K-means vs. liquid clustering. Dlaczego „sztywne szuflady" zawodzą w sprzedaży maszyn B2B?

W środowisku sprzedaży drogich maszyn i linii technologicznych dla przemysłu spożywczego precyzja doboru pary Klient-Handlowiec (tzw. *smart routing*) staje się kluczowym czynnikiem przewagi konkurencyjnej. Niniejsze opracowanie analizuje fundamentalny problem niestabilności klasyfikacji handlowców w systemach opartych na sztucznej inteligencji.

Porównujemy dwa dominujące paradygmaty matematyczne: klasyczny, geometryczny algorytm K-średnich (K-means) oraz zaawansowany, statystyczny model Mieszanin Gaussowskich (GMM). Analiza dowodzi, że w przypadku profili hybrydowych, stanowiących znaczną część populacji handlowców, podejście deterministyczne (K-means) generuje błędy kategoryzacji, wynikające ze sztywnych granic decyzyjnych, co prowadzi do fluktuacji w procesie sprzedaży. Podejście probabilistyczne (GMM), oferujące tzw. miękką klasteryzację, zapewnia stabilność operacyjną niezbędną do wdrożenia przewidywalnego systemu wsparcia decyzji.

Inżynieria relacji w przemyśle maszynowym

Sektor produkcji zaawansowanych maszyn dla przetwórstwa spożywczego charakteryzuje się specyficzną dynamiką handlową. W odróżnieniu od sprzedaży detalicznej, gdzie transakcje są szybkie i impulsywne, proces zakupu linii pakującej lub systemu dozowania trwa tygodniami i angażuje wiele osób decyzyjnych. Wartość pojedynczego kontraktu liczona jest często w setkach tysięcy lub milionach złotych. W takim ekosystemie rola call center i działu sprzedaży wykracza daleko poza prostą obsługę zamówień. Staje się strategicznym węzłem, w którym kompetencje inżynieryjne muszą zostać idealnie zsynchronizowane z umiejętnościami negocjacyjnymi.

Celem nadrzędnym analityki danych w tym obszarze jest maksymalizacja prawdopodobieństwa konwersji (zamknięcia sprzedaży) poprzez inteligentne kierowanie przychodzących zapytań do najodpowiedniejszych handlowców. Jest to proces optymalizacyjny, który wymaga stworzenia modelu matematycznego opisującego rzeczywistość – profilowania klientów i handlowców.

Fundamentalnym wyzwaniem, przed którym stają działy Data Science, nie jest samo pozyskanie danych (korzystamy z transkrypcji Whisper-1, analizy sentymentu, modeli leksykalnych), lecz ich interpretacja w czasie. Kluczowym parametrem oceny jakości każdego modelu wdrożonego w środowisku produkcyjnym jest stabilność. System, który jednego tygodnia klasyfikuje pracownika jako „Eksperta Technicznego", a w kolejnym jako „Agresywnego Negocjatora" tylko na podstawie marginalnej zmiany w jego zachowaniu, jest bezużyteczny z punktu widzenia zarządzania procesem. Takie fluktuacje uniemożliwiają budowę wiarygodnych macierzy prawdopodobieństwa sprzedaży, a co za tym idzie – uniemożliwiają prognozowanie przychodów.

Niniejszy artykuł stawia tezę, że źródłem tej niestabilności nie jest jakość danych wejściowych, lecz niewłaściwy dobór metody matematycznej służącej do grupowania (klasteryzacji) obiektów. Przeanalizujemy, dlaczego standard rynkowy (K-means) zawodzi w kontakcie z ludzką naturą i dlaczego metody probabilistyczne (GMM/LCA) stanowią konieczną ewolucję w systemach klasy Enterprise.

Natura danych: problem ciągłości spektrum behawioralnego

Aby zrozumieć istotę problemu, musimy przyjrzeć się strukturze danych, na których operujemy. Modele psychometryczne, takie jak Wielka Piątka (OCEAN) czy HEXACO, oraz behawioralne analizy technik wywierania wpływu (według Cialdiniego), dostarczają nam zmiennych o charakterze ciągłym, a nie dyskretnym.

W inżynierii materiałowej rzadko spotykamy czyste pierwiastki. Podobnie w psychologii sprzedaży rzadko spotykamy „czyste archetypy". Handlowiec nie jest zmienną binarną. Nie jest albo „Doradcą" (1), albo „Sprzedawcą" (0). W rzeczywistości każdy pracownik działu handlowego jest unikalną mieszanką różnych stylów i kompetencji. Możemy wyobrazić sobie przestrzeń wielowymiarową, w której każdy wymiar reprezentuje inną cechę (np. asertywność, empatię, wiedzę techniczną).

Większość populacji nie znajduje się w ekstremach tej przestrzeni (w narożnikach), lecz w jej centrum. Są to tzw. profile

hybrydowe. Pracownik taki może w 30% przejawiać zachowania analityczne, w 30% zachowania relacyjne, a w 40% zachowania dyrektywne. Jest to naturalne zjawisko – ludzie dostosowują się do sytuacji, a ich styl jest wypadkową osobowości i bieżącego kontekstu.

Problem analityczny pojawia się w momencie, gdy biznes żąda kategoryzacji. Dyrektor Handlowy potrzebuje raportu, w którym pracownicy są przypisani do konkretnych grup, aby móc zarządzać zespołami. Data Scientist staje więc przed zadaniem: jak zredukować tę złożoną, wielowymiarową i ciągłą rzeczywistość do skończonej liczby grup (klastrów), nie tracąc przy tym kluczowych informacji? Tu właśnie dochodzi do kolizji dwóch światów matematycznych: geometrii euklidesowej i statystyki probabilistycznej.

Podejście deterministyczne: K-means i iluzja twardych granic

Metoda K-średnich (K-means), często wspierana redukcją wymiarowości PCA (*Principal Component Analysis*), jest najczęściej stosowanym algorytmem w biznesie. Jej popularność wynika z prostoty implementacji i szybkości działania. Jest to jednak metoda deterministyczna, oparta na twardej logice geometrycznej.

Mechanizm działania: voronoi i podział przestrzeni

Algorytm K-means traktuje problem klasyfikacji jako zadanie podziału przestrzeni na rozłączne obszary. Działa on w sposób iteracyjny, dążąc do wyznaczenia punktów centralnych (centroidów), które minimalizują sumę kwadratów odległości wszystkich punktów danych od najbliższego im środka.

Można to porównać do sortowania fizycznych obiektów, np. kamieni o różnych odcieniach, do kilku wiader. Każdy kamień musi trafić do jednego i tylko jednego wiadra. Algorytm bierze kamień, mierzy jego „odległość" (podobieństwo) do wzorca w każdym wiadrze i wrzuca go tam, gdzie pasuje najlepiej. Nie ma możliwości, aby kamień znalazł się w dwóch wiadrach jednocześnie.

W rezultacie przestrzeń cech zostaje pocięta na tzw. komórki Voronoia. Są to wielościenne obszary oddzielone od siebie ostrymi granicami matematycznymi. Wewnątrz komórki algorytm traktuje wszystkie punkty jako „takie same". Punkt znajdujący się w samym centrum klastra i punkt znajdujący się na jego absolutnym brzegu otrzymują tę samą etykietę.

Problem obszarów granicznych

Główną słabością K-means w zastosowaniach psychometrycznych jest sposób traktowania przypadków granicznych. Wróćmy do naszego handlowca hybrydowego, którego profil składa się w 33% z cech typu A, w 33% z cech typu B i w 34% z cech typu C. Z geometrycznego punktu widzenia, ten punkt danych znajduje się niemal w równej odległości od trzech centroidów, ale jest minimalnie bliżej centroidu C. Algorytm K-means, działając w trybie „Hard Voting" (twardego głosowania), podejmuje decyzję binarną: przypisuje handlowca w 100% do klastra C. Indykator przynależności przyjmuje postać wektora: .

W tym momencie następuje drastyczna utrata informacji. System „zapomina", że ten pracownik posiada 66% cech innych

typów. Co gorsza, system zaczyna traktować go identycznie jak pracownika, który jest w 95% typem C. Jest to błąd generalizacji, który w inżynierii nazywamy błędem kwantyzacji.

Konsekwencje: niestabilność systemowa

Najpoważniejszą konsekwencją podejścia twardego jest wrażliwość na szum. Ludzkie zachowanie jest z natury stochastyczne (losowe). Handlowiec w jednym tygodniu może być nieco bardziej energiczny, a w innym nieco bardziej wycofany. Te naturalne wahania powodują, że punkt reprezentujący handlowca w przestrzeni cech nieustannie drga.

Dla handlowca hybrydowego (33/33/34), minimalne przesunięcie parametrów, np. wynikające z jednej gorszej rozmowy, może zmienić jego proporcje na 34/33/33. Dla człowieka to zmiana niezauważalna (1%). Jednak dla algorytmu K-means oznacza to przekroczenie matematycznej granicy komórki Voronoia. W rezultacie, handlowiec z tygodnia na tydzień zmienia etykietę z typu C na typ A. Z perspektywy systemu routingu jest to trzęsienie ziemi. Algorytm sterujący, który do tej pory kierował do niego klientów pasujących do profilu C, nagle przestaje to robić i zaczyna kierować klientów profilu A.

Prowadzi to do zjawiska „migotania klastrów". Macierz prawdopodobieństwa sprzedaży, która powinna być fundamentem strategii, staje się obrazem chaotycznym. Dyrektorzy techniczni widzą zmieniające się słupki wydajności, które nie wynikają ze zmian rynkowych, lecz z artefaktów obliczeniowych metody K-means.

Podejście probabilistyczne: GMM i rzeczywistość mieszana

Odpowiedzią na ograniczenia geometrii euklidesowej jest statystyka bayesowska i modelowanie gęstości. Podejście to reprezentują Modele Mieszanin Gaussowskich (*Gaussian Mixture Models* – GMM) oraz, w przypadku zmiennych kategorycznych, Analiza Klas Ukrytych (LCA).

Mechanizm działania: rozkłady i miękka klasteryzacja

GMM nie dzieli przestrzeni na sztywne kawałki. Zamiast tego zakłada, że dane obserwowane są wynikiem działania kilku ukrytych procesów generujących, z których każdy ma charakter rozkładu normalnego (krzywej Gaussa). Używając analogii przemysłowej: zamiast sortować kamienie do wiader, GMM działa jak spektrometr analizujący skład stopu metalu. Jeśli badamy próbkę stali, urządzenie nie mówi nam „to jest żelazo" albo „to jest węgiel". Mówi nam: „ta próbka składa się w 98% z żelaza, w 1,5% z węgla i w 0,5% z domieszek".

Algorytm GMM, wykorzystując metodę Oczekiwania-Maksymalizacji (*Expectation-Maximization*, EM), dopasowuje do chmury danych wielowymiarowe kształty (elipsoidy) reprezentujące gęstość prawdopodobieństwa. Kluczową różnicą

jest wynik końcowy. Dla każdego punktu danych GMM nie zwraca etykiety, lecz wektor prawdopodobieństw przynależności do każdego z klastrów. Jest to tzw. *soft clustering*.

Absorpcja wariancji

Wróćmy do naszego problematycznego handlowca hybrydowego (33% A, 33% B, 34% C). Algorytm GMM „zrozumie" jego naturę. Wynikiem działania modelu będzie wektor: . Co się stanie w kolejnym tygodniu, gdy nastąpi owa 1-procentowa fluktuacja zachowania i proporcje zmienią się na 34/33/33? GMM zaktualizuje wektor wyjściowy na: .

Różnica jest fundamentalna. W podejściu K-means nastąpiła 100% zmiana kategorii (skokowa zmiana etykiety). W podejściu GMM nastąpiła jedynie marginalna korekta wag prawdopodobieństwa. System odnotował zmianę, ale zachował stabilność. Handlowiec nadal jest postrzegany jako ta sama, złożona jednostka, a nie jako ktoś zupełnie inny.

Implikacje dla routingu i predykcji

Zastosowanie podejścia probabilistycznego zmienia sposób budowania macierzy decyzyjnej. Zamiast prostej tabeli „Jeśli Klaster A to Klient X", budujemy zaawansowaną funkcję ważącą.

Prawdopodobieństwo sukcesu sprzedaży dla danego handlowca obliczamy jako sumę iloczynów jego przynależności do klastrów i skuteczności tych klastrów. Jeśli handlowiec jest „mieszanką", jego przewidywana skuteczność będzie średnią ważoną skuteczności składowych archetypów. Dzięki temu system routingu staje się odporny na szum. Drobne wahania nastroju handlowca nie powodują rewolucji w przydzielaniu mu klientów. Zyskujemy system, który ewoluuje płynnie, a nie skokowo. Jest to cecha krytyczna dla systemów sterowania procesami przemysłowymi i biznesowymi.

Analiza komparatywna: zestawienie parametrów

Poniższa analiza punktuje kluczowe różnice techniczne i biznesowe między omawianymi metodami.

Granice decyzyjne

- **K-means** – tworzy granice liniowe (w przypadku metryki euklidesowej) i ostre. Obiekt leżący mikrometr od granicy jest traktowany tak samo, jak obiekt leżący w centrum i zupełnie inaczej niż obiekt leżący mikrometr po drugiej stronie granicy.
- **GMM** – tworzy granice płynne. Prawdopodobieństwo przynależności zmienia się w sposób ciągły (gradientowy). Model rozumie pojęcie „niepewności" przypisania.

Kształt klastrów

- **K-means** – zakłada sferyczność klastrów (koła w 2D, kule w 3D). Oznacza to, że wariancja cech we wszystkich kierunkach musi być podobna. W danych sprzedażowych jest to rzadkie, często jedna cecha (np. „techniczność") ma znacznie większy rozrzut niż inna (np. „uprzejmość"). K-means radzi sobie z tym słabo, często sztucznie dzieląc naturalne, wydłużone grupy.
- **GMM** – dzięki macierzy kowariancji, klastry mogą przyjmować kształt elipsoid o dowolnej orientacji i spłaszczeniu. Model może wykryć, że grupa „Handlowców Technicznych" jest bardzo zróżnicowana pod względem „Ekstrawersji", ale bardzo spójna pod względem „Skrupulatności".

Wymagania danych i koszt obliczeniowy

- **K-means** jest algorytmem niezwykle szybkim i zbieżnym. Wymaga mniej danych do uzyskania wyniku, choć wynik ten może być obarczony błędem uproszczenia (bias).
- **GMM** jest algorytmem bardziej złożonym obliczeniowo. Estymacja parametrów (szczególnie pełnej macierzy kowariancji) wymaga większej liczby próbek danych, aby uniknąć problemu osobliwości lub nadmiernego dopasowania (*overfitting*). Wymaga bardziej zaawansowanej wiedzy inżynierskiej do kalibracji (np. kryteria AIC/BIC do doboru liczby komponentów).

Asymetria wolumenu danych: wpływ ekstremalnej dysproporcji (500+ vs. 1)

W praktyce operacyjnej mierzymy się z fundamentalną asymetrią, która stawia pod znakiem zapytania zasadność stosowania jakiejkolwiek klasycznej metody klasteryzacji. Mamy do czynienia z dwoma populacjami o skrajnie różnej gęstości informacji:

- **handlowcy** – populacja stabilna, gdzie każdy obiekt opisany jest przez **500+ rozmów**. Prawo wielkich liczb działa tu na naszą korzyść, wariancja średniej jest niska;
- **klienci** – populacja wysoce dynamiczna, gdzie **90% obiektów posiada tylko 1 rozmowę**, a reszta zazwyczaj nie przekracza 3.

Ta dysproporcja rodzi uzasadnioną wątpliwość: Czy profilowanie psychologiczne na podstawie jednej rozmowy jest w ogóle metodologicznie poprawne? Czy próba przypisania klienta do klastra po 15 minutach kontaktu nie jest wróżeniem z fusów?

Wada metody K-means: halucynacja pewności

Algorytm K-means jest „ślepy" na liczebność próby. Traktuje punkt wyznaczony z 1 rozmowy tak samo poważnie, jak punkt wyznaczony z 500 rozmów. Jeśli w tej jednej rozmowie klient był zdenerwowany awarią, K-means bez wahania przypisze go do klastra „Klient Roszczeniowy" (100% pewności). Jest to **fałszywa pewność**. System buduje profil na szumie, a nie na sygnale. Prowadzi to do błędnych decyzji routingowych, które mogą być gorsze niż losowy przydział.

Rola GMM: detekcja braku wiedzy (*safety brake*)

Metoda GMM również nie wyczaruje informacji, której nie ma. Przy próbie N=1 nie jesteśmy w stanie policzyć wariancji dla konkretnego klienta. Jednakże, GMM pozwala nam modelować tę sytuację poprzez **priory (rozkłady a priori)**.

Dla klienta z 1 rozmową, model probabilistyczny zwróci wynik o maksymalnej entropii (niepewności). Zamiast „Klient

typu A", otrzymamy wynik: 25% typ A, 25% typ B, 25% typ C, 25% typ D. Jest to matematyczny odpowiednik stwierdzenia: **„Nie mam pojęcia, kim jest ten facet"**.

Wniosek krytyczny. W warunkach, gdy 90% klientów to „Cold Start", GMM pełni rolę bezpiecznika. Nie służy do precyzyjnego celowania (bo nie ma danych), ale do powstrzymania systemu przed podejmowaniem głupich decyzji opartych na złudzeniach K-means. Sugeruje to, że dla tej grupy 90% klientów routing powinien opierać się na regułach uniwersalnych (bezpiecznych), a zaawansowana klasteryzacja powinna być aktywowana dopiero po przekroczeniu progu np. 3-4 rozmów.

W sytuacji „zimnego startu" trzeba wybierać GMM, ale stosować hybrydy.

Ze względu na specyfikę „zimnego startu" (mało danych o klientach), K-means jest niebezpieczne, bo „kłamie", że zna profil klienta. GMM jest bezpieczniejsze, bo „przyznaje się", że nie wie.

Prosta strategia wdrożenia:

- Dla klientów nowych (1-3 rozmowy):
  - metoda: brak zaawansowanej klasteryzacji;
  - akcja: routing do „Handlowca Uniwersalnego" (bezpiecznik). Nie zgaduj profilu na siłę;
  - dlaczego: jedna rozmowa to za mało na rzetelny profil.
- dla klientów powracających (4+ rozmowy) i wszystkich handlowców:
  - metoda: GMM;
  - akcja: precyzyjny routing oparty na prawdopodobieństwie sprzedaży;
  - dlaczego: masz już dość danych, by statystyka zadziałała. GMM najlepiej poradzi sobie z niejednoznacznymi profilami („trochę techniczny, trochę negocjator"), które dominują w rzeczywistości.

W skrócie: K-means odrzuć całkowicie (zbyt ryzykowne). GMM wdróż jako silnik docelowy, ale aktywuj go dopiero, gdy masz minimum danych.

To jest kluczowy wniosek: dla Handlowców (którzy mają historię setek rozmów) K-means jest metodologicznie gorsze i powinno zostać całkowicie zastąpione przez GMM. Prawo wielkich liczb działa tu na naszą korzyść, więc model probabilistyczny jest stabilny i precyzyjny.

Studium przypadku: symulacja stabilności macierzy

Rozważmy hipotetyczną sytuację w firmie produkującej maszyny pakujące. Mamy grupę 50 handlowców.

Scenariusz A (K-means). W tygodniu 1 algorytm identyfikuje grupę „Domykaczy Transakcji" liczącą 10 osób. Ich konwersja wynosi 15%. W tygodniu 2, na skutek drobnych zmian w rozmowach, 3 handlowców „granicznych" wypada z tej grupy, a 2 innych do niej wpada. Grupa liczy teraz 9 osób, ale jej skład osobowy zmienił się w 50% (jeśli patrzymy na małą podgrupę). Konwersja grupy spada nagle do 12%. **Wniosek Zarządu:** „Grupa Domykaczy przestała być skuteczna. Musimy zmienić skrypty rozmów". **Błąd:** Skuteczność grupy nie spadła. To algorytm zmienił definicję grupy, mieszając ludzi skutecznych z nieskutecznymi. Decyzja zarządcza jest błędna.

Scenariusz B (GMM). W tygodniu 1 handlowcy mają przypisane wagi „Domykacza". Suma wag wynosi 10,0 jednostek. W tygodniu 2 wagi handlowców zmieniają się nieznacznie (np. z 0,8 na 0,75). Suma wag wynosi 9,8 jednostek. Średnia ważona konwersja zmienia się z 15% na 14,8%. **Wniosek Zarządu:** „Sytuacja jest stabilna, brak anomalii". **Decyzja:** System kontynuuje pracę, delikatnie korygując routing.

W tym przykładzie widać wyraźnie, że metoda probabilistyczna działa jak amortyzator w zawieszeniu samochodu – wygładza nierówności drogi (danych), zapewniając płynną jazdę (decyzje biznesowe), podczas gdy metoda sztywna przenosi każde drganie na karoserię.

Wnioski i rekomendacje wdrożeniowe

Wybór między K-means a GMM nie jest wyborem czysto akademickim. Jest to decyzja o architekturze ryzyka w systemie sprzedaży.

Dla Data Scientista wniosek jest następujący: **K-means jest narzędziem eksploracyjnym, ale GMM jest narzędziem operacyjnym.** K-means świetnie nadaje się do szybkiego podglądu danych („zobaczmy, jakie mamy z grubsza grupy"), ale nie nadaje się do sterowania automatycznym procesem, w którym stabilność jest krytyczna.

Rekomendujemy wdrożenie Modelu Mieszanin Gaussowskich (GMM) jako silnika segmentacyjnego dla systemu *smart routing* w firmie maszynowej. Mimo wyższego progu wejścia (trudniejsza implementacja, większe wymagania obliczeniowe), metoda ta oferuje:

1. **wierność odwzorowania** – lepiej oddaje rzeczywistą, hybrydową naturę kompetencji handlowych,
2. **stabilność temporalną** – jest odporna na szum i drobne fluktuacje w zachowaniu pracowników,
3. **precyzję predykcyjną** – pozwala na budowanie bardziej zniuansowanych modeli scoringowych, wykorzystujących pełną informację o profilu, a nie tylko uproszczoną etykietę.

Z powodu małej ilości rozmów klientów takie przyporządkowanie GMM można zastosować wyłącznie dla handlowców.

W świecie maszyn precyzyjnych nie używamy młotka tam, gdzie potrzebny jest skalpel. W analityce sprzedaży nie powinniśmy używać twardej klasteryzacji tam, gdzie mamy do czynienia z delikatną materią ludzkich zachowań i relacji. Przejście na probabilistykę jest naturalnym krokiem ewolucyjnym dla każdej organizacji dążącej do doskonałości operacyjnej opartej na danych (*Data-Driven Excellence*).

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.