Pięć obszarów nauki o danych

Nauka o danych wpływa na nasze współczesne życie na znacznie więcej sposobów, niż myślisz. Kiedy używasz Google, Bing lub DuckDuckGo, używasz bardzo wyrafinowanej aplikacji do nauki o danych. Sugestie dotyczące innych wyszukiwanych haseł, które pojawiają się podczas pisania? Te pochodzą z nauki o danych. Diagnozy medyczne oraz interpretacje obrazów i objawów to przykłady nauki o danych. W dzisiejszych czasach lekarze coraz częściej polegają na interpretacjach danych. Podobnie jak w przypadku większości tematów w tej książce, nauka o danych może wydawać się onieśmielająca dla niewtajemniczonych. Wnioski, wykresy danych i statystyki, o rany! Jednakże, podobnie jak w naszych poprzednich rozdziałach poświęconych sztucznej inteligencji, jeśli zagłębisz się i spojrzysz na kilka przykładów, naprawdę zrozumiesz, czym jest data science, a czym nie. Omówimy tylko tyle statystyk i „zadawania pytań dotyczących danych", abyś mógł zacząć i uzyskać proste wyniki. Celem jest zapoznanie Cię z wykorzystaniem Pythona w nauce o danych i omówienie wystarczającej ilości teorii, aby zacząć. Jeśli nic więcej, chcemy zostawić Cię z procesem nauki o danych i dać Ci wyższy poziom zrozumienia tego, co kryje się za niektórymi gadającymi głowami w telewizji i różnymi komunikatami prasowymi pochodzącymi z uniwersytetów. Ci ludzie zawsze cytują wyniki pochodzące z analizy dużych zbiorów danych i często wyolbrzymiają to, co tak naprawdę mają na myśli. Przykładem tego jest sytuacja, gdy jedno badanie mówi, że kawa jest dla ciebie zła, a w następnym miesiącu pojawia się badanie, że kawa jest dla ciebie dobra - i czasami badania opierają się na tych samych danych! Określenie, co oznaczają twoje wyniki, poza prostą interpretacją, to miejsce, w którym spotykają się naprawdę trudne części nauki o danych i statystyki, które są warte osobnej książki. Pod koniec naszej podróży do nauki o danych dowiesz się więcej o procesach związanych z odpowiadaniem na niektóre z tych pytań. Nauka o danych jest tajemnicą, ale przy odrobinie wiedzy i odrobinie Pythona możemy przeniknąć przez zasłonę i zająć się nauką o danych. Python oraz niezliczone dostępne narzędzia i biblioteki mogą sprawić, że nauka o danych stanie się znacznie bardziej dostępna. Należy pamiętać, że większość naukowców (w tym analityków danych) niekoniecznie jest ekspertami w dziedzinie informatyki. Lubią korzystać z narzędzi, które upraszczają kodowanie i pozwalają skupić się na uzyskaniu odpowiedzi i przeprowadzeniu analizy potrzebnych im danych.

Praca z dużymi, dużymi danymi

Media lubią rzucać wokół pojęcia „big data" i tego, jak ludzie mogą uzyskać wgląd w zachowanie konsumentów (i Twoje). Big data to termin używany w odniesieniu do dużych i złożonych zbiorów danych, które są zbyt duże, aby mogły obsłużyć tradycyjne oprogramowanie do przetwarzania danych (bazy danych do odczytu, arkusze kalkulacyjne i tradycyjne pakiety statystyczne, takie jak SPSS). Branża mówi o dużych zbiorach danych, używając trzech różnych koncepcji, zwanych „trzema V": wolumen (volume), różnorodność (variety) i prędkość (velocity).

Wolumen

Wolumen odnosi się do tego, jak duży jest zestaw danych, który rozważamy. Może być naprawdę, naprawdę duży - prawie trudny do uwierzenia. Na przykład Facebook ma więcej użytkowników niż populacja Chin. Na Facebooku znajduje się ponad 250 miliardów zdjęć i 2,5 biliona postów. To dużo danych. Naprawdę duża ilość danych. A co z nadchodzącym światem IOT (Internet of Things)? Gartner, jedna z wiodących na świecie firm analitycznych, szacowała, że do 2022 roku będzie 22 miliardy urządzeń. To 22 miliardy urządzeń wytwarzających tysiące danych. Wyobraź sobie, że raz na minutę przez rok mierzysz temperaturę w swojej kuchni. To ponad ½ miliona punktów danych. Dodaj wilgotność do pomiarów i masz 1 milion punktów danych. Pomnóż to przez pięć pokoi i garaż, wszystkie z pomiarami temperatury i wilgotności, a Twój dom generuje 6 milionów danych z jednego małego urządzenia IOT na pokój. Bardzo szybko robi się to szaloney. I spójrz na swój smartfon. Wyobraź sobie,

ile danych generuje w ciągu dnia. Lokalizacja, użycie, poziomy mocy, łączność z telefonem komórkowym są nieustannie wyrzucane z telefonu do baz danych oraz aplikacji i pulpitów nawigacyjnych aplikacji, takich jak Blynk. Czasami (jak niedawno dowiedzieliśmy się od firm telefonii komórkowej) informacje o lokalizacji są gromadzone i sprzedawane nawet bez Twojej zgody lub zgody. Dane, dane i jeszcze raz dane. Nauka o danych to sposób, w jaki to wykorzystujemy.

Różnorodność

Pamiętaj, że zdjęcia to bardzo różne typy danych od temperatury i wilgotności lub informacji o lokalizacji. Czasami idą razem, a czasami nie. Zdjęcia to bardzo wyrafinowane struktury danych, trudne do interpretacji i trudne do sklasyfikowania przez maszyny. Wrzuć do tego nagrania audio i masz dość zróżnicowany zestaw typów danych. Porozmawiajmy chwilę o głosie. Mówiłem o Alexie, która jest bardzo dobra w tłumaczeniu głosu na tekst, ale nie tak dobrze w przypisywaniu znaczenia tekstowi. Jednym z powodów jest brak kontekstu, ale innym powodem jest wiele różnych sposobów, w jakie ludzie proszą o rzeczy, komentują i tak dalej. Wyobraź sobie więc, że Alexa (i Amazon) śledzą wszystkie zapytania, a następnie przeprowadzają na ich podstawie analizę danych, aby dowiedzieć się, o jakie rzeczy proszą ludzie i na różne sposoby o nie proszą. To dużo danych i wiele informacji, które można zebrać. Nie tylko z nikczemnych powodów, ale aby zbudować system, który lepiej służy konsumentowi. To działa w obie strony. Nauka o danych ma znacznie większe szanse na zidentyfikowanie wzorców, jeśli głos został przetłumaczony na tekst. Jest to o wiele łatwiejsze. Jednak w tym tłumaczeniu tracisz wiele informacji na temat tonu głosu, akcentów i tak dalej.

Szybkość

Szybkość odnosi się do tego, jak szybko dane się zmieniają i jak szybko są dodawane do stosów danych. Użytkownicy Facebooka przesyłają około 1 miliarda zdjęć dziennie, więc w ciągu najbliższych kilku lat Facebook będzie miał ponad 1 bilion zdjęć. Facebook to zestaw danych o dużej prędkości. Zbiór danych o niskiej prędkości (w ogóle się nie zmienia) może być zbiorem odczytów temperatury i wilgotności z twojego domu w ciągu ostatnich pięciu lat. Nie trzeba dodawać, że zestawy danych o dużej szybkości wymagają innych technik niż zbiory danych o małej prędkości.

RÓŻNICA MIĘDZY DATA SCIENCE A ANALITYKĄ DANYCH

W prawdziwym sensie analiza danych jest podzbiorem nauki o danych - konkretnie krokami 3-5 na naszej liście nauki o danych. Jest wielu ludzi, którzy wciąż lubią rozróżniać te dwa typy naukowców, ale z biegiem czasu różnica ta staje się coraz mniej zauważalna. Opracowuje się coraz więcej technik umożliwiających analizę danych w oparciu o duże zbiory danych (co nie dziwi, że nazwano je „analizą dużych zbiorów danych"). Obecnie nauka o danych ogólnie odnosi się do procesu opracowywania spostrzeżeń z dużych zbiorów danych nieustrukturyzowanych. Oznacza to wykorzystanie analiz predykacyjnych, statystyk i uczenia maszynowego do przedzierania się przez masę danych. Analityka danych koncentruje się przede wszystkim na wykorzystaniu i tworzeniu analiz statystycznych dla istniejących zestawów danych w celu uzyskania wglądu w te dane. Dzięki tym nieco niejasnym opisom możesz zobaczyć, jak te dwa obszary zbliżają się do siebie. Ryzykując wyśmiewanie się ze strony moich kolegów akademickich, zdecydowanie nazwałbym analitykę danych podzbiorem nauki o danych.

Zarządzanie wolumenem, różnorodnością i szybkością

To bardzo złożony temat. Analitycy danych opracowali wiele metod przetwarzania danych z odmianami trzech V. Trzy V opisują zestaw danych i dają wyobrażenie o parametrach konkretnego zestawu danych. Proces uzyskiwania wglądu w dane nazywa się analizą danych. W kolejnych rozdziałach skupimy się na zdobyciu wiedzy na temat analityki oraz nauczeniu się zadawania niektórych

pytań z zakresu analizy danych za pomocą Pythona. Po kilku latach nauki o danych będziesz bardzo dobry w zarządzaniu nimi.

Gotowanie na gazie: pięcioetapowy proces analizy danych

Zasadniczo możemy podzielić proces prowadzenia badań naukowych na danych (zwłaszcza big data) na pięć etapów. Zakończę ten rozdział wprowadzający, omawiając każdy z tych kroków, aby dać nam pojęcie o przebiegu procesu nauki o danych i wyczuciu złożoności zadań. Te kroki są

1. Przechwyć dane

2. Przetwarzaj dane

3. Przeanalizuj dane

4. Przekaż wyniki

5. Zachowaj dane

Przechwytywanie danych

Aby mieć coś do analizy, musisz przechwycić pewne dane. W każdej rzeczywistej sytuacji prawdopodobnie masz wiele potencjalnych źródeł danych. Zinwentaryzuj je i zdecyduj, co uwzględnić. Wiedza o tym, co uwzględnić, wymaga starannego zdefiniowania warunków biznesowych i celów na nadchodzącą analizę. Czasami twoje cele mogą być niejasne, ponieważ czasami „chcesz po prostu zobaczyć, co możesz uzyskać" z danych. Jeśli możesz, zintegruj swoje źródła danych, aby łatwo było uzyskać informacje potrzebne do uzyskania wglądu i stworzenia tych wszystkich fajnych raportów, którymi po prostu nie możesz się doczekać, aby pochwalić się kierownictwu.

Przetwarzanie danych

Moim skromnym zdaniem jest to część nauki o danych, która powinna być łatwa, ale prawie nigdy taka nie jest. Widziałem, jak analitycy danych spędzają miesiące masując swoje dane, aby móc je przetwarzać i im ufać. Musisz zidentyfikować anomalie i wartości odstające, wyeliminować duplikaty, usunąć brakujące wpisy i dowiedzieć się, które dane są niespójne. A wszystko to musi być wykonane odpowiednio, aby nie usuwać danych, które są ważne dla przyszłych prac analitycznych. W wielu przypadkach nie jest to łatwe. Jeśli masz temperaturę w domu, która wynosi 170 stopni C, łatwo zauważyć, że te dane są błędne i niespójne. (No chyba, że twój dom się pali.) Czyszczenie i przetwarzanie twoich danych musi być wykonywane ostrożnie, bo inaczej doprowadzisz do stronniczości i być może zniszczysz zdolność do wyciągania właściwych wniosków lub uzyskiwania dobrych odpowiedzi w dalszej kolejności. W prawdziwym świecie spodziewaj się spędzić dużo czasu na wykonaniu tego kroku. Aha, i jeszcze jedna sprawa do sprzątania, o którą należy się martwić, początkujący konsumenci zajmujący się analizą danych udostępniają w Internecie coraz więcej fałszywych i wprowadzających w błąd danych. Według Marketing Week w 2015 roku 60 procent konsumentów celowo podaje nieprawdziwe informacje podczas przesyłania danych online. Pokornie przyznajemy, że robimy to cały czas w internetowych formularzach marketingowych, a nawet w sondażach politycznych, zwłaszcza gdy wyczuwamy w pytaniach program polityczny. Jesteśmy złymi chłopcami Zrozum, że wystarczy bardzo niewielka ilość nieproporcjonalnych informacji, aby radykalnie zdewaluować bazę danych. Więcej materiału do przemyśleń.

Analiza danych

Zanim zużyjesz całą energię, aby faktycznie spojrzeć na dane i zobaczyć, co możesz znaleźć, możesz pomyśleć, że zadawanie pytań powinno być stosunkowo proste. Nie jest. Analizowanie dużych

zestawów danych pod kątem spostrzeżeń i wniosków, a nawet zadawanie złożonych pytań to najtrudniejsze wyzwanie, które wymaga najbardziej ludzkiej intuicji w całej nauce o danych. Niektóre pytania, takie jak „Jakie są średnie pieniądze wydawane na zboża w 2017 roku?" można łatwo zdefiniować i obliczyć, nawet w przypadku ogromnych ilości danych. Ale wtedy masz naprawdę, naprawdę przydatne pytania, takie jak: „Jak mogę zachęcić więcej ludzi do kupowania płatków Sugar Frosted Flakes?" To jest pytanie za 64 000 $. W bezczelnej próbie bycia bardziej naukowymi będziemy nazywać Sugar Frosted Flakes akronimem SFF. Pytanie takie jak to ma za sobą warstwy i warstwy złożoności. Chcesz mieć punkt odniesienia, ile SFF kupują obecnie Twoi klienci. To powinno być całkiem łatwe. Następnie musisz zdefiniować, co rozumiesz przez większą liczbę osób. Czy naprawdę masz na myśli więcej ludzi, czy masz na myśli większe przychody? Zmień cenę na 0,01 USD za pudełko, a będziesz mieć znacznie więcej osób kupujących SFF. Naprawdę chcesz więcej przychodów, a dokładniej, większej marży (marża = cena – koszt). Pytanie jest już bardziej złożone. Ale najtrudniejszą częścią pytania jest to, jak zmotywujemy ludzi do kupowania większej ilości SFF? I czy odpowiedź jest zawarta w naszych danych, które zebraliśmy? To jest najtrudniejsza część analizy: upewnienie się, że zadajemy właściwe pytanie we właściwy sposób z właściwego rodzaju danych. Analiza danych wymaga umiejętności i doświadczenia w technikach statystycznych, takich jak regresja liniowa i logistyczna oraz znajdowanie korelacji między różnymi typami danych za pomocą różnych algorytmów prawdopodobieństwa i wzorów, takich jak niezwykle fajnie nazwane formuły i koncepcje „naiwnego Bayesa". Chociaż pełne omówienie tych technik wykracza poza zakres tej książki, później przejdziemy do kilku przykładów.

Komunikowanie wyników

Po zgnieceniu i zniekształceniu danych do wymaganego formatu, a następnie przeanalizowaniu danych w celu uzyskania odpowiedzi na pytania, należy przedstawić wyniki kierownictwu lub klientowi. Większość ludzi lepiej i szybciej wizualizuje informacje, gdy widzą je w formie graficznej, a nie tylko w formie tekstowej. Istnieją dwa główne pakiety Pythona, którymi zajmuje się nauka o danych: język „R" i atPlotLib. Używamy MatPlotLib do wyświetlania naszej „grafiki Big Data". (Jeśli przeczytałeś Części o sztucznej inteligencji, to znasz już MatPlotLib z pierwszej ręki.)

Utrzymanie danych

Jest to krok w nauce o danych, który wszyscy ignorują. Po zadaniu pierwszej rundy pytań i uzyskaniu pierwszej rundy odpowiedzi wielu profesjonalistów po prostu po prostu się zamknie i przejdzie do następnego projektu. Problem z takim sposobem myślenia polega na tym, że istnieje bardzo duże prawdopodobieństwo, że będziesz musiał zadać więcej pytań dotyczących tych samych danych, czasem w bardzo odległej przyszłości. Ważne jest, aby zarchiwizować i udokumentować następujące informacje, aby można było szybko wznowić projekt, a co bardziej prawdopodobne, że w przyszłości napotkasz podobny zestaw problemów i będziesz mógł szybko odkurzyć modele i szybciej uzyskać odpowiedzi.

* Poświęć trochę czasu na zachowanie:

* Dane i źródła

* Modele, których użyłeś do modyfikacji danych (w tym wszelkie dane wyjątków i zastosowane „kryteria odrzucania danych")

* Zapytania i wyniki otrzymane z zapytań