzać dane i oceniać, które atrybuty zawierają informacje o wielkości docelowej.

Innym sposobem wprowadzania wiedzy fachowej do obliczeń podobieństwa jest dostrajanie funkcji podobieństwa/odległości ręcznie. Możemy na przykład wiedzieć, że atrybut *Liczba kart kredytowych* powinien mieć silny wpływ na to, czy klient zaakceptuje ofertę kolejnej karty. Badacz danych może dostroić funkcję odległości, przypisując różne wagi różnym atrybutom (np. przyznając większą wagę atrybutowi *Liczba kart kredytowych*). Z wiedzy fachowej możemy korzystać, nie tylko będąc przekonani, że wiemy, jaki atrybut będzie miał w danym przypadku większą wartość predykcyjną, ale także w znaczeniu bardziej ogólnym, ponieważ wiemy coś o podobnych jednostkach, które chcemy znaleźć. Jeżeli aromat „torfowy" ma dla nas istotne znaczenie przy poszukiwaniu podobnie smakujących whisky, to możemy przypisać mu wyższą wagę przy obliczaniu podobieństwa. Jeśli inna zmienna smaku jest nieistotna, to możemy ją usunąć lub po prostu przyznać jej niską wagę.

Wydajność obliczeniowa

Jedną z zalet metod najbliższego sąsiedztwa jest to, że uczenie jest w tym przypadku bardzo szybkie, ponieważ wymaga zwykle tylko zmagazynowania wystąpień. Nie wkładamy żadnego wysiłku w tworzenie modelu. Głównym kosztem obliczeniowym metody najbliższego sąsiedztwa jest etap predykcji/klasyfikacji, kiedy należy złożyć zapytanie do bazy danych, aby znaleźć najbliższych sąsiadów nowego wystąpienia. Może to być bardzo kosztowne, wobec czego należy się zastanowić nad wydatkami związanymi z klasyfikacją. Niektóre aplikacje wymagają bardzo szybkich prognoz; na przykład w sferze targetowania reklam internetowych

[4] Okazuje się na przykład, że przy dużych ilościach cech z przyczyn technicznych pewne konkretne wystąpienia pojawiają się wyjątkowo często w zbiorach *k* najbliższych sąsiadów innych wystąpień. Te konkretne wystąpienia mają tym samym bardzo znaczący wpływ na wiele klasyfikacji.

może istnieć konieczność podejmowania decyzji w czasie kilkudziesięciu milisekund. Dla takich zastosowań metody najbliższych sąsiadów mogą okazać się niepraktyczne.

Istnieją techniki przyspieszenia wyszukiwań sąsiadów. W celu poprawy efektywności zapytań o najbliższych sąsiadów w niektórych komercyjnych systemach baz danych i eksploracji danych stosowane są wyspecjalizowane struktury danych, takie jak drzewa kd i metody haszowania (Shakhnarovich, Darrell i Indyk, 2005; Papadopoulos i Manolopoulos, 2005). Należy jednak pamiętać, że wiele prostszych narzędzi eksploracji danych zwykle nie wykorzystuje tych technik i nadal opiera się na naiwnym wyszukiwaniu siłowym.

Kilka istotnych szczegółów technicznych dotyczących podobieństw i sąsiadów

Atrybuty heterogeniczne

Do tej pory posługiwaliśmy się odległością euklidesową, wykazując, że jest ona łatwa do obliczenia. Jeżeli atrybuty są liczbowe i są bezpośrednio porównywalne, to obliczanie odległości jest naprawdę proste. Kiedy jednak przykłady zawierają złożone, niejednorodne atrybuty, sprawy zaczynają się komplikować. Rozważmy kolejny przykład z tej samej domeny, ale z nieco większą liczbą atrybutów:

Atrybut	osoba A	osoba B
Płeć	Mężczyzna	Kobieta
Wiek	23	40
Liczba lat pod aktualnym adresem	2	10
Status rezydenta (1 = Właściciel, 2 = Najemca, 3 = Inny)	2	1
Dochód	50 000	90 000

Pojawia się tutaj szereg komplikacji. Po pierwsze, równanie odległości euklidesowej jest liczbowe, a płeć jest atrybutem kategorycznym (symbolicznym). Musi on zostać wyrażony liczbowo. Dla zmiennych binarnych wystarczyć może proste kodowanie, M = 0, K = 1, ale jeśli istnieje wiele wartości dla atrybutu kategorycznego, nie będzie to wystarczające.

Równie ważny jest fakt, że dysponujemy zmiennymi, które, choć są liczbowe, to mają bardzo różne skale i zakresy. Atrybut wiek może mieć zakres od 18 do 100, a dochód może mieścić się w zakresie od 10 dolarów do 10 milionów dolarów. Bez skalowania nasza miara odległości uznałaby 10 dolarów różnicy dochodu za tak znaczące, jak dziesięć lat różnicy wieku, a to jest całkowicie nieprawidłowe. Z tego powodu systemy bazujące na metodach najbliższego sąsiedztwa często wykorzystują interfejsy umożliwiające skalowanie zmiennych. Mierzą one zakresy zmiennych i odpowiednio do nich skalują wartości lub rozkładają wartości do skończonej liczby koszyków. Obowiązuje tutaj ogólna zasada, że należy zachować ostrożność, bo obliczenia podobieństwa/odległości mają zasadnicze znaczenie dla każdego konkretnego zastosowania.

* Inne funkcje odległości

Przed nami szczegóły techniczne

Do tej pory dla uproszczenia wykorzystaliśmy tylko jedną miarę, odległość euklidesową. W tym miejscu przedstawiamy bardziej szczegółowy opis funkcji odległości i pewne miary alternatywne.

Należy zwrócić uwagę, że przedstawione tutaj miary podobieństwa to tylko niewielka część wszystkich miar podobieństwa, które bywają wykorzystywane. Miary opisane przez nas są najbardziej popularne, ale zarówno badacze danych, jak i analitycy biznesowi powinni pamiętać, że istotne jest korzystanie z sensownej miary podobieństwa w kontekście badanego problemu biznesowego. Ten podrozdział może zostać pominięty bez utraty ciągłości wywodu.

Jak zaznaczyliśmy wcześniej, odległość euklidesowa jest chyba miarą odległości najczęściej stosowaną w sferze nauki o danych. Jest ogólna, intuicyjna i bardzo szybka pod względem obliczeniowym. Ponieważ wykorzystuje *kwadraty* odległości w poszczególnych wymiarach, to jest czasami nazywana **normą L2** i przedstawiana jako $\| \bullet \|_2$. Równanie 6.2 przedstawia jej wygląd formalny.

Równanie 6.2. Odległość euklidesowa (norma L2)

$$d_E(\mathbf{X},\mathbf{Y}) = \left\|\mathbf{X}-\mathbf{Y}\right\|_2 = \sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\ldots}$$

Chociaż odległość euklidesowa jest powszechnie wykorzystywana, to odległość można obliczać na wiele innych sposobów. W książce *The Dictionary of Distances* autorstwa Michela i Eleny Deza (Elsevier Science, 2006) przedstawiono ich kilkaset, a kilkanaście z nich jest regularnie stosowanych w eksploracji danych. Powodem tak znacznej liczby miar jest fakt, że w przypadku metody najbliższego sąsiedztwa funkcja odległości ma zasadnicze znaczenie. Redukuje ona zasadniczo porównywanie dwóch (potencjalnie złożonych) przykładów do jednej liczby. Typy danych i specyfika obszaru zastosowania w znaczącym stopniu wpływają na to, w jaki sposób różnice w ramach poszczególnych atrybutów powinny być łączone.

Odległość Manhattan lub **norma L1** to suma (*niepodniesionych do kwadratu*) par odległości, jak pokazano w równaniu 6.3.

Równanie 6.3. Odległość Manhattan (norma L1)

$$d_{Manh}(\mathbf{X},\mathbf{Y}) = \left\|\mathbf{X}-\mathbf{Y}\right\|_1 = \left|x_1-y_1\right|+\left|x_2-y_2\right|+\ldots$$

Jest to po prostu suma różnic pomiędzy X i Y w poszczególnych wymiarach. Nazywa się ją odległością Manhattan (lub taksówkową), ponieważ przedstawia całkowity dystans, który należałoby pokonać ulicami miejsca przypominającego centrum Manhattanu (przedstawionego w postaci siatki), aby pokonać całkowitą odległość ze wschodu na zachód oraz z północy na południe.

Naukowcy badający przedstawiony wcześniej problem analityki whisky stosowali jeszcze inną popularną miarę odległości[5]. Jest nią **odległość Jaccarda**. Odległość Jaccarda traktuje dwa

[5] Patrz Lapointe i Legendre (1994); rozdział 3. („Classification of Pure Malt Scotch Whiskies") zawiera szczegółowe omówienie opracowanego przez nich sformułowania problemu: *http://www.dcs.ed.ac.uk/home/jhb/whisky/lapointe/text.html*.

obiekty jako zbiory cech. Postrzeganie dwóch obiektów w postaci zbiorów cech pozwala myśleć o łącznej wielkości wszystkich cech dwóch obiektów X i Y, $|X \cup Y|$ i wielkości zbioru cech wspólnych dla dwóch obiektów (części wspólnej) $|X \cap Y|$. W przypadku dwóch obiektów, X i Y, odległość Jaccarda to udział wszystkich cech (posiadanych przez każdą z nich), które są wspólne dla obu. Sprawdza się ona w przypadku problemów, w których istotne jest posiadanie przez dwa obiekty wspólnej cechy charakterystycznej, ale wspólny *brak* danej cechy istotny nie jest. Na przykład, gdy poszukujemy dwóch podobnych whisky, istotne jest, czy obie są torfowe, lecz może nie być istotne to, czy obie nie są *słone*. W notacji zbiorów miarę odległości Jaccarda przedstawia równanie 6.4.

Równanie 6.4. Odległość Jaccarda

$$d_{Jacc}(\mathbf{X},\mathbf{Y}) = 1 - \frac{|X \cap Y|}{|X \cup Y|}$$

Odległość kosinusowa jest często wykorzystywana w klasyfikacji tekstu do mierzenia podobieństwa dwóch dokumentów. Definiuje ją równanie 6.5.

Równanie 6.5. Odległość kosinusowa

$$d_{cos}(\mathbf{X},\mathbf{Y}) = 1 - \frac{\mathbf{X} \cdot \mathbf{Y}}{\|\mathbf{X}\|_2 \cdot \|\mathbf{Y}\|_2}$$

gdzie $\| \bullet \|_2$ ponownie odpowiada normie L2, czyli długości euklidesowej każdego wektora cech (dla wektora jest to po prostu odległość od początku układu współrzędnych).

W literaturze z zakresu wyszukiwania informacji bardziej powszechny jest termin **podobieństwo kosinusowe**, które jest po prostu ułamkiem z równania 6.5. Alternatywnie jest to 1–odległość kosinusowa.

W sferze klasyfikacji tekstu każdemu wyrazowi lub tokenowi odpowiada wymiar, a położenie dokumentu w ramach każdego wymiaru to liczba wystąpień wyrazu w tym dokumencie. Załóżmy na przykład — pomijając kwestię odmiany słów — że dokument A zawiera siedem wystąpień wyrazu *skuteczność*, trzy wystąpienia wyrazu *transformacja* i dwa wystąpienia wyrazu *pieniężny*. Dokument B zawiera dwa wystąpienia wyrazu *skuteczność*, trzy wystąpienia wyrazu *transformacja* i nie zawiera wystąpień wyrazu *pieniężny*. Oba dokumenty zostałyby przedstawione jako wektory o współrzędnych odpowiadających wystąpieniom tych trzech wyrazów, A = <7,3,2> i B = <2,3,0>. Odległość kosinusowa tych dwóch dokumentów to:

$$\begin{aligned} d_{cos}(A,B) &= 1 - \frac{<7,3,2> \cdot <2,3,0>}{\|<7,3,2>\|_2 \cdot \|<2,3,0>\|_2} \\ &= 1 - \frac{7 \cdot 2 + 3 \cdot 3 + 2 \cdot 0}{\sqrt{49+9+4} \cdot \sqrt{4+9}} \\ &= 1 - \frac{23}{28{,}4} \approx 0{,}19 \end{aligned}$$

Odległość kosinusowa jest szczególnie przydatna, gdy chcemy zignorować różnice skali pomiędzy wystąpieniami — ujmując rzecz technicznie, chcemy zignorować wielkość wektorów. Tytułem konkretnego przykładu, w klasyfikacji tekstu moglibyśmy chcieć zignorować fakt, że

dany dokument jest znacznie dłuższy niż inny, i po prostu skoncentrować się na ich zawartości tekstowej. Załóżmy wobec tego, że w naszym wcześniejszym przykładzie dysponujemy trzecim dokumentem, C, w którym jest siedemdziesiąt wystąpień wyrazu *skuteczność*, trzydzieści wystąpień wyrazu *transformacja* i dwadzieścia wystąpień wyrazu *pieniężny*. Wektor przedstawiający C miałby współrzędne C = <70, 30, 20>. Jeśli to przeliczymy, to okaże się, że odległość kosinusowa między dokumentami A i C wynosi zero, ponieważ C to po prostu A pomnożone przez 10.

Ostatni przykład, ilustrujący zróżnicowanie miar odległości, ponownie związany jest z tekstem, ale w zupełnie inny sposób. Może się zdarzyć, że zechcemy zmierzyć odległość pomiędzy dwoma ciągami znaków. W ramach zastosowań biznesowych często musimy być w stanie ocenić, czy dwa rekordy danych dotyczą tej samej osoby. Mogą się oczywiście zdarzyć błędy. Chcielibyśmy umieć określić, na ile podobne są dwa pola tekstowe. Załóżmy, że mamy dwa ciągi:

1. ul. Poleska 11
2. ul. Podlaska 1

Chcemy ustalić, na ile są one podobne. Do tego celu przydatny jest inny typ funkcji odległości, zwany **odległością edycyjną** lub **metryką Lewensztejna**. Ta miara oblicza minimalną liczbę operacji edycyjnych niezbędnych do skonwertowania danego ciągu na inny, a operacja edycyjna to wstawianie, usuwanie lub zastępowanie znaków (można także wybrać inne operatory edycji). W przypadku naszych dwóch ciągów pierwszy można przekształcić w drugi za pomocą następującej sekwencji operacji:

1. Usuń `1`.
2. Wstaw `d`.
3. Zastąp literę `e` literą `a`.

Te dwa ciągi mają więc odległość edycyjną o wartości trzy. Podobnie możemy przeprowadzić obliczenie odległości edycyjnej w innych polach, takich jak imię i nazwisko (radząc sobie na przykład w ten sposób z brakującymi inicjałami), a następnie obliczyć podobieństwo wyższego poziomu, łączące różne podobieństwa odległości edycyjnej.

Odległość edycyjna jest również powszechnie stosowana w biologii, w której wykorzystuje się ją do pomiaru odległości genetycznej między ciągami alleli. Ogólnie rzecz biorąc, odległość edycyjna jest popularnym wyborem, gdy elementy danych składają się z ciągów lub sekwencji, w przypadku których bardzo istotna jest kolejność.

* Funkcje łączące: obliczanie wskaźników na podstawie sąsiadów

Przed nami szczegóły techniczne

W celu uzyskania pełnego obrazu omówmy także pokrótce „funkcje łączące" — wzory stosowane do obliczania predykcji wystąpienia ze zbioru najbliższych sąsiadów tego wystąpienia.

Zaczęliśmy od większości głosów, która jest prostą strategią. Tę zasadę decyzyjną możemy zobaczyć w równaniu 6.6:

Równanie 6.6. Klasyfikacja większością głosów

$$c(\mathbf{x}) = \arg\max_{c \in \text{klasy}} \text{wskaźnik}(c, sąsiedzi_k(\mathbf{x}))$$

W tym równaniu czynnik $sąsiedzi_k(\mathbf{x})$ oznacza k najbliższych sąsiadów wystąpienia $\mathbf{x}$, *arg max* to argument (w tym przypadku c), który maksymalizuje wielkość następującą po nim, a funkcja scoringowa definiowana jest jak w równaniu 6.7.

Równanie 6.7. Funkcja scoringowa większości głosów

$$\text{wskaźnik}(c, N) = \sum_{y \in N} [klasa(\mathbf{y}) = c]$$

W tym równaniu wyrażenie $[klasa(y) = c]$ ma wartość jeden, jeżeli $klasa(y) = c$, albo zero, jeżeli tak nie jest.

Głosowanie moderowane podobieństwem, które omawialiśmy w podrozdziale „Ilu sąsiadów i jak duży wpływ?", możemy uzyskać poprzez modyfikację równania 6.6, włączając do niego wagę, co widać w równaniu 6.8.

Równanie 6.8. Klasyfikacja moderowana podobieństwem

$$\text{wskaźnik}(c, N) = \sum_{y \in N} w(\mathbf{x}, \mathbf{y}) \times [klasa(\mathbf{y}) = c]$$

gdzie w jest funkcją ważącą opartą na podobieństwie pomiędzy przykładami x i y. Powszechnie wykorzystuje się odwrotność kwadratu odległości:

$$w(\mathbf{x}, \mathbf{y}) = \frac{1}{dist^2(\mathbf{x}, \mathbf{y})}$$

gdzie *dist* jest funkcją odległości wykorzystywaną w danej domenie.

Łatwo jest przekształcić równania 6.6 i 6.8, aby uzyskać wynik, który może zostać wykorzystany jako oszacowanie prawdopodobieństwa. Równanie 6.8 daje już wskaźnik, musimy więc po prostu podzielić go przez całkowite wskaźniki wniesione przez wszystkich sąsiadów, tak aby znalazł się w przedziale od zera do jednego, co widać w równaniu 6.9.

Równanie 6.9. Scoring moderowany podobieństwem

$$p(c \mid \mathbf{x}) = \frac{\sum_{y \in \text{sąsiedzi}(\mathbf{x})} w(\mathbf{x}, \mathbf{y}) \times [klasa(\mathbf{y}) = c]}{\sum_{y \in \text{sąsiedzi}(\mathbf{x})} w(\mathbf{x}, \mathbf{y})}$$

I wreszcie, za pomocą jeszcze jednego kroku możemy dokonać generalizacji tego równania, aby móc przeprowadzić regresję. Przypomnijmy, że w problemach regresji zamiast próbować oszacować klasę nowego wystąpienia x, staramy się oszacować jakąś wartość $f(\mathbf{x})$, dysponując wartościami f sąsiadów $\mathbf{x}$. Możemy po prostu zamienić umieszczoną w nawiasie część równania 6.9, odnoszącą się do konkretnej klasy, na wartości liczbowe. Pozwoli to oszacować wartość regresji jako średnią ważoną wartości wielkości docelowych sąsiadów (choć w zależności od zastosowania rozsądne może być posłużenie się alternatywnymi funkcjami łączącymi, takimi jak mediana).

Równanie 6.10. Regresja moderowana podobieństwem

$$f(\mathbf{x}) = \frac{\sum_{y \in \text{sąsiedzi}(\mathbf{x})} w(\mathbf{x},\mathbf{y}) \times t(\mathbf{y})}{\sum_{y \in \text{sąsiedzi}(\mathbf{x})} w(\mathbf{x},\mathbf{y})}$$

gdzie $t(\mathbf{y})$ jest wartością wielkości docelowej dla przykładu $\mathbf{y}$.

Zatem, szacując przykładowo oczekiwany poziom wydatków potencjalnego klienta, posiadającego określony zbiór cech, równanie 6.10 określiłoby tę kwotę jako średnią ważoną odległości historycznych kwot wydatków sąsiadów.

Klastrowanie

Jak wspomnieliśmy na początku tego rozdziału, pojęcia podobieństwa i odległości są jedną z podstaw nauki o danych. Aby w pełni je docenić, przyjrzyjmy się zupełnie innemu zadaniu. Przypomnijmy pierwsze zastosowanie nauki o danych, któremu dokładnie się przyglądaliśmy: nadzorowaną segmentację — wyszukiwanie grup obiektów, które różnią się pod względem jakiejś interesującej z naszego punktu widzenia cechy docelowej. Może to być na przykład znajdowanie grup klientów, które różnią się pod względem skłonności tychże klientów do rezygnacji z usług firmy po wygaśnięciu umowy. Dlaczego, mówiąc o nadzorowanej segmentacji, zawsze używamy określnika „nadzorowana"?

W ramach innych zastosowań mogłoby zależeć nam na znalezieniu grup obiektów, na przykład grup klientów, ale nie na podstawie jakiejś określonej z góry cechy docelowej. Czy nasi klienci w naturalny sposób należą do różnych grup? Może to być użyteczne z wielu powodów. Na przykład może chcielibyśmy spojrzeć wstecz i staranniej zastanowić się nad naszymi działaniami marketingowymi. Czy rozumiemy, kim są nasi klienci? Czy możemy wytworzyć lepsze produkty, lepsze kampanie marketingowe, lepsze metody sprzedaży albo poprawić jakość obsługi klienta poprzez uwzględnienie naturalnie istniejących podgrup? Ta koncepcja wyszukiwania naturalnych grup w danych jest nazywana nienadzorowaną segmentacją, lub prościej **klastrowaniem**.

Klastrowanie jest kolejnym zastosowaniem naszej podstawowej koncepcji podobieństwa. Zasadnicza idea polega na tym, że chcemy znaleźć grupy obiektów (klientów, firm, gatunków whisky itp.), w których obiekty w ramach grupy są do siebie podobne, ale obiekty należące do różnych grup już nie tak bardzo.

Modelowanie nadzorowane wiąże się z odkrywaniem wzorców w celu dokonania prognozy wartości określonej zmiennej docelowej na podstawie danych, w przypadku których znamy wartości zmiennej docelowej. Modelowanie nienadzorowane nie koncentruje się na zmiennej docelowej. Zamiast tego poszukuje innych rodzajów prawidłowości w zbiorze danych.

Przykład: analityka whisky — nowe spojrzenie

Zanim przejdziemy do szczegółów, przyjrzyjmy się ponownie naszemu przykładowemu problemowi analityki whisky. Omówiliśmy wykorzystywanie miar podobieństwa do znajdowania

podobnych szkockich whisky single malt. Dlaczego miałoby zależeć nam na pójściu o krok dalej i znalezieniu klastrów podobnych whisky?

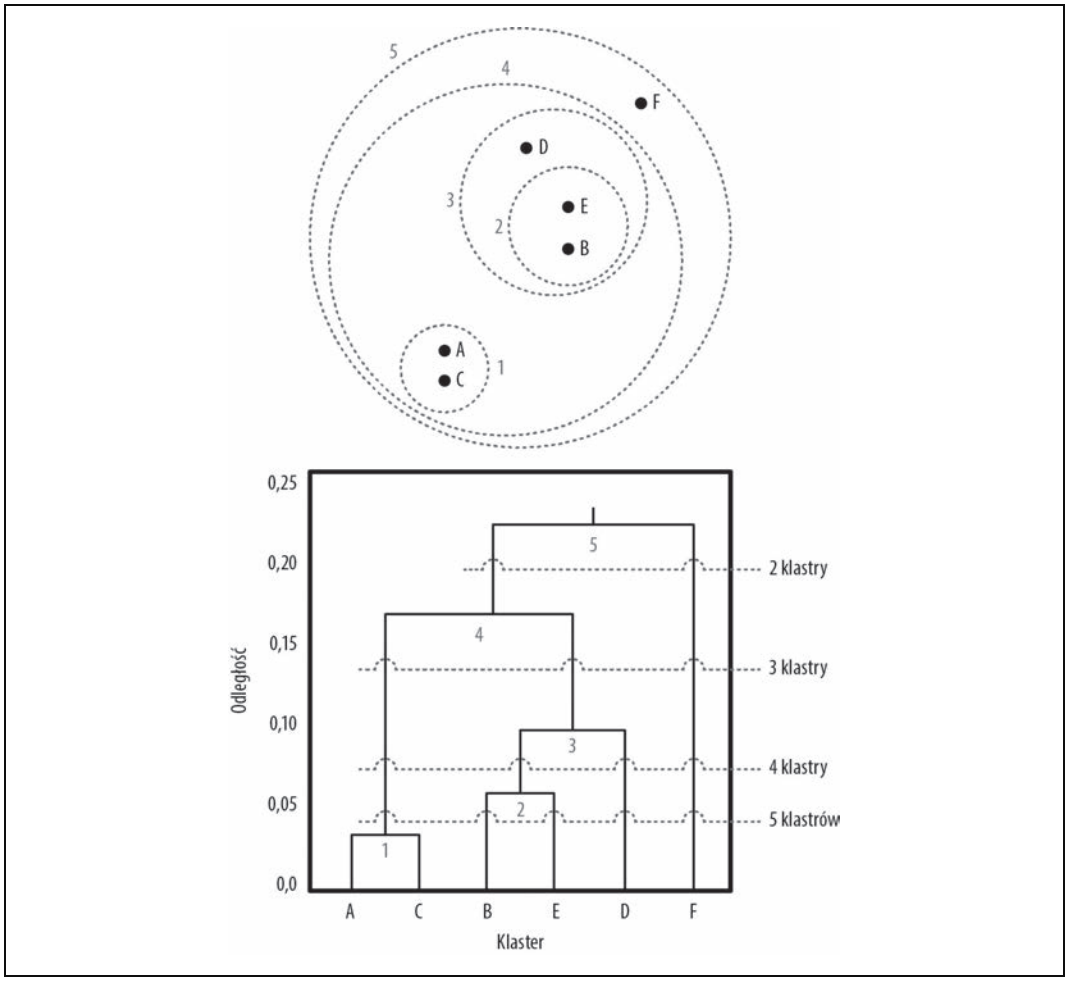
Jednym z powodów, dla których moglibyśmy poszukiwać klastrów whisky, jest po prostu chęć lepszego zrozumienia problemu. Oto przykład eksploracyjnej analizy danych, na którą firmy dysponujące bogatymi zasobami danych powinny stale poświęcać nieco energii i zasobów, bo taka eksploracja może prowadzić do pożytecznych i dochodowych odkryć. W naszym przykładzie, jeśli interesują nas szkockie whisky, to może po prostu chcemy zrozumieć ich naturalne grupowania ze względu na smak, bo chcemy zrozumieć naszą „firmę", co może doprowadzić do stworzenia lepszego produktu lub usługi. Powiedzmy, że prowadzimy mały sklep w zamożnej dzielnicy i, jako część naszej strategii biznesowej, chcemy być postrzegani jako odpowiednie miejsce do nabywania szkockiej whisky single malt. Być może nie jesteśmy w stanie zapewnić największego wyboru, ze względu na ograniczoną ilość miejsca i ograniczoną kwotę, którą możemy zainwestować w zapasy, ale moglibyśmy obrać strategię posiadania znaczącej i eklektycznej kolekcji. Gdybyśmy rozumieli, w jaki sposób whisky single malt są grupowane według smaku, to moglibyśmy (na przykład) wybrać z każdej grupy smakowej jedną popularną i jedną mniej znaną whisky. Albo drogą i mającą bardziej przystępną cenę. Każdy z tych wyborów oparty jest na dobrym zrozumieniu tego, jak whisky grupują się pod względem smaku.

Porozmawiajmy teraz o klastrowaniu bardziej ogólnie. Zaprezentujemy dwa główne rodzaje klastrowania, ilustrując koncepcję podobieństwa na rzeczywistych przykładach. W czasie trwania tego procesu będziemy mogli zbadać rzeczywiste klastry whisky.

Klastrowanie hierarchiczne

Zacznijmy od bardzo prostego przykładu. W górnej części rysunku 6.6 widzimy sześć punktów, od A do F, umieszczonych na płaszczyźnie (tj. w dwuwymiarowej przestrzeni wystąpień). Przy wykorzystaniu odległości euklidesowej uzyskujemy punkty, które są tym bardziej podobne do innych, im bliżej nich znajdują się na płaszczyźnie. Okręgi oznaczone 1 – 5 są umieszczone wokół punktów, wskazując *klastry*. Diagram przedstawia kluczowe aspekty tak zwanego klastrowania „hierarchicznego". Jest to *klastrowanie*, ponieważ grupuje punkty według ich podobieństwa. Zwróćmy uwagę, że jedynymi przypadkami zachodzenia na siebie klastrów są sytuacje, kiedy jeden klaster zawiera w sobie inne klastry. Ze względu na tę strukturę okręgi przedstawiają faktycznie hierarchię klastrowań. Najbardziej ogólne klastrowanie (najwyższy poziom) to po prostu pojedynczy klaster, który zawiera wszystko — w przykładzie to klaster 5. Klastrowanie najniższego poziomu ma miejsce, gdy usuniemy wszystkie okręgi i same punkty będą sześcioma (trywialnymi) klastrami. Usuwanie okręgów w kolejności malejącej według cyfr, którymi są one oznaczone na rysunku, tworzy zestaw różnych klastrowań, z których każde ma większą liczbę klastrów.

Wykres na dole rysunku nazywa się **dendrogram** i wyraźnie ukazuje hierarchię klastrów. Wzdłuż osi x umieszczone są (w przypadkowej kolejności, ale tak, aby uniknąć krzyżowania się linii) poszczególne punkty danych. Oś y oznacza odległość między klastrami (opowiemy o tym szerzej niebawem). Na dole ($y = 0$) każdy punkt znajduje się w oddzielnym klastrze. Ze wzrostem wartości y różne grupy klastrów zaczynają się łączyć; najpierw grupują się A i C, potem B i E, a następnie klaster BE łączy się z D i tak dalej, aż wszystkie klastry połączą się na górze. Cyfry w spojeniach dendrogramów odpowiadają ponumerowanym okręgom w górnym diagramie.



Rysunek 6.6. Sześć punktów i ich możliwe klastrowania. Na górze widać sześć punktów, od A do F, z okręgami 1 – 5 ukazującymi różne możliwe grupowania oparte na odległości. Grupy te tworzą ukrytą hierarchię. Na dole znajduje się dendrogram odpowiadający grupowaniom, dzięki któremu hierarchia staje się wyraźna

Obie części rysunku 6.6 pokazują, że hierarchiczne klastrowanie nie tworzy po prostu „zgrupowania", czy też pojedynczego zbioru grup obiektów. Tworzy ono zestaw sposobów grupowania punktów. Aby zobaczyć to wyraźnie, spróbujmy „przeciąć" dendrogram poziomą linią, ignorując wszystko powyżej niej. Przesuwając linię w dół, otrzymujemy, co widać na rysunku, różne klastrowania z rosnącą liczbą klastrów. Przecinając dendrogram linią oznaczoną „2 klastry", widzimy poniżej dwie różne grupy; tutaj zbiór jednoelementowy w postaci punktu F i grupę zawierającą wszystkie pozostałe punkty. Wracając do górnej części rysunku, dostrzegamy, że punkt F rzeczywiście odstaje od pozostałych. Przecięcie dendrogramu na poziomie 2 klastrów odpowiada usunięciu okręgu numer 5. Jeśli przeniesiemy się niżej, do poziomej linii oznaczonej jako „3 klastry", i tam przetniemy dendrogram, to zauważymy, że w dendrogramie poniżej linii pozostaną trzy grupy (AC, BED, F), co na wykresie odpowiada usunięciu okręgów 5 i 4, po czym zobaczymy tam te same trzy klastry. Klastry są intuicyjnie zrozumiałe.

Punkt F nadal pozostaje samotny. Punkty A i C tworzą ścisłą grupę. Punkty B, E i D także tworzą ścisłą grupę.

Zaletą klastrowania hierarchicznego jest to, że umożliwia ono analitykowi danych zobaczenie zgrupowań — „krajobrazu" podobieństwa danych — zanim podejmie on decyzję o liczbie klastrów, które chce uzyskać. Jak widać na przykładzie poziomych przerywanych linii, diagram może zostać przecięty na każdym poziomie, w wyniku czego otrzymujemy taką liczbę klastrów, na jakiej nam zależy. Należy również zauważyć, że gdy dwa klastry zostaną połączone na określonym poziomie, to pozostają połączone na wszystkich wyższych poziomach hierarchii.

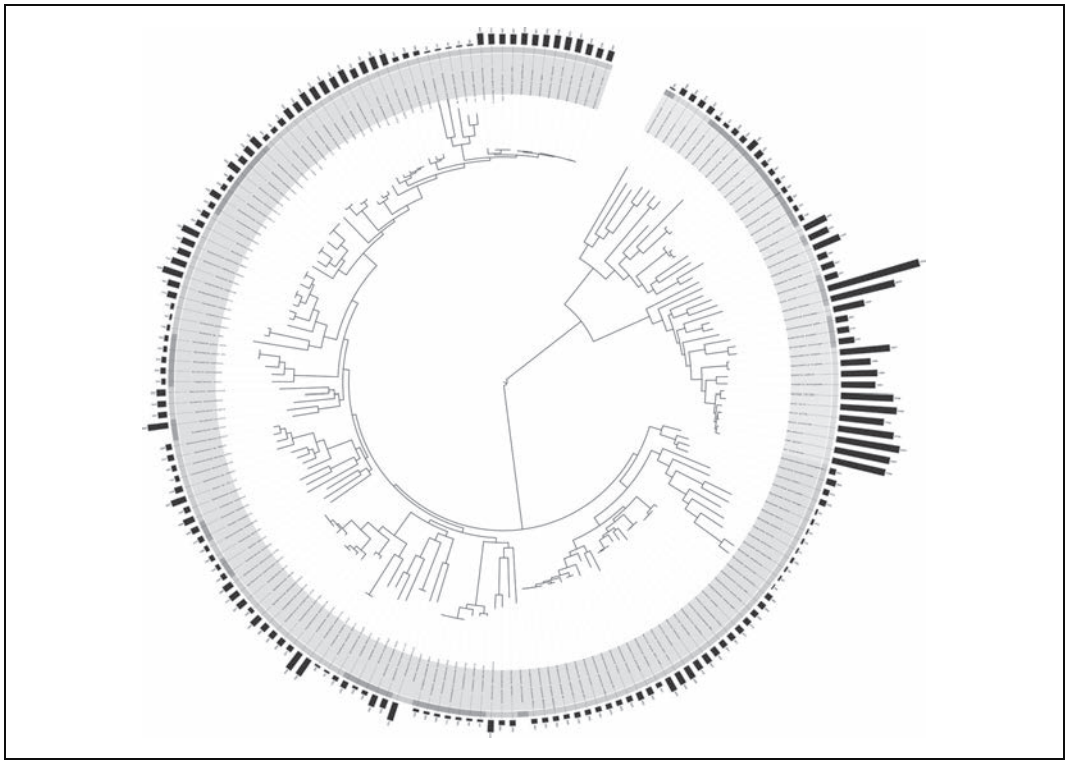
Hierarchiczne klastrowania są zwykle tworzone, poczynając od każdego węzła, traktowanego jako jego własny klaster. Następnie klastry są łączone iteracyjnie, do czasu gdy pozostanie tylko jeden klaster. Klastry są łączone na podstawie podobieństwa czy też wybranej funkcji odległości. Do tej pory omawialiśmy odległość pomiędzy wystąpieniami. Dla klastrowania hierarchicznego będziemy potrzebowali funkcji odległości pomiędzy klastrami traktującej pojedyncze wystąpienia jako najmniejsze klastry. Nazywa się to czasem funkcją **powiązania**. Funkcją powiązania mogłaby więc być na przykład „odległość euklidesowa pomiędzy najbliższymi punktami w każdym z klastrów", co odnosiłoby się do dowolnych dwóch klastrów.

Uwaga: dendrogramy

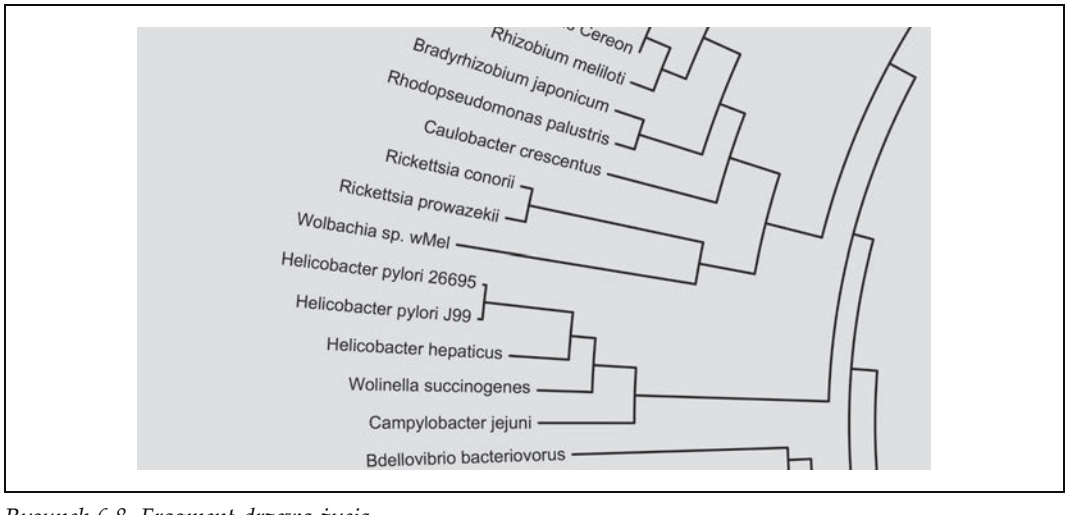
W dendrogramie można zazwyczaj zauważyć dwie kwestie. Ponieważ oś *y* oznacza odległość między klastrami, dendrogram może dać wyobrażenie o tym, gdzie mogą pojawić się naturalne klastry. Zauważmy, że w dendrogramie z rysunku 6.6 mamy do czynienia z relatywnie dużą odległością pomiędzy klastrem 3 (na wysokości mniej więcej 0,10) oraz klastrem 4 (około 0,17). Sugeruje to, że segmentacja danych, w wyniku której otrzymujemy trzy klastry, może być dobrym podziałem. Zwróćmy również uwagę na punkt F dendrogramu. Zawsze, gdy pojedynczy punkt łączy się z innymi wysoko w dendrogramie, mamy do czynienia ze wskazówką, że wydaje się on różnić od pozostałych; moglibyśmy więc nazwać go „elementem odstającym", który warto zbadać.

Jednym z najbardziej znanych zastosowań hierarchicznego klastrowania jest „drzewo życia" (Sugden i in., 2003; Pennisi, 2003), hierarchiczny, filogenetyczny wykres wszelkich form życia na Ziemi. Wykres ten jest oparty na hierarchicznym klastrowaniu sekwencji RNA. Część drzewa ze strony *Interactive Tree of Life* (*http://itol.embl.de/index.shtml*) widoczna jest na rysunku 6.7 (Letunic i Bork, 2006). Wielkie hierarchiczne drzewa są często przedstawiane w postaci kolistej, w celu zaoszczędzenia miejsca, tak jak tutaj. Diagram przedstawia globalną filogenezę (taksonomię) w pełni zsekwencjonowanych genomów, automatycznie zrekonstruowaną przez Francescę Ciccarelli i jej współpracowników (2006). W centrum znajduje się „ostatni uniwersalny wspólny przodek" wszelkich form życia na Ziemi, od którego pochodzą trzy domeny organizmów żywych (eukarionty, bakterie i archeony). Rysunek 6.8 przedstawia powiększoną część tego drzewa, odnoszącą się do konkretnej bakterii, *Helicobacter pylori*, która powoduje powstawanie wrzodów.

Wracając do naszego przykładu z początku tego rozdziału, górna część rysunku 6.9 pokazuje w formie dendrogramu 50 gatunków szkockiej whisky single malt, zgrupowanych przy użyciu metodologii opisanej przez Lapointe'a i Legendre'a (1994). Przecinając dendrogram, możemy uzyskać dowolną liczbę klastrów, na której nam zależy, a więc na przykład usuwając 11 najwyżej położonych segmentów łączących, otrzymamy 12 klastrów.

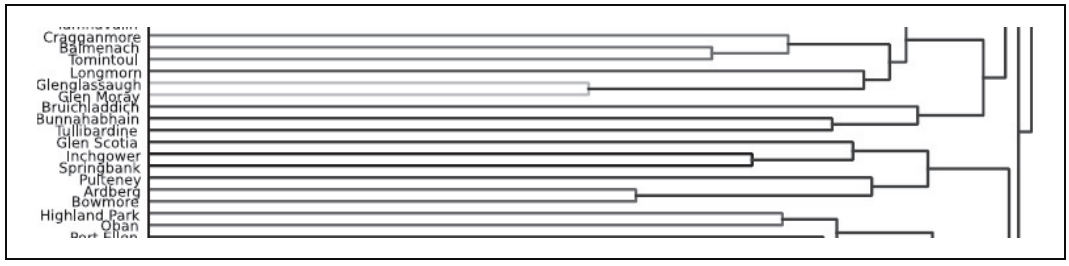


Rysunek 6.7. Filogenetyczne drzewo życia, ogromne hierarchiczne klastrowanie gatunków, przedstawione w formie koła



Rysunek 6.8. Fragment drzewa życia

W dolnej części rysunku 6.9 widać zbliżenie fragmentu hierarchii, w którym koncentrujemy się na nowej ulubionej whisky Fostera, Bunnahabhain. Poprzednio, w podrozdziale „Przykład: analityka whisky", znaleźliśmy whisky do niej podobne. Ten fragment pokazuje, że większość



Rysunek 6.9. Klastrowanie hierarchiczne szkockich whisky. Niewielki wycinek hierarchii, ukazujący Bunnahabhain i jej sąsiadów

jej najbliższych sąsiadów (Tullibardine, Glenglassaugh itd.) rzeczywiście znajduje się w jej pobliżu w hierarchii. (Być może zastanawiasz się, dlaczego klastry nie odpowiadają *dokładnie* uszeregowaniu podobieństwa. Powodem jest to, że chociaż pięć whisky, które znaleźliśmy, jest najbardziej podobnych do Bunnahabhain, to niektóre z nich są bardziej podobne do innych whisky w zbiorze danych, a więc są klastrowane z tymi bliższymi sąsiadami, przed zestawieniem z Bunnahabhain).

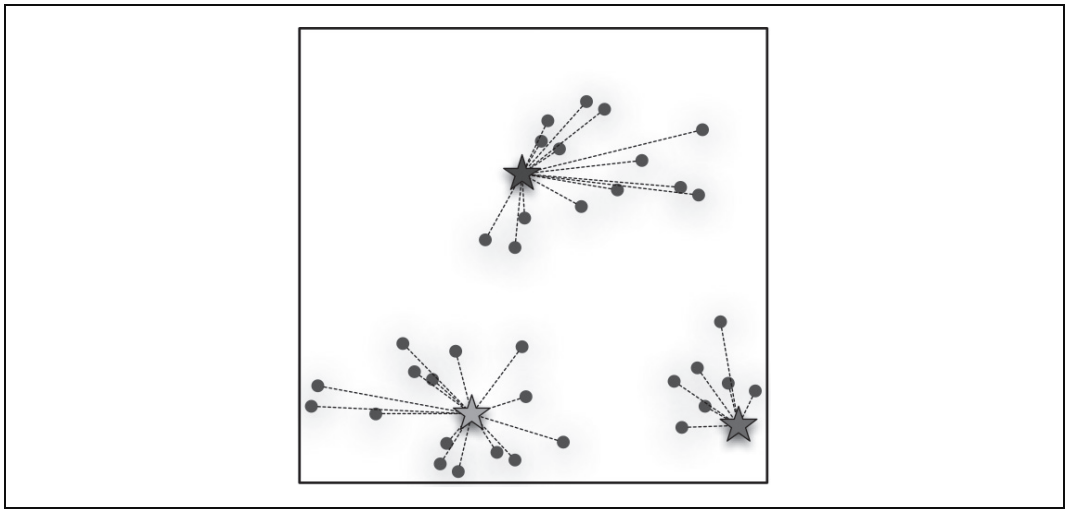
Z punktu widzenia klasyfikacji whisky interesujący jest fakt, że poszczególne grupy whisky single malt, wynikające z tego klastrowania opartego na smaku, nie odpowiadają dokładnie regionom Szkocji — będącym zazwyczaj podstawą kategoryzacji szkockich whisky. Jak wykazują Lapointe i Legendre (1994), istnieje tutaj jednak korelacja.

Zamiast więc mieć na składzie najbardziej rozpoznawalne szkockie whisky albo po kilka marek z regionów Highland, Lowland i Islay, nasz właściciel specjalistycznego sklepu z alkoholami może wybrać słodowe whisky z różnych klastrów. Można byłoby także stworzyć przewodnik po szkockich whisky, który mógłby wspomagać jej amatorów w wyborze[6]. Skoro Foster przepada na przykład za Bunnahabhain, którą polecił mu kiedyś kolega w restauracji, to klastrowanie sugeruje zbiór innych „najbardziej podobnych" whisky (Bruichladdich, Tullibardine itp.). Najbardziej niecodziennie smakującą w tym zestawieniu wydaje się być Aultmore, znajdująca się na samej górze, będąca ostatnią whisky, która dołącza do pozostałych.

Najbliżsi sąsiedzi na nowo: klastrowanie wokół centroidów

Klastrowanie hierarchiczne koncentruje się na podobieństwach pomiędzy poszczególnymi wystąpieniami i na tym, w jaki sposób te podobieństwa je łączą. Innym sposobem myślenia o klastrowaniu danych jest koncentrowanie się na klastrach jako takich — na grupach wystąpień. Najpopularniejszą metodą koncentrowania się na klastrach jako takich jest reprezentacja każdego klastra za pomocą jego „centrum", zwanego **centroidem**. Rysunek 6.10 ilustruje tę koncepcję w dwóch wymiarach: mamy tutaj trzy klastry, których wystąpienia przedstawione są jako kółka. Każdy klaster posiada centroid, przedstawiony jako gwiazdka o ciągłym obwodzie. Gwiazdka nie musi być jednym z wystąpień; to geometryczny środek grupy wystąpień. Ta sama koncepcja ma zastosowanie do dowolnej liczby wymiarów, o ile dysponujemy liczbową przestrzenią wystąpień i miarą odległości (w przestrzeni wysoko wymiarowej nie będziemy oczywiście w stanie odwzorować klastrów tak dokładnie, o ile w ogóle uda nam się to zrobić).

[6] Już to zrobiono: patrz książka Davida Wisharta *Whisky. Leksykon smakosza*.

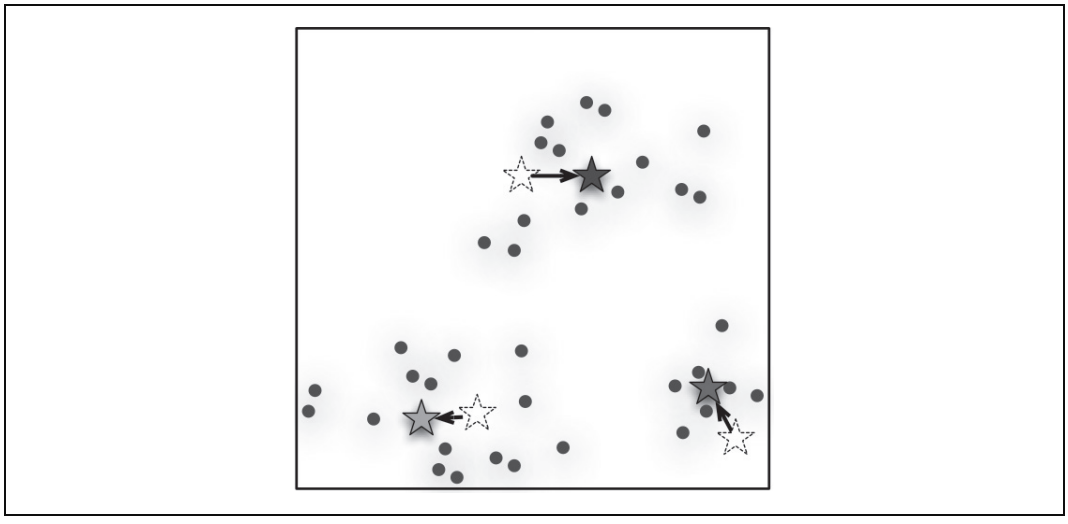


Rysunek 6.10. Drugi etap algorytmu k-średnich: znajdowanie rzeczywistych centrów klastrów określonych na pierwszym etapie

Najpopularniejszy algorytm klastrowania z wykorzystaniem centroidów to algorytm ***k*-średnich** (MacQueen, 1967; Lloyd, 1982; MacKay, 2003), a koncepcja stanowiąca jego podstawę zasługuje na szersze omówienie, bo o klastrowaniu algorytmem *k*-średnich często wspomina się w sferze nauki o danych. W algorytmie *k*-średnich „średnie" to centroidy reprezentowane przez średnie arytmetyczne wartości współrzędnych każdego wymiaru dla wystąpień zawartych w klastrze. Zatem na rysunku 6.10, aby obliczyć wartość centroidu dla każdego klastra, powinniśmy uśrednić wszystkie wartości współrzędnej x punktów znajdujących się w klastrze w celu określenia współrzędnej x centroidu oraz uśrednić wszystkie wartości y, aby określić jego współrzędną y. Mówiąc ogólnie, centroid to średnia wartości każdej cechy każdego przykładu w klastrze. Wynik przedstawiono na rysunku 6.10.

Litera k w *k*-średnich to po prostu liczba klastrów, które chcielibyśmy wyszukać w danych. W przeciwieństwie do klastrowania hierarchicznego, *k*-średnich rozpoczyna się od żądanej liczby klastrów k. Wobec tego na rysunku 6.11 analityk określiłby $k = 3$, a w wyniku zastosowania metody klastrowania *k*-średnich otrzymalibyśmy (i) trzy centroidy klastrów po jej zakończeniu (trzy gwiazdki o ciągłym obwodzie na rysunku 6.10), oraz (ii) informację dotyczącą tego, które z punktów danych należą do każdego klastra. Czasami określa się to jako klastrowanie najbliższego sąsiedztwa, ponieważ odpowiedzią na pytanie (ii) jest po prostu stwierdzenie, że każdy klaster zawiera te punkty, które znajdują się najbliżej jego centroidu (a nie któregoś z pozostałych centroidów).

Algorytm wyszukiwania klastrów *k*-średnich jest prosty i elegancki, więc warto o nim wspomnieć. Przedstawiony został na rysunku 6.11 i rysunku 6.10. Algorytm zaczyna się od utworzenia k początkowych centrów klastrów, zwykle losowo, ale czasami wybierając k spośród rzeczywistych punktów danych albo wykorzystując punkty początkowe określone przez użytkownika lub wstępne przetwarzanie danych, w celu wyznaczenia dobrego zbioru początkowych centrów (Arthur i Vassilvitskii, 2007). Potraktujmy gwiazdki na rysunku 6.11 jako te początkowe ($k = 3$) centra klastrów. Następnie algorytm postępuje w następujący sposób. Jak widać na rysunku 6.11, tworzone są klastry odpowiadające tym centrom klastrów na podstawie określenia które centrum jest najbliższe któremu punktowi.



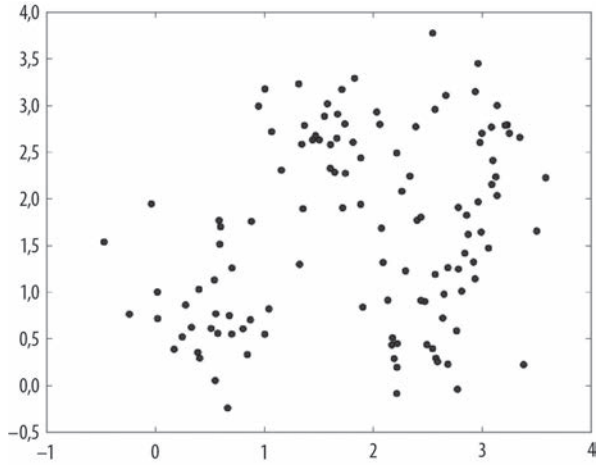
Rysunek 6.11. Pierwszy etap algorytmu k-średnich: znajdowanie punktów najbliższych wybranym (zwykle losowo) centrom. Powstaje pierwszy zbiór klastrów

Następnie dla każdego z tych klastrów jego centrum jest przeliczane poprzez wyszukanie rzeczywistego centroidu punktów w klastrze. Jak widać na rysunku, centra klastrów zwykle się przemieszczają; z diagramu widzimy, że nowe gwiazdki o ciągłym obwodzie znajdują się rzeczywiście bliżej punktów, które intuicyjnie wydawały się być centrami każdego klastra. I to w zasadzie wszystko. Proces po prostu się powtarza: skoro centra klastrów przesunęły się, to musimy przeliczyć, które punkty należą do każdego klastra (jak na rysunku 6.11). Gdy zostaną one przypisane, to być może będziemy musieli ponownie przesunąć centra klastrów. Procedura *k*-średnich powtarza się dopóty, dopóki w klastrach przestaną zachodzić zmiany (lub ewentualnie do czasu, aż spełnione zostanie jakieś inne kryterium zatrzymania).

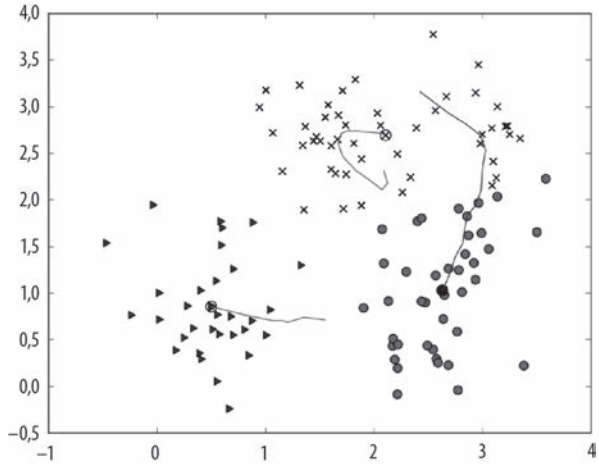
Rysunki 6.12 i 6.13 przedstawiają przykładowy przebieg algorytmu *k*-średnich na podstawie 90 punktów danych, przy $k = 3$. Ten zbiór danych jest nieco bardziej realistyczny, bo nie ma w nim tak doskonale zdefiniowanych klastrów, jak w poprzednim przykładzie. Rysunek 6.12 przedstawia początkowe punkty danych, przed klastrowaniem. Na rysunku 6.13 widać ostateczne wyniki klastrowania po 16 iteracjach. Trzy (nieregularne) linie ukazują drogę od początkowej (losowej) lokalizacji każdego z centroidów do jego lokalizacji ostatecznej. Punkty w trzech klastrach są oznaczone różnymi symbolami (kółkami, iksami i trójkątami).

Nie ma gwarancji, że jednokrotne uruchomienie algorytmu *k*-średnich doprowadzi do powstania odpowiednich klastrów. W wyniku jednokrotnego cyklu powstanie lokalne optimum — lokalnie najlepsze klastrowanie — ale będzie to uzależnione od początkowych lokalizacji centroidów. Z tego powodu *k*-średnich zwykle powtarza się wielokrotnie, rozpoczynając za każdym razem od szeregu losowych centroidów. Wyniki można porównać, badając klastry (więcej na ten temat za chwilę), lub za pomocą miary liczbowej, takiej jak **dystorsja** klastra, która jest sumą kwadratów różnic między każdym punktem danych i odpowiadającym mu centroidem. W tym ostatnim przypadku klastrowanie o najniższej wartości dystorsji może zostać uznane za najlepsze.

Z punktu widzenia czasu pracy algorytm *k*-średnich jest wydajny. Nawet w przypadku wielu powtórzeń jest on ogólnie stosunkowo szybki, ponieważ w ramach każdej iteracji oblicza tylko



Rysunek 6.12. Przykład klastrowania algorytmem k-średnich przy wykorzystaniu 90 punktów na płaszczyźnie i k = 3 centroidów. Ten rysunek przedstawia początkowy zbiór punktów



Rysunek 6.13. Przykład klastrowania metodą k-średnich, przy wykorzystaniu 90 punktów na płaszczyźnie i k = 3 centroidów. Rysunek przedstawia drogi centroidów (każda z trzech linii), w trakcie 16 iteracji algorytmu klastrowania. Znaczniki określające daną klasę punktów identyfikują klastry, do których zostały one ostatecznie przypisane

odległości między każdym punktem danych i centrami klastrów. Klastrowanie hierarchiczne jest zasadniczo wolniejsze, ponieważ algorytm musi znać odległości pomiędzy wszystkimi parami klastrów w każdej iteracji, co początkowo oznacza wszystkie pary punktów danych.

Powszechnym problemem związanym z algorytmami opartymi na centroidach, takimi jak *k*-średnich, jest określenie, w jaki sposób wyznaczyć właściwe wartości dla *k*. Jednym ze sposobów jest po prostu eksperymentowanie z różnymi wartościami *k* i sprawdzanie, które z nich generują dobre wyniki. Ponieważ *k*-średnich jest często wykorzystywany w eksploracyjnej

ekstrakcji danych, to analityk i tak musi badać wyniki klastrowania, aby ustalić, czy klastry są odpowiednie. Zazwyczaj można dzięki temu stwierdzić, czy właściwa jest liczba klastrów. Wartość *k* można zmniejszyć, jeżeli część klastrów jest zbyt mała i nadmiernie skupiona, oraz zwiększyć, jeżeli część z nich jest zbyt duża i rozproszona.

Aby znaleźć bardziej obiektywną miarę, analityk może eksperymentować ze zwiększaniem wartości *k* i obserwować wykresy różnych wskaźników (czasami nazywanych **indeksami**) jakości uzyskanych klastrowań. Gdy wartość *k* się zwiększa, wskaźniki jakości powinny się ostatecznie ustabilizować, czy też wypłaszczyć, osiągając albo poziom najniższy, jeżeli wskaźnik miał być zminimalizowany, albo najwyższy, jeśli miał on zostać zmaksymalizowany. Niezbędny jest tu pewien element oceny, ale często dobrym wyborem bywa minimalna wartość *k* w miejscu, w którym wykres zaczyna się stabilizować. Różnorodne miary oceny zbiorów potencjalnie użytecznych klastrów przedstawiono w Wikipedii, w artykule *Determining the number of clusters in a data set* (*https://en.wikipedia.org/w/index.php?title=Determining_the_number_of_clusters_in_a_data_set&oldid=526596002*).

Przykład: klastrowanie wiadomości biznesowych

Jako konkretny przykład klastrowania opartego na centroidach, rozważmy zadanie identyfikacji pewnych naturalnych zgrupowań wiadomości biznesowych, podawanych przez agregator wiadomości. Celem tego przykładu jest określenie w sposób nieformalny różnych ugrupowań wiadomości o danej firmie. Może to być przydatne do konkretnego zastosowania, na przykład aby szybko uzyskać zrozumiałe informacje o firmie, bez konieczności czytania całych artykułów, by móc dokonywać kategoryzacji nadchodzących wiadomości na potrzeby procesu priorytetyzacji wiadomości lub po prostu zrozumieć dane przed rozpoczęciem konkretnego projektu z zakresu eksploracji danych, takiego jak łączenie wiadomości biznesowych z notowaniami akcji.

W tym przykładzie wykorzystaliśmy ogromny zestaw wiadomości (tekstowych): Thomson Reuters Text Research Collection (TRC2), *http://trec.nist.gov/data/reuters/reuters.html*, czyli korpus wiadomości wytworzonych przez Agencję Reutera i udostępniony badaczom. Cały korpus obejmuje 1 800 370 wiadomości, zebranych od stycznia 2008 r. do lutego 2009 r. (w ciągu 14 miesięcy). Aby przykład był przejrzysty, ale realistyczny, zamierzamy wyodrębnić tylko te wiadomości, które odnoszą się do konkretnej firmy, w tym przypadku Apple (którego symbolem giełdowym jest AAPL).

Przygotowanie danych

Do celów tego przykładu przydatne będzie bardziej szczegółowe omówienie procesu przygotowywania danych, ponieważ będziemy traktować tekst jako dane, a wcześniej tej kwestii nie omawialiśmy. W rozdziale 10. znajdziesz więcej szczegółów dotyczących eksploracji tekstu.

W ramach tego korpusu o dużych firmach wspomina się zawsze, gdy są one głównym tematem wiadomości, na przykład w raportach o dochodach i informacjach dotyczących fuzji; często jednak są one także wymieniane peryferyjnie, w cotygodniowych podsumowaniach aktywności biznesowej, zestawieniach aktywnych akcji i wiadomościach o znaczących wydarzeniach w sektorach ich działalności. W wielu wiadomościach z branży informatycznej wspomina się na przykład o tym, w jaki sposób zareagowały w danym dniu ceny akcji HP i Della, nawet jeśli żadna z tych firm nie brała udział w wydarzeniu. Z tego powodu wyodrębniliśmy wia-

domości, w których tytułach wymieniony został konkretnie Apple — zyskując tym samym graniczące z pewnością przekonanie, że dana wiadomość z dużą dozą prawdopodobieństwa dotyczy rzeczywiście Apple. Takich wiadomości było 312, ale, jak się przekonamy, obejmowały one szeroki zakres tematów.

Przed klastrowaniem wiadomości przeszły proces podstawowego przetwarzania tekstu, usunięto znaczniki HTML i adresy URL oraz ujednolicono sposób formatowania tekstu. Słowa, które w ramach korpusu występowały rzadko (w mniej niż dwóch dokumentach) lub zbyt często (w ponad 50% dokumentów), zostały wyeliminowane, a pozostałe utworzyły *słownik* niezbędny na kolejnym etapie. Każdy dokument został następnie przedstawiony w postaci liczbowego wektora cech z wykorzystaniem „punktacji TFIDF", oceniającej każde słowo ze słownika w dokumencie. Punktacja TFIDF (*Term Frequency times Inverse Document Frequency*) określa częstotliwość występowania danego słowa w dokumencie, karaną przez częstotliwość występowania tego słowa w korpusie. TFIDF opisujemy szczegółowo w dalszej części książki, w rozdziale 10.

Posłużyliśmy się miarą podobieństwa w postaci podobieństwa kosinusowego, opisanego w podrozdziale „* Inne funkcje odległości" (równanie 6.5). Jest ono powszechnie używane w zastosowaniach związanych z tekstem do pomiaru podobieństwa dokumentów.

Klastry wiadomości

Postanowiliśmy przeprowadzić klastrowanie wiadomości na dziewięć grup (a więc $k = 9$ dla k-średnich). Poniżej przedstawiamy opis klastrów, wraz z nagłówkami zawartych w nich wiadomości. Należy pamiętać, że w procesie klastrowania wykorzystywane były całe wiadomości, a nie tylko nagłówki.

Klaster 1. Te wiadomości to informacje analityków dotyczące zmian notowań i korekt kursów:

- `RBC PODNOSI WYCENĘ APPLE <AAPL.O> DO $ 200 Z $ 190; PODTRZYMANIE REKOMENDACJI AKUMULUJ`
- `THINKPANMURE PRZYZNAJE APPLE <AAPL.O> REKOMENDACJA: KUPUJ; WYCENA $ 225`
- `AMERICAN TECHNOLOGY PODNOSI RATING APPLE <AAPL.O> Z NEUTRALNEGO DO KUPUJ`
- `CARIS PODNOSI WYCENĘ APPLE <AAPL.O> DO $ 200 Z $ 170; REKOMENDACJA: POWYŻEJ ŚREDNIEJ`
- `CARIS OBNIŻA WYCENĘ APPLE <AAPL.O> DO $ 155 Z $ 165; UTRZYMUJE REKOMENDACJĘ POWYŻEJ ŚREDNIEJ`

Klaster 2. Ten klaster zawiera wiadomości o ruchach cen akcji Apple w czasie i po zakończeniu każdego dnia obrotu:

- `akcje Apple odrabiają straty, nadal zniżka o 5 pkt.`
- `Apple rośnie o 5 pkt. w reakcji na dobre wyniki`
- `akcje Apple rosną wobec optymizmu odnośnie popytu na iPhone'a`
- `akcje Apple spadają przed wtorkowym wydarzeniem`
- `akcje Apple szybują, inwestorzy zadowoleni z wyceny`

Klaster 3. W roku 2008 pojawiało się wiele wiadomości o Steve'ie Jobsie, charyzmatycznym prezesie firmy Apple, i o jego walce z rakiem trzustki. Pogarszający się stan zdrowia Jobsa był częstym tematem dyskusji, a w wielu wiadomościach biznesowych spekulowano na temat tego, jak Apple poradzi sobie bez niego. Tego rodzaju wiadomości widać poniżej:

- `ANALYSIS – sukces Apple powiązany nie tylko ze Steve'm Jobsem`
- `NEWSMAKER – jako twarz Apple Jobs był brawurowo charyzmatyczny`

- COLUMN – Co Apple straci bez Steve'a: Eric Auchard
- Apple może stanąć w obliczu pozwów w związku ze zdrowiem Jobsa
- INSTANT VIEW 1 – Prezes Apple Jobs ma wziąć urlop zdrowotny
- ANALYSIS – Inwestorzy obawiają się Apple bez Jobsa

Klaster 4. Ten klaster zawiera różne komunikaty i informacje o nowych produktach Apple. Te wiadomości były na pierwszy rzut oka podobne, choć konkretne tematy się różniły:

- Apple wprowadza oprogramowanie „push" na iPhone'a
- Dyrektor finansowy Apple prognozuje w II kwartale marżę na poziomie około 32%
- Apple twierdzi, że jest pewne celów sprzedażowych iPhone'a w 2008r.
- Dyrektor finansowy Apple oczekuje niskiej marży brutto w 3. kwartale
- Apple ma przedstawić plany dotyczące oprogramowania na iPhone'y w dniu 6 marca

Klaster 5. W tym klastrze znalazły się wiadomości o iPhone'ie i informacje o kontraktach na jego sprzedaż w innych krajach:

- MegaFon mówi o sprzedaży Apple iPhone'a w Rosji
- Umowa Apple z tajskim True Move dotycząca sprzedaży iPhone'a 3G
- rosyjscy detaliści mają zacząć sprzedaż iPhone'a firmy Apple 3 października
- Rozmowy tajskiego AIS z Apple dotyczące premiery iPhone'a
- Softbank informuje o sprzedaży iPhone'a firmy Apple w Japonii

Klaster 6. Jedna z klas wiadomości dotyczy ruchów cen akcji poza normalnymi godzinami obrotu (tzw. Before the Bell i After the Bell):

- Before the Bell – Apple nieco w górę na skutek działań brokerów
- Before the Bell – Akcje Apple w górę o 1,6 pkt przed rozpoczęciem notowań
- Before the Bell – Apple w dół wobec niskich ocen brokerów
- After the Bell – Akcje Apple spadają
- After the Bell – Akcje Apple kontynuują spadek

Klaster 7. Ten klaster nie był spójny tematycznie:

- ANALYSIS – Mniej radości w obliczu niepewnego dla Apple w roku 2009
- TAKE A LOOK – Konwencja Apple MacWorld
- TAKE A LOOK – Konwencja Apple MacWorld
- Apple i płaski laptop, wypożyczalnie filmów online
- Jobs z Apple kończy przemówienie, ogłaszając plany filmowe

Klaster 8. W tym klastrze znalazły się wiadomości o iTunes oraz o pozycji Apple w branży muzyki dystrybuowanej cyfrowo:

- PluggedIn – Nokia w bitwie z Apple o muzykę dystrybuowaną cyfrowo
- Apple iTunes drugim w USA detalistą w sferze muzyki dystrybuowanej cyfrowo
- Apple wyprzedza konkurencję dzięki iTunes
- Nokia przyjmuje muzyczne wyzwanie Apple, telefon z ekranem dotykowym
- Rozmowy Apple z wytwórniami płytowymi o braku ograniczeń dotyczących muzyki

Klaster 9. Szczególnym rodzajem wiadomości Agencji Reutera są tzw. News Briefs, które zwykle obejmują kilka linijek bardzo lakonicznego tekstu (np. „`x potwierdza dostępność nowych filmów na iTunes w dniu premiery DVD`"). Zawartość tych krótkich wiadomości była różna, ale ze względu na bardzo podobną formę zostały one umieszczone w jednym klastrze:

- `BRIEF – Apple wypuszcza Safari 3.1`
- `BRIEF – Apple wprowadza iLife 2009`
- `BRIEF – Apple zapowiada iPhone 2.0, oprogramowanie beta`
- `BRIEF – Apple ma oferować filmy na iTunes w dniu premiery DVD`
- `BRIEF – Apple informuje o sprzedaży miliona iPhone'ów 3G w ciągu pierwszego weekendu`

Jak widać, niektóre z tych klastrów są interesujące i tematycznie spójne, a inne nie. Niektóre z nich są tylko zbiorami powierzchownie podobnych tekstów. W dziedzinie statystyki popularna jest stara maksyma: *Korelacja nie jest przyczynowością*, która oznacza, że fakt współwystępowania dwóch elementów nie oznacza, że jeden powoduje drugi. Podobne zastrzeżenie w sferze klastrowania mogłoby brzmieć: *Podobieństwo syntaktyczne nie jest podobieństwem semantycznym*. To, że dwa obiekty — a w szczególności fragmenty tekstu — mają wspólne cechy zewnętrzne, nie oznacza, że koniecznie muszą być one powiązane semantycznie. Nie powinniśmy oczekiwać, że każdy klaster będzie istotny i ciekawy. Tym niemniej, klastrowanie często bywa użytecznym narzędziem do odkrywania w naszych danych struktury, której nie przewidzieliśmy. Klastry mogą sugerować nowe i interesujące możliwości eksploracji danych.

Zrozumienie wyników klastrowania

Skoro już sformułowaliśmy wystąpienia i przeprowadziliśmy klastrowanie, to co dalej? Jak wspomnieliśmy powyżej, wynik klastrowania to albo dendrogram, albo zbiór centrów klastrów oraz odpowiadające im punkty danych dla każdego klastra. Jak mamy rozumieć klastrowanie? To szczególnie istotne, ponieważ jest ono często wykorzystywane w analizie eksploracyjnej. Chodzi więc o to, żeby zrozumieć, czy coś zostało odkryte, a jeśli tak, to co?

Sposób rozumienia klastrowania i klastrów zależy od rodzaju danych podlegających klastrowaniu i od obszaru zastosowania, ale istnieje kilka metod, które stosowane są szeroko. Widzieliśmy już niektóre z nich w akcji.

Zastanówmy się nad naszym przykładem z whisky. Badacze whisky, Lapointe i Legendre, przecięli swój dendrogram na poziomie 12 klastrów; oto dwa z nich:

Grupa A
: **Whisky:** Aberfeldy, Glenugie, Laphroaig, Scapa

Grupa H
: **Whisky:** Bruichladdich, Deanston, Fettercairn, Glenfiddich, Glen Mhor, Glen Spey, Glentauchers, Ladyburn, Tobermory

Aby zbadać klastry, możemy więc po prostu przyjrzeć się gatunkom whisky w każdym z nich. Wydaje się to dość proste, ale należy pamiętać, że przykład z whisky został wybrany do celów ilustracyjnych. Co takiego jest akurat w tym obszarze zastosowania, co pozwoliło na stosunkowo łatwe badanie klastrów (i przez to sprawiło, że stał się on dobrym przykładem do książki)? Moglibyśmy stwierdzić: „No cóż, liczba wszystkich gatunków whisky jest stosunkowo niewielka; dzięki temu możemy przyjrzeć się wszystkim". To prawda, ale faktycznie

nie jest to aż tak istotne. Gdybyśmy mieli ogromne ilości różnych whisky, to wciąż moglibyśmy pobrać próbki z każdego klastra, aby określić ich skład.

Ważniejszym czynnikiem dla zrozumienia tych klastrów — przynajmniej dla kogoś, kto wie co nieco o słodowych whisky — jest fakt, że elementy klastra mogą być reprezentowane przez nazwy whisky. W tym przypadku *nazwy* punktów danych są istotne jako takie i zawierają określone znaczenia dla eksperta w dziedzinie whisky.

Zyskujemy więc dzięki temu wytyczne, które możemy wykorzystać do innych zastosowań. Jeżeli na przykład dokonujemy klastrowania klientów wielkiej sieci detalicznej, to lista nazwisk klientów w klastrze prawdopodobnie będzie miała niewielkie znaczenie, więc ta technika nie byłaby przydatna z punktu widzenia zrozumienia wyników klastrowania. Z drugiej jednak strony, jeżeli IBM przeprowadza klastrowanie swoich klientów biznesowych, to może być tak, że nazwy firm (a przynajmniej części z nich) będą miały istotne znaczenie dla kierownika lub personelu działu sprzedaży.

Co możemy zrobić w przypadku, gdy nie jesteśmy w stanie po prostu wskazać nazw naszych punktów danych lub gdy wskazanie nazw nie przynosi wystarczającego zrozumienia? Spójrzmy jeszcze raz na nasze klastry whisky, ale tym razem z większą ilością informacji:

Grupa A

- Whisky: Aberfeldy, Glenugie, Laphroaig, Scapa
- Najlepsza w klasie: Laphroaig (Islay), 10 lat, 86 punktów
- Uśrednione cechy: złoty; owocowy, słony; średnie; oleisty, słony, sherry; wytrawny

Grupa H

- Whisky: Bruichladdich, Deanston, Fettercairn, Glenfiddich, Glen Mhor, Glen Spey, Glentauchers, Ladyburn, Tobermory
- Najlepsza w klasie: Bruichladdich (Islay) 10 lat, 76 punktów
- Uśrednione cechy: białe wino, blady; słodki; gładkie, lekkie; słodki, wytrawny, owocowy, dymny; wytrawny, lekki

Tutaj widzimy dwie dodatkowe informacje, przydatne dla zrozumienia wyników klastrowania. Po pierwsze, poza składem poszczególnych klastrów wyodrębniony został element „wzorcowy". W tym przypadku to „najlepsza w klasie" whisky, na podstawie książki Jacksona (1989) (te dodatkowe informacje nie zostały podane algorytmowi klastrowania). Alternatywnie mogłaby to być najbardziej znana lub najlepiej się sprzedająca whisky w klastrze. Te techniki mogą być szczególnie użyteczne w przypadku znacznej liczby wystąpień w każdym klastrze, gdy losowe próbkowanie może okazać się nie tak skuteczne jak staranne wybieranie egzemplarzy wzorcowych. Stale jednak zakładamy, że nazwy przypadków mają znaczenie. Inny nasz przykład, klastrowanie wiadomości biznesowych, ujmuje tę ogólną ideę nieco inaczej: pokazuje wzorcowe wiadomości i ich nagłówki, ponieważ nagłówki mogą w znaczący sposób streszczać te wiadomości.

Przykład ilustruje również inny sposób rozumienia wyników klastrowania: ukazuje uśrednione cechy członków klastra — a więc faktycznie centroid klastra. Ukazanie centroidu można zastosować do każdego klastrowania, a to, czy będzie ono znaczące, zależy od tego, czy znaczące są dane jako takie.

* Wykorzystywanie uczenia nadzorowanego do generowania opisów klastrów

Przed nami szczegóły techniczne

W tym podrozdziale opisujemy sposób automatycznego generowania opisów klastrów. To bardziej skomplikowana kwestia niż nasze dotychczasowe rozważania. Wykorzystujemy tutaj łączenie uczenia nienadzorowanego (klastrowania) z uczeniem nadzorowanym w celu stworzenia różnicowych charakterystyk klastrów. Jeśli w tym rozdziale po raz pierwszy stykasz się z klastrowaniem i uczeniem nienadzorowanym, to poniższy podrozdział może wprawić Cię w pewną dezorientację, oznaczyliśmy go więc gwiazdką, jako materiał dla zaawansowanych. Można go pominąć bez utraty ciągłości wywodu.

Niezależnie od tego, jak przeprowadzone zostało klastrowanie, otrzymujemy listę przypisań wskazujących, które przykłady należą do których klastrów. W rezultacie uśredniony członek klastra zostaje opisany przez centroid klastra. Problemem jest to, że opisy mogą być bardzo szczegółowe i nie mówią nam, w jaki sposób klastry się różnią. W przypadku każdego klastra zależy nam na ustaleniu, *co odróżnia ten klaster od wszystkich innych*. Tym właśnie zasadniczo zajmują się metody uczenia nadzorowanego, możemy więc się nimi posłużyć.

Ogólna strategia wygląda następująco: wykorzystujemy przypisania do klastrów do etykietowania przykładów. Każdy przykład otrzyma etykietę klastra, do którego należy. Może ona być traktowana jako etykieta klasy. Dysponując etykietowanym zbiorem przykładów, uruchamiamy na zbiorze przykładów algorytm uczenia nadzorowanego, aby wygenerować klasyfikator dla każdej klasy / każdego klastra. Następnie możemy sprawdzić opisy klasyfikatorów, aby otrzymać (miejmy nadzieję) zrozumiały i zwięzły opis odpowiadającego każdemu z nich klastra. Należy pamiętać o tym, że będą to opisy **różnicowe**, określające, co odróżnia dany klaster od innych.

W tym podrozdziale, poczynając od tego miejsca, zrównujemy klastry z klasami. Będziemy używać obu określeń zamiennie.

Zasadniczo moglibyśmy wykorzystać do tego celu jakąkolwiek predykcyjną (nadzorowaną) metodę uczenia, ale istotna jest tutaj *zrozumiałość*: mamy zamiar wykorzystać definicję nauczonego klasyfikatora jako opis klastra, potrzebujemy więc modelu, który będzie służył temu celowi. W rozdziale 3., w podrozdziale „Drzewa jako zbiory reguł", pokazaliśmy, w jaki sposób możliwe jest wydobywanie reguł z drzew klasyfikacyjnych, więc jest to użyteczna metoda do celów tego zadania.

Istnieją dwa sposoby ujmowania zadania klasyfikacji. Mamy k klastrów, możemy więc potraktować zadanie jako k-klas (po jednej klasie na klaster). Można też skonfigurować k odrębnych zadań uczących i w ramach każdego z nich starać się odróżnić jeden klaster od wszystkich pozostałych (k–1) klastrów.

Wykorzystamy to drugie podejście w ramach zadania klastrowania whisky, posługując się przypisaniami klastrów w ujęciu Lapointe'a i Legendre'a (Appendix A, *A Classification of Pure Malt Scotch Whiskies http://www.dcs.ed.ac.uk/home/jhb/whisky/lapointe/text.html*). Otrzymujemy 12 klastrów whisky, oznaczonych literami od A do L. Wracamy do naszych surowych danych i łączymy opis każdej whisky z jej przypisaniem do klastra. Wykorzystamy podejście binarne: wybierzemy kolejno każdy klaster i dokonamy jego klasyfikacji na tle innych. Wybierzemy klaster J, który Lapointe i Legendre opisują następująco:

Grupa J

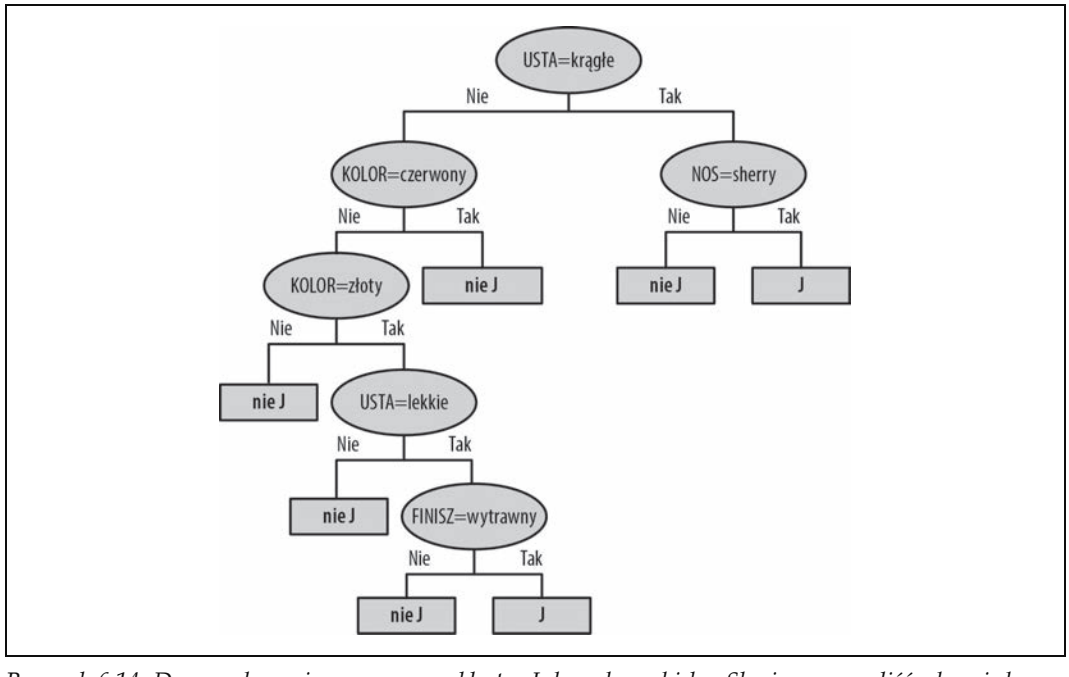
- Whisky: Glen Albyn, Glengoyne, Glen Grant, Glenlossie, Linkwood, North Port, Saint Magdalene, Tamdhu.
- Najlepsza w klasie: Linkwood (Speyside), 12 lat, 83 punkty.
- Uśrednione cechy: złoty; wytrawny, torfowy, sherry; lekkie do średnich, okrągłe; słodkie; wytrawny.

Z podrozdziału „Przykład: analityka whisky" pamiętamy, że każda whisky została opisana za pomocą 68 cech binarnych. Zestaw danych posiada teraz dla każdej whisky etykietę (*J* lub *nie_J*) wskazującą, czy należy ona do klastra J. Fragment zbioru danych wygląda następująco:

```
0,0,0 , ... , 0,0,0,0,0,1,0,0,0,0,0,0,1,0,0,0,0,0,0,0 , 0,0,0 , J     % Glen Grant
0,0,0 , ... , 0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0 , 1,0,0 , nie_J % Glen Keith
0,0,0 , ... , 0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,0 , 0,0,0 , nie_J % Glen Mhor
```

Tekst po oznaczeniu „%" to komentarz wskazujący nazwę whisky.

Ten zbiór danych przekazywany jest do algorytmu uczącego drzewa klasyfikacyjnego[7]. Wynik przedstawia rysunek 6.14.



Rysunek 6.14. Drzewo decyzyjne nauczone z klastra J danych o whisky. Skrajny prawy liść odpowiada segmentowi populacji z krągłymi ustami i nosem sherry. Whisky w tym segmencie należą głównie do klastra J

W ramach tego drzewa koncentrujemy się wyłącznie na liściach oznaczonych *J* (pomijając te, które są oznaczone *nie_J*). Są tylko dwa takie liście. Śledząc ścieżkę od korzenia do tych liści, możemy wyodrębnić dwie reguły:

[7] Konkretnie jest to procedura J48 Weka (*http://www.cs.waikato.ac.nz/ml/weka/*), z wyłączonym przycinaniem.

1. `(USTA=krągłe) I (NOS=sherry = 1) → J`
2. `(USTA= krągłe) I (KOLOR= czerwony) I (KOLOR = złoty) I (USTA = lekkie) I (FINISZ = wytrawny) → J`

W wolnym tłumaczeniu na polski, klaster **J** obejmuje whisky posiadające albo:

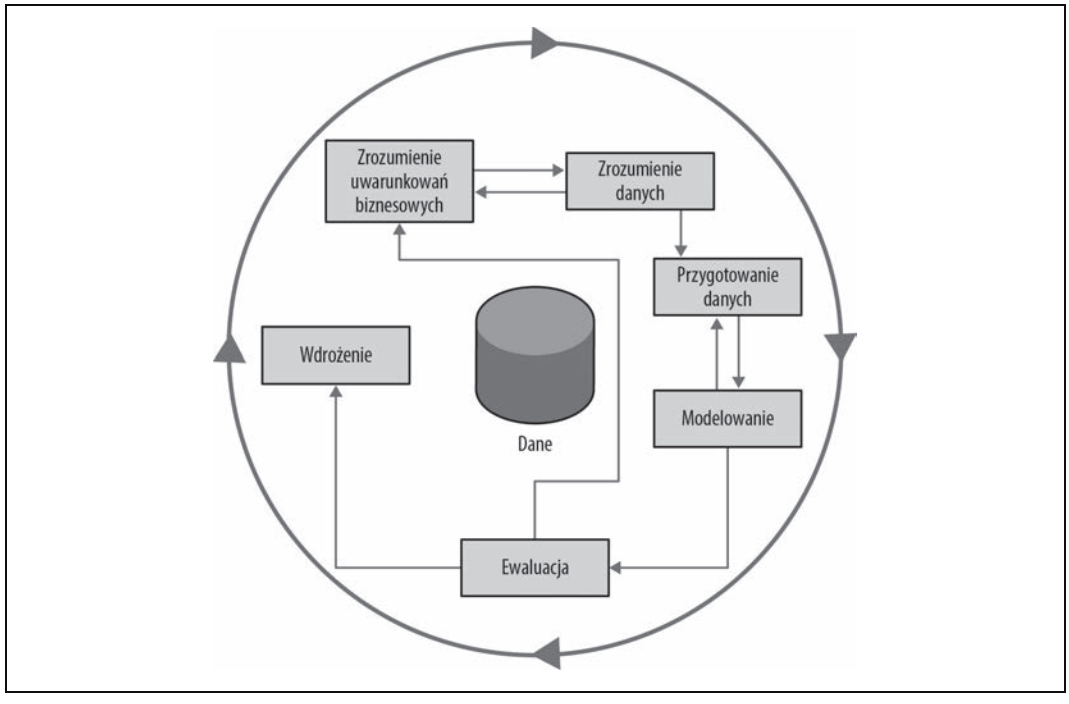
1. Krągłe usta i nos sherry, albo
2. Złoty (nie czerwony) kolor oraz lekkie (ale nie krągłe) usta i wytrawny finisz.

Czy ten opis klastra J jest lepszy niż zaproponowany przez Lapointe'a i Legendre'a? Możesz zdecydować, który bardziej Ci odpowiada, należy jednak podkreślić, że istnieją różne *typy* opisów. Opis Lapointe'a i Legendre'a jest opisem **cech**; opisuje to, co jest typowe czy też **charakterystyczne** dla klastra, pomijając to, czy inne klastry mogłyby także posiadać niektóre z tych cech. Opis wygenerowany przez drzewo decyzyjne jest opisem **różnicowym**; opisuje tylko to, co odróżnia ten klaster od innych, ignorując cechy, które mogą być wspólne dla znajdujących się w nim whisky. Ujmując to w inny sposób: opisy cech koncentrują się na podobieństwach wewnątrz grupy, natomiast opisy różnicowe koncentrują się na różnicach międzygrupowych. Żaden z nich nie jest z natury lepszy — zależy to od tego, do czego je wykorzystujemy.

Krok wstecz: rozwiązywanie problemu biznesowego kontra eksploracja danych

Przyjrzeliśmy się różnym przykładom naszych podstawowych pojęć nauki o danych w akcji. Być może zauważyłeś, że przykłady klastrowania wydają się w pewien sposób inne od przykładów modelowania predykcyjnego, a nawet przykładów związanych z wyszukiwaniem podobnych obiektów. Zbadajmy, dlaczego tak jest.

W naszych przykładach modelowania predykcyjnego oraz w przykładach z bezpośrednim wykorzystywaniem podobieństwa skupiliśmy się na rozwiązywaniu bardzo konkretnego problemu biznesowego. Jak już podkreślaliśmy, jednym z podstawowych pojęć nauki o danych jest to, że powinniśmy dołożyć wszelkich starań, żeby jak najdokładniej określić cel każdej eksploracji danych. Przypomnijmy tutaj proces eksploracji danych CRISP, powtórzony na rysunku 6.15. W ramach minicyklu zrozumienia uwarunkowań biznesowych/zrozumienia danych powinniśmy poświęcić jak najwięcej czasu, aby uzyskać konkretną, jednoznaczną definicję problemu, który staramy się rozwiązać. Podczas stosowania modelowania predykcyjnego wspomaga nas potrzeba precyzyjnego zdefiniowania zmiennej docelowej, a w rozdziale 7. przekonamy się, że możemy osiągać coraz większy stopień precyzji w definiowaniu problemu wraz z coraz większym zaawansowaniem naszego poziomu zrozumienia nauki o danych. W przykładach z dopasowywaniem podobieństwa także mieliśmy bardzo konkretne pojęcie o tym, czego dokładnie szukaliśmy: chcieliśmy znaleźć podobne firmy w celu optymalizacji naszych działań i dokładnie określaliśmy, co to znaczy „podobne". Chcemy znaleźć podobne whisky — zwłaszcza pod względem smaku — i znów podejmujemy działania, aby zgromadzić i przedstawić dane tak, byśmy mogli znaleźć właśnie te gatunki. W dalszej części książki omówimy częstą kwestię poświęcania znaczącego wysiłku, aby zastosować platformy nauki o danych do rozkładania problemów biznesowych na wiele jasno zdefiniowanych komponentów, do których rozwiązania możemy zastosować metody nauki o danych.



Rysunek 6.15. Proces eksploracji danych CRISP

Jednak nie wszystkie problemy są tak dobrze zdefiniowane. Co zrobimy, gdy w fazie zrozumienia uwarunkowań biznesowych stwierdzimy, że *chcielibyśmy dokonać eksploracji naszych danych, mając jednak tylko mgliste pojęcie, jaki dokładnie problem rozwiązujemy*? Problemy, do rozwiązywania których wykorzystujemy klastrowanie, często należą do tej kategorii. Chcemy przeprowadzić *nienadzorowaną* segmentację: znaleźć grupy, które występują „naturalnie" (oczywiście w zależności od tego, jak definiujemy nasze miary podobieństwa).

W celu zachowania jasności wywodu zastosujmy pewne uproszczenie i podzielmy nasze problemy na nadzorowane (np. modelowanie predykcyjne) i nienadzorowane (np. klastrowanie). Świat nie jest jednoznaczny i prawie każdą z technik eksploracji danych, które przedstawiliśmy, można byłoby wykorzystać do eksploracji danych, ale nasze rozważania będą znacznie klarowniejsze, jeśli po prostu dokonamy podziału na metody nadzorowane oraz nienadzorowane. Istnieje bezpośredni kompromis, dotyczący tego, gdzie i jak prowadzimy działania w ramach procesu eksploracji danych. W przypadku problemów nadzorowanych, skoro poświęciliśmy tyle czasu na precyzyjne zdefiniowanie problemu, który chcemy rozwiązać, to na etapie ewaluacji procesu eksploracji danych dysponujemy już jednoznacznie sformułowanym pytaniem ewaluacyjnym: czy wygląda na to, że wyniki modelowania rozwiązują problem, który zdefiniowaliśmy? Gdybyśmy na przykład zdefiniowali nasz cel jako poprawę skuteczności predykcji rezygnacji klienta w czasie, gdy zbliża się termin wygaśnięcia jego umowy, to moglibyśmy ocenić, czy nasz model tego dokonał.

W przeciwieństwie do problemów nadzorowanych, problemy nienadzorowane mają znacznie bardziej eksploracyjny charakter. Możemy być przekonani, że gdybyśmy mogli dokonać klastrowania firm, wiadomości lub whisky, to lepiej zrozumielibyśmy naszą firmę, a więc bylibyśmy w stanie coś poprawić. Możemy jednak nie dysponować precyzyjnym sformułowaniem

problemu. Nie powinniśmy pozwalać, aby dążenie do konkretu i precyzji odwodziło nas od dokonywania istotnych odkryć na podstawie danych. Istnieje tutaj jednak kompromis. Polega on na tym, że problemom, których nie udało nam się precyzyjnie sformułować na wczesnych etapach procesu eksploracji danych, musimy poświęcić więcej czasu w dalszej części tego procesu, na etapie ewaluacji.

Zwłaszcza w przypadku klastrów często trudno jest zrozumieć, co ujawnia klastrowanie (o ile w ogóle coś ujawnia). Nawet wtedy, gdy wydaje się, że klastrowanie ujawnia ciekawe informacje, to często nie jest jasne, w jaki sposób powinniśmy wykorzystać je do podejmowania lepszych decyzji. Dlatego w przypadku klastrów konieczne jest wykorzystywanie dodatkowych zasobów kreatywności i znajomości branży w fazie ewaluacyjnej procesu eksploracji danych.

Ira Haimowitz i Henry Schwartz (1997) przedstawiają konkretny przykład wykorzystania klastrowania do poprawy jakości decyzji dotyczących linii kredytowych dla nowych klientów. Przeprowadzili oni klastrowanie istniejących klientów banku GE Capital na podstawie podobieństwa w sposobie użytkowania przez nich kart, płacenia rachunków i ich rentowności dla firmy. Wyodrębnili pięć klastrów, określających bardzo zróżnicowane zachowania klientów (byli np. tacy, którzy dużo wydawali, ale spłacali zadłużenie na swoich kartach w całości co miesiąc, oraz tacy, którzy dużo wydawali, a ich salda stale pozostawały w okolicach limitu kredytowego). Dla tak różnych klientów różna powinna być też wartość linii kredytowych (w drugim przypadku należy zachować szczególną ostrożność, aby uniknąć utraty zdolności do spłaty zadłużenia). Problem z bezpośrednim wykorzystaniem wyników tego klastrowania do podejmowania decyzji polegał na tym, że dane nie były dostępne w czasie określania pierwotnej wartości linii kredytowych. Mówiąc krótko, Haimowitz i Schwarz posłużyli się tą nową wiedzą, wracając do początku procesu eksploracji danych. Wykorzystali tę wiedzę, aby dokładnie zdefiniować problem modelowania predykcyjnego: jak za pomocą danych dostępnych w momencie przyznawania kredytów przewidzieć stopień prawdopodobieństwa znalezienia się klienta w poszczególnych klastrach. Ten model predykcyjny mógł następnie zostać wykorzystany do poprawy trafności decyzji o przyznawaniu linii kredytowych.

Podsumowanie

Podstawowe pojęcie podobieństwa elementów danych pojawia się w ramach całego obszaru eksploracji danych. W tym rozdziale najpierw omówiliśmy szeroki zakres zastosowań podobieństwa, poczynając od znajdowania podobnych jednostek (lub obiektów) na podstawie opisujących je danych, poprzez modelowanie predykcyjne, na klastrowaniu jednostek kończąc. Omówiliśmy te różnorodne zastosowania i zilustrowaliśmy je przykładami.

Bardzo częstym przybliżeniem podobieństwa dwóch jednostek jest odległość między nimi w przestrzeni wystąpień, zdefiniowanej przez ich wektory cech. Zaprezentowaliśmy sposoby obliczania podobieństwa i odległości, zarówno ogólnie, jak i za pomocą szczegółów technicznych. Przedstawiliśmy także rodzinę metod, zwanych metodami najbliższego sąsiedztwa, które wykonują zadania predykcyjne, obliczając jednoznacznie podobieństwo między nowym przykładem a zbiorem przykładów uczących (o znanych wartościach wielkości docelowej). Mogąc pobrać zbiór najbliższych sąsiadów (najbardziej podobnych przykładów), możemy użyć ich do różnych zadań z zakresu eksploracji danych: klasyfikacji, regresji i scoringu wystąpień. Wskazaliśmy wreszcie, że to samo podstawowe pojęcie — podobieństwo — jest fundamentem najbardziej rozpowszechnionej nienadzorowanej metody eksploracji danych: klastrowania.

Omówiliśmy także inną ważną koncepcję, która pojawia się, ilekroć zaczynamy bardziej szczegółowo przyglądać się metodom (takim jak klastrowanie) wykorzystywanym w analizie danych o bardziej eksploracyjnym charakterze. Podczas eksploracji danych, zwłaszcza metodami nienadzorowanymi, zwykle zaczynamy poświęcać mniej czasu na początkową fazę zrozumienia uwarunkowań biznesowych, a więcej na etap ewaluacji i powtarzanie całego cyklu procesu eksploracji danych. Aby to zilustrować, omówiliśmy różnorodne metodologie prowadzące do zrozumienia wyników klastrowań.

Skorowidz

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

Ś

T

U

W

Z

Przeanalizuj posiadane dane i podejmij trafne decyzje!

Posiadanie zbiorów danych to połowa sukcesu. Druga połowa to umiejętność ich skutecznej analizy i wyciągania wniosków! Dopiero na tej podstawie będziesz w stanie właściwie ocenić kondycję Twojej firmy oraz podjąć słuszne decyzje. Wiedza zawarta w tej książce może zadecydować o sukcesie lub porażce Twojego biznesu. Nie ryzykuj i sięgnij po to doskonałe źródło wiedzy poświęcone nauce o danych.

To unikalny podręcznik, który pomoże Ci sprawnie opanować nawet najtrudniejsze zagadnienia związane z analizą danych. Dowiesz się, jak zbudowany jest proces eksploracji danych, z jakich narzędzi możesz skorzystać oraz jak stworzyć model predykcyjny i dopasować go do danych. W kolejnych rozdziałach przeczytasz o tym, czym grozi nadmierne dopasowanie modelu i jak tego unikać oraz jak wyciągać wnioski metodą najbliższych sąsiadów. Na koniec zaznajomisz się z możliwościami wizualizacji skuteczności modelu oraz odkryjesz związek pomiędzy nauką o danych a strategią biznesową. To obowiązkowa lektura dla wszystkich osób chcących podejmować świadome decyzje na podstawie posiadanych danych!

Dzięki tej książce:

- poznasz model predykcyjny
- dowiesz się, jak dopasować model do danych
- zwizualizujesz skuteczność zbudowanego modelu
- zwiększysz swoje szanse na osiągnięcie sukcesu w biznesie!

onepress

Helion

helion.pl

HELION SA
ul. Kościuszki 1c
44-100 Gliwice
tel.: 32 230 98 63
helion@helion.pl

KOD KORZYŚCI
Sięgnij po więcej!

ISBN 978-83-8322-580-7

9 788383 225807

Cena: 79,00 zł

O'REILLY