

Jak wydobyć sens z nadmiaru danych.

PCA, SVD i LDA w zarządzaniu nowoczesnym koncernem spożywczym



Współczesny przemysł spożywczy funkcjonuje dziś w rzeczywistości, w której informacja stała się jednym z najcenniejszych zasobów strategicznych. Wielkie koncerny spożywcze nie konkurują już wyłącznie jakością produktu, ceną, logistyką czy siłą marki. Coraz częściej o ich przewadze decyduje zdolność do obserwowania, rozumienia i odpowiednio szybkiego interpretowania ogromnych strumieni danych napływających z wielu kierunków jednocześnie.



PCA



SVD



LDA

Są to informacje o działaniach konkurencji, zachowaniach klientów, zmianach cen surowców, sytuacji na rynkach surowców, notowaniach na giełdach towarowych, oczekiwaniach sieci handlowych, a także o zjawiskach makroekonomicznych i mikroekonomicznych, które mogą wpływać na cały łańcuch dostaw.

Skala tego wyzwania jest dziś ogromna. Firmy muszą nieustannie obserwować rynek, analizować dynamikę popytu ze strony supermarketów, przewidywać skutki zmian kursów walut, cen energii, kosztów transportu oraz wahań podaży wynikających z pogody, sytuacji geopolitycznej czy lokalnych zaburzeń produkcji. Do tego dochodzą informacje mniej oczywiste, ale równie istotne: zmiana preferencji konsumentów, rosnące zainteresowanie określonym typem żywności, presja na produkty tańsze, zdrowsze lub bardziej ekologiczne, a także sygnały płynące z mediów, raportów branżowych i danych sprzedażowych. Wszystko to tworzy złożony obraz rynku, którego nie da się już skutecznie analizować wyłącznie intuicją, doświadczeniem menedżera czy prostym arkuszem kalkulacyjnym.

Największy problem nie polega jednak wyłącznie na tym, że danych jest dużo. Problemem jest to, że jest ich za dużo, są

rozproszone, niejednorodne i często wzajemnie sprzeczne. Część informacji wymaga natychmiastowej reakcji, ponieważ może oznaczać ryzyko strat, zakłóceń dostaw albo utraty pozycji wobec konkurencji. Inne okazują się szumem informacyjnym, który spowalnia proces decyzyjny i zaciemnia obraz rzeczywistości. Co więcej, na początku często nie wiadomo, które sygnały są naprawdę ważne, a które jedynie pozornie przyciągają uwagę. Właśnie dlatego kluczową kompetencją nowoczesnego przedsiębiorstwa spożywczego staje się nie tylko zbieranie danych, lecz przede wszystkim ich mądre porządkowanie, filtrowanie, łączenie i redukowanie do takiej formy, która pozwala dostrzec to, co naprawdę istotne.

W praktyce oznacza to konieczność budowy systemów analitycznych zdolnych do oddzielania informacji ważnych od nieistotnych, wyłapywania ukrytych zależności oraz wskazywania, kiedy potrzebne są dodatkowe dane, aby lepiej zrozumieć daną sytuację. Jedna pozornie błaha zmiana w zachowaniu odbiorców detalicznych może być zapowiedzią większego trendu. Niewielkie przesunięcie cen na giełdzie towarowej może stać się początkiem silniejszej tendencji. Z kolei pojedyncza informacja o działaniach konkurencji może zyskać znaczenie

dopiero wtedy, gdy zostanie zestawiona z danymi o kosztach surowców, poziomie zapasów i popycie w kanale nowoczesnym.

W cyfrowym świecie przewagę uzyskują więc nie ci, którzy mają najwięcej danych, lecz ci, którzy potrafią z tej nadmiarowości wydobyć sens. To właśnie w tym miejscu na znaczeniu zyskują metody agregacji i redukcji danych, które pomagają uprościć bardzo złożony obraz rzeczywistości bez utraty jego najważniejszej treści. W świecie przemysłu spożywczego, gdzie liczba zmiennych bywa ogromna, a szybkość decyzji ma bezpośrednie przełożenie na sprzedaż, marżę i bezpieczeństwo operacyjne, narzędzia takie jak PCA (*Principal Component Analysis*), SVD (*Singular Value Decomposition*) i LDA (*Linear Discriminant Analysis*) stają się nie tylko zagadnieniem akademickim, lecz praktycznym wsparciem dla biznesu. Pozwalają one porządkować dane, redukować ich wymiar, wydobywać najważniejsze wzorce oraz wspierać klasyfikację i interpretację zjawisk rynkowych.

Znaczenie takich metod rośnie wraz z cyfryzacją całego sektora. Nowoczesny koncern spożywczy działa dziś jak organizm nerwowy podłączony do setek źródeł informacji. Dane spływają z systemów ERP, platform sprzedażowych, raportów handlowych, źródeł giełdowych, systemów planowania produkcji, monitoringu cen, analiz konsumenckich i wielu innych kanałów. Bez zaawansowanych metod analitycznych łatwo utonąć w tym oceanie danych, zamiast potraktować go jako źródło przewagi konkurencyjnej. Właśnie dlatego zagadnienie inteligentnej agregacji informacji przestaje być domeną wyłącznie specjalistów od data science, a staje się realnym elementem strategii przedsiębiorstw spożywczych.

W niniejszym artykule przyjrzymy się trzem ważnym podejściom do porządkowania i interpretacji złożonych zbiorów danych: PCA, SVD oraz LDA. Choć ich nazwy mogą brzmieć technicznie, ich rola jest bardzo praktyczna: pomagają zobaczyć w danych mniej chaosu, a więcej struktury. W branży, w której każda decyzja musi uwzględniać jednocześnie rynek surowców, klientów, konkurencję, ceny, logistykę i dynamikę popytu, takie metody stają się narzędziami wspierającymi rozsądek, szybkość reakcji i trafność decyzji. A to właśnie one coraz częściej decydują o tym, kto w cyfrowym świecie potrafi nie tylko przetrwać, ale także budować trwałą przewagę.

PCA, czyli *Principal Component Analysis*

PCA, czyli analiza głównych składowych, to jedna z tych metod, które na pierwszy rzut oka brzmią technicznie i akademicko, ale w praktyce odpowiadają na bardzo proste pytanie: jak spojrzeć na ogromną masę danych tak, aby przestać widzieć chaos, a zacząć widzieć porządek. W dużej firmie spożywczej, która codziennie obserwuje ceny towarów, koszty transportu, zachowania supermarketów, opóźnienia dostaw, poziomy zapasów, promocje, reklamacje i ruchy konkurencji, bardzo szybko pojawia się problem nadmiaru informacji. Tych danych jest tak dużo, że człowiek zaczyna się w nich gubić. I właśnie wtedy wchodzi PCA. Ono nie przewiduje jeszcze wyniku, nie mówi od razu, ile zarobimy ani czy sprzedaż wzrośnie. Ono najpierw robi coś innego i bardzo cennego: porządkuje obraz świata.

Najprościej mówiąc, PCA szuka takich grup informacji, które poruszają się razem. I tutaj pojawia się bardzo ważne

słowo: kowariancja. Brzmi poważnie, ale sens jest prosty. Kowariancja mówi nam czy dwie rzeczy mają tendencję do wspólnego ruchu. Jeżeli rośnie jedna i zwykle rośnie też druga, to mówimy, że te dane idą w tym samym kierunku. Jeżeli jedna rośnie, a druga spada, to znaczy, że idą w przeciwnych kierunkach. A jeśli zachowują się niezależnie, to znaczy, że nie tworzą wspólnej historii. Dla przedsiębiorcy nie jest ważne samo słowo „kowariancja”, tylko to, co się za nim kryje: odpowiedź na pytanie, które wskaźniki tak naprawdę opowiadają o tym samym zjawisku.

Wyobraźmy sobie prostą sytuację. Rośnie cena pszenicy. Za chwilę rośnie koszt mąki. Potem pojawia się większa presja na marżę. Jednocześnie zakład zaczyna ostrożniej planować produkcję, bo boi się, że źle policzy koszty. Sieci handlowe mocniej naciskają na cenę. Niby patrzymy na kilka różnych liczb, ale one wszystkie mogą opowiadać jedną historię. PCA potrafi to zauważyć. Zamiast zostawić menedżera z pięcioma osobnymi wykresami, pokazuje mu, że pod spodem działa jeden większy mechanizm, jedna główna siła, jeden wspólny nurt zdarzeń. I właśnie ten nurt nazywa się główną składową.

Dobrze widać to na przykładzie transportu. Wyobraźmy sobie firmę, która wozi produkty spożywcze do sieci handlowych. W pewnym okresie rośnie liczba opóźnień ciężarówek, rośnie liczba pilnych przeplanowań, czyli zmian w awizacji dostaw, rośnie liczba reklamacji od odbiorców, a jednocześnie spada terminowość dostaw do sklepów. Ktoś może powiedzieć: mamy cztery osobne problemy. Ale PCA może pokazać coś innego. Może pokazać, że to nie są cztery niezależne awarie, tylko jeden większy problem logistyczny, który przejawia się na kilka sposobów naraz. I to jest ogromnie ważne, bo firma przestaje wtedy gasić pożary osobno. Zaczyna rozumieć, gdzie naprawdę jest źródło napięcia. Oczywiście doświadczeni menadżerowie to rozumieją. Tylko nie o to chodzi. Chodzi o to, aby systemy automatyczne wyręczały ich w pracy. Aby każdy kto zaczyna swą karierę wiedział i był skuteczny jak doświadczony menadżer. Właśnie o to chodzi.

Gdy mówi się, że pewne dane idą w jednym kierunku, nie chodzi o żaden fizyczny kierunek, tylko o to, że zachowują się podobnie. Jeżeli zamówienia z supermarketów rosną wtedy, gdy rośnie liczba promocji i jednocześnie szybciej schodzą zapasy z magazynu, to te trzy rzeczy tworzą wspólny ruch. Jeżeli natomiast koszty paliwa rosną, ale liczba promocji w sklepach raz rośnie, raz spada i nie widać tu związku, to znaczy, że to są już inne historie. PCA właśnie po to istnieje, aby oddzielić jedną historię od drugiej. Ono bierze dziesiątki zmiennych i próbuje odkryć, ile naprawdę dużych, sensownych procesów działa w tle.

Można to porównać do sytuacji, w której dyrektor operacyjny stoi nad stołem zasypanym raportami. Na jednym arkuszu ma surowce, na drugim transport, na trzecim sprzedaż, na czwartym magazyny, na piątym reklamacje. Każdy raport coś mówi, ale razem tworzą hałas. PCA działa jak ktoś, kto podchodzi do tego stołu i mówi: dobrze, z tych wszystkich papierów wynika tak naprawdę, że dzieją się tu trzy najważniejsze rzeczy. Pierwsza to presja kosztowa. Druga to rozchwiany popyt. Trzecia to napięcie logistyczne. I nagle z chaosu robi się mapa. A kiedy pojawia się mapa, można już podejmować decyzje.

I właśnie tu zaczyna się wartość biznesowa. PCA samo nie podejmuje decyzji za człowieka, ale bardzo dobrze pokazuje, gdzie człowiek powinien patrzeć. Jeżeli z danych wynika, że wiele wskaźników jednocześnie składa się na jeden silny komponent związany z kosztami surowców, energii i transportu, to zarząd może wcześniej zabezpieczać kontrakty, renegocjować ceny albo ograniczać asortyment, który ma zbyt słabą rentowność. Jeżeli inna składowa pokazuje, że napięcie tworzy niestabilność zamówień ze strony sieci, to można zwiększać bufor bezpieczeństwa, inaczej planować promocje lub poprawiać wymianę danych z odbiorcami. PCA nie mówi: zrób dokładnie to i to. Ono mówi: tu jest prawdziwa oś problemu, tu dane układają się w mocny wzór, tutaj firma powinna być szczególnie czujna.

Bardzo ważne jest też to, że PCA jest metodą bezwzorcową, czyli nienadzorowaną. To oznacza, że ona nie potrzebuje gotowej odpowiedzi. Nie trzeba jej wcześniej mówić, co jest sukcesem, co jest porażką, co jest dobrą sprzedażą, a co złą. Ona nie uczy się na zasadzie: pokażę ci historyczne dane, dawne wyniki i ty na podstawie obecnych danych odgadniesz następny wynik. PCA patrzy wyłącznie na same dane i pyta: jakie są ich główne układy, jakie zmienne się pokrywają, które ruchy są wspólne, a które odrębne. To jest trochę tak, jakby w ciemnym pokoju najpierw zapalić światło, zanim zaczniesz się szukać konkretnego przedmiotu. PCA właśnie zapala światło. Dopiero później można zdecydować, do czego tę wiedzę wykorzystać.

Dlatego PCA bywa tak użyteczne także przed klasteryzacją. Kiedy firma chce podzielić produkty, klientów, regiony albo sklepy na sensowne grupy, bardzo często nie warto robić tego od razu na setkach surowych kolumn. Najpierw dobrze jest uprościć obraz i wydobyć najważniejsze wymiary różnic. PCA często świetnie się do tego nadaje, bo redukuje nadmiar i pozwala zobaczyć, które obserwacje naprawdę są do siebie podobne. Dzięki temu klasteryzacja nie odbywa się już na szumie informacyjnym, tylko na bardziej uporządkowanym obrazie rzeczywistości. Innymi słowy, zanim zaczniesz dzielić świat na grupy, dobrze jest najpierw zrozumieć, jakie są główne osie tego świata.

Warto też pamiętać, że PCA nie powstało wczoraj. To nie jest chwilowa moda z ery sztucznej inteligencji. Jej korzenie sięgają początku XX wieku. Za początki tej metody uznaje się prace Karla Pearsona, a później rozwinięcia Harolda Hotellinga. To pokazuje coś ważnego: choć dziś korzystają z niej analitycy danych i zespoły data science, sama idea jest stara i bardzo solidna. Ludzie od dawna mieli ten sam problem co współczesne firmy: jak z wielu liczb wydobyć kilka głównych prawidłowości. Dzisiejszy świat cyfrowy tylko z wielokrotną skalę tego wyzwania.

W przedsiębiorstwie spożywczym PCA jest więc czymś więcej niż statystyczną sztuczką. To narzędzie porządkowania świata. Pozwala zobaczyć, że za setkami wskaźników często stoi zaledwie kilka głównych sił. Jedne dotyczą kosztów, inne popytu, jeszcze inne logistyki lub zachowań odbiorców. Z perspektywy menedżera to niezwykle cenna zmiana. Zamiast patrzeć na dziesiątki tabel, można zacząć patrzeć na kilka prawdziwych mechanizmów rządzących biznesem. A kiedy firma lepiej rozumie, które informacje poruszają się razem i tworzą wspólny

kierunek, łatwiej odróżnia sygnał od szumu, szybciej reaguje i łatwiej podejmuje decyzje, które nie są chaotyczne, lecz oparte na strukturze ukrytej w danych.

SVD, czyli *Singular Value Decomposition*

SVD też ma bardzo ciekawą historię. Jej korzenie są starsze niż wiele współczesnych narzędzi data science. Początki tej idei sięgają XIX wieku, kiedy matematycy badali rozkład macierzy i szukali sposobów opisywania złożonych układów liniowych. Potem metoda rozwijała się dalej i z czasem stała się jednym z filarów nowoczesnej algebry liniowej, analizy danych, przetwarzania sygnałów i uczenia maszynowego. To metoda głęboko zakorzeniona w matematyce, która po prostu dziś znalazła ogromnie praktyczne zastosowanie w świecie cyfrowego biznesu.

Singular Value Decomposition, po polsku rozkład według wartości osobliwych, to kolejna metoda, która pomaga uporządkować ogromny bałagan informacyjny i wydobyć z niego kilka najważniejszych układów. Jeżeli wcześniej mówiliśmy, że PCA, czyli analiza głównych składowych, szuka głównych kierunków zmienności w danych i bardzo mocno opiera się na idei kowariancji, to SVD można potraktować jako metodę jeszcze bardziej ogólną i jeszcze bardziej techniczną. PCA patrzy na dane trochę jak analityk, który chce zrozumieć, które cechy poruszają się razem. SVD patrzy na dane bardziej jak inżynier matematyczny, który rozbiera wielką tabelę na kilka prostszych części i pyta z jakich głównych wzorów ten cały obraz właściwie się składa.

W praktyce przedsiębiorcy różnica nie musi być od razu odczuwalna, bo obie metody często prowadzą do podobnego celu: zmniejszenia złożoności i pokazania, co naprawdę rządzi danymi. Ale sposób myślenia jest trochę inny. PCA mówi: poszukajmy głównych osi zmienności i zobaczmy, które informacje idą razem, czyli mają wspólny ruch, wspólną historię. SVD mówi: weźmy wielką macierz danych, czyli ogromną tabelę, i rozłóżmy ją na kilka prostszych warstw, tak aby zobaczyć, które ukryte wzory są najmocniejsze, które słabsze, a które można już uznać za szum lub drobne szczegóły.

Brzmi to może abstrakcyjnie, ale wyobraźmy sobie wielki koncern spożywczy, który ma tysiące produktów, setki sklepów, wiele regionów sprzedaży i ogromne tabele pokazujące ile którego produktu sprzedano w którym miejscu, w którym tygodniu, przy jakiej cenie, promocji i sytuacji rynkowej. Taka tabela jest tak wielka, że człowiek nie widzi w niej nic poza ścianą liczb. I wtedy właśnie SVD działa jak narzędzie do rozmontowywania ogromnej maszyny. Ono mówi: spójrzmy spokojnie. W tej tabeli nie ma tysiąca niezależnych historii. Są pewne powtarzające się wzory. Jedne produkty zachowują się podobnie w wielu sklepach. Niektóre regiony kupują podobnie. Niektóre tygodnie sprzedaży mają wspólny charakter, bo są promocyjne, sezonowe albo kryzysowe. SVD pomaga rozłożyć ten chaos na kilka głównych warstw znaczenia.

Można to sobie wyobrazić tak, jakby firma miała przed sobą ogromną mapę ruchu ciężarówek, zamówień i sprzedaży. Na pierwszy rzut oka wszystko się miesza. Jedna ciężarówka jedzie do Warszawy, druga do Łodzi, trzecia do Krakowa. Tu rośnie popyt na mąkę, tam na olej, gdzie indziej na gotowe

pieczywo. Menedżer patrzy i widzi tłok. A SVD mówi: w tym tłoku są pewne układy. Na przykład pierwszy, najsilniejszy wzór może pokazywać ogólny rytm dużych sieci handlowych. Drugi wzór może oddzielać regiony bardziej wrażliwe cenowo od mniej wrażliwych. Trzeci może pokazywać sezonowość. Czwarty może ujawniać zachowania związane z promocjami. I nagle z pozornego chaosu zaczyna się wyłaniać struktura.

Warto tu wyjaśnić coś bardzo ważnego. W PCA dużo mówiliśmy o tym, że dane idą w jednym kierunku, czyli że na przykład koszt zboża, mąki i presja na cenę mogą rosnąć razem. W PCA był język wspólnego ruchu i kowariancji. W SVD też szukamy układu, ale trochę inaczej. Tu nie chodzi wyłącznie o to, czy dwie kolumny rosną razem. Tu chodzi o to, czy cała wielka tabela da się opisać przez kilka głównych wzorów. To jest tak, jakby ktoś powiedział: nie analizujemy każdej ciężarówki osobno, tylko sprawdzimy czy cała flota działa według kilku powtarzalnych schematów. Może jeden schemat dotyczy dostaw do dyskontów, drugi do hipermarketów, trzeci do małych sklepów regionalnych. SVD jest świetne właśnie do takiego odkrywania ukrytej architektury danych.

Bardzo prosty przykład. Firma ma tabelę, w której w wierszach są produkty, a w kolumnach sklepy. W komórkach jest sprzedaż. Patrząc zwyczajnie, widzimy tylko tysiące liczb. Ale SVD może ujawnić, że pod spodem działają na przykład trzy wielkie wzory. Pierwszy wzór mówi, że część produktów dobrze sprzedaje się prawie wszędzie, bo są to towary podstawowe. Drugi wzór pokazuje, że pewne produkty szczególnie dobrze idą w sklepach miejskich, a słabiej poza miastem. Trzeci wzór pokazuje, że część asortymentu żyje głównie promocją. Czyli zamiast patrzeć na każdy produkt i każdy sklep osobno, firma zaczyna widzieć kilka głównych logik rynku. A to już jest bardzo cenna wiedza do decyzji.

I właśnie tutaj SVD robi się szczególnie praktyczne. Bo kiedy firma widzi, że dane da się sprowadzić do kilku dominujących wzorów, może podejmować bardziej sensowne decyzje. Może inaczej budować grupy produktów, może szybciej wykrywać odstępstwa od normalnego rytmu rynku, może sprawdzać, czy nagły spadek sprzedaży to problem jednego sklepu, jednego regionu czy może pęknięcie całego większego wzoru. Może też odróżnić, co jest dużym zjawiskiem systemowym, a co tylko drobnym lokalnym zakłóceniem. W wielkim koncercie spożywczym to ma ogromne znaczenie, bo nie każda anomalia wymaga alarmu na poziomie zarządu. Czasem to tylko mały hałas. Ale czasem to początek większej zmiany i właśnie SVD pomaga zobaczyć, czy coś należy do głównego obrazu, czy tylko do tła.

Jeżeli porównamy SVD z PCA jeszcze bardziej po ludzku, to można powiedzieć tak: PCA przypomina analityka, który siada przy stole i mówi: zobaczmy, które wskaźniki poruszają się razem, które tworzą wspólne kierunki i jakie są główne osie zmienności w firmie. Natomiast SVD przypomina mechanika precyzyjnego, który bierze ogromną maszynę danych, rozkłada ją na części i pokazuje, z jakich najważniejszych warstw jest zbudowana. PCA bardzo dobrze tłumaczy świat przez zmienność i kowariancję. SVD bardzo dobrze tłumaczy świat przez rozkład całej tabeli na główne wzory. Dla przedsiębiorcy najważniejsze jest to, że obie metody pomagają zobaczyć mniej chaosu, a więcej sensu. Tyle że PCA częściej tłumaczy

się językiem statystycznym, a SVD częściej językiem struktury danych.

Co ciekawe, te dwie metody są ze sobą bardzo blisko związane. W praktyce bardzo często PCA oblicza się właśnie z pomocą SVD. Można więc powiedzieć, że PCA i SVD to nie są dwie odległe planety. Raczej dwa różne sposoby opowiedzenia o podobnym porządkowaniu świata. PCA mówi bardziej o głównych składowych i wariancji. SVD mówi bardziej o rozkładzie macierzy na podstawowe elementy. Z punktu widzenia biznesu efekt bywa podobny: ogromna liczba danych przestaje przysygniać i zaczyna układać się w kilka głównych linii interpretacyjnych.

W firmie spożywczej SVD może być szczególnie użyteczne wtedy, gdy dane mają postać bardzo dużych tabel i relacji. Na przykład wtedy, gdy chcemy zrozumieć układ sprzedaży produktów w kanałach handlowych, relacje między regionami a asortymentem, wzorce zamówień supermarketów albo strukturę podobieństw między okresami sprzedażowymi. Tam, gdzie człowiek widzi tylko ogromną matrycę liczb, SVD potrafi pokazać główne osie ukrytego porządku. Dzięki temu łatwiej zauważyć, że pewne grupy produktów zachowują się podobnie, że część sklepów należy do tej samej logiki popytu albo, że rynek właśnie zaczyna przesuwać się w nowy wzór zachowań.

Można to też porównać do słuchania orkiestry. Kiedy słyszymy cały zespół naraz, początkujący słuchacz odbiera tylko głośny, złożony dźwięk. Ale ktoś doświadczony potrafi wychwycić główne warstwy: melodię, rytm, tło, wejścia instrumentów. SVD robi coś podobnego z danymi. Zamiast patrzeć na tysiące liczb jako na niezrozumiały hałas, rozdziela ten hałas na główne warstwy wzoru. I właśnie dlatego bywa tak cenne w biznesie. Nie dlatego, że daje magiczną odpowiedź, ale dlatego, że pozwala lepiej usłyszeć, co naprawdę gra w danych.

Jeżeli więc PCA pomaga zrozumieć, które informacje poruszają się razem i tworzą główne kierunki zmienności, to SVD pomaga rozłożyć całą złożoną tabelę świata na jej najważniejsze ukryte wzory. Jedno i drugie służy temu samemu celowi: zmniejszeniu chaosu i wydobywaniu sensu. W nowoczesnym koncercie spożywczym, gdzie decyzje trzeba podejmować szybko, a danych jest więcej, niż człowiek jest w stanie ogarnąć, takie metody nie są już akademicką ciekawostką. Stają się narzędziem przetrwania, porządkowania rzeczywistości i budowania przewagi nad tymi, którzy nadal patrzą tylko na pojedyncze tabelki, zamiast widzieć większy wzór.

Teraz poznamy metodę nadzorowaną. Metodę, której wskazujemy cele.

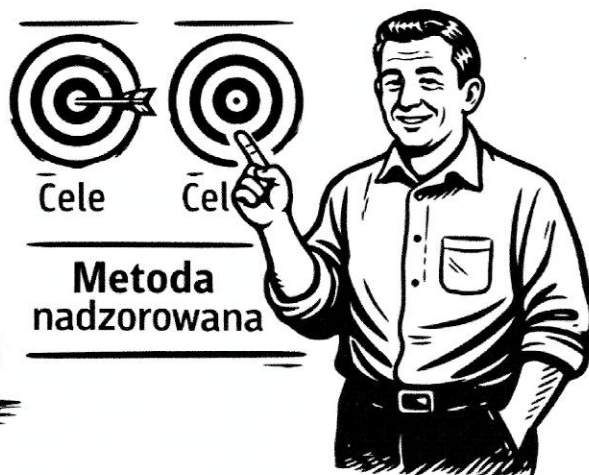
LDA, czyli *Linear Discriminant Analysis*

LDA, to po polsku liniowa analiza dyskryminacyjna, to moment, w którym z metod patrzących na sam chaos danych przechodzimy do metody, której pokazujemy cel. I właśnie dlatego mówi się, że jest to metoda nadzorowana. W metodzie nienadzorowanej, takiej jak PCA albo SVD, algorytm dostaje tylko tabelę liczb i ma sam odkryć, jakie są w niej główne układy, wzory i kierunki. Natomiast w LDA dostaje coś więcej: do danych dokładamy informację, co jest czym. Mówimy na przykład: te sklepy są stabilne, a te niestabilne. Te partie dostaw kończyły się reklamacjami, a te były bezproblemowe.



PCA

SVD



LDA

Te produkty należą do segmentu premium, a te do segmentu bardzo wrażliwego cenowo. LDA nie patrzy już więc na świat niewinnie i bez celu. Od początku wie, jakiego podziału szukamy, i właśnie dlatego działa inaczej niż PCA i SVD. LDA jest opisywana wprost jako metodę do nadzorowanej redukcji wymiaru, która rzutuje dane na takie kierunki, które maksymalizują separację między klasami. Jednocześnie to także klasyfikator z liniową granicą decyzyjną. Zabrzmiało to zdecydowanie zbyt skomplikowanie.

Najprościej można to wyjaśnić tak. Wyobraźmy sobie duży koncern spożywczy i setki sklepów. Mamy o nich mnóstwo danych: wielkość zamówień, liczbę promocji, opóźnienia dostaw, marżę, liczbę reklamacji, tempo rotacji towaru, udział produktów świeżych, wrażliwość klientów na cenę. Jeśli użyjemy PCA, to metoda zapyta: które cechy w tej ogromnej tabeli poruszają się razem i gdzie jest największa zmienność? Czyli będzie porządkowała sam krajobraz danych. Jeśli użyjemy SVD, to rozłoży wielką tabelę na kilka głównych ukrytych wzorów, jakby rozbiierała ogromny mechanizm sprzedaży na podstawowe warstwy. Ale jeśli użyjemy LDA, to najpierw musimy powiedzieć: oto sklepy, które chcemy uznać za „bezpieczne biznesowo”, a oto sklepy, które uznajemy za „ryzykowne”. I wtedy LDA zaczyna szukać nie tego, co w danych jest po prostu największe, tylko tego, co najlepiej oddziela jedną klasę od drugiej. To właśnie jest sens słowa „nadzorowana”: człowiek daje metodzie kierunek, a metoda szuka najlepszego rozdzielania według tego wskazanego celu. To bardzo ważna różnica, bo ludzie często myślą, że wszystkie te metody robią mniej więcej to samo. A tu mamy ogromne różnice.

PCA może odkryć, że w danych największą rolę gra ogólna skala działalności sklepu: duży sklep ma większy obrót, większe zatowarowanie, większą liczbę dostaw i większą liczbę promocji. Tyle, że ta informacja nie musi jeszcze pomagać w odróżnieniu sklepu zdrowego od problematycznego. Może po prostu opisywać rozmiar działalności.

LDA działa inaczej. Ono powie: mnie nie interesuje wyłącznie największa zmienność. Mnie interesuje taki kierunek

patrzenia na dane, który sprawi, że sklepy zdrowe i sklepy problematyczne odsuną się od siebie jak najmocniej, a jednocześnie sklepy wewnątrz każdej z tych grup pozostaną możliwie zwarte. W tym sensie LDA jest jak dyrektor handlowy, który pyta: „po czym najłatwiej odróżnić biznes dobry od złego?”. Matematycznie ta intuicja bierze się z idei maksymalizacji rozrzutu między klasami przy jednoczesnym ograniczaniu rozrzutu wewnątrz klas; współczesne ujęcie LDA jest też oparte na modelowaniu rozkładów klas i użyciu reguły Bayesa, czyli najprościej mówiąc prawdopodobieństwie warunkowym.

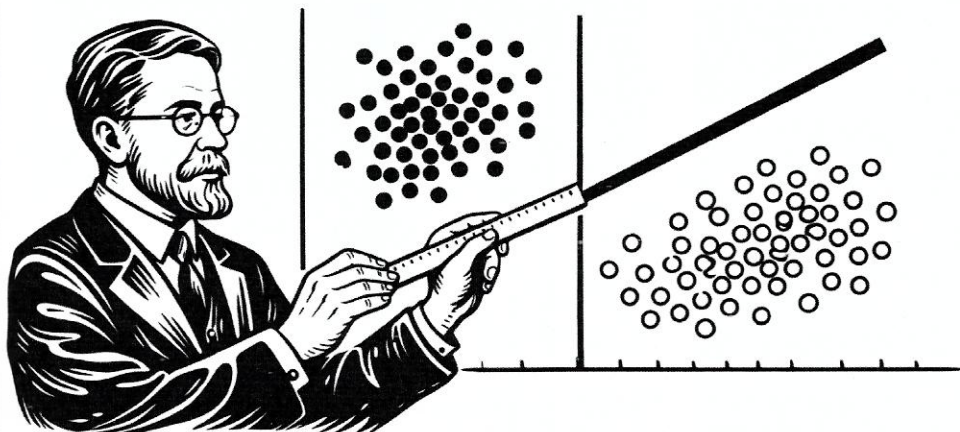
Można to zobaczyć na bardzo prostym przykładzie z ciężarówkami. Załóżmy, że firma codziennie ocenia trasy dostaw i po czasie już wie, które trasy kończyły się terminowo, a które kończyły się opóźnieniami i reklamacjami. Do każdej trasy mamy dane: długość przejazdu, liczbę punktów rozładunku, porę dnia, natężenie ruchu, typ klienta, średni czas postoju, liczbę zmian w zamówieniu.

PCA spojrzaloby na to i próbowało odkryć, które z tych cech zwykle poruszają się razem. SVD próbowałoby rozłożyć całą wielką tabelę tras na kilka głównych wzorów logistycznych. Natomiast LDA dostałoby już gotową wiedzę: te trasy były dobre, a te złe. I zaczęłoby szukać takiej kombinacji cech, która najczyściej oddziela jedno od drugich. Może się okazać, że nie sama długość trasy jest tu najważniejsza, ani nie sama liczba punktów rozładunku, lecz połączenie kilku rzeczy naraz: długi przejazd, wiele punktów, długi czas postoju i klient często zmieniający zamówienie. LDA potrafi z takich pozornie zwykłych kolumn złożyć nową oś, która bardzo dobrze mówi: tu zaczyna się strefa ryzyka. I to jest już wiedza operacyjna, a nie tylko opisowa.

Właśnie dlatego LDA jest tak cenne tam, gdzie firma nie chce już tylko „rozumieć danych”, ale chce oddzielać klasy zjawisk. W branży spożywczej może to być oddzielanie sklepów dobrze rokujących od słabo rokujących, partii towaru wysokiego i niskiego ryzyka, klientów stabilnych i niestabilnych, produktów regularnych i silnie promocyjnych, dostaw podatnych i niepodatnych na zakłócenia. Jeżeli PCA jest trochę jak

RONALD A. FISHER - 1936

O wykorzystaniu wielu pomiarów do rozróżniania grup



Warto też powiedzieć uczciwie, że LDA nie jest po prostu „lepszym PCA”. To byłoby nieporozumienie. Ono jest inne. PCA i SVD są świetne, gdy firma jeszcze nie chce albo nie może narzucić jednej etykiety i chce najpierw zrozumieć strukturę świata. LDA robi się mocne wtedy, gdy firma ma już jakąś wiedzę historyczną i może powiedzieć: te przypadki były dobre, a te złe; te należą do klasy A, a te do klasy B. Wtedy LDA zaczyna pracować znacznie bardziej celowo. Co ważne, scikit-learn podaje, że LDA redukuje wymiar do najbardziej dyskryminujących kierunków, a liczba tych nowych osi jest ograniczona przez liczbę

spokojny kartograf, który rysuje mapę świata danych, a SVD jak mechanik rozkładający wielką maszynę na najważniejsze moduły, to LDA jest jak doświadczony kierownik jakości, który mówi: „dobrze, ale pokażmy wreszcie, co odróżnia przypadki dobre od złych”. Ono nie zadowala się samym porządkiem. Ono szuka porządku użytecznego do rozróżniania.

Tu dobrze widać, dlaczego metoda nadzorowana bywa dla biznesu tak kusząca. W PCA i SVD można odkryć bardzo mądre wzory, ale nie ma gwarancji, że będą one najważniejsze z punktu widzenia konkretnego celu. Na przykład największa zmienność w danych może dotyczyć sezonowości albo wielkości sklepów, choć zarząd najbardziej chce zrozumieć reklamacje albo spadki rentowności. LDA od razu ustawia świat pod wskazany cel. To tak, jakby dwie osoby patrzyły na ten sam magazyn. Jedna mówi: „pokażmy główne układy ruchu towaru”. To jest sposób myślenia bliższy PCA i SVD. Druga mówi: „pokażmy, co odróżnia zamówienia bezpieczne od tych, które kończą się problemem”. To jest sposób myślenia LDA. Jedno podejście jest odkrywcze, drugie celowane.

Bardzo praktyczny przykład można wziąć z relacji z supermarketami. Firma ma dane o setkach promocji i po czasie wie, które promocje były opłacalne, a które tylko pompowały wolumen bez sensownego zysku. Do każdej promocji ma dane o poziomie rabatu, czasie trwania, typie sklepu, regionie, sezonie, wielkości ekspozycji, udziale marki własnej konkurenta, kosztach dostaw i skali zwrotów. Gdyby zastosować PCA, firma mogłaby odkryć główne osie zmienności promocji. Gdyby zastosować SVD, mogłaby zobaczyć główne ukryte wzory całej tabeli promocyjnej. Ale LDA dostaje etykiety: promocja opłacalna, promocja nieopłacalna. I wtedy może wskazać kombinację cech, która najlepiej rozdziela jedne przypadki od drugich. W praktyce taka oś może ujawnić, że najbardziej niebezpieczne są promocje długie, głębokie cenowo, prowadzone w określonym typie sklepów i przy jednoczesnym wzroście kosztu dostaw. To już jest materiał do konkretnej decyzji: które promocje przepuszczać, które ciąć, a które zatwierdzać tylko warunkowo.

bę klas, więc jest to z natury dość mocna redukcja wymiaru.

Historycznie korzenie tej metody wywodzą się od ojca współczesnej statystyki Ronalda Fishera i jego pracy z 1936 roku o wykorzystaniu wielu pomiarów do rozróżniania grup. Jest starym, klasycznym narzędziem statystycznym, które bardzo wcześniej próbowało odpowiedzieć na pytanie: jak spojrzeć na wiele cech jednocześnie, aby jak najlepiej odróżnić jedną grupę od drugiej. Znaleźć takie spojrzenie na dane, które nie tylko upraszcza obraz, lecz przede wszystkim wyodrębnia różnice ważne z punktu widzenia celu.

Podsumowanie

I właśnie dlatego LDA domyka całą tę historię o trzech sposobach. PCA mówiło nam: zobacz, które informacje poruszają się razem i tworzą główne kierunki zmienności. SVD mówiło: rozłóżmy wielką tabelę świata na kilka najważniejszych ukrytych wzorów. A LDA mówi: skoro już wiemy, jaki podział nas interesuje, znajdziemy takie osie patrzenia, które najlepiej oddzielą jedną klasę od drugiej. W nowoczesnym koncernie spożywczym każda z tych metod ma swoje miejsce. Jedne pomagają zrozumieć chaos, inne pomagają go rozebrać, a LDA pomaga wreszcie powiedzieć: tu przebiega granica między przypadkiem bezpiecznym a ryzykownym, między promocją zdrową a szkodliwą, między klientem stabilnym a klientem wymagającym szczególnej uwagi. I właśnie dlatego metoda nadzorowana jest dla biznesu tak cenna: nie tylko porządkuje świat, ale porządkuje go pod konkretną decyzję.

Wojciech Moszczyński

Wojciech Moszczyński – absolwent Katedry Ekonometrii i Statystyki Uniwersytetu Mikołaja Kopernika w Toruniu; specjalista z zakresu ekonometrii, finansów, 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 machine learning oraz data science w środowiskach biznesowych.