Nauczanie maszynowe . Wydobywanie spostrzeżeń z danych.

Chociaż Katrina Lake lubiła robić zakupy online, wiedziała, że może być znacznie lepiej. Główny problem: trudno było znaleźć modę spersonalizowaną. Tak zaczęła się inspiracja dla Stitch Fix, którą Katrina uruchomiła w swoim mieszkaniu w Cambridge podczas nauki w Harvard Business School w 2011 roku (swoją drogą, oryginalna nazwa firmy brzmiała mniej chwytliwie „Rack Habit"). Witryna zawierała pytania i odpowiedzi dla swoich użytkowników - pytając o rozmiar i style ubioru, żeby wymienić tylko kilka czynników - a doświadczeni styliści składali następnie wyselekcjonowane pudełka z ubraniami i akcesoriami, które były wysyłane co miesiąc. Koncepcja szybko się przyjęła, a wzrost był solidny. Trudno było jednak pozyskać kapitał, ponieważ wielu inwestorów venture capital nie widziało potencjału w tym biznesie. Jednak Katrina uparła się i była w stanie stworzyć dochodową operację - dość szybko. Po drodze program Stitch Fix zbierał ogromne ilości cennych danych, takich jak rozmiary ciała i preferencje dotyczące stylu. Katrina zdała sobie sprawę, że byłoby to idealne rozwiązanie do uczenia maszynowego. Aby to wykorzystać, zatrudniła Erica Colsona, który był wiceprezesem Data Science and Engineering w Netflix, a jego nowym tytułem jest dyrektor ds. algorytmów. Ta zmiana strategii była kluczowa. Modele uczenia maszynowego stawały się coraz lepsze dzięki ich przewidywaniom, ponieważ Stitch Fix zebrał więcej danych - nie tylko z początkowych ankiet, ale także z bieżących informacji zwrotnych. Dane zostały również zakodowane w jednostkach SKU. Rezultat: Stitch Fix zapewnił stałą poprawę lojalności klientów i współczynników konwersji. Nastąpiła również poprawa w obrocie zapasami, co pomogło obniżyć koszty. Ale nowa strategia nie oznaczała zwolnienia stylistów. Uczenie maszynowe znacznie zwiększyło ich produktywność i skuteczność. Dane dostarczyły również informacji o tym, jakie rodzaje odzieży należy tworzyć. Doprowadziło to do uruchomienia Hybrid Designs w 2017 roku, który jest marką prywatną Stitch Fix. Okazało się to skuteczne w radzeniu sobie z lukami w zapasach. Do listopada 2017 r. Katrina upubliczniła Stitch Fix, zdobywając 120 milionów dolarów. Firma została wyceniona na niezłe 1,63 miliarda dolarów, co czyni ją jedną z najbogatszych kobiet w Stanach Zjednoczonych. Aha, w tym czasie miała 14-miesięcznego syna! Przechodząc do dzisiaj, Stitch Fix ma 2,7 miliona klientów w Stanach Zjednoczonych i generuje ponad 1,2 miliarda dolarów przychodów. Zatrudnionych jest również ponad 100 naukowców zajmujących się danymi, a większość z nich ma doktoraty w dziedzinach takich jak neuronauka, matematyka, statystyka i sztuczna inteligencja. Według zgłoszenia firmy 10-K:

Nasze możliwości w zakresie analizy danych napędzają naszą działalność. Na te możliwości składa się nasz bogaty i stale powiększający się zestaw szczegółowych danych o klientach i towarach oraz nasze autorskie algorytmy. Korzystamy z analizy danych w całej naszej działalności, w tym do stylizowania naszych klientów, przewidywania zachowań zakupowych, prognozowania popytu, optymalizacji zapasów i projektowania nowej odzieży.

Bez wątpienia historia Stitch Fix wyraźnie pokazuje niesamowitą moc uczenia maszynowego i to, jak może ono zrewolucjonizować branżę. W wywiadzie dla digiday.com Lake zauważyła:

Historycznie istniała przepaść między tym, co dajesz firmom, a tym, jak bardzo poprawia się ich doświadczenie. Wielkie zbiory danych śledzą Cię w całej sieci, a największą korzyścią, jaką teraz z tego czerpiesz, jest: jeśli klikniesz parę butów, zobaczysz tę parę ponownie za tydzień. Zobaczymy, jak ta przepaść zacznie się zamykać. Oczekiwania są bardzo różne w odniesieniu do personalizacji, ale co ważne, jej autentycznej wersji. Nie: „Porzuciłeś swój koszyk, a my to rozpoznajemy". Będzie to autentyczne rozpoznanie, kim jesteś jako wyjątkowy człowiek. Jedynym sposobem na zrobienie tego w sposób skalowalny jest uwzględnienie nauki o danych i tego, co można zrobić dzięki innowacjom.

No dobrze, na czym tak naprawdę polega uczenie maszynowe? Dlaczego może mieć taki wpływ? A jakie zagrożenia należy wziąć pod uwagę? Odpowiemy na te pytania i nie tylko.

Co to jest uczenie maszynowe?

Po przepracowaniu w MIT i Bell Telephone Laboratories, Arthur L. Samuel dołączył do IBM w 1949 roku w Poughkeepsie Laboratory. Jego wysiłki pomogły zwiększyć moc obliczeniową maszyn firmy, na przykład przy opracowaniu 701 (był to pierwszy skomercjalizowany system komputerowy IBM). Ale programował też aplikacje. I była taka, która zapisze się w historii, to znaczy jego komputerowa gra w warcaby. Był to pierwszy przykład systemu uczenia maszynowego (Samuel opublikował wpływowy artykuł na ten temat w 1959 roku). CEO IBM, Thomas J. Watson, Sr., powiedział, że innowacja zwiększy cenę akcji o 15 punktów! Dlaczego więc artykuł Samuela był tak istotny? Patrząc na warcaby, pokazał, jak działa uczenie maszynowe - innymi słowy, komputer może uczyć się i ulepszać, przetwarzając dane bez konieczności wyraźnego programowania. Było to możliwe dzięki wykorzystaniu zaawansowanych koncepcji statystyki, zwłaszcza analizy prawdopodobieństwa. W ten sposób komputer można wytrenować, aby dokonywał dokładnych prognoz. Było to rewolucyjne, ponieważ tworzenie oprogramowania w tym czasie było głównie o liście poleceń, które następowały po przepływie pracy logiki. Aby zrozumieć, jak działa uczenie maszynowe, użyjmy przykładu z serialu komediowego HBO Dolina Krzemowa. Inżynier Jian-Yang miał stworzyć Shazam do jedzenia. Aby wytrenować aplikację, musiał dostarczyć ogromny zestaw danych ze zdjęciami jedzenia. Niestety, z powodu ograniczeń czasowych aplikacja nauczyła się dopiero rozpoznawać hot dogi. Innymi słowy, jeśli użyjesz aplikacji, odpowie tylko „hot dog” i „nie hot dog”. Choć zabawny, odcinek wykonał całkiem niezłą robotę, demonstrując uczenie maszynowe. W istocie jest to proces przyjmowania oznaczonych danych i znajdowania relacji. Jeśli wytrenujesz system z hot dogami – na przykład z tysiącami obrazów – będzie coraz lepiej je rozpoznawał. Tak, nawet programy telewizyjne mogą dostarczyć cennych lekcji na temat sztucznej inteligencji! Ale oczywiście nadal potrzebujesz znacznie więcej. W następnej części rozdziału przyjrzymy się dokładniej podstawowym statystykom, które musisz wiedzieć o uczeniu maszynowym. Obejmuje to odchylenie standardowe, rozkład normalny, twierdzenie Bayesa, korelację i ekstrakcję cech. Następnie omówimy takie tematy, jak przypadki użycia uczenia maszynowego, ogólny proces i popularne algorytmy.

Odchylenie standardowe

Odchylenie standardowe mierzy średnią odległość od średniej. W rzeczywistości nie ma potrzeby uczenia się, jak to obliczyć (proces obejmuje wiele kroków), ponieważ Excel lub inne oprogramowanie może to zrobić z łatwością. Aby zrozumieć odchylenie standardowe, weźmy przykład wartości domu w Twojej okolicy. Załóżmy, że średnia wynosi 145 000 USD, a odchylenie standardowe to 24 000 USD. Oznacza to, że jedno odchylenie standardowe poniżej średniej wyniosłoby 133 000 USD (145 000 USD - 12 000 USD), a jedno odchylenie standardowe powyżej średniej wyniosłoby 157 000 USD (145 000 USD + 12 000 USD). Daje nam to możliwość ilościowego określenia zmienności danych. Oznacza to, że istnieje spread w wysokości 24 000 USD od średniej. Następnie spójrzmy na dane, jeśli, no cóż, Mark Zuckerberg przenosi się do twojego sąsiedztwa i w rezultacie średnia skacze do 850 000 USD, a odchylenie standardowe wynosi 175 000 USD. Ale czy te dane statystyczne odzwierciedlają wyceny? Nie bardzo. Zakup Zuckerberga jest czymś odstającym. W takiej sytuacji najlepszym rozwiązaniem może być zamiast tego wykluczenie jego domu.

Rozkład normalny

Na wykresie rozkład normalny wygląda jak dzwon (dlatego inna nazwa to „krzywa dzwonowa”). Reprezentuje sumę prawdopodobieństw dla zmiennej. Co ciekawe, krzywa normalna jest powszechna w świecie przyrody, ponieważ odzwierciedla rozkład takich rzeczy, jak wzrost i waga. Ogólnym

podejściem do interpretacji rozkładu normalnego jest zastosowanie reguły 68-95-99,7. Szacuje się, że 68% pozycji danych będzie mieścić się w obrębie jednego odchylenia standardowego, 95% w obrębie dwóch odchyleń standardowych, a 99,7% w obrębie trzech odchyleń standardowych. Sposobem na zrozumienie tego jest użycie wyników IQ. Załóżmy, że średni wynik to 100, a odchylenie standardowe to 15. Otrzymalibyśmy to dla trzech odchyleń standardowych Zauważ, że szczyt na tym wykresie jest średnią. Tak więc, jeśli dana osoba ma IQ 145, to tylko 0,15% będzie miało wyższy wynik. Teraz krzywa może mieć różne kształty, w zależności od zmienności danych. Na przykład, jeśli nasze dane IQ mają dużą liczbę geniuszy, rozkład będzie przekrzywiony w prawo.

Twierdzenie Bayesa

Jak sama nazwa wskazuje, statystyki opisowe dostarczają informacji o Twoich danych. Widzieliśmy już to z takimi rzeczami jak średnie i odchylenia standardowe. Ale oczywiście możesz wyjść daleko poza to - w zasadzie, używając twierdzenia Bayesa. Takie podejście jest powszechne w analizie chorób medycznych, w których przyczyna i skutek są kluczowymi słowami dla badań FDA (Federal Drug Administration). Aby zrozumieć, jak działa twierdzenie Bayesa, weźmy przykład. Badacz wymyśla test na określony rodzaj nowotworu i okazuje się, że jest dokładny w 80% przypadków. Nazywa się to prawdziwym pozytywem. Jednak w 9,6% przypadków test zidentyfikuje osobę jako chorą na raka, nawet jeśli go nie ma, co jest znane jako fałszywie dodatni. Pamiętaj, że w niektórych testach narkotykowych odsetek ten może być wyższy niż wskaźnik dokładności! I wreszcie 1% populacji ma raka. W świetle tego wszystkiego, jeśli lekarz użyje testu na tobie i wykaże, że masz raka, jakie jest prawdopodobieństwo, że naprawdę masz raka? Cóż, twierdzenie Bayesa wskaże drogę. To obliczenie wykorzystuje takie czynniki, jak wskaźniki dokładności, wyniki fałszywie dodatnie, a wskaźnik populacji do wymyślenia prawdopodobieństwa:

• Krok #1: 80% wskaźnik dokładności × prawdopodobieństwo zachorowania na raka (1%) = 0,008.

• Krok #2: Szansa na brak raka (99%) × 9,6% fałszywie dodatnich = 0,09504.

• Krok #3: Następnie wstaw powyższe liczby do następującego równania: 0,008 / (0,008 + 0,09504) = 7,8%.

Brzmi trochę dziwnie, prawda? Zdecydowanie. W końcu, jak to się dzieje, że test, który jest dokładny w 90%, ma tylko 7,8% prawdopodobieństwa poprawności? Pamiętaj jednak, że wskaźnik dokładności opiera się na pomiarze tych, którzy mają grypę. A to niewielka liczba, ponieważ tylko 1% populacji ma grypę. Co więcej, test nadal daje fałszywie pozytywne wyniki. Twierdzenie Bayesa jest więc sposobem na lepsze zrozumienie wyników – co ma kluczowe znaczenie dla systemów takich jak sztuczna inteligencja.

Korelacja

Algorytm uczenia maszynowego często obejmuje pewien rodzaj korelacji między danymi. Ilościowym sposobem opisania tego jest wykorzystanie korelacji Pearsona, która pokazuje siłę związku między dwiema zmiennymi w zakresie od 1 do –1 (jest to współczynnik). Oto jak to działa:

• Większe niż 0: W tym przypadku wzrost jednej zmiennej prowadzi do wzrostu innej. Na przykład: Załóżmy, że istnieje korelacja 0,9 między dochodem a wydatkami. Jeśli dochód wzrośnie o 1000 USD, wydatki wzrosną o 900 USD (1000 USD x 0,9).

• 0: Nie ma korelacji między dwiema zmiennymi.

• Mniej niż 0: Każdy wzrost zmiennej oznacza spadek innej i odwrotnie. To opisuje odwrotną zależność.

Czym zatem jest silna korelacja? Ogólna zasada mówi, że współczynnik wynosi około +0,7. A jeśli jest poniżej 0,3, to korelacja jest słaba. Wszystko to nawiązuje do starego powiedzenia: „Korelacja niekoniecznie jest przyczyną". Jednak jeśli chodzi o uczenie maszynowe, tę koncepcję można łatwo zignorować i prowadzić do mylących wyników. Na przykład istnieje wiele korelacji, które są po prostu losowe. W rzeczywistości niektóre mogą być wręcz komiczne. Sprawdź następujące informacje z Tylervigen.com:

• Wskaźnik rozwodów w Maine wykazuje 99,26% korelacji ze spożyciem margaryny na mieszkańca.

• Wiek Miss America ma 87,01% korelacji z morderstwami parą wodną, gorącymi oparami i gorącymi tropikami.

• Import ropy naftowej do USA z Norwegii wykazuje 95,4% korelacji z kierowcami zabitymi w zderzeniu z pociągiem kolejowym.

Jest na to nazwa: wzorzystość. Jest to tendencja do znajdowania wzorców w bezsensownym szumie.

Ekstrakcja funkcji

W Części 2 przyjrzeliśmy się doborowi zmiennych do modelu. Proces ten jest często nazywany wyodrębnianiem funkcji lub inżynierią funkcji.

Przykładem może być model komputerowy, który identyfikuje mężczyznę lub kobietę na zdjęciu. Dla ludzi jest to dość łatwe i szybkie. To coś intuicyjnego. Ale gdyby ktoś poprosił Cię o opisanie różnic, czy byłbyś w stanie to zrobić? Dla większości ludzi byłoby to trudne zadanie. Jeśli jednak chcemy zbudować skuteczny model uczenia maszynowego, musimy prawidłowo wyodrębnić funkcje - a to może być subiektywne. To tylko zarysowuje powierzchnię, ponieważ jestem pewien, że masz własne pomysły lub podejścia. I to jest normalne. Ale właśnie dlatego takie rzeczy jak rozpoznawanie twarzy są bardzo złożone i obarczone błędami. Ekstrakcja funkcji ma również pewne niuanse. Jednym z nich jest potencjał stronniczości. Na przykład, czy masz uprzedzenia dotyczące tego, jak wygląda mężczyzna lub kobieta? Jeśli tak, może to spowodować, że modele dadzą błędne wyniki. Z tego powodu dobrze jest mieć grupę ekspertów, którzy potrafią określić odpowiednie cechy. A jeśli inżynieria funkcji okaże się zbyt złożona, uczenie maszynowe prawdopodobnie nie jest dobrym rozwiązaniem. Ale jest jeszcze inne podejście do rozważenia: głębokie uczenie. Obejmuje to wyrafinowane modele, które znajdują funkcje w danych. Właściwie jest to jeden z powodów, dla których uczenie głębokie stało się wielkim przełomem w sztucznej inteligencji.

Co możesz zrobić dzięki uczeniu maszynowemu?

Ponieważ uczenie maszynowe istnieje od dziesięcioleci, ta potężna technologia ma wiele zastosowań. Pomaga również w uzyskaniu wyraźnych korzyści pod względem oszczędności kosztów, możliwości uzyskania przychodów i monitorowania ryzyka. Aby przybliżyć niezliczone zastosowania, przyjrzyjmy się kilku przykładom:

• Konserwacja predykcyjna: Monitoruje czujniki, aby przewidzieć, kiedy sprzęt może ulec awarii. Pomaga to nie tylko obniżyć koszty, ale także skraca przestoje i zwiększa bezpieczeństwo. W rzeczywistości firmy takie jak PrecisionHawk faktycznie używają dronów do zbierania danych, co jest znacznie bardziej wydajne. Technologia okazała się dość skuteczna w branżach takich jak energetyka, rolnictwo i budownictwo. Oto, co PrecisionHawk zauważa na temat własnego systemu konserwacji predykcyjnej opartego na dronach: „Jeden klient przetestował wykorzystanie dronów z wizualną linią wzroku (VLOS), aby sprawdzić klaster 10 studzienek w promieniu trzech mil. Nasz klient ustalił, że

użycie dronów obniżyło koszty inspekcji o około 66%, z 80-90 USD za podkładkę od tradycyjnej metodologii inspekcji do 45–60 USD za studnię przy użyciu misji dronów VLOS".

• Rekrutacja pracowników: może to być żmudny proces, ponieważ wiele życiorysów jest często zróżnicowanych. Oznacza to, że łatwo jest pominąć świetnych kandydatów. Jednak uczenie maszynowe może w dużym stopniu pomóc. Spójrz na CareerBuilder, który zebrał i przeanalizował ponad 2,3 miliona ofert pracy, 680 milionów unikalnych profili, 310 milionów unikalnych życiorysów, 10 milionów tytułów pracy, 1,3 miliarda umiejętności i 2,5 miliona weryfikacji przeszłości, aby stworzyć Hello to Hire. Jest to platforma, która wykorzystała uczenie maszynowe, aby zmniejszyć liczbę podań o pracę - w celu pomyślnego zatrudnienia - średnio do 75. Z drugiej strony średnia w branży wynosi około 150,10. System automatyzuje również tworzenie opisów stanowisk, co uwzględnia nawet niuanse oparte na branży i lokalizacji!

• Doświadczenie klienta: W dzisiejszych czasach klienci chcą spersonalizowanych doświadczeń. Przyzwyczaili się do tego, korzystając z usług takich jak Amazon.com i Uber. Dzięki uczeniu maszynowemu firma może wykorzystać swoje dane, aby uzyskać wgląd — dowiedzieć się, co naprawdę działa. Jest to tak ważne, że skłoniło Krogera do zakupu firmy w przestrzeni o nazwie 84,51°. Zdecydowanie kluczowe jest to, że posiada dane dotyczące ponad 60 milionów gospodarstw domowych w USA. Oto krótki przypadek badania: W większości swoich sklepów Kroger miał duże ilości awokado, a tylko kilka miało 4-paki. Powszechnie uważano, że 4-paki należy przecenić ze względu na różnice w wielkości w stosunku do produktów masowych. Jednak przy zastosowaniu analizy uczenia maszynowego okazało się to niepoprawne, ponieważ 4-paki przyciągnęły nowe i różne gospodarstwa domowe, takie jak Millenialsi i kupujący na ClickList. Dzięki rozszerzeniu 4-paków w całej sieci nastąpił ogólny wzrost sprzedaży awokado.

• Finanse: uczenie maszynowe może wykrywać rozbieżności, na przykład z rozliczeniami. Ale istnieje nowa kategoria technologii, zwana RPA (Robotic Process Automation), która może w tym pomóc. Automatyzuje rutynowe procesy w celu zmniejszenia liczby błędów. ZAP może również wykorzystywać uczenie maszynowe do wykrywania nietypowych lub podejrzanych transakcji.

• Obsługa klienta: W ciągu ostatnich kilku lat nastąpił wzrost liczby chatbotów, które wykorzystują uczenie maszynowe do automatyzacji interakcji z klientami.

• Randki: Uczenie maszynowe może pomóc w znalezieniu bratniej duszy! Tinder, jedna z największych aplikacji randkowych, wykorzystuje tę technologię do ulepszania dopasowań. Na przykład ma system, który automatycznie oznacza ponad 10 miliardów zdjęć przesyłanych codziennie.

Proces uczenia maszynowego

Aby odnieść sukces w zastosowaniu uczenia maszynowego do problemu, ważne jest, aby przyjąć systematyczne podejście. Jeśli nie, wyniki mogą być dalekie od podstaw. Przede wszystkim musisz przejść przez proces przetwarzania danych, który omówiliśmy wcześniej. Po zakończeniu dobrze jest wykonać wizualizację danych. Czy jest w większości rozproszony? A może są jakieś wzory? Jeśli odpowiedź brzmi tak, dane mogą być dobrym kandydatem do uczenia maszynowego. Celem procesu uczenia maszynowego jest stworzenie modelu, który opiera się na jednym lub kilku algorytmach. Rozwijamy to, szkoląc go. Celem jest, aby model powinien zapewniać wysoki stopień przewidywalności. Teraz przyjrzyjmy się temu bliżej (nawiasem mówiąc, będzie to również miało zastosowanie do głębokiego uczenia się.

Krok #1-Zamówienie danych

Jeśli twoje dane są posortowane, może to zniekształcić wyniki. Oznacza to, że algorytm uczenia maszynowego może wykryć to jako wzorzec! Dlatego dobrym pomysłem jest losowanie kolejności danych.

Krok #2-Wybierz model

Musisz wybrać algorytm. Będzie to przypuszczenie oparte na wiedzy, które będzie wymagało procesu prób i błędów. W tym rozdziale przyjrzymy się różnym dostępnym algorytmom.

Krok #3 - Trenuj model

Dane uczące, które będą stanowić około 70% całego zbioru danych, zostaną wykorzystane do stworzenia relacji w algorytmie. Załóżmy na przykład, że budujesz system uczenia maszynowego, aby znaleźć wartość używanego samochodu. Niektóre cechy obejmują rok produkcji, markę, model, przebieg i stan. Przetwarzając te dane treningowe, algorytm obliczy wagi dla każdego z tych czynników. Przykład: Załóżmy, że używamy algorytmu regresji liniowej, który ma następujący format:

y = m * x + b

W fazie uczenia system wygeneruje wartości dla m (co jest nachyleniem na wykresie) i b (co jest punktem przecięcia z osią y).

Krok #4 - Oceń model

Zbierzecie dane testowe, które stanowią pozostałe 30% zbioru danych. Powinien być reprezentatywny dla zakresów i rodzaju informacji w danych treningowych. Dzięki danym testowym możesz sprawdzić, czy algorytm jest dokładny. Czy w naszym przykładzie używanego samochodu wartości rynkowe są zgodne z tym, co dzieje się w prawdziwym świecie?

Uwaga: Z danymi treningowymi i testowymi nie może być żadnego przemieszania. Może to łatwo prowadzić do zniekształconych wyników. Co ciekawe, jest to częsty błąd.

Teraz dokładność jest jedną z miar sukcesu algorytmu. Ale w niektórych przypadkach może to być mylące. Rozważ sytuację z odliczeniem oszustwa. W porównaniu do zbioru danych zazwyczaj występuje niewielka liczba funkcji. Ale pominięcie jednego może być katastrofalne i kosztować firmę miliony dolarów strat. Dlatego możesz chcieć użyć innych podejść, takich jak twierdzenie Bayesa.

Krok #5 - Dostrój model

W tym kroku możemy dostosować wartości parametrów w algorytmie. Ma to na celu sprawdzenie, czy możemy uzyskać lepsze wyniki. Podczas dostrajania modelu mogą również występować hiperparametry. Są to parametry, których nie można nauczyć się bezpośrednio z procesu szkolenia.

Stosowanie algorytmów

Niektóre algorytmy są dość łatwe do obliczenia, podczas gdy inne wymagają skomplikowanych kroków i matematyki. Dobrą wiadomością jest to, że zwykle nie musisz obliczać algorytmu, ponieważ istnieje wiele języków, takich jak Python i R, które upraszczają ten proces. Jeśli chodzi o uczenie maszynowe, algorytm zazwyczaj różni się od tradycyjnego. Powodem jest to, że pierwszym krokiem jest przetwarzanie danych - a następnie komputer zacznie się uczyć. Mimo że dostępne są setki algorytmów uczenia maszynowego, można je w rzeczywistości podzielić na cztery główne kategorie: uczenie nadzorowane, uczenie nienadzorowane, uczenie ze wzmacnianiem i uczenie częściowo nadzorowane. Przyjrzymy się każdemu.

Nadzorowana nauka

Uczenie nadzorowane wykorzystuje dane oznaczone etykietami. Załóżmy na przykład, że mamy zestaw zdjęć tysięcy psów. Dane uważa się za oznaczone, jeśli każde zdjęcie identyfikuje każdą z ras. W większości ułatwia to analizę, ponieważ możemy porównać nasze wyniki z poprawną odpowiedzią. Jednym z kluczy w uczeniu nadzorowanym jest to, że powinny być duże ilości danych. Pomaga to udoskonalić model i uzyskać dokładniejsze wyniki. Ale jest duży problem: w rzeczywistości wiele dostępnych danych nie jest oznaczonych. Ponadto dostarczenie etykiet może być czasochłonne, jeśli istnieje ogromny zbiór danych. Istnieją jednak kreatywne sposoby radzenia sobie z tym, takie jak finansowanie społecznościowe. Tak powstał system ImageNet, który był przełomem w innowacjach AI. Ale stworzenie go zajęło kilka lat. Lub, w niektórych przypadkach, mogą istnieć zautomatyzowane podejścia do etykietowania danych. Weźmy za przykład Facebooka. W 2018 roku firma ogłosiła - na konferencji deweloperów F8 - że wykorzystała swoją ogromną bazę zdjęć z Instagrama, które zostały oznaczone hashtagami. To prawda, to podejście miało swoje wady. Hashtag może zawierać niewizualny opis zdjęcia, powiedzmy #tbt (co oznacza „wspomnienie w czwartek)" lub może być zbyt niejasny, jak #impreza. Dlatego Facebook nazwał swoje podejście „słabo nadzorowanymi danymi".

Ale utalentowani inżynierowie w firmie znaleźli sposoby na poprawę jakości, na przykład przez zbudowanie wyrafinowanego modelu przewidywania hashtagów. W sumie wszystko ułożyło się całkiem dobrze. Model uczenia maszynowego Facebooka, który obejmował 3,5 miliarda zdjęć, miał wskaźnik dokładności 85,4%, co oparto na benchmarku rozpoznawania ImageNet. W rzeczywistości była to najwyższa odnotowana w historii, bo aż o 2%. Ten projekt AI wymagał również innowacyjnych podejść do budowy infrastruktury. Zgodnie z wpisem na blogu na Facebooku: Ponieważ jedna maszyna zajęłaby ponad rok, aby ukończyć szkolenie modelu, stworzyliśmy sposób na rozłożenie zadania na maksymalnie 336 procesorów graficznych, skracając całkowity czas szkolenia do zaledwie kilku tygodni. Przy coraz większych rozmiarach modelu - największy w tym badaniu jest ResNeXt 101-32x48d z ponad 861 milionami parametrów – takie rozproszone szkolenie jest coraz bardziej istotne. Ponadto opracowaliśmy metodę usuwania duplikatów, aby upewnić się, że przypadkowo nie przeszkolimy naszych modeli na obrazach, na których chcemy je ocenić, co jest problemem, który nęka podobne badania w tej dziedzinie.

Wybiegając w przyszłość, Facebook dostrzega potencjał w wykorzystaniu swojego podejścia do różnych obszarów, w tym:

• Ulepszony ranking w kanale informacyjnym

• Lepsze wykrywanie kontrowersyjnych treści

• Automatyczne generowanie napisów dla osób niedowidzących

Nauka nienadzorowana

Uczenie nienadzorowane ma miejsce wtedy, gdy pracujesz z danymi nieoznaczonymi. Oznacza to, że do wykrywania wzorców użyjesz algorytmów głębokiego uczenia. Zdecydowanie najczęstszym podejściem do uczenia nienadzorowanego jest grupowanie, które polega na pobieraniu danych nieoznaczonych i używa algorytmów do grupowania podobnych elementów. Proces zwykle rozpoczyna się od zgadywania, a następnie wykonuje się iteracje obliczeń, aby uzyskać lepsze wyniki. Sednem tego jest znajdowanie elementów danych, które są blisko siebie, co można osiągnąć za pomocą różnych metod ilościowych:

• Metryka Euklidesa: Jest to linia prosta pomiędzy dwoma punktami danych. Metryka euklidesowa jest dość powszechna w przypadku uczenia maszynowego.

• Metryka podobieństwa cosinusa: Jak sama nazwa wskazuje, do pomiaru kąta użyjesz cosinusa. Chodzi o to, aby znaleźć podobieństwa między dwoma punktami danych pod względem orientacji.

• Manhattan Metric: Obejmuje sumę bezwzględnych odległości dwóch punktów na współrzędnych wykresu. Nazywa się „Manhattanem", ponieważ nawiązuje do układu ulic miasta, który pozwala na krótsze odległości podróży.

Jeśli chodzi o przypadki użycia klastrowania, jednym z najczęstszych jest segmentacja klientów, która ma pomóc w lepszym ukierunkowaniu komunikatów marketingowych. W większości grupa, która ma podobne cechy, prawdopodobnie podziela zainteresowania i preferencje. Inną aplikacją jest analiza sentymentu, w której wydobywasz społecznościowe dane medialne i znajdź trendy. Dla firmy modowej może to mieć kluczowe znaczenie dla zrozumienia, jak dostosować styl nadchodzącej linii ubrań. Obecnie istnieją inne podejścia niż tylko grupowanie. Oto spojrzenie na trzy kolejne:

• Asocjacja: Podstawowa koncepcja jest taka, że jeśli zdarzy się X, to prawdopodobnie wydarzy się Y. Tak więc, jeśli kupisz moją książkę o AI, prawdopodobnie będziesz chciał kupić inne tytuły z tego gatunku. Dzięki skojarzeniu algorytm głębokiego uczenia może rozszyfrować tego rodzaju relacje. Może to skutkować potężnymi silnikami rekomendacji

• Wykrywanie anomalii: Identyfikuje wartości odstające lub anomalne wzorce w zestawie danych, co może być pomocne w aplikacjach cyberbezpieczeństwa. Według Asafa Cidona, który jest wiceprezesem ds. bezpieczeństwa poczty e-mail w Barracuda Networks: „Odkryliśmy, że łącząc wiele różnych sygnałów, takich jak treść wiadomości e-mail, nagłówek, wykres komunikacji społecznościowej, loginy IP, reguły przekazywania skrzynki odbiorczej itp. .- jesteśmy w stanie osiągnąć niezwykle wysoką precyzję w wykrywaniu ataków socjotechnicznych, nawet jeśli ataki są wysoce spersonalizowane i stworzone z myślą o konkretnej osobie w określonej organizacji. Uczenie maszynowe umożliwia nam wykrywanie ataków pochodzących z wewnątrz organizacji, których źródłem jest legalna skrzynka pocztowa pracownika, co byłoby niemożliwe w przypadku statycznego, uniwersalnego silnika reguł".

• Autokodery: Dzięki temu dane zostaną umieszczone w skompresowanej formie, a następnie zostaną zrekonstruowane. Z tego mogą wyłonić się nowe wzorce. Jednak użycie autokoderów jest rzadkie. Można jednak wykazać, że może być przydatny w aplikacjach, takich jak redukcja szumu w danych.

Weź pod uwagę, że wielu badaczy sztucznej inteligencji uważa, że nienadzorowane uczenie się będzie prawdopodobnie miało kluczowe znaczenie dla następnego poziomu osiągnięć. Według artykułu w Nature autorstwa Yanna LeCuna, Geoffreya Hintona i Yoshuy Bengio: „Oczekujemy, że nauka bez nadzoru stanie się znacznie ważniejsza w dłuższej perspektywie. Uczenie się ludzi i zwierząt jest w dużej mierze nienadzorowane: odkrywamy strukturę świata, obserwując go, a nie wymawiając nazwę każdego przedmiotu

Nauka przez wzmacnianie

Kiedy byłeś dzieckiem i chciałeś uprawiać nowy sport, prawdopodobnie nie czytałeś instrukcji. Zamiast tego obserwowałeś, co robią inni ludzie i próbowałeś to rozgryźć. W niektórych sytuacjach popełniłeś błędy i straciłeś piłkę, gdy twoi koledzy z drużyny okazywali niezadowolenie. Ale w innych przypadkach wykonałeś właściwe ruchy i zdobyłeś punkty. Dzięki temu procesowi opartemu na próbach i błędach Twoja nauka została ulepszona w oparciu o pozytywne i negatywne wzmocnienie. Na wysokim poziomie jest to analogiczne do uczenia się przez wzmacnianie. Było to kluczowe dla niektórych z najbardziej znaczących osiągnięć w sztucznej inteligencji, takich jak:

• Gry: są idealne do uczenia się przez wzmacnianie, ponieważ istnieją jasne zasady, wyniki i różne ograniczenia (takie jak plansza do gry). Budując model, możesz go przetestować za pomocą milionów symulacji, co oznacza, że system szybko stanie się coraz mądrzejszy. W ten sposób program może nauczyć się pokonać mistrza świata Go lub szachów.

• Robotyka: Kluczem jest umiejętność poruszania się w przestrzeni – a to wymaga oceny środowiska w wielu różnych punktach. Jeśli robot chce się przenieść, powiedzmy, do kuchni, będzie musiał omijać meble i inne przeszkody. Jeśli wpadnie na różne rzeczy, nastąpi negatywna akcja wzmacniająca.

Nauka częściowo nadzorowana

Jest to połączenie nauki nadzorowanej i nienadzorowanej. Dzieje się tak, gdy masz niewielką ilość nieoznakowanych danych. Można jednak użyć systemów uczenia głębokiego, aby przetłumaczyć nienadzorowane dane na dane nadzorowane - proces ten nazywa się pseudo-etykietowaniem. Następnie możesz zastosować algorytmy. Ciekawym przypadkiem użycia częściowo nadzorowanego uczenia się jest interpretacja MRI. Radiolog może najpierw oznaczyć skany, a następnie system głębokiego uczenia może znaleźć pozostałe wzorce.

Typowe typy algorytmów uczenia maszynowego

Nie mamy wystarczająco dużo miejsca, aby omówić wszystkie algorytmy uczenia maszynowego! Zamiast tego lepiej skupić się na najczęstszych. W dalszej części przyjrzymy się tym w następujących przypadkach:

• Uczenie nadzorowane: Algorytmy można sprowadzić do dwóch wariantów. Jednym z nich jest klasyfikacja, która dzieli zbiór danych na wspólne etykiety. Przykłady algorytmów obejmują naiwny klasyfikator Bayesa i k-najbliższy sąsiad. Następnie następuje regresja, która znajduje w danych ciągłe wzorce. W tym celu przyjrzymy się regresji liniowej, modelowaniu zespołowemu i drzewom decyzyjnym.

• Nienadzorowane uczenie się: W tej kategorii przyjrzymy się grupowaniu. W tym celu omówimy klastrowanie k-średnich.

Naiwny klasyfikator Bayesa (nadzorowane uczenie się/klasyfikacja)

Wcześniej przyjrzeliśmy się twierdzeniu Bayesa. Jeśli chodzi o uczenie maszynowe, zostało to zmodyfikowane w coś, co nazywa się Naïve Bayes Classifier. Jest to „naiwne", ponieważ zakłada się, że zmienne są od siebie niezależne, to znaczy występowanie jednej zmiennej nie ma nic wspólnego z innymi. To prawda, może się to wydawać wadą. Ale faktem jest, że klasyfikator Naïve Bayes okazał się dość skuteczny i szybki w rozwoju. Należy również zwrócić uwagę na inne założenie: założenie a priori. To mówi, że prognozy będą błędne, jeśli dane się zmienią. Istnieją trzy odmiany klasyfikatora Naïve Bayes:

• Bernoulli: Dzieje się tak, jeśli masz dane binarne (prawda/fałsz, tak/nie).

• Wielomianowy: Dzieje się tak, gdy dane są dyskretne, takie jak liczba stron książki.

• Gaussian: Dzieje się tak, jeśli pracujesz z danymi, które są zgodne z rozkładem normalnym.

Typowym przypadkiem użycia klasyfikatorów naiwnych Bayesa jest analiza tekstu. Przykłady obejmują wykrywanie spamu w wiadomościach e-mail, segmentację klientów, analizę sentymentu, diagnostykę medyczną i prognozy pogody. Powodem jest to, że takie podejście jest przydatne w klasyfikowaniu danych na podstawie kluczowych cech i wzorców. Aby zobaczyć, jak to się robi, weźmy przykład: załóżmy, że prowadzisz witrynę e-commerce i masz dużą bazę danych transakcji klientów. Chcesz

zobaczyć, jak zmienne, takie jak oceny produktów, rabaty i pora roku, wpływają na sprzedaż. Tabela jak wygląda zbiór danych.

Rabat : Recenzja produktu : Zakup

Tak : Wysoki : Tak

Tak : Niski : Tak

Nie : Niski : Nie

Nie : Niski : Nie

Nie : Niski : Nie

Nie : Wysoki : Tak

Tak : Wysoki : Nie

Tak : Niski : Tak

Nie : Wysoki : Tak

Tak : Wysoki : Tak

Nie : Wysoki : Nie

Nie : Niski : Tak

Tak : Wysoki : Tak

Tak : Niski : Nie

Następnie zorganizujesz te dane w tabelach częstości, jak pokazano w tabelach 3 i 4.

		Purchase	
		Yes	No
Discount	Yes	19	1
	Yes	5	5

		Purchase		
		Yes	No	Total
Product Review	**High**	21	2	11
	Low	3	4	8
	Total	24	6	19

Patrząc na to, zakup nazywamy wydarzeniem, a rabaty i recenzje produktów jako zmienne niezależne. Następnie możemy stworzyć tabelę prawdopodobieństwa dla jednej ze zmiennych niezależnych, powiedzmy, recenzje produktów.

		Purchase		
		Yes	No	
Product Reviews	High	9/24	2/6	11/30
	Low	7/24	1/6	8/30
		24/30	6/30	

Korzystając z tego wykresu, widzimy, że prawdopodobieństwo zakupu przy niskiej recenzji produktu wynosi 7/24 lub 29%. Innymi słowy, naiwny klasyfikator Bayesa umożliwia bardziej szczegółowe prognozy w zbiorze danych. Jest również stosunkowo łatwy do nauczenia i może dobrze współpracować z małymi zestawami danych.

K-Nearest Neighbor (nadzorowane uczenie się/klasyfikacja)

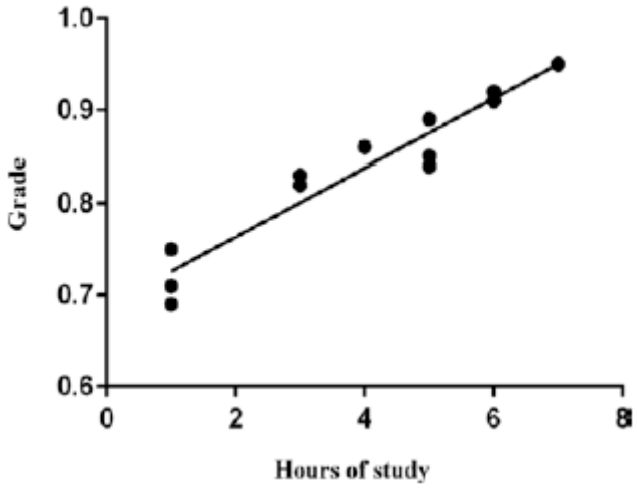
K-Nearest Neighbor (k-NN) to metoda klasyfikacji zbioru danych (k oznacza liczbę sąsiadów). Teoria mówi, że te wartości, które są blisko siebie, prawdopodobnie będą dobrymi predyktorami dla modelu. Pomyśl o tym jako „Ptaki ze stada piór razem". Przypadkiem użycia k-NN jest ocena kredytowa, która opiera się na różnych czynnikach, takich jak dochód, historia płatności, lokalizacja, własność domu i tak dalej. Algorytm podzieli zbiór danych na różne segmenty klientów. Następnie, gdy do bazy zostanie dodany nowy klient, zobaczysz, do jakiego klastra się zalicza – i będzie to ocena kredytowa. K-NN jest w rzeczywistości prosty do obliczenia. W rzeczywistości nazywa się to leniwym uczeniem się, ponieważ nie ma procesu uczenia się z danymi. Aby użyć k-NN, musisz wymyślić odległość między najbliższymi wartościami. Jeśli wartości są liczbowe, można je oprzeć na odległości euklidesowej, co wymaga skomplikowanej matematyki. Lub, jeśli istnieją dane kategoryczne, możesz użyć metryki nakładania się (w tym przypadku dane są takie same lub bardzo podobne). Następnie musisz określić liczbę sąsiadów. Chociaż posiadanie większej ilości wygładzi model, może również oznaczać zapotrzebowanie na ogromną ilość zasobów obliczeniowych. Aby temu zaradzić, możesz przypisać wyższe wagi do danych, które są bliżej swoich sąsiadów.

Regresja liniowa (nadzorowane uczenie/regresja)

Regresja liniowa pokazuje związek między pewnymi zmiennymi. Równanie — zakładając, że istnieje wystarczająca ilość danych wysokiej jakości — może pomóc w przewidywaniu wyników na podstawie danych wejściowych. Przykład: Załóżmy, że mamy dane dotyczące liczby godzin spędzonych na nauce do egzaminu i oceny.

Hours of Study	Grade Percentage
1	0.75
1	0.69
1	0.71
3	0.82
3	0.83
4	0.86
5	0.85
5	0.89
5	0.84
6	0.91
6	0.92
7	0.95

Jak widać, ogólna zależność jest pozytywna (opisuje to tendencję, w której wyższa ocena jest skorelowana z większą liczbą godzin nauki). Za pomocą algorytmu regresji możemy wykreślić linię, która ma najlepsze dopasowanie (odbywa się to za pomocą obliczenia zwanego „najmniejszymi kwadratami", które minimalizuje błędy).



Z tego otrzymujemy następujące równanie:

Ocena = liczba godzin nauki × 0,03731 + 0,6889

Następnie załóżmy, że uczysz się 4 godziny do egzaminu. Jaka będzie Twoja szacunkowa ocena? Równanie mówi nam, jak:

0,838 = 4 × 0,03731 + 0,6889

Jak dokładne jest to? Aby pomóc odpowiedzieć na to pytanie, możemy użyć obliczenia o nazwie R-kwadrat. W naszym przypadku jest to 0,9180 (zakres od 0 do 1). Im wartość jest bliższa 1, tym lepsze dopasowanie. Tak więc 0,9180 jest dość wysokie. Oznacza to, że godziny nauki wyjaśniają 91,8% oceny z egzaminu. Teraz to prawda, że ten model jest uproszczony. Aby lepiej odzwierciedlić rzeczywistość,

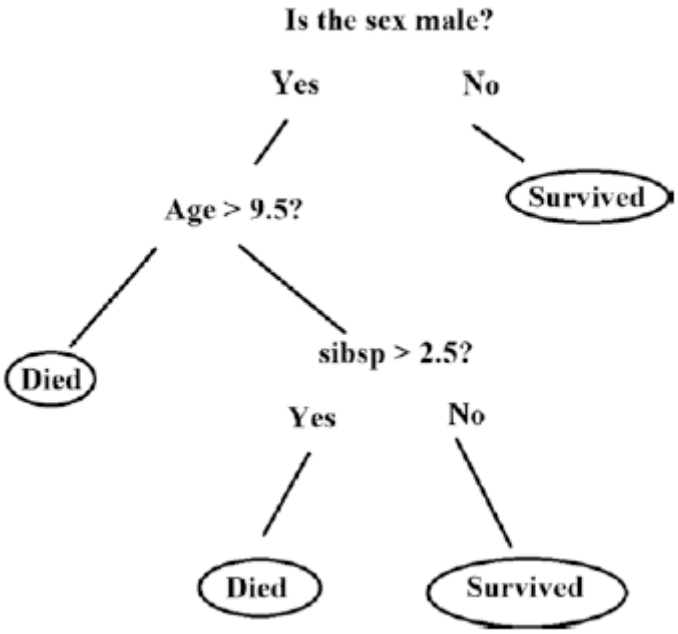
możesz dodać więcej zmiennych, aby wyjaśnić ocenę na egzaminie - powiedz obecność ucznia. Robiąc to, użyjesz czegoś, co nazywa się regresją wielowymiarową.

* Uwaga Jeśli współczynnik dla zmiennej jest dość mały, dobrym pomysłem może być nieuwzględnianie go w modelu.

Czasami dane mogą również nie być w linii prostej, w którym to przypadku algorytm regresji nie zadziała. Ale możesz użyć bardziej złożonej wersji, zwanej regresją wielomianową.

Drzewo decyzyjne (nadzorowane uczenie się/regresja)

Bez wątpienia klastrowanie może nie działać w przypadku niektórych zestawów danych. Ale dobrą wiadomością jest to, że istnieją alternatywy, takie jak drzewo decyzyjne. To podejście ogólnie działa lepiej w przypadku danych nienumerycznych. Początkiem drzewa decyzyjnego jest węzeł główny, który znajduje się na górze schematu blokowego. Od tego momentu powstanie drzewo ścieżek decyzyjnych, które nazywane są podziałami. W tych punktach użyjesz algorytmu do podjęcia decyzji i zostanie obliczone prawdopodobieństwo. Na końcu drzewa będzie liść (lub wynik). Znanym przykładem w kręgach uczenia maszynowego jest wykorzystanie drzewa decyzyjnego do tragicznego zatonięcia Titanica. Model przewiduje przeżycie pasażera na podstawie trzech cech: płci, wieku i liczby współmałżonków lub dzieci (sibsp). Oto jak to wygląda na rysunku.



Drzewa decyzyjne mają wyraźne zalety. Są łatwe do zrozumienia, dobrze współpracują z dużymi zestawami danych i zapewniają przejrzystość modelu. Jednak drzewa decyzyjne mają również wady. Jednym z nich jest propagacja błędów. Jeśli jeden z podziałów okaże się błędny, błąd ten może kaskadowo rozprzestrzenić się na resztę modelu! Następnie, w miarę wzrostu drzew decyzyjnych, będzie wzrastać złożoność, ponieważ będzie duża liczba algorytmów. Może to ostatecznie skutkować niższą wydajnością modelu.

Modelowanie zespołowe (nadzorowane uczenie się/regresja)

Modelowanie zespołowe oznacza używanie więcej niż jednego modelu do prognoz. Mimo że zwiększa to złożoność, wykazano, że takie podejście generuje dobre wyniki. Aby zobaczyć to w akcji, spójrz na „Nagrodę Netflix”, która rozpoczęła się w 2006 roku. Firma ogłosiła, że zapłaci milion dolarów każdemu

lub dowolnemu zespołowi, który może poprawić dokładność swojego systemu rekomendacji filmów o 10% lub więcej. Netflix dostarczył również zbiór danych zawierający ponad 100 milionów ocen 17 770 filmów od 480 189 użytkowników16. Ostatecznie liczba pobrań przekroczyłaby 30 000. Dlaczego Netflix to wszystko zrobił? Dużym powodem jest to, że inżynierowie firmy mieli problemy z postępem. Dlaczego więc nie dać tego tłumowi, aby się zorientował? Okazało się to dość pomysłowe — a wypłata 1 miliona dolarów była naprawdę skromna w porównaniu z potencjalnymi korzyściami. Konkurs z pewnością wywołał dużą aktywność ze strony programistów i naukowców zajmujących się danymi, od studentów po pracowników firm takich jak AT&T. Netflix również uprościł konkurs. Głównym wymogiem było ujawnienie przez zespoły swoich metod, co pomogło zwiększyć wyniki (był nawet dashboard z rankingami zespołów). Jednak dopiero w 2009 roku nagrodę zdobył zespół – Pragmatic Chaos firmy BellKor. Z drugiej strony pojawiły się spore wyzwania. Jak więc udało się to zwycięskiej drużynie? Pierwszym krokiem było stworzenie modelu bazowego, który wygładził trudne problemy z danymi. Na przykład niektóre filmy miały tylko kilka ocen, podczas gdy inne miały tysiące. Następnie pojawił się drażliwy problem polegający na tym, że byli użytkownicy, którzy zawsze oceniali film jedną gwiazdką. Aby poradzić sobie z tymi sprawami, BellKor wykorzystał uczenie maszynowe do przewidywania ocen w celu wypełnienia luk. Po zakończeniu planu bazowego pojawiły się trudniejsze wyzwania, takie jak:

• System może zakończyć się polecaniem tych samych filmów wielu użytkownikom.

• Niektóre filmy mogą nie pasować do gatunków. Na przykład Alien to tak naprawdę skrzyżowanie science fiction i horroru.

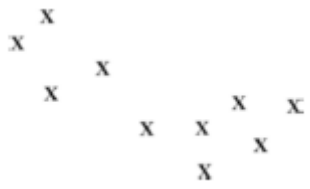
• Były filmy, takie jak Napoleon Dynamite, które okazały się niezwykle trudne do zrozumienia dla algorytmów.

• Oceny filmu często zmieniały się z czasem.

Zwycięski zespół wykorzystał modelowanie zespołowe, które obejmowało setki algorytmów. Wykorzystali również coś, co nazywa się boostingiem, czyli gdzie budujesz kolejne modele. Dzięki temu wagi w algorytmach są dostosowywane w oparciu o wyniki poprzedniego modelu, co pomaga z czasem poprawiać przewidywania (inne podejście, zwane baggingiem, polega na równoległym budowaniu różnych modeli, a następnie wybraniu najlepszego). Ale w końcu BellKor znalazł rozwiązanie. Jednak mimo to Netflix nie wykorzystał tego modelu! Teraz nie jest jasne, dlaczego tak było. Być może chodziło o to, że Netflix i tak odchodził od pięciogwiazdkowych ocen i był bardziej skoncentrowany na przesyłaniu strumieniowym. Konkurs spotkał się również z reakcją osób, które sądziły, że mogło dojść do naruszenia prywatności. Niezależnie od tego konkurs podkreślał siłę uczenia maszynowego i znaczenie współpracy.

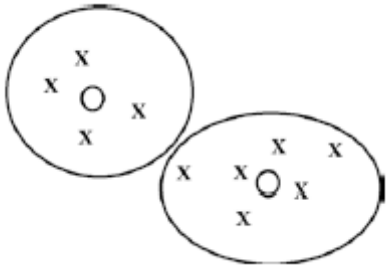
Klastrowanie metodą K-Means (nienadzorowane/klastrowanie)

Algorytm grupowania k-średnich, który jest skuteczny w przypadku dużych zestawów danych, umieszcza podobne, nieoznakowane dane w różnych grupach. Pierwszym krokiem jest wybranie k, czyli liczby klastrów. Aby w tym pomóc, możesz wykonać wizualizacje tych danych, aby sprawdzić, czy istnieją zauważalne obszary grupowania. Oto spojrzenie na przykładowe dane na rysunku:

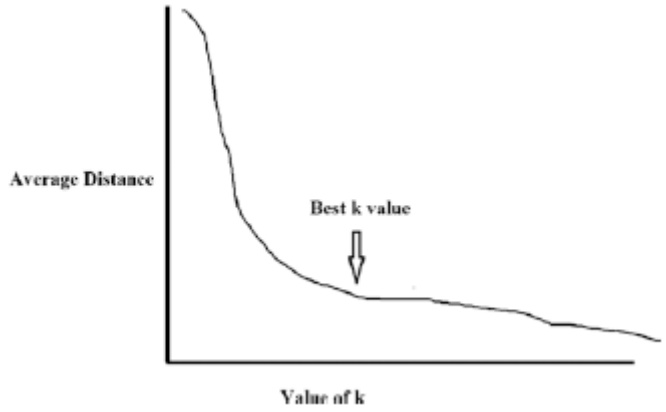


W tym przykładzie zakładamy, że będą dwa klastry, a to oznacza, że będą też dwa centroidy. Centroid to środek klastra. Każdy przypiszemy losowo, co widać na rysunku.

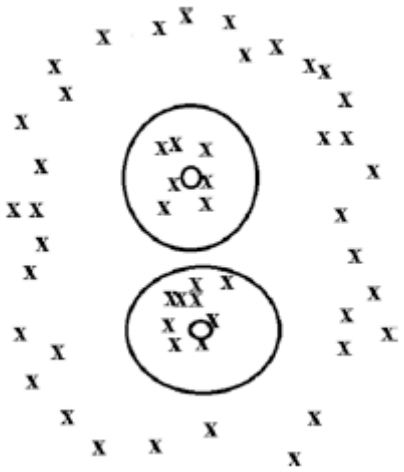
Jak widać, centroid w lewym górnym rogu wygląda daleko, ale ten po prawej stronie jest lepszy. Algorytm k-średnich obliczy następnie średnie odległości centroidów, a następnie zmieni ich położenie. Będzie to powtarzane, aż błędy będą dość minimalne - punkt zwany zbieżnością, który można zobaczyć na rysunku.



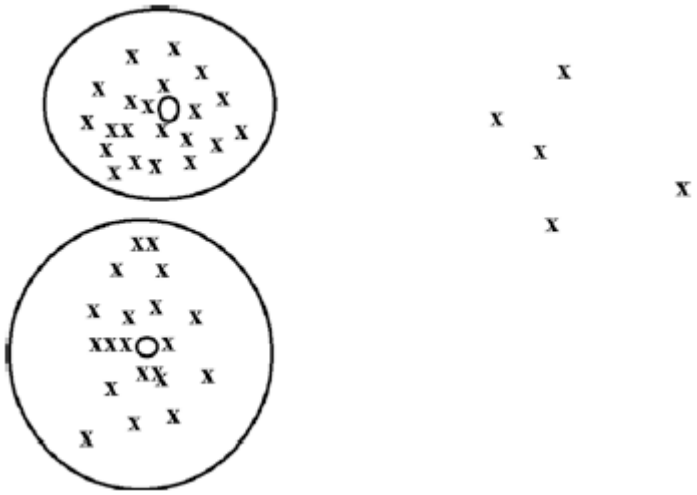
To prawda, to prosta ilustracja. Ale oczywiście przy złożonym zbiorze danych trudno będzie określić liczbę początkowych klastrów. W tej sytuacji możesz poeksperymentować z różnymi wartościami k, a następnie zmierzyć średnie odległości. Robiąc to wiele razy, powinno być więcej dokładności. Dlaczego więc nie mieć po prostu wysokiej liczby dla k? Z pewnością możesz to zrobić. Ale kiedy obliczysz średnią, zauważysz, że będą tylko przyrostowe ulepszenia. Tak więc jedną metodą jest zatrzymanie się w punkcie, w którym to się zaczyna. Widać to na rysunku.



Jednak k-Means ma swoje wady. Na przykład nie działa dobrze z danymi niesferycznymi, co ma miejsce na rysunku.



Dzięki temu algorytm k-średnich prawdopodobnie nie wykryłby otaczających danych, mimo że ma wzór. Ale istnieją pewne algorytmy, które mogą pomóc, takie jak DBScan (przestrzenne klastrowanie aplikacji z szumem oparte na gęstości), które ma obsługiwać mieszankę zestawów danych o bardzo różnych rozmiarach. Chociaż DBScan może wymagać dużej mocy obliczeniowej. Następnie mamy do czynienia z sytuacją, w której istnieją klastry z dużą ilością danych, a inne z niewielką ilością. Co może się stać? Istnieje szansa, że algorytm k-średnich nie wykryje tego lekkiego. Tak jest w przypadku rysunku.



Wniosek

Algorytmy te mogą się skomplikować i wymagają dużych umiejętności technicznych. Ale ważne jest, aby nie ugrzęznąć zbytnio w technologii. W końcu skupiamy się na znalezieniu sposobów wykorzystania uczenia maszynowego do osiągnięcia jasnych celów. Ponownie, Stich Fix to dobre miejsce, aby uzyskać wskazówki na ten temat. W listopadowym numerze Harvard Business Review, dyrektor ds. algorytmów firmy, Eric Colson, opublikował artykuł „Curiosity-Driven Data Science”17. Przedstawił w nim swoje doświadczenia w tworzeniu organizacji opartej na danych. Sednem tego jest umożliwienie analitykom danych odkrywania nowych pomysłów, koncepcji i podejść. Doprowadziło to do wdrożenia sztucznej inteligencji w podstawowych funkcjach firmy, takich jak zarządzanie zapasami, zarządzanie relacjami, logistyka i kupowanie towarów. Przekształciło to, czyniąc organizację bardziej zwinną i usprawnioną. Colson uważa również, że stanowi „barierę ochronną przed konkurencją”. Jego artykuł zawiera również inne przydatne porady dotyczące analizy danych:

• Naukowcy zajmujący się danymi: Nie powinni być częścią innego działu. Raczej powinni mieć swój własny, który podlega bezpośrednio prezesowi. Pomaga to skoncentrować się na kluczowych priorytetach, a także mieć holistyczne spojrzenie na potrzeby organizacji.

• Eksperymenty: Kiedy analityk danych ma nowy pomysł, powinien zostać przetestowany na małej próbce klientów. Jeśli jest przyczepność, można ją rozwinąć na resztę podstawy.

• Zasoby: Analitycy danych potrzebują pełnego dostępu do danych i narzędzi. Powinny też odbywać się ciągłe szkolenia.

• Generaliści: Zatrudnij naukowców zajmujących się danymi, którzy zajmują się różnymi dziedzinami, takimi jak modelowanie, uczenie maszynowe i analityka (Colson określa tych ludzi jako „pełnostackowych analityków danych"). Prowadzi to do powstania małych zespołów, które często są bardziej wydajne i produktywne.

• Kultura: Colson poszukuje wartości, takich jak „uczenie się przez działanie, bycie komfortowym w niejednoznaczności, równoważenie długo- i krótkoterminowych zysków".

Kluczowe dania na wynos

• Uczenie maszynowe, którego korzenie sięgają lat pięćdziesiątych, to miejsce, w którym komputer może uczyć się bez wyraźnego programowania. Raczej pozyskuje i przetwarza dane przy użyciu zaawansowanych technik statystycznych.

• Wartość odstająca to dane, które znacznie wykraczają poza pozostałe liczby w zbiorze danych.

• Odchylenie standardowe mierzy średnią odległość od średniej.

• Rozkład normalny, który ma kształt dzwonu, reprezentuje sumę prawdopodobieństw dla zmiennej.

• Twierdzenie Bayesa to zaawansowana technika statystyczna, która zapewnia głębsze spojrzenie na prawdopodobieństwa.

• Prawdziwie pozytywne jest to, gdy model dokonuje prawidłowej prognozy. Z drugiej strony, fałszywie dodatni wynik ma miejsce, gdy prognoza modelu pokazuje, że wynik jest prawdziwy, mimo że tak nie jest.

• Korelacja Pearsona pokazuje siłę związku między dwiema zmiennymi w zakresie od 1 do -1.

• Ekstrakcja cech lub inżynieria cech opisuje proces wyboru zmiennych do modelu. Jest to krytyczne, ponieważ nawet jedna niewłaściwa zmienna może mieć duży wpływ na wyniki.

• Dane uczące są używane do tworzenia relacji w algorytmie. Z drugiej strony dane testowe służą do oceny modelu.

• Uczenie nadzorowane wykorzystuje oznakowane dane do stworzenia modelu, podczas gdy uczenie nienadzorowane nie. Istnieje również nauka częściowo nadzorowana, która wykorzystuje mieszankę obu podejść.

• Uczenie się ze wzmacnianiem to sposób na trenowanie modelu poprzez nagradzanie trafnych prognoz i karanie tych, które nie są.

• K-Nearest Neighbor (k-NN) to algorytm oparty na założeniu, że wartości, które są blisko siebie, są dobrymi predyktorami modelu.

• Regresja liniowa szacuje związek między pewnymi zmiennymi. R-kwadrat wskaże siłę związku.

- Drzewo decyzyjne to model oparty na przepływie pracy decyzji tak/nie.
- Model zespołowy wykorzystuje do prognoz więcej niż jeden model.
- Algorytm grupowania k-średnich umieszcza podobne nieoznakowane dane w różnych grupach.

Regresja liniowa (nadzorowane uczenie/regresja)

Regresja liniowa pokazuje związek między pewnymi zmiennymi. Równanie — zakładając, że istnieje wystarczająca ilość danych wysokiej jakości — może pomóc w przewidywaniu wyników na podstawie danych wejściowych. Przykład: Załóżmy, że mamy dane dotyczące liczby godzin spędzonych na nauce do egzaminu i oceny. Zobacz Tabela 3-6.