

Zastosowanie reguł asocjacyjnych algorytmu Apriori



Dzisiaj praktycznie każdy przedmiot i każdą usługę można kupić w internecie. Skala sprzedaży internetowej znacznie przewyższyła wielkość sprzedaży tradycyjnej. Rewolucja retail nie ominęła również branży cukierniczej i piekarskiej.

Czym są reguły asocjacyjne?

Na początku sklepy internetowe prowadziły dość paśywną politykę sprzedaży. Towary były wystawione na stronach internetowych, były promocje, wyprzedaże i ceny podstawowe, słowem było tam to wszystko, co widzimy w tradycyjnych sklepach. Z biegiem czasu internet zapełnił się sklepami, które intensywnie zaczęły między sobą konkurować.

Konkurowanie w sprzedaży internetowej nie polega wyłącznie na wyścigu cenowym lub kolejności na wyświetlanych wynikach wyszukiwarki. W internecie najważniejsze jest zdobycie zainteresowania kupującego, ponieważ to zwiększa popularność. Relacje z kupującym i jego zainteresowanie można zbudować również przez sprytne podpowiadanie mu co ma kupić i co jest dla niego najlepszym wyborem. Zazwyczaj kupujący w internecie wchodzi z sprzedającym (który nota bene nie jest człowiekiem lecz maszyną) w głębszą relację niż gdyby kupowali w zwykłym sklepie. Kupujący internauta porównuje, czyta opisy, charakterystyki i opinie, szuka okazji i jest bardzo podatny na wszelkie sugestie i podpowiedzi ze strony wirtualnych asystentów.

Kiedy kupujemy produkt w internecie pojawia się sugestia, abyśmy kupili również inne towary. Podpowiedzi te są najczęściej zdumiewająco trafne. Często odkrywamy, że zapomnieliśmy o czymś o czym teraz przypominał nam wirtualny asystent sklepu. Za tym zdawałoby się prostym mechanizmem stoją bardzo wyrafinowane algorytmy, który teraz pokrótce opiszę.

Grupowanie klientów

Grupowanie klientów jest najczęściej używaną metodą służącym do polecania produktów klientom. Algorytm zbiera dane z historii wyszukiwania prowadzonych przez klientów. Następnie dane są filtrowane. Na podstawie dokonywanych wyborów system tworzy grupy klientów zachowujących się podobnie.

Kiedyś w klasycznym marketingu mieliśmy do czynienia z segmentacją klientów, grupowaniem ich według wieku, wykształcenia lub na przykład statusu społecznego. W tym wypadku wybór klientów do klastrów opierał się wyłącznie na charakterystyce zachowania klientów.

W przypadku zakupów internetowych zachowanie klientów jest zapisane w rejestrach stron internetowych. Widać tam co robił klient, ile czasu poświęcił na czytanie opinii. Gdzie zaglądał i jakimi produktami się interesował. Na podstawie ogromnej ilości danych stosunkowo prosty algorytm grupowania tworzy kilka lub kilkanaście grup zwanych klastrami. Osoby w klastrach mają zwykle bardzo podobne cechy osobowości i sposób postępowania i co najważniejsze szukają podobnych rzeczy.

Maszyna dostosowuje się do zachowania każdej grupy stosując ogromny wachlarz technik i trików marketingowych. Jednym podpowiada opinie, innym proponuje dodatkowe produkty.

Mechanizm grupuje podobnych klientów szukając wspólnych wzorców zachowań, podobnych preferencji i temperamentów. Jednocześnie algorytm filtruje zakupy dokonywane w klastrach w celu wyszukania podobnych produktów, aby polecić je potencjalnym klientom.

Algorytm Apriori do grupowania par produktów

Szukanie reguł asocjacyjnych jest powszechną techniką przygotowywania oferty dla klienta. Posłużmy się przykładem sieci cukiernio-kafejek.

Sieć kawiarenek postanowiła proponować klientom drugie ciastko za 80% ceny. Dzięki temu cukiernia miała nadzieję na zmniejszenie ilości zwrotów poprzez zachęcenie klientów do większych zakupów.

Nie chciano proponować każdego ciastka każdemu klientowi. Chciano robić to pozornie sporadycznie. Ten trik miał spowodować, że klient czułby się wyróżniony. Problem polegał jednak na tym, że nie bardzo wiadomo było jakie „wyjątkowe” ciastko zaproponować „wyjątkowemu” klientowi. Oczywiście można zaproponować takie, które aktualnie jest w promocji. Jednak klienci cukierni zazwyczaj mają swoje gusta i taka propozycja mogłaby do nich nie trafić. Poza tym ciastko w promocji powinno być sprzedawane przynajmniej z 50% rabatem, a oni chcieli sprzedawać drugie ciastko za 80% ceny podstawowej. Poza tym chodziło o zmniejszenie wielkości zwrotów, a ciastka z promocji tego problemu nie rozwiążą.

Aby dowiedzieć się co najczęściej kupowali klienci wystarczyło przejrzeć wszystkie operacje sprzedaży z ostatnich 3 lat. Było ich 37 tysięcy. Na szczęście sieć cukierni miała komputer z arkuszem kalkulacyjnym. Księgową ściągnęła z programu rejestr sprzedaży i otworzyła go w Excelu.

Za pomocą prostych tabel przedstawnych utworzono zestawienie zawarte w tab. 2. Aby zestawienie było czytelne nazwy towarów zostały zastąpione ich skrótami (np. skrót CH oznacza kawę, a MK oznacza makowiec).

Tab. 2 przedstawia macierz zakupów, gdzie kolumny oznaczają sprzedaż 14 określonych towarów, a wiersze – poszczególne paragony. Gdy klient kupował kawę (oznaczoną skrótem CH) oraz sernik (oznaczony jako SP) w tabeli w kolumnach CH i SP pojawiają się jedynki. Przy pozostałych produktach w tej transakcji pojawiają się zera, ponieważ klient nie kupił tych produktów.

Praktyczne zastosowanie algorytmu Apriori

W naszym przykładzie właściciel cukierni chciał sprzedawać kolejne ciastko za 80% jego wartości. Chciał jednak sprzedawać po obniżonej cenie konkretne ciastka, które byłyby proponowane kupującemu w trakcie zakupów. Celem było zmniejszenie wielkości dziennych zwrotów oraz przyzwyczajanie klientów do większych zakupów w cukierni. Aby akcja się powiodła konieczne jest proponowanie klientom tych ciastek, które oni będą chcieli kupić. Tak więc trzeba wyszukać takie pary produktów, które sprzedają się dobrze razem. Ekspedientka w cukierni po nabiciu na kasę pierwszego ciastka powinna otrzymać podpowiedź z systemu, jakie ciastko zaproponować jako następne.

Aby uruchomić algorytm Apriori wchodzimy do darmowego środowiska Jupyter notebook i importujemy odpowiednie biblioteki.

```
from mlxtend.frequent_patterns import apriori
from mlxtend.frequent_patterns import association_rules
```

Następnie wprowadzamy komendę tworzenia tabeli reguł asocjacyjnych. Jak widać całe zadanie zajęło nam cztery linijki kodu Python.

```
rules = association_rules(frequent_itemsets, metric="lift", min_threshold=0.01)
rules
```

Tabela 1. Zestawienie sprzedaży cukierni.

Data paragonu		Towar	Cena sprzedaży	Numer transakcji
2348822	2019-01-02	WZka	2.70	1
2103627	2019-01-02	Galaretka	2.90	2
897304	2019-01-02	WZka	2.70	3
1477219	2019-01-02	WZka	2.70	4
1175184	2019-01-02	WZka	2.70	5
...	...	...	...	...
1467610	2021-06-30	WZka	2.70	1599624
1856004	2021-06-30	Makowiec	3.10	1599625
246603	2021-06-30	WZka	2.70	1599626
253321	2021-06-30	Szarlotka	2.40	1599627
2085801	2021-06-30	Szarlotka	2.40	1599628

Tabela 2. Macierz zakupów.

Nr. paragonu	CH	DA	HV	HY	IN	IT	MK	NA	PD	SP	ST	SZ	TM	WE
0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
1	1	0	1	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	1	0	0
3	1	0	0	1	0	0	0	0	0	0	1	0	0	0
4	1	0	0	0	0	0	0	0	0	0	0	1	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
35467	1	0	0	0	0	0	0	0	1	0	0	0	0	0
35468	1	0	0	0	0	0	0	0	0	0	0	0	0	0
35469	0	0	0	0	0	0	0	0	0	0	0	1	0	0
35470	0	0	0	0	0	0	0	0	1	1	0	1	0	0
35471	1	0	0	0	0	1	0	0	0	1	1	1	0	1

Algorytm Apriori znalazł 47 reguł asocjacyjnych, 12 z nich znajduje się w tab. 3.

Druga kolumna w pierwszym wierszu oznacza, że algorytm wybrał kawę (CH) i zauważył, że do kawy często kupowane są również szarlotki (oznaczoną skrótem SZ) oraz strucle serowe (oznaczoną jako ST). Tak więc z kawą klienci kupowali często jeszcze dwa określone ciastka.

Kolumna *antecedents* oznacza zakup poprzedzający, natomiast kolumna *consequents* oznacza kolejny zakup. Warto zauważyć, że na paragonach nie ma kolejności dokonywanych zakupów. Algorytm dokonuje symulacji prawdopodobieństw kolejności zakupów.

Reguły asocjacyjne

Poniżej omówimy niektóre reguły asocjacyjne jakie znalazły się w tab. 3.

Antecedent support – mierzy prawdopodobieństwo zakupu wybranego produktu. W pierwszym wierszu znajduje się kawa (oznaczona jako CH). Pojawia się ona w 31% wszystkich zakupów dokonanych w kafejkach.

Consequent support – mierzy prawdopodobieństwo zakupu wybranego produktu lub pary produktów. W pierwszym wierszu w kolumnie *consequents*

Tabela 3. Tablica reguł asocjacyjnych.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
39	(CH)	(SZ, ST)	0.313402	0.035972	0.020439	0.065215	1.812948	0.009165	1.031284
47	(SZ)	(CH, WE)	0.304522	0.048658	0.024583	0.080726	1.659041	0.009765	1.034884
23	(SZ)	(PD)	0.304522	0.055706	0.026218	0.086095	1.545530	0.009254	1.033252
10	(CH)	(TM)	0.313402	0.059286	0.027825	0.088783	1.497531	0.009244	1.032371
33	(SZ)	(TM)	0.304522	0.059286	0.020100	0.066006	1.113350	0.002046	1.007195
45	(CH)	(WE, SZ)	0.313402	0.065291	0.024583	0.078438	1.201368	0.004120	1.014267
40	(SZ)	(CH, ST)	0.304522	0.065601	0.020439	0.067117	1.023112	0.000462	1.001625
14	(IT)	(HV)	0.099853	0.066278	0.026556	0.265951	4.012688	0.019938	1.272017
21	(ST)	(MK)	0.140167	0.068251	0.021087	0.150442	2.204253	0.011521	1.096746
2	(CH)	(MK)	0.313402	0.068251	0.028699	0.091571	1.341687	0.007309	1.025671
15	(HV)	(IT)	0.066278	0.099853	0.026556	0.400681	4.012688	0.019938	1.501948
46	(WE)	(CH, SZ)	0.149216	0.107606	0.024583	0.164746	1.531010	0.008526	1.068410

znajduje się szarlotka i strucla. Paragonów, w których jednocześnie klient kupił te dwa ciastka jest 3,5%. Rzadko klienci kupują jednocześnie te dwa ciastka.

Support – mówi, jak popularny jest zestaw produktów w kolumnach *antecedents* i *consequents*. Jednoczesny zakup kawy, szarlotki i strucli stanowi zaledwie 2% wszystkich transakcji. Przy tej regule asocjacyjnej nie ma znaczenia kolejność w jakiej kupowane były produkty.

Confidence – to reguła asocjacyjna, która mówi jakie jest prawdopodobieństwo, że jak ktoś kupi jeden przedmiot to kupi również drugi. W pierwszym wierszu gdy ktoś kupi kawę, kupi też szarlotkę i struclę. Prawdopodobieństwo tego zjawiska wynosi 6,5%. Proszę zauważyć inną parę z tab. 3. Jeżeli klient kupi galaretkę (oznaczone jako IT) to zgodnie z regułą *confidence* na 27% kupi też sorbet owocowy (oznaczony jako HV). Z kolei, gdy klient zamówi sorbet owocowy jako pierwszy to na 40% zamówi też galaretkę. Oba produkty są ze sobą mocno związane, mimo że są rzadko kupowane przez klientów (co widać w kolumnie *antecedent support*).

Lift – mierzy zależność między prawdopodobieństwem wystąpienia danej reguły a jej prawdopodobieństwem oczekiwanym, gdyby pozycje były od siebie niezależne. Lift określa kategorię możliwości analizowanego zakupu. Gdy lift jest większy od zera prawdopodobnie zostanie kupiony drugi przedmiot.

Jak posługiwać się regułami asocjacyjnymi?

Wykryte reguły asocjacyjne określają różne cechy. Wysoki *support* oznacza, że reguła występuje bardzo często, podczas gdy towarzysząca jej wysoka reguła *confidence* wskazuje, że ma ona dużą moc predykcyjną (co oznacza, że gdy występuje zakup pierwszego produktu to zwykle pojawia się kolejny produkt z reguły asocjacyjnej). Niestety, nie ma żadnych określonych progów ani wartości wzorcowych, które mogłyby nam powiedzieć, czy więz

skojarzeniowa jest względnie mocna czy nie. Ocena na więzi zależy od wielu rzeczy, w tym od specyfiki przedmiotu badań. Ocena ta będzie bazowała w znacznej mierze na subiektywnych odczuciach badacza.

Klasycznym przykładem oceny jest związek pomiędzy kawiozem i szampanem. Oba produkty pojedynczo kupowane są stosunkowo rzadko, jednak często kupowane są razem. Oczekuje się więc, że reguła asocjacji wykaże wysoki poziom *lift* oraz niską poziom *support*. Podobną sytuację mieliśmy przy produktach galaretki (IT) i sorbet owocowy (HV).

Podsumowanie

Reguły asocjacji dostarczają obiektywnych informacji o wzajemnych relacjach między produktami. Niemniej ostatnie słowo leży po stronie badacza, który musi powiedzieć na ile reguła jest przydatna. Subiektywne odczucia i wnioski bazujące na doświadczeniu badacza są ważnym elementem oceny wzorców i podobieństw.

Wojciech Moszczyński

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