Typy Danych

Teraz, gdy mamy podstawowe wprowadzenie do świata nauki o danych i rozumiemy, dlaczego ta dziedzina jest tak ważna, przyjrzyjmy się różnym sposobom tworzenia danych. W szczególności w tej części przyjrzymy się następującym tematom:

• Dane ustrukturyzowane a nieustrukturyzowane

• Dane ilościowe a dane jakościowe

• Cztery poziomy danych

Zagłębimy się w każdy z tych tematów, pokazując przykłady tego, jak naukowcy zajmujący się danymi patrzą na dane i jak z nimi pracują. Ten rozdział ma na celu zapoznanie się z podstawowymi ideami nauki o danych.

Smaki danych

W terenie ważne jest, aby zrozumieć różne rodzaje danych z kilku powodów. Nie tylko rodzaj danych będzie dyktować metody stosowane do analizy i wyodrębniania wyników, wiedza o tym, czy dane są nieustrukturyzowane, czy może ilościowe, może również wiele powiedzieć o mierzonym zjawisku w świecie rzeczywistym. Przyjrzymy się trzem podstawowym klasyfikacjom danych:

• Strukturalny a nieustrukturyzowany (czasami nazywany zorganizowanym lub niezorganizowanym)

• Ilościowe a jakościowe

• Cztery poziomy danych

Pierwszą rzeczą, na którą należy zwrócić uwagę, jest użycie przeze mnie słowa data. W ostatniej części określiłem dane jako jedynie zbiór informacji. Ta niejasna definicja istnieje, ponieważ możemy podzielić dane na różne kategorie i potrzebujemy, aby nasza definicja była luźna. Następną rzeczą do zapamiętania podczas przechodzenia przez tą część jest to, że w większości przypadków, kiedy mówię o rodzaju danych, będę odnosić się do określonej cechy zestawu danych lub do całego zestawu danych jako całości. Będę bardzo jasno określał, do którego w danym momencie się odwołuję.

Po co patrzeć na te rozróżnienia?

Zatrzymanie się i zastanowienie nad rodzajem danych może wydawać się bezwartościowe. Mamy przed sobą zabawne rzeczy, takie jak statystyki i uczenie maszynowe, ale jest to prawdopodobnie jeden z najważniejszych kroków, które musisz podjąć, aby przeprowadzić analizę danych. Rozważmy przykład, w którym patrzymy na wyniki wyborów dla hrabstwa. W zbiorze danych osób znajduje się kolumna „rasa", która jest oznaczona numerem identyfikacyjnym, aby zaoszczędzić miejsce. Na przykład być może kaukaski jest oznaczany przez 7, podczas gdy azjatycki amerykański to 2. Bez zrozumienia, że te liczby nie są w rzeczywistości liczbami uporządkowanymi, jak o nich myślimy (gdzie 7 jest większe niż 2, a zatem kaukaski jest „większy niż" azjatycki Amerykanin) zrobimy straszne błędy w naszej analizie.

Ta sama zasada dotyczy nauki o danych. Po otrzymaniu zestawu danych kuszące jest, aby przejść od razu do eksploracji, stosowania modeli statystycznych i badania zastosowań uczenia maszynowego w celu szybszego uzyskania wyników. Jeśli jednak nie rozumiesz typu danych, z którymi pracujesz, możesz stracić dużo czasu na stosowanie modeli, o których wiadomo, że są nieefektywne w przypadku tego konkretnego typu danych. Kiedy otrzymujesz nowy zestaw danych, zawsze zalecam poświęcenie około godziny (zwykle mniej) na dokonanie rozróżnień wymienionych w kolejnych sekcjach.

Dane ustrukturyowane i nieustrukturyzowane

Rozróżnienie między danymi ustrukturyzowanymi i nieustrukturyzowanymi jest zwykle pierwszym pytaniem, jakie chcesz zadać sobie na temat całego zbioru danych. Odpowiedź na to pytanie może oznaczać różnicę między potrzebą trzech dni lub trzech tygodni na wykonanie prawidłowej analizy. Podstawowy podział jest następujący (jest to przerobiona definicja uporządkowanych i niezorganizowanych danych w pierwszej Części):

• Dane ustrukturyzowane (zorganizowane): Są to dane, które można traktować jako obserwacje i cechy. Zwykle jest zorganizowany przy użyciu metody tabelarycznej (wiersze i kolumny).

• Dane nieustrukturyzowane (niezorganizowane): Dane te istnieją jako wolna jednostka i nie podlegają żadnej standardowej hierarchii organizacyjnej. Oto kilka przykładów, które mogą pomóc w rozróżnieniu między nimi:

• Większość danych, które istnieją w formie tekstowej, w tym dzienniki serwera i posty na Facebooku, jest nieustrukturyzowana

• Obserwacje naukowe, zarejestrowane przez uważnych naukowców, są przechowywane w bardzo schludnym i zorganizowanym (ustrukturyzowanym) formacie

• Sekwencja genetyczna nukleotydów chemicznych (na przykład ACGTATTGCA) jest pozbawiona struktury, nawet jeśli kolejność nukleotydów ma znaczenie, ponieważ nie możemy utworzyć deskryptorów sekwencji przy użyciu formatu wiersza/kolumny bez dalszego przyjrzenia się

Ogólnie uważa się, że ustrukturyzowane dane są znacznie łatwiejsze w pracy i analizie. Większość modeli statystycznych i uczenia maszynowego została zbudowana z myślą o danych strukturalnych i nie może działać na luźnej interpretacji danych niestrukturalnych. Naturalna struktura rzędów i kolumn jest łatwa do przyswojenia dla oczu człowieka i maszyny. Po co więc w ogóle mówić o danych nieustrukturyzowanych? Bo to takie powszechne! Większość szacunków umieszcza dane nieustrukturyzowane jako 80-90% danych światowych. Dane te istnieją w wielu formach iw większości pozostają niezauważone przez ludzi jako potencjalne źródło danych. Tweety, e-maile, literatura i logi serwera to ogólnie nieustrukturyzowane formy danych. Chociaż naukowcy zajmujący się danymi prawdopodobnie preferują dane ustrukturyzowane, muszą być w stanie poradzić sobie z ogromnymi ilościami nieustrukturyzowanych danych na świecie. Jeśli 90% światowych danych jest nieustrukturyzowanych, oznacza to, że około 90% światowych informacji jest uwięzionych w trudnym formacie. Tak więc, ponieważ większość naszych danych istnieje w tym dowolnym formacie, musimy zwrócić się do technik analizy wstępnej, zwanej przetwarzaniem wstępnym, aby nadać strukturę przynajmniej części danych do dalszej analizy. Następny rozdział zajmie się bardzo szczegółowo przetwarzaniem wstępnym; na razie rozważymy część przetwarzania wstępnego, w której próbujemy zastosować przekształcenia, aby przekonwertować nieustrukturyzowane dane na ustrukturyzowany odpowiednik.

Przykład wstępnego przetwarzania danych

Patrząc na dane tekstowe (które prawie zawsze uważane są za nieustrukturyzowane), mamy wiele możliwości przekształcenia zbioru w format ustrukturyzowany. Możemy to zrobić, stosując nowe cechy opisujące dane. Oto kilka takich cech:

• Liczba słów/fraz

• Istnienie pewnych znaków specjalnych

• Względna długość tekstu

• Wybieranie tematów

Użyję poniższego tweeta jako szybkiego przykładu nieustrukturyzowanych danych, ale możesz użyć dowolnego nieustrukturyzowanego tekstu w dowolnej formie, który Ci się podoba, w tym tweetów i postów na Facebooku. Czy w ten środowy poranek wstajesz wcześnie? Następnie spójrz na wschód. Półksiężyc dołącza do Wenus i Saturna. Unosząc się na niebie o świcie. Ważne jest, aby powtórzyć, że dla tego tweeta konieczne jest wstępne przetwarzanie, ponieważ zdecydowana większość algorytmów uczenia wymaga danych liczbowych (do których omówimy po tym przykładzie). Wstępne przetwarzanie nie tylko wymaga określonego rodzaju danych, ale pozwala nam zbadać funkcje, które zostały utworzone na podstawie istniejących funkcji. Na przykład ze wspomnianego tweeta możemy wyodrębnić takie funkcje, jak liczba słów i znaki specjalne. Przyjrzyjmy się teraz kilku funkcjom, które możemy wyodrębnić z tekstu.

Liczba słów/fraz

Możemy podzielić tweeta na liczbę słów/fraz. Słowo to pojawia się w tweecie raz, podobnie jak każde inne słowo. Możemy przedstawić ten tweet w ustrukturyzowanym formacie w następujący sposób, konwertując w ten sposób nieustrukturyzowany zestaw słów na format wiersza/kolumny:

	this	wednesday	morn	are	this wednesday
Word Count	1	1	1	1	1

Zauważ, że aby uzyskać ten format, możemy wykorzystać scikit-learn's CountVectorizer, który widzieliśmy w poprzedniej części.

Obecność niektórych znaków specjalnych

Możemy również przyjrzeć się obecności znaków specjalnych, takich jak znak zapytania i wykrzyknik. Pojawienie się tych znaków może sugerować pewne koncepcje dotyczące danych, które w inny sposób są trudne do poznania. Na przykład fakt, że ten tweet zawiera znak zapytania, może silnie sugerować, że zawiera on pytanie do czytelnika. Możemy dołączyć do poprzedniej tabeli nową kolumnę, jak pokazano:

	this	wednesday	morn	are	this wednesday	?
Word Count	1	1	1	1	1	1

Względna długość tekstu

Ten tweet ma 121 znaków.

len("This Wednesday morn, are you early to rise? Then look East. The Crescent Moon joins Venus & Saturn. Afloat in the dawn skies.")

get the length of this text (number of characters for a string)

121

Przeciętny tweet, jak odkryli analitycy, ma około 30 znaków. Możemy więc narzucić nową cechę, zwaną długością względną (która jest długością tweeta podzieloną przez średnią długość), informującą nas o

długości tego tweeta w porównaniu z przeciętnym tweetem. Ten tweet jest w rzeczywistości 4,03 razy dłuższy niż przeciętny tweet, jak pokazano:

121/30 = 4,03

Możemy dodać kolejną kolumnę do naszej tabeli za pomocą tej metody:

	this	wednesday	morn	are	this wednesday	?	Relative length
Word Count	1	1	1	1	1	1	4.03

Wybieranie tematów

Możemy wybrać niektóre tematy tweeta, aby dodać je jako kolumny. Ten tweet dotyczy astronomii, więc możemy dodać kolejną kolumnę, jak pokazano na ilustracji:

	this	wednesday	morn	are	this wednesday	?	Relative length	Topic
Word Count	1	1	1	1	1	1	4.03	astronomy

I tak po prostu, możemy przekonwertować fragment tekstu na ustrukturyzowane/zorganizowane dane gotowe do użycia w naszych modelach i analizie eksploracyjnej. Temat jest jedyną wyodrębnioną funkcją, którą przyjrzeliśmy się, która nie jest automatycznie wyprowadzana z tweeta. Sprawdzanie liczby słów i długości tweetów w Pythonie jest łatwe; jednak bardziej zaawansowane modele (zwane modelami tematycznymi) są w stanie wyprowadzać i przewidywać tematy jako tekst naturalny. Możliwość szybkiego rozpoznania, czy Twoje dane są ustrukturyzowane czy nieustrukturyzowane, może zaoszczędzić godziny, a nawet dni pracy w przyszłości. Gdy już jesteś w stanie rozróżnić organizację prezentowanych danych, następne pytanie dotyczy indywidualnych cech zestawu danych.

Dane ilościowe a dane jakościowe

Kiedy pytasz analityka danych „co to za dane?", zazwyczaj zakładają, że pytasz go, czy są to dane głównie ilościowe, czy jakościowe. Jest to prawdopodobnie najczęstszy sposób opisywania specyficznych cech zestawu danych. W większości przypadków, gdy mówimy o danych ilościowych, zwykle (nie zawsze) mówimy o ustrukturyzowanym zbiorze danych o ścisłej strukturze wierszy/kolumn (ponieważ nie zakładamy, że nieustrukturyzowane dane mają nawet jakiekolwiek cechy). Tym bardziej, że etap przetwarzania wstępnego jest tak ważny. Te dwa typy danych można zdefiniować w następujący sposób:

• Dane ilościowe: Dane te można opisać za pomocą liczb, a podstawowe procedury matematyczne, w tym dodawanie, są możliwe na zestawie.

• Dane jakościowe: Tych danych nie można opisać za pomocą liczb i podstawowej matematyki. Te dane są ogólnie uważane za opisane przy użyciu kategorii i języka „naturalnego".

Przykład - dane kawiarni

Powiedzmy, że przetwarzaliśmy obserwacje kawiarni w dużym mieście przy użyciu następujących pięciu deskryptorów (cech):

Dane: kawiarnia

• Nazwa kawiarni

• Przychody (w tysiącach złotych)

• Kod pocztowy

• Przeciętni klienci miesięcznie

• Kraj pochodzenia kawy

Każdą z tych cech można sklasyfikować jako ilościową lub jakościową, a to proste rozróżnienie może wszystko zmienić. Przyjrzyjmy się każdemu z nich:

• Nazwa kawiarni - Jakościowa

Nazwa kawiarni nie jest wyrażona jako liczba i nie możemy wykonać obliczeń matematycznych na nazwie sklepu.

• Przychody - ilościowe

Ile pieniędzy przynosi kawiarnia, z pewnością można opisać za pomocą liczby. Możemy również wykonywać podstawowe operacje, takie jak sumowanie przychodów przez 12 miesięcy, aby uzyskać roczny przychód.

• Kod pocztowy - jakościowy

Ten jest trudny. Kod pocztowy jest zawsze przedstawiany za pomocą liczb, ale to, co czyni go jakościowym, to fakt, że nie pasuje do drugiej części definicji ilościowej - nie możemy wykonać podstawowych operacji matematycznych na kodzie pocztowym. Jeśli dodamy do siebie dwa kody pocztowe, jest to bezsensowny pomiar. Niekoniecznie otrzymujemy nowy kod pocztowy i zdecydowanie nie otrzymujemy „podwójnego kodu pocztowego".

• Przeciętni miesięczni klienci - ilościowo

Ponownie, opisanie tego czynnika za pomocą liczb i dodawania ma sens. Dodaj wszystkich swoich miesięcznych klientów, a otrzymasz rocznych klientów.

• Kraj pochodzenia kawy – jakościowy

Załóżmy, że jest to bardzo mała kawiarnia z kawą jednego pochodzenia. Ten kraj jest opisany nazwą (etiopską, kolumbijską), a nie cyframi.

Kilka ważnych rzeczy do zapamiętania:

• Mimo że kod pocztowy jest opisywany za pomocą liczb, nie jest on ilościowy. Dzieje się tak, ponieważ nie można mówić o sumie wszystkich kodów pocztowych lub średnim kodzie pocztowym. To są bezsensowne opisy.

• Prawie zawsze, gdy słowo jest używane do opisania cechy, jest to czynnik jakościowy.

Jeśli masz problem z określeniem, który z nich jest zasadniczo, próbując zdecydować, czy dane są jakościowe, czy ilościowe, zadaj sobie kilka podstawowych pytań dotyczących cech danych:

• Czy możesz to opisać za pomocą liczb?

- Nie? To jest jakościowe.

- Tak? Przejdź do następnego pytania.

• Czy po dodaniu ich do siebie nadal ma to sens?

- Nie? Są jakościowe.

- Tak? Prawdopodobnie masz dane ilościowe.

Ta metoda pomoże Ci zaklasyfikować większość, jeśli nie wszystkie, dane do jednej z tych dwóch kategorii. Różnica między tymi dwiema kategoriami określa rodzaje pytań, które możesz zadać w każdej kolumnie. W przypadku kolumny ilościowej możesz zadawać pytania takie jak:

• Jaka jest średnia wartość?

• Czy ta ilość rośnie czy maleje z upływem czasu (jeśli czas jest czynnikiem)?

• Czy istnieje próg, dla którego wzrost liczby powyżej lub zbyt niski będzie sygnalizował kłopoty dla firmy?

W przypadku kolumny jakościowej nie można odpowiedzieć na żadne z poprzednich pytań; jednak poniższe pytania dotyczą tylko wartości jakościowych:

• Która wartość występuje najczęściej, a która najmniej?

• Ile jest unikalnych wartości?

• Jakie są te wyjątkowe wartości?

Przykład - dane dotyczące spożycia alkoholu na świecie

Światowa Organizacja Zdrowia opublikowała zestaw danych opisujący przeciętne nawyki związane z piciem ludzi w krajach na całym świecie. Użyjemy Pythona i narzędzia do eksploracji danych Pandas, aby uzyskać lepszy wygląd:

```
import pandas as pd

# read in the CSV file from a URL

drinks = pd.read_csv('https://raw.githubusercontent.com/sinanuozdemir/

principles_of_data_science/master/data/chapter_2/drinks.csv')

# examine the data's first five rows

drinks.head() # print the first 5 rows
```

Te trzy wiersze wykonały następujące czynności:

• Importowane pandy, które w przyszłości będą określane jako pd

• Wczytaj plik CSV (wartości oddzielone przecinkami) jako zmienną o nazwie napoje

• Wywołano metodę, head, która ujawnia pierwsze pięć wierszy zbioru danych

	country	beer_servings	spirit_servings	wine_servings	total_litres_of_pure_alcohol	continent
0	Afghanistan	0	0	0	0.0	AS
1	Albania	89	132	54	4.9	EU
2	Algeria	25	0	14	0.7	AF
3	Andorra	245	138	312	12.4	EU
4	Angola	217	57	45	5.9	AF

W tym przykładzie pracujemy z sześcioma różnymi kolumnami:

• kraj: Jakościowy

• beer_servings: ilościowe

• spirit_servings: ilościowo

• wine_servings: Ilościowe

• total_litres_of_pure_alcohol: ilościowe

• kontynent: jakościowy

Spójrzmy na jakościowy kontynent kolumnowy. Możemy użyć Pand, aby uzyskać podstawowe statystyki podsumowujące dotyczące tej nienumerycznej cechy. Stosowana jest tutaj metoda description(), która najpierw określa, czy kolumna jest prawdopodobnie ilościowa, czy jakościowa, a następnie podaje podstawowe informacje o kolumnie jako całości. Jest to pokazane w następujący sposób:

drinks['continent'].describe()

>> count 193

>> unique 5

>> top AF

>> freq 53

Wynika z niego, że WHO zgromadziła dane dotyczące pięciu unikalnych kontynentów, z których najczęstszym jest AF (Afryka), który wystąpił 53 razy w 193 obserwacjach. Jeśli spojrzymy na jedną z kolumn ilościowych i wywołamy tę samą metodę, zobaczymy różnicę w wynikach, jak pokazano:

drinks['beer_servings'].describe()

>> mean 106.160622

>> min 0.000000

>> max 376.000000

Teraz możemy przyjrzeć się średniej (średniej) porcji piwa na osobę w danym kraju (106,2 porcji), jak również najniższej porcji piwa, zero, i najwyższej zarejestrowanej porcji piwa, 376 (to więcej niż jedno piwo dziennie).

Kopać głębiej

Dane ilościowe można podzielić krok dalej na wielkości dyskretne i ciągłe. Można je zdefiniować w następujący sposób:

• Dane dyskretne: Opisuje zliczane dane. Może przyjmować tylko określone wartości.

Przykładami dyskretnych danych ilościowych są rzut kostką, ponieważ może przyjąć tylko sześć wartości, oraz liczbę klientów w kawiarni, ponieważ nie można mieć prawdziwego zakresu osób.

• Dane ciągłe: Opisuje dane, które są mierzone. Istnieje na nieskończonym zakresie wartości.

Dobrym przykładem danych ciągłych może być waga osoby, ponieważ może ona wynosić 150 funtów lub 197,66 funtów (zwróć uwagę na liczby dziesiętne). Wysokość osoby lub budynku jest liczbą ciągłą, ponieważ możliwa jest nieskończona liczba miejsc dziesiętnych. Innymi przykładami danych ciągłych są czas i temperatura.

Droga do tej pory…

Do tej pory przyjrzeliśmy się różnicom między danymi ustrukturyzowanymi i nieustrukturyzowanymi, a także między cechami jakościowymi i ilościowymi. Te dwa proste rozróżnienia mogą mieć drastyczny wpływ na przeprowadzaną analizę. Pozwólcie, że podsumuję, zanim przejdę do drugiej połowy. Dane jako całość mogą być ustrukturyzowane lub nieustrukturyzowane, co oznacza, że dane mogą mieć zorganizowaną strukturę wierszy/kolumn z odrębnymi cechami opisującymi każdy wiersz zestawu danych lub istnieć w stanie swobodnym, który zwykle musi być wstępnie przetworzony forma, która jest łatwo przyswajalna. Jeśli dane są ustrukturyzowane, możemy spojrzeć na każdą kolumnę (cechę) zbioru danych jako ilościową lub jakościową. Zasadniczo, czy kolumnę można opisać za pomocą matematyki i liczb, czy nie? W kolejnej części dane zostaną podzielone na cztery bardzo szczegółowe i szczegółowe poziomy. Przy każdym zamówieniu będziemy stosować bardziej skomplikowane reguły matematyki, a dzięki temu możemy uzyskać bardziej intuicyjne i wymierne zrozumienie danych.

Cztery poziomy danych

Ogólnie przyjmuje się, że konkretną cechę (cecha/kolumnę) danych strukturalnych można podzielić na jeden z czterech poziomów danych. Poziomy to:

• Poziom nominalny

• Poziom porządkowy

• Poziom interwału

• Poziom proporcji

W miarę przesuwania się w dół listy zyskujemy większą strukturę, a tym samym więcej zwrotów z naszej analizy. Każdy poziom ma własną przyjętą praktykę pomiaru środka danych. Zwykle myślimy o średniej/średniej jako o akceptowalnej formie centrum, jednak dotyczy to tylko określonego typu danych.

Poziom nominalny

Pierwszy poziom danych, poziom nominalny (który również brzmi jak słowo nazwa) składa się z danych, które są opisane wyłącznie nazwą lub kategorią. Podstawowe przykłady obejmują płeć, narodowość, gatunek lub szczep drożdży w piwie. Nie są one opisane liczbami i dlatego są jakościowe. Oto kilka przykładów:

• Rodzaj zwierzęcia znajduje się na nominalnym poziomie danych. Możemy również powiedzieć, że jeśli jesteś szympansem, to należysz również do klasy ssaków.

• Część mowy jest również uwzględniana na nominalnym poziomie danych. Słowo ona jest zaimkiem, a także rzeczownikiem.

Oczywiście, będąc jakościowymi, nie możemy wykonywać żadnych ilościowych operacji matematycznych, takich jak dodawanie czy dzielenie. To nie miałoby sensu.

Dozwolone operacje matematyczne

Nie możemy wykonywać matematyki na nominalnym poziomie danych, z wyjątkiem podstawowych funkcji równości i przynależności zbioru, jak pokazano w następujących dwóch przykładach:

• Bycie przedsiębiorcą technologicznym jest tym samym, co bycie w branży technologicznej, ale nie odwrotnie

• Figura opisana jako kwadrat podpada pod opis bycia prostokątem, ale nie odwrotnie

Miary centrum

Miarą środka jest liczba opisująca tendencję do danych. Czasami określa się go jako punkt równowagi danych. Typowe przykłady to średnia, mediana i tryb.

Aby znaleźć centrum danych nominalnych, zazwyczaj zwracamy się do trybu (najczęstszy element) zbioru danych. Spójrz na przykład wstecz na dane WHO dotyczące spożycia alkoholu. Najczęściej badanym kontynentem była Afryka, co czyni go możliwym wyborem na środek kolumny kontynentu. Miary środka, takie jak średnia i mediana, nie mają sensu na tym poziomie, ponieważ nie możemy uporządkować obserwacji ani nawet dodać ich do siebie.

Jak wyglądają dane na poziomie nominalnym

Dane na poziomie nominalnym mają głównie charakter kategoryczny. Ponieważ generalnie możemy używać tylko słów do opisania danych, mogą one zostać utracone w tłumaczeniu między krajami, a nawet mogą zostać błędnie napisane. Chociaż dane na tym poziomie z pewnością mogą być przydatne, musimy uważać na to, jakie wnioski możemy z nich wyciągnąć. Mając tylko tryb jako podstawową miarę środka, nie jesteśmy w stanie wyciągnąć wniosków na temat średniej obserwacji. Ta koncepcja nie istnieje na tym poziomie. Dopiero na następnym poziomie możemy zacząć wykonywać prawdziwą matematykę na naszych obserwacjach.

Poziom porządkowy

Poziom nominalny nie zapewniał nam dużej elastyczności w zakresie działań matematycznych ze względu na jeden pozornie nieistotny fakt - nie mogliśmy w naturalny sposób uporządkować obserwacji. Dane na poziomie porządkowym zapewniają nam porządek rangowy lub sposób na umieszczenie jednej obserwacji przed drugą; jednak nie zapewnia nam względnych różnic między obserwacjami, co oznacza, że chociaż możemy uporządkować obserwacje od pierwszej do ostatniej, nie możemy ich dodawać ani odejmować, aby uzyskać jakiekolwiek rzeczywiste znaczenie.

Przykłady

Likert jest jedną z najpopularniejszych skal porządkowych. Za każdym razem, gdy otrzymujesz ankietę z prośbą o ocenę zadowolenia w skali od 1 do 10, podajesz dane na poziomie porządkowym. Twoja odpowiedź, która musi zawierać się w przedziale od 1 do 10, może być uporządkowana: osiem jest

lepsze niż siedem, a trzy gorsze niż dziewięć. Jednak różnice między liczbami nie mają większego sensu. Różnica między siódemką a szóstką może być inna niż różnica między dwójką a jedynką.

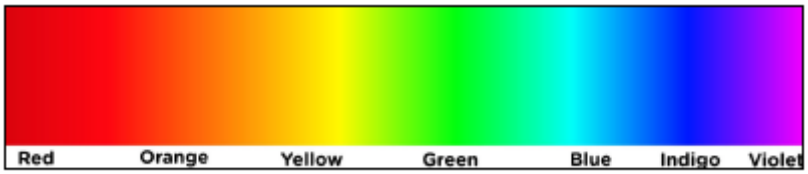
Dozwolone operacje matematyczne

Na tym poziomie mamy dużo więcej swobody w operacjach matematycznych. Dziedziczymy całą matematykę z poziomu porządkowego (równość i przynależność do zbioru), a do listy operacji dozwolonych na poziomie nominalnym możemy również dodać:

• Porządkowanie

• Porównanie

Zamówienie odnosi się do naturalnego porządku, jaki zapewniają nam dane; jednak czasami może to być trudne do rozszyfrowania. Mówiąc o widmie światła widzialnego, możemy odnieść się do nazw kolorów - czerwony, pomarańczowy, żółty, zielony, niebieski, indygo i fioletowy. Naturalnie, gdy poruszamy się od lewej do prawej, światło nabiera energii i inne właściwości. Możemy to nazwać porządkiem naturalnym.



Jednak w razie potrzeby artysta może narzucić innym porządek danych, np. sortować kolory w oparciu o koszt materiału, z którego dany kolor będzie wykonany. Może to zmienić kolejność danych, ale dopóki jesteśmy konsekwentni w tym, co definiuje kolejność, nie ma znaczenia, co ją definiuje. Porównania to kolejna nowa operacja dozwolona na tym poziomie. Na poziomie porządkowym nie miałoby sensu mówienie, że jeden kraj był z natury lepszy od drugiego lub że jedna część mowy jest gorsza od drugiej. Na poziomie porządkowym możemy dokonać tych porównań. Na przykład możemy porozmawiać o tym, że umieszczenie „7” w ankiecie jest gorsze niż umieszczenie „10”.

Miary centrum

Na poziomie porządkowym mediana jest zwykle odpowiednim sposobem zdefiniowania centrum danych. Środek byłby jednak niemożliwy, ponieważ podział nie jest dozwolony na tym poziomie. Możemy też korzystać z trybu, tak jak moglibyśmy na poziomie nominalnym. Przyjrzymy się teraz przykładowi użycia mediany: Wyobraź sobie, że przeprowadziłeś ankietę wśród swoich pracowników, pytając „jak bardzo jesteś zadowolony, że tu pracujesz w skali od 1 do 5”, a Twoje wyniki są następujące:

5, 4, 3, 4, 5, 3, 2, 5, 3, 2, 1, 4, 5, 3, 4, 4, 5, 4, 2, 1, 4, 5, 4,

3, 2, 4, 4, 5, 4, 3, 2, 1

Użyjmy Pythona, aby znaleźć medianę tych danych. Warto zauważyć, że większość ludzi twierdzi, że średnia z tych wyników działałaby dobrze. Powodem, dla którego średnia nie byłaby tak matematycznie realna, jest to, że jeśli odejmiemy/dodamy dwie oceny, powiedzmy, że cztery minus dwa, różnica dwóch tak naprawdę nic nie znaczy. Jeśli dodawanie/odejmowanie wyników nie ma sensu, średnia też nie ma sensu.

import numpy

```
results = [5, 4, 3, 4, 5, 3, 2, 5, 3, 2, 1, 4, 5, 3, 4, 4, 5, 4, 2, 1,

4, 5, 4, 3, 2, 4, 4, 5, 4, 3, 2, 1]

sorted_results = sorted(results)

print sorted_results

'''

[1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 4,

4, 4, 5, 5, 5, 5, 5, 5, 5]

'''

print numpy.mean(results) # == 3.4375

print numpy.median(results) # == 4.0
```

Znak „” (potrójny apostrof) oznacza dłuższy (ponad dwie linie) komentarz. Działa w sposób podobny do #.

Okazuje się, że mediana jest nie tylko bardziej dźwięczna, ale sprawia, że wyniki ankiety wyglądają znacznie lepiej.

Szybkie podsumowanie i sprawdzenie

Do tej pory widzieliśmy połowę poziomów danych:

• Poziom nominalny

• Poziom porządkowy

Na poziomie nominalnym mamy do czynienia z danymi zwykle opisywanymi za pomocą słownictwa (ale czasami za pomocą liczb), bez uporządkowania i z niewielkim wykorzystaniem matematyki. Na poziomie porządkowym mamy dane, które można opisać za pomocą liczb, a także mają „naturalny” porządek, co pozwala nam umieścić jedno przed drugim. Spróbujmy sklasyfikować następujący przykład jako porządkowy lub nominalny :

• Pochodzenie ziaren w filiżance kawy

• Miejsce, które ktoś otrzymuje po ukończeniu wyścigu pieszego

• Metal użyty do wykonania medalu, który otrzymują po zajęciu miejsca we wspomnianym wyścigu

• Numer telefonu klienta

• Ile filiżanek kawy wypijasz dziennie

Poziom interwału

Teraz docieramy do czegoś ciekawego. Na poziomie przedziału zaczynamy patrzeć na dane, które można wyrazić za pomocą bardzo wymiernych środków i gdzie dozwolone są znacznie bardziej skomplikowane formuły matematyczne. Podstawowa różnica między poziomem porządkowym a poziomem interwału to po prostu różnica. Dane na poziomie interwału umożliwiają znaczące odejmowanie między punktami danych.

Przykład

Temperatura jest doskonałym przykładem danych na poziomie interwału. Jeśli jest 100 stopni Fahrenheita w Teksasie i 80 stopni Fahrenheita w Stambule w Turcji, to w Teksasie jest o 20 stopni cieplej niż w Stambule. Ten prosty przykład pozwala na o wiele więcej manipulacji na tym poziomie niż poprzednie przykłady.

(Nie) Przykład

Wydaje się, że przykład na poziomie porządkowym (przy użyciu ankiety od jednego do pięciu) pasuje do rachunku poziomu przedziału. Pamiętaj jednak, że różnica między wynikami (gdy je odejmiesz) nie ma sensu, dlatego danych tych nie można wywołać na poziomie interwału.

Dozwolone operacje matematyczne

Możemy użyć wszystkich operacji dozwolonych na niższych poziomach (porządkowanie, porównania itp.) wraz z dwoma innymi ważnymi operacjami:

• Dodawanie

• Odejmowanie

Uwzględnienie tych dwóch operacji pozwala nam mówić o danych na tym poziomie w zupełnie nowy sposób.

Miary centrum

Na tym poziomie możemy użyć mediany i trybu do opisania tych danych; jednak zwykle najdokładniejszym opisem centrum danych byłaby średnia arytmetyczna, częściej nazywana po prostu „średnią". Przypomnijmy, że definicja średniej wymaga zsumowania wszystkich pomiarów. Na poprzednich poziomach dodawanie było bez znaczenia; dlatego średnia straciłaby na wartości ekstremalnej. Dopiero na poziomie interwału i powyżej średnia arytmetyczna ma sens. Przyjrzymy się teraz przykładowi użycia średniej. Załóżmy, że przyjrzymy się temperaturze lodówki zawierającej nową szczepionkę firmy farmaceutycznej. Mierzymy temperaturę umiarkowaną co godzinę za pomocą następujących punktów danych (w stopniach Fahrenheita):

31, 32, 32, 31, 28, 29, 31, 38, 32, 31, 30, 29, 30, 31, 26

Używając ponownie Pythona, znajdźmy średnią i medianę danych:

```
import numpy

temps = [31, 32, 32, 31, 28, 29, 31, 38, 32, 31, 30, 29, 30, 31, 26]

print numpy.mean(temps) # == 30.73

print numpy.median(temps) # == 31.0
```

Zauważ, że średnia i mediana są dość blisko siebie i obie wynoszą około 31 stopni. Pytanie jak średnio zimna jest lodówka?, około 31, jednak do szczepionki dołączone jest ostrzeżenie: Nie trzymaj tej szczepionki w temperaturze poniżej 29 stopni. Zauważ, że co najmniej dwa razy temperatura spadła poniżej 29 stopni, ale ostatecznie założyłeś, że to nie wystarczy, aby była szkodliwa. W tym miejscu miara zmienności może pomóc nam zrozumieć, jak zła może być sytuacja w lodówce.

Miary zmienności

To jest coś nowego, o czym jeszcze nie rozmawialiśmy. Jedną rzeczą jest mówienie o centrum danych, ale w nauce o danych bardzo ważne jest również, aby wspomnieć, jak „rozprzestrzeniają się" dane.

Miary opisujące to zjawisko nazywane są miarami zmienności. Prawdopodobnie słyszałeś już o „odchyleniu standardowym” i teraz doświadczasz łagodnego PTSD na zajęciach statystycznych. Ten pomysł jest niezwykle ważny i chciałbym się nim pokrótce omówić. Miarą zmienności (jak odchylenie standardowe) jest liczba, która próbuje opisać, jak rozłożone są dane. Wraz z miarą środka, miara zmienności może prawie całkowicie opisywać zbiór danych zawierający tylko dwie liczby.

Odchylenie standardowe

Prawdopodobnie odchylenie standardowe jest najczęstszą miarą zmienności danych na poziomie interwału i poza nim. Odchylenie standardowe można traktować jako „średnią odległość punktu danych od średniej”. Chociaż ten opis jest technicznie i matematycznie niepoprawny, jest to dobry sposób na myślenie o tym. Wzór na odchylenie standardowe można podzielić na następujące kroki:

1. Znajdź średnią danych.

2. Dla każdej liczby w zbiorze danych odejmij ją od średniej, a następnie podnieś ją do kwadratu.

3. Znajdź średnią każdego kwadratu różnicy.

4. Wyciągnij pierwiastek kwadratowy z liczby otrzymanej w kroku trzecim. To jest odchylenie standardowe.

Zwróć uwagę, jak w tych krokach faktycznie bierzemy średnią arytmetyczną jako jeden z kroków. Na przykład spójrz wstecz na zestaw danych temperatury. Znajdźmy odchylenie standardowe zbioru danych za pomocą Pythona:

```
import numpy

temps = [31, 32, 32, 31, 28, 29, 31, 38, 32, 31, 30, 29, 30, 31, 26]

mean = numpy.mean(temps) # == 30.73

squared_differences = []

# empty list o squared differences

for temperature in temps:

difference = temperature - mean

# how far is the point from the mean

squared_difference = difference**2

# square the difference

squared_differences.append(squared_difference)

# add it to our list

average_squared_difference = numpy.mean(squared_differences)

# This number is also called the "Variance"

standard_deviation = numpy.sqrt(average_squared_difference)

# We did it!

print standard_deviation # == 2.5157
```

Cały ten kod doprowadził nas do ustalenia, że odchylenie standardowe zestawu danych wynosi około 2,5, co oznacza, że „średnio” punkt danych jest oddalony o 2,5 stopnia od średniej temperatury wynoszącej około 31 stopni, co oznacza, że temperatura może prawdopodobnie spaść poniżej 29 stopni ponownie w najbliższej przyszłości. Powodem, dla którego chcemy „kwadratowej różnicy” między każdym punktem a średnią, a nie „rzeczywistą różnicą”, jest to, że podniesienie do kwadratu wartości faktycznie kładzie nacisk na wartości odstające - punkty danych, które są nienormalnie odległe.

Miary zmienności dają nam bardzo jasny obraz tego, jak rozproszone lub rozproszone są nasze dane. Jest to szczególnie ważne, gdy zajmujemy się zakresami danych i ich wahaniami (pomyśl o procentowym zwrocie z akcji). Duża różnica między danymi na tym a kolejnym poziomie polega na czymś, co nie jest oczywiste. Dane na poziomie interwału nie mają „naturalnego punktu startowego ani naturalnego zera”. Jednak przebywanie w temperaturze zerowej nie oznacza, że nie masz „temperatury”.

Poziom wskaźnika

Na koniec przyjrzymy się poziomowi wskaźnika. Po przejściu przez trzy różne poziomy z różnymi poziomami dozwolonych operacji matematycznych, poziom współczynnika okazuje się najsilniejszy z czterech. Nie tylko możemy zdefiniować kolejność i różnicę, ale także poziom proporcji pozwala nam mnożyć i dzielić. Może się wydawać, że nie jest to zbyt wielkie zamieszanie, ale zmienia prawie wszystko w sposobie, w jaki patrzymy na dane na tym poziomie.

Przykłady

Podczas gdy Fahrenheit i Celsjusz utknęli na poziomie interwału, skala temperatury Kelvina może pochwalić się naturalnym zerem. Pomiar zero Kelvin dosłownie oznacza brak ciepła. Jest to niearbitralne początkowe zero. W rzeczywistości możemy naukowo powiedzieć, że 200 kelwinów to dwa razy więcej ciepła niż 100 kelwinów. Pieniądze w banku są na poziomie ratio. Możesz „nie mieć pieniędzy w banku” i ma sens, że 200 000 dolarów to „dwa razy więcej niż” 100 000 dolarów.

Wiele osób może argumentować, że stopnie Celsjusza i Fahrenheita również mają punkt wyjścia (głównie dlatego, że możemy nawrócić się z Kelvina na jeden z nich). Prawdziwa różnica może wydawać się głupia, ale ponieważ przeliczenie na stopnie Celsjusza i Fahrenheita powoduje, że obliczenia przechodzą w ujemną wartość, nie definiuje ona wyraźnego i „naturalnego” zera.

Miary centrum

Średnia arytmetyczna nadal ma znaczenie na tym poziomie, podobnie jak nowy typ średniej, zwany średnią geometryczną. Miara ta na ogół nie jest tak często stosowana nawet na poziomie wskaźnika, ale warto o tym wspomnieć. Jest to pierwiastek kwadratowy iloczynu wszystkich wartości.

Na przykład w naszych danych dotyczących temperatury lodówki możemy obliczyć średnią geometryczną, jak pokazano tutaj:

```
import numpy

temps = [31, 32, 32, 31, 28, 29, 31, 38, 32, 31, 30, 29, 30, 31, 26]

num_items = len(temps)

product = 1.

for temperature in temps:
```

```
product *= temperature

geometric_mean = product**(1./num_items)

print geometric_mean # == 30.634
```

Zwróć uwagę, jak bardzo zbliża się do średniej arytmetycznej i mediany, jak obliczono wcześniej. Nie zawsze tak jest i będzie szczegółowo omówione w części poświęconej statystykom

Problemy z poziomem wskaźnika

Nawet przy całej tej dodanej funkcjonalności na tym poziomie, musimy ogólnie przyjąć bardzo duże założenie, które w rzeczywistości sprawia, że poziom współczynnika jest nieco restrykcyjny. Dane na poziomie wskaźnika są zwykle nieujemne. Już z tego powodu wielu analityków danych woli poziom interwału od poziomu współczynnika. Powodem tej restrykcyjnej właściwości jest to, że gdybyśmy dopuścili wartości ujemne, stosunek może nie zawsze mieć sens. Weźmy pod uwagę, że na przykładzie banku pozwoliliśmy, aby w naszych pieniądzach pojawił się dług. Gdybyśmy mieli saldo w wysokości 50 000 USD, następujący stosunek nie miałby żadnego sensu:

50 000$ / -50 000$ = -1

Dane są w oku patrzącego

Istnieje możliwość narzucenia struktury na dane. Na przykład, chociaż powiedziałem, że technicznie nie można użyć średniej dla danych od jednego do pięciu w skali porządkowej, wielu statystyków nie miałoby problemu z użyciem tej liczby jako deskryptora zbioru danych. Poziom interpretacji danych jest ogromnym założeniem, które należy przyjąć na początku każdej analizy. Jeśli patrzysz na dane, o których ogólnie myśli się na poziomie porządkowym, i stosujesz narzędzia, takie jak średnia arytmetyczna i odchylenie standardowe, to jest to coś, czego naukowcy zajmujący się danymi muszą być świadomi. Dzieje się tak głównie dlatego, że jeśli nadal będziesz uważać te założenia za ważne w swojej analizie, możesz napotkać problemy. Na przykład, jeśli przez pomyłkę zakładasz podzielność na poziomie porządkowym, narzucasz strukturę, w której struktura może nie istnieć.

Podsumowanie

Typ danych, z którymi pracujesz, to bardzo duża część nauki o danych. Musi poprzedzać większość twojej analizy, ponieważ rodzaj danych, które posiadasz, wpływa na typ analizy, który jest nawet możliwy! Za każdym razem, gdy masz do czynienia z nowym zbiorem danych, pierwsze trzy pytania, które powinieneś zadać, są następujące:

• Czy dane są zorganizowane lub niezorganizowane?

Na przykład, czy nasze dane istnieją w ładnej, czystej strukturze wierszy/kolumn?

• Czy każda kolumna jest ilościowa czy jakościowa?

Na przykład, czy wartości są liczbami, ciągami, czy też reprezentują ilości?

• Na jakim poziomie danych znajduje się każda kolumna?

Na przykład, czy wartości są na poziomie nominalnym, porządkowym, przedziałowym czy ilorazowym?

Odpowiedzi na te pytania nie tylko wpłyną na twoją znajomość danych na końcu, ale także będą dyktować kolejne etapy analizy. Będą dyktować rodzaje wykresów, których możesz użyć, i sposób ich interpretacji w przyszłych modelach danych. Czasami będziemy musieli przejść z jednego poziomu na drugi, aby zyskać większą perspektywę. W kolejnych częściach przyjrzymy się znacznie głębiej, jak

radzić sobie z danymi i badać je na różnych poziomach. Pod koniec tej książki będziemy w stanie nie tylko rozpoznawać dane na różnych poziomach, ale także będziemy wiedzieć, jak sobie z nimi radzić na tych poziomach.