Algorytm prognostyczny Naive Bayes

Algorytm Naive Bayes nie wymaga dużych zasobów obliczeniowych. Jest to wielka zaleta tego narzędzia, kiedy zmuszani jesteśmy do przeanalizowania ogromnej ilości danych w stosunkowo krótkim czasie. Algorytm ten jest bardzo szybki przy zachowaniu wysokiej jakości prognoz. Niskie wymagania wynikają z tego, że algorytm ten bazuje głównie na średnich. Opiera się na twierdzeniu Bayesa opisującym prawdopodobieństwo występowania niezależnych klas.

Dlaczego założenia Bayesa nazwano naiwnymi?

Angielski matematyk Thomas Bayes wynalazł metodę określania prawdopodobieństwa subiektywnego hipotez w oparciu o dotychczasowe prawdopodobieństwo oraz nowe dane. Technika ta zakłada bezwzględną niezależność predyktorów. Innymi słowy mówiąc, zmienne opisujące muszą być od siebie całkowicie niezależne. Zmienne predykcyjne nazywane są wprawdzie zmiennymi niezależnymi, nikt jednak aż tak restrykcyjnie nie zakłada ich całkowitej, wzajemnej niezależność. Takie założenie zostało ogólnie uzna za wysoce naiwne. Dzięki temu sam algorytm zyskał ten nieco pejoratywny przydomek.

Amerykanie mówią – jak coś chodzi jak kaczka, kwacze jak kaczka i ma dziób jak kaczka to musi być to kaczka. Według Bayesa każda pojedyncza cecha kaczki wystarczy, aby uznać ptaka za kaczkę. Zakładał, że wszystkie te cechy niezależnie przyczyniają się do prawdopodobieństwa, że ptak jest lub nie jest kaczką.

Podstawowe założenia modelu Naive Bayes

Model Naive Bayes wbrew niechlubnemu przydomkowi „naiwnego" potrafi znacznie przewyższać dokładnością analiz niejeden bardziej wyrafinowany model prognostyczny. Jest przy tym niezwykle oszczędny, co czyni go bardzo przydatnym w sytuacjach, gdy niezbędne jest ciągłe uzyskiwanie potoku informacji na bieżąco. Wzór prawdopodobieństwa Thomasa Bayesa można zapisać jako:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \quad (1)$$

$$P(c|x) = P(x_1|c) \times P(x_2|c) \times \ldots \times P(x_n|c) \times P(c) \quad (2)$$

P (c|x) – prawdopodobieństwo klasy na podstawie danego predyktora (x)
P (c) – prawdopodobieństwo wystąpienia klasy
P (x|c) – prawdopodobieństwo wystąpienia predyktora danej klasy
P (x) – prawdopodobieństwo wystąpienia predyktora

Aby zrozumieć ten wzór użyjemy prostego przykładu.

Naive Bayes w przewidywaniu zmiany cen pszenicy

Wyobraźmy sobie analityka giełdowego, który na podstawie doświadczenia i obserwacji doszedł do wniosku, że ceny pszenicy na giełdzie wzrastają gdy jest ładna pogoda i nie wzrastają gdy niebo jest zachmurzone. Aby zbadać postawioną przez siebie hipotezę, postanowił on zapisywać swoje obserwacje. W tabeli zapisał zbiór 21 obserwacji.

Należy zauważyć, że obserwacje mają postać dyskretną. Pogoda została opisana w postaci trzech stanów, natomiast wzrost ceny w postaci dwóch stanów (tak/nie). Pogoda jest predyktorem, zaś wzrost ceny jest wynikiem. W tej sytuacji nie możemy określić zależności predyktora i wyniku na podstawie zwykłej korelacji. Dla zmiennych dyskretnych należy posłużyć się wskaźnikiem zależności ch-kwadrat lub wskaźnikiem pojemności informacyjnej. Oba wymienione tu wskaźniki służą do pomiaru wzajemnej zależności zjawisk kategorycznych.

Tabela 1. Zależność wzrostu cen pszenicy od pogody

Pogoda	Wzrost ceny
Brak zachmurzenia	tak
Wysokie zachmurzenie	nie
Umiarkowane zachmurzenie	nie
Wysokie zachmurzenie	nie
Wysokie zachmurzenie	tak
Brak zachmurzenia	tak
Brak zachmurzenia	tak
Brak zachmurzenia	nie
Umiarkowane zachmurzenie	tak
Wysokie zachmurzenie	nie
Wysokie zachmurzenie	tak
Wysokie zachmurzenie	tak
Wysokie zachmurzenie	nie
Umiarkowane zachmurzenie	nie
Brak zachmurzenia	tak
Umiarkowane zachmurzenie	tak
Umiarkowane zachmurzenie	tak
Umiarkowane zachmurzenie	nie
Brak zachmurzenia	tak
Brak zachmurzenia	tak
Wysokie zachmurzenie	nie

Aby zastosować algorytm Bayesa należy przeformatować nieco dane do postaci przedstawionej w tabeli poniżej.

Analityk zauważył, że gdy dzień był pochmurny 3 razy ceny pszenicy rosły,

Tabela 2. Prawdopodobieństwo zmiany cen w zależności od pogody

	tak	nie			
Brak zachmurzenia	6	1	7	0,33	7/21
Umiarkowane zachmurzenie	3	3	6	0,29	6/21
Wysokie zachmurzenie	3	5	8	0,38	8/21
	12	9			
	0,57	0,43			
	12/21	9/21			

zaś 5 razy ceny nie rosły. Analityk wykonał 21 obserwacji wzrostu cen, z czego 8 było dokonanych w dni pochmurne. Tak więc prawdopodobieństwo, że cena pszenicy wzrośnie w dzień pochmurny wynosi 3/8 = 37,5%.

Porównajmy teraz liczbę wzrostów ceny pszenicy w dni słoneczne. Na 7 obserwacji dokonanych w dni słoneczne zaobserwowano 6 wzrostów cen. Prawdopodobieństwo wzrostu ceny pszenicy w dzień słoneczny był więc wysoki i wynosił 6/7 = 85,7%.

Odpowiednie ustawieniu obserwacji w tab. 2 pozwala określić jakie było prawdopodobieństwo poszczególnych stanów pogody oraz jaki jest prawdopodobieństwo, że cena pszenicy będzie wzrastała.

Jest słonecznie, czy cena pszenicy wzrośnie?

Czy postawiona w ten sposób hipoteza jest prawdziwa? Gdyby okazało się, że taka zależność jest prawdziwa nasz analityk giełdowy mógłby na tym nieźle zarobić.

Możemy sprawdzić czy ta hipoteza jest prawdziwa za pomocą omówionej powyżej metody prawdopodobieństwa.

P(tak | słonecznie) = P(słonecznie | tak) x P(tak) : P (słonecznie)

więc

P (słonecznie |tak) = 6/12 = 0,50

W powyższym wzorze na 12 zaobserwowanych wzrostów cen pszenicy, 6 razy wzrosty miały miejsce w dzień słoneczny.

P (słonecznie) = 7/21 = 0,33

Prawdopodobieństwo pojawienia się dnia słonecznego wynosi 33% zaś prawdopodobieństwo, że cena pszenicy wzrośnie wynosi 57%.

P(tak) = 12/21 = 0,57

Teraz jeżeli podstawimy nasze obliczenia do wzoru (1) otrzymamy następującą wartość:

$$P(c|x) = \frac{0{,}50 \times 0{,}57}{0{,}33} = 0{,}86$$

Okazało się więc, że nasz analityk odkrył coś, co pozwoliłoby mu zarabiać na giełdzie więcej niż jego koledzy. Prawdopodobieństwo wzrostu ceny akcji w dzień słoneczny wynosi 86%. Analityk przeprowadził zaledwie 21 obserwacji. Tak mała liczba pomiarów może powodować pewne wątpliwości. Z drugiej strony odkryliśmy kolejną zaletę modelu Naive Bayes – bardzo niskie wymagania co do długości szeregu czasowego. Żaden inny model klasyfikacji nie mógłby działać poprawnie z tak krótkim szeregiem czasowym zmiennych.

Zalety i wady modelu Thomasa Bayesa

Niektóre osoby mogą być zirytowane, że ten prosty wzór prawdopodobieństwa nazywamy modelem. Inni mogą czuć niepokój budową samego wzoru (1), gdzie *de facto* obliczana jest średnia ze średnich, co jak wiadomo wydaje się praktyką nieprawidłową.

Model Naive Bayes jest szeroko stosowany do klasyfikacji i regresji dla bardzo dużych i złożonych zestawów danych. Jest on traktowany na równi z bardzo złożonymi modelami *machine learning* i wielokrotnie okazuje się od nich lepszy. Dzieje się tak często wtedy, gdy spełnione jest założenie o niezależności zmiennych predykcyjnych. Model Bayesa działa lepiej dla zmiennych dyskretnych, nie tylko dychotomicznych lecz również dla zmiennych wieloklasowych. W przypadku danych ciągłych przyjmuje się założenie o rozkładzie normalnym zmiennych, co, w połączeniu z małym zbiorem danych (co jest możliwe w tym algorytmie), jest warunkiem trudnym do spełnienia.

Model ten ma również wady. Pierwszą z nich jest tzw. „częstotliwość zerowa", jeżeli do wymienionych przez nas kategorii dodalibyśmy np. czwarty stan pogody np. „pogoda mglista". Model nie znając tego stanu pogody z zestawu treningowego przypisze mu prawdopodobieństwo zerowe. Wywróci to cały algorytm (1), ponieważ jak wiadomo nie można dzielić przez zero. Do rozwiązywania tego problemu stosuje się technikę wygładzania metodą Laplace'a.

Kolejną wadą jest wspomniany wcześniej postulat niezależności zmiennych niezależnych. Zmienne predykcyjne musza być niezależne dla większości modeli. Nigdzie jednak nie podchodzi się do tej kwestii tak bezwzględnie. Ten problem wydaje się nieistotny, jest jednak poważny. Odwrócimy nasz przykład i spróbujmy oszacować, jaka jest pogoda na podstawie wzrostów i braku wzrostów cen pszenicy. A więc zmienną predykcyjną byłby w tym wypadku wzrost lub brak wzrostu. Tę zmienną można potraktować jako dwie zmienne predykcyjne: „wzrost ceny": 1 – tak, 0 – nie oraz „brak wzrostu ceny": 1- tak, 0 – nie. Łatwo zauważyć, że obie zmienne są bardzo silnie ze sobą współzależne. Taka sytuacja uniemożliwia skuteczne zastosowanie modelu Thomasa Bayesa.

Zastosowanie Naive Bayes

Model Naive Bayes stosowany jest tam, gdzie wymagana jest błyskawiczna reakcja. Działa doskonale w układach klasyfikacji nawet tam, gdzie jest bardzo niewielki zbiór danych treningowych lub tam, gdzie istnieje bardzo dużo mało liczebnych klas. Te cechy powodują, że algorytm ten doskonale nadaje się do filtrowania spamów i ogólnie klasyfikacji cech kategorycznych w procesach o dużej liczbie klas.

Wojciech Moszczyński
Współpraca Ewa Moszczyńska

Wojciech Moszczyński – absolwent Katedry Ekonometrii i Statystyki Uniwersytetu Mikołaja Kopernika w Toruniu, specjalista z zakresu ekonometrii, *data science* oraz rachunkowości zarządczej. Specjalizuje się w optymalizacji procesów produkcyjnych i logistycznych. Prowadzi badania w obszarze rozwoju i zastosowania sztucznej inteligencji. Od lat zaangażowany w popularyzację ekonometrii oraz *data science* w środowiskach biznesowych