Zaawansowana analityka biznesowa – siła modeli predykcyjnych 226160-1380

Adrianna Wołowiec
e-mail: *awolow1@sgh.waw.pl*

Podstawy teoretyczne segmentacji i modeli wartości w czasie

Co to wartość klienta w czasie?

Wartość klienta w czasie (ang. Customer Lifetime Value - CLV) to zdyskontowana suma przyszłych wpływów, które można przypisać do relacji z klientem. Pozwala oszacować zysk, jaki firma osiągnie w przyszłości dzięki klientowi. Maksymalizacja CLV jest jednym z głównych celów organizacji nastawionych na osiąganie zysku.

$$CLV = \sum_{t=1}^{\infty} \frac{E(V_t)}{(1+d)^{t-1}}$$

V_t - wpływ pieniężny netto od klienta w okresie t

d – stopa dyskontowa

Wartość klienta w czasie

Jednym z zastosowań CLV jest wspieranie decyzji dotyczących ustalania wydatków na przyciąganie nowych klientów. Koszty akwizycji są uzasadnione, jeśli wynoszą one mniej niż CLV klienta. Celem jest identyfikacja potencjalnych klientów z wysoką wartością CLV i unikanie tych o niskiej wartości w czasie.

(np. dostawcy mediów, operatorzy telekomunikacyjni)

Wartość klienta w czasie stanowi najlepsze podstawy do wyznaczania zasobów marketingowych: organizacje powinny inwestować zasoby przeznaczone na marketing jedynie w te działania, które zwiększają CLV, tak, aby CLV było wyższe niż koszty.

3 sposoby na zwiększenie CLV obecnych klientów:

- Dłuższe ich utrzymanie
- Zwiększenie wpływów od klientów
- Zmniejszenie kosztów obsługi klientów, marketingu lub obu

Wysokość środków ponoszonych na te taktyki jest wyznaczana przez zmianę CLV, którą wywołają.

Modelowanie CLV

Modelowanie CLV stosowane jest szczególnie w branżach, w których występuje kontraktowa relacja z klientem, tzn. kiedy można określić dokładną datę końca relacji firma-klient.

Przykłady:

- Klienci operatora komórkowego kończą obowiązującą ich umowę i przestają opłacać rachunki
- Subskrypcje treści medialnych kończą się, kiedy użytkownicy rezygnują z nich lub nie decydują się na ich odnowienie
- Członkostwo w programie zdrowotno-sportowym wygasa w określonym terminie

Rodzaje modeli CLV

Wyróżnia się dwa typy modeli CLV:

1. Modele typu „gone-for-good”
   - zakładają, że klienci, którzy zrezygnowali z usługi, już nie wrócą
   - w tym przypadku najważniejsze jest zatrzymanie klienta przez jak najdłuższy czas
   - do analizy czasu do odejścia klienta stosuje się modele przeżycia
   - przykłady: prosty i ogólny model retencji
2. Modele typu „always-a-share”
   - nie zakładają, że brak aktywności ze strony klienta oznacza jego stałe odejście
   - klient, który nie dokonał zakupu w bieżącym miesiącu, może wrócić w następnym
   - przykłady: model migracji i podejście data mining do wartości w czasie

Modele segmentacyjne

Heterogeniczność klientów

W większości przypadków klienci mają różne pragnienia, potrzeby, preferencje itd. W przypadku występowania takiej heterogeniczności firmy mogą osiągnąć przewagę konkurencyjną poprzez jej rozpoznanie i dostosowanie się do niej. Nie wszystkie potrzeby heterogenicznych klientów mogą zostać zaspokojone poprzez jedną ofertę. Konkurencja może zaoferować lepiej stargetowany/dostosowany produkt i w ten sposób przyciągnąć klientów.

Jednym z podjeść będących odpowiedzią na heterogeniczność jest segmentacja. Klienci z podobnymi pragnieniami i potrzebami są grupowani w segmenty tak, aby firma mogła lepiej zaspokoić różne potrzeby. W tym celu wdrażane są strategie segmentacji klientów oraz metody klastrujące.

Zastosowania biznesowe

1. **Segmentacja rynku.**

W pierwszej kolejności segmentacji na homogeniczne grupy poddawany jest cały rynek. Firma zwykle celuje w jeden segment i rozwija produkt lub usługę dla tego konkretnego segmentu.

2. **Customizacja i personalizacja w obrębie podsegmentów.**

W obrębie jednego segmentu rynkowego nadal będzie występować pewien poziom heterogeniczności. Customizacja polega na tym, że firma pozwala swoim klientom dostosować oferowany produkt według ich uznania, aby w jak największym stopniu zaspokoić ich różnorodne potrzeby. Firma umożliwia dokonanie customizacji produktu poprzez udostępnienie wielu opcji, na które klient może się zdecydować.

Wprowadzenie do modeli segmentacyjnych

Mamy losową próbę n jednostek i zakładamy, że każda jednostka należy dokładnie do jednej z K niezaobserwowanych grup, które mają etykiety $1, \dots, K$. Wartość K jest ustalana jeszcze przed estymacją modelu. Niezaobserwowane tzn. że nie jest znane faktyczne przyporządkowanie jednostek do grup. Zadaniem modelu jest znalezienie tych grup i oszacowanie, jakie jest prawdopodobieństwo, że każda z jednostek pochodzi z poszczególnych grup. Grupy te nazywane są klastrami, typami, segmentami lub podsegmentami. Niech $g_i \epsilon \{1, \dots, K\}$ będzie prawdziwym przyporządkowaniem jednostki $i = 1, \dots, n$ do grupy. Zamiast zaobserwować przynależność do grupy, dokonano pomiaru p zmiennych, które wskazują na przyporządkowanie do segmentu. Niech x_{ij} będzie zaobserwowaną wartością zmiennej j dla jednostki i.

Zadanie 1: Czytelnictwo prasy

Dysponujemy losową próbą $n = 2\,939$ mieszkańców Chicago i tabelą krzyżową przedstawiającą rozkład łącznych ich dwóch cech – ilości czasu spędzanego na czytaniu gazet w ciągu tygodnia oraz liczbą czytanych sekcji w prasie.

Time	Number of Sections								Total
	0	1	2	3	4	5	6	**7**	
0	370	0	0	0	0	0	0	0	370
1	9	127	119	55	34	55	34	21	454
2	3	72	94	68	48	74	48	29	436
3	1	43	107	48	40	63	52	38	392
4	0	21	58	39	32	47	30	25	252
5	0	15	40	23	25	51	38	32	224
6	0	21	67	54	39	90	59	71	401
7	0	15	47	31	51	103	67	96	410
Total	383	314	532	318	269	483	328	312	2 939

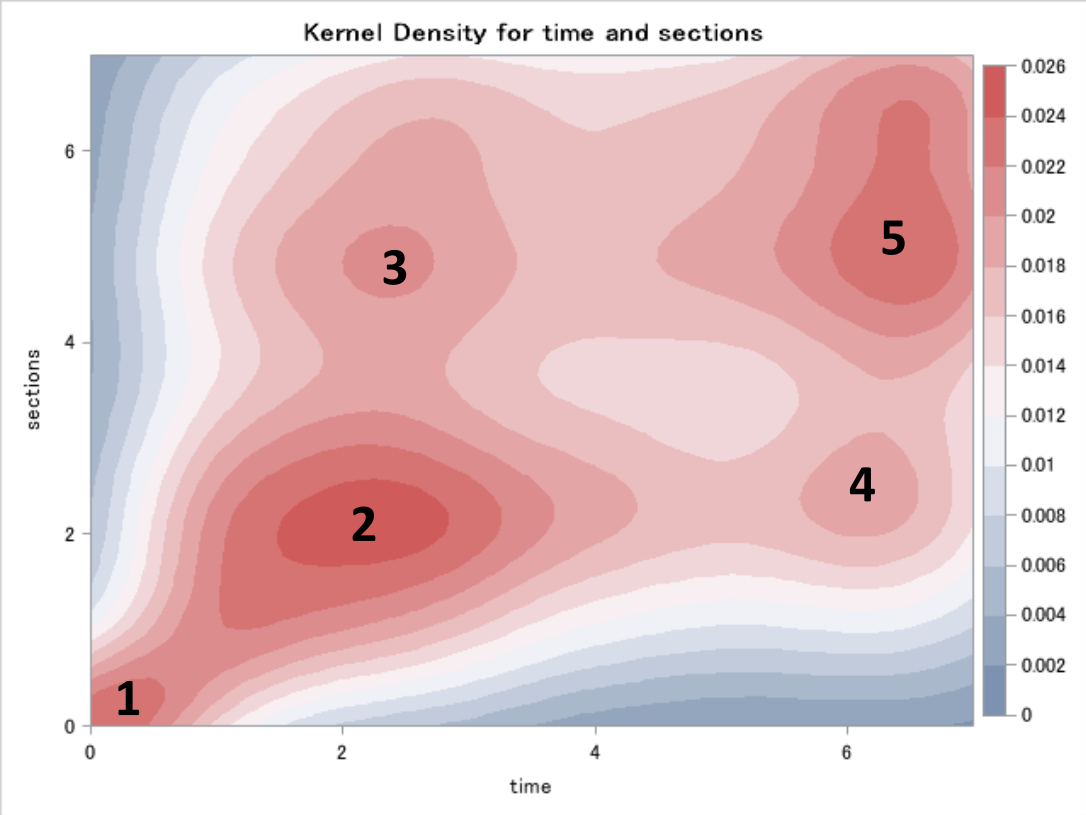
Czy można wyróżnić typy czytelników prasy?

Rozwiązanie

W przypadku występowania tylko jednej lub dwóch zmiennych wskazujących na przynależność do grup, nie jest konieczne stosowanie analizy klastrów czy innych modeli grupujących do identyfikacji segmentów. Wystarczy analiza rozkładów częstotliwości, histogramów, tabeli krzyżowych lub innych statystyk opisowych do identyfikacji grup naturalnych.

Jeśli występują więcej niż dwie zmienne, metody wizualne są w mniejszym stopniu możliwe.

Rozwiązanie – estymator jądrowy gęstości



Wykres estymatora jądrowego gęstości stanowi dwuwymiarowy histogram, pokazujący łączny rozkład czasu czytania i liczby sekcji, przy czym obszary z wyższym prawdopodobieństwem są oznaczone ciemnym czerwonym kolorem, a z małym – kolorem niebieskim.

Segmenty:

1 – Osoby nieczytające (nie spędzają w ogóle czasu na czytaniu pracy i nie czytają żadnych sekcji)

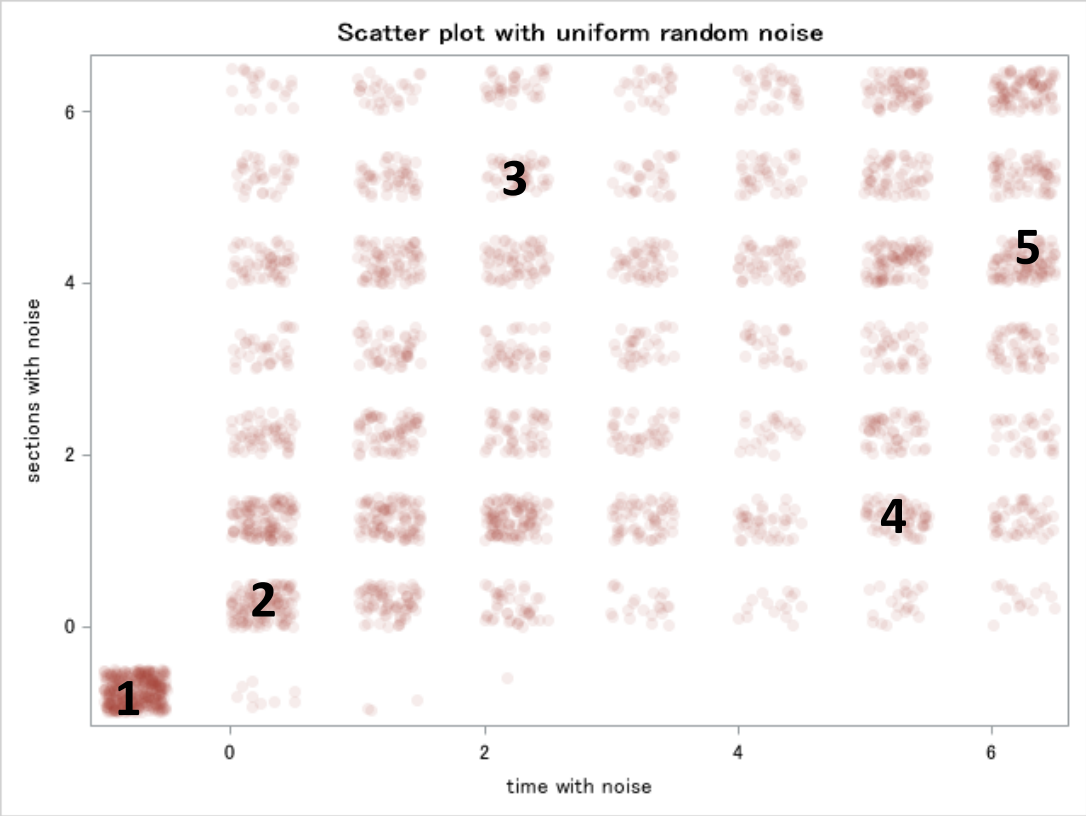
2 – Amatorzy czytania (spędzają mało czasu na czytaniu pracy i czytają mało sekcji)

3 – Przeglądający (przeglądają wiele sekcji w prasie w ciągu małej ilości czasu)

4 - Czytelnicy wybiórczy (spędzają dużo czasu na czytaniu, ale wybierają nieliczne sekcje)

5 – Zaawansowani czytelnicy (spędzają dużo czasu na czytaniu pracy i czytają większość sekcji)

Rozwiązanie – wykres punktowy z elementem losowości



Wykres punktowy z elementami losowości dodaje małą ilość jednakowego losowego szumu poprzez funkcję RANUNI. W przeciwnym przypadku widoczny byłby tylko jeden punkt dla każdej kombinacji wartości dwóch zmiennych. Ciemniejsze obszary wskazują na większą liczebność..

Segmenty:

1 – Osoby nieczytające (nie spędzają w ogóle czasu na czytaniu pracy i nie czytają żadnych sekcji)

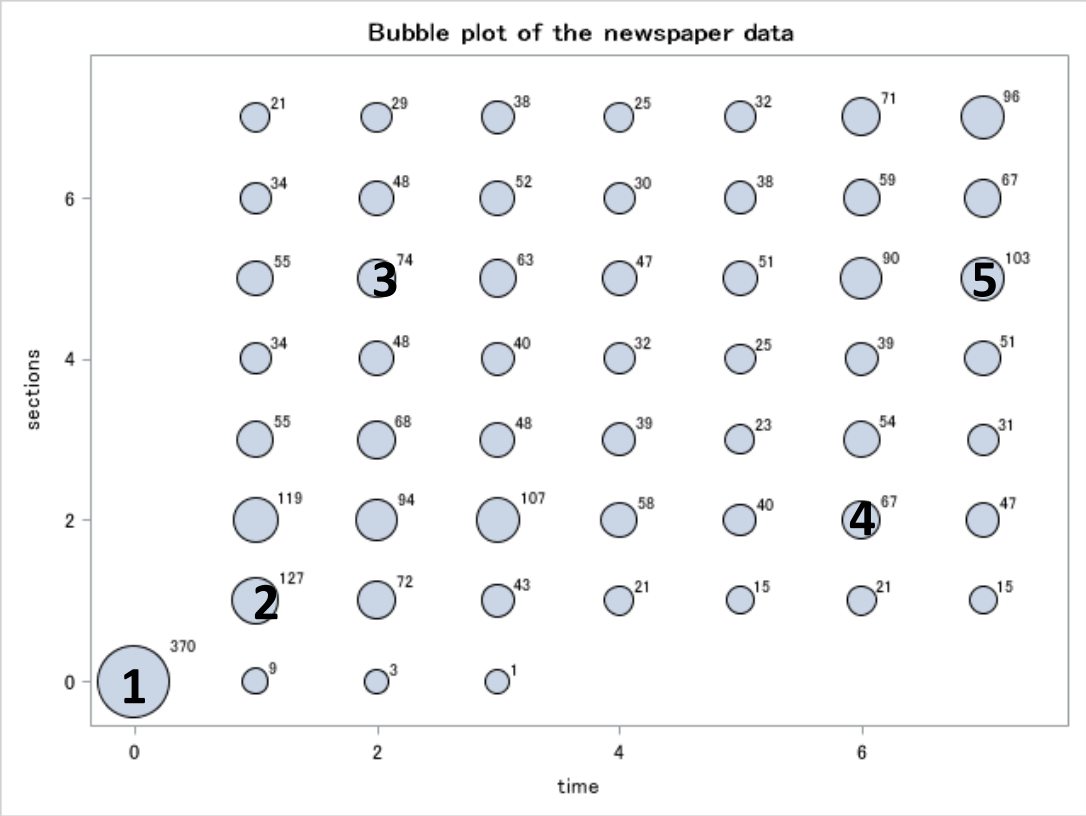
2 – Amatorzy czytania (spędzają mało czasu na czytaniu pracy i czytają mało sekcji)

3 – Przeglądający (przeglądają wiele sekcji w prasie w ciągu małej ilości czasu)

4 - Czytelnicy wybiórczy (spędzają dużo czasu na czytaniu, ale wybierają nieliczne sekcje)

5 – Zaawansowani czytelnicy (spędzają dużo czasu na czytaniu pracy i czytają większość sekcji)

Rozwiązanie – wykres bąbelkowy



Pole bąbelków wskazuje na wysokość liczebności.

Segmenty:

1 – Osoby nieczytające (nie spędzają w ogóle czasu na czytaniu pracy i nie czytają żadnych sekcji)

2 – Amatorzy czytania (spędzają mało czasu na czytaniu pracy i czytają mało sekcji)

3 – Przeglądający (przeglądają wiele sekcji w prasie w ciągu małej ilości czasu)

4 - Czytelnicy wybiórczy (spędzają dużo czasu na czytaniu, ale wybierają nieliczne sekcje)

5 – Zaawansowani czytelnicy (spędzają dużo czasu na czytaniu pracy i czytają większość sekcji)

Pytania poruszane przez modele segmentacji

1. Ile segmentów występuje? W praktyce to analityk musi zadecydować, jaka liczba segmentów będzie odpowiednia.
2. Dla określonej liczby segmentów K i zbioru zmiennych, jakie naturalne grupy można wyróżnić? Jak można je scharakteryzować (np. średnie wartości cech dla obserwacji należących do segmentu)
3. Jak duży jest każdy segment? Jaki jest jego udział w populacji?
4. Do którego segmentu należy każda z jednostek, biorąc pod uwagę zebrane informacje na jej temat?
5. Jakie działania powinny zostać podjęte, w oparciu o wyróżnione segmenty? (np. akcje marketingowe, kontakt z klientem, spersonalizowane oferty)

Dwie klasy modeli segmentacyjnych

Oparte na metodzie k-średnich

- szczególny przypadek bardziej ogólnych modeli opartych na modelach mieszanych
- relatywnie łatwe do estymacji
- dostępne w ramach większości pakietów statystycznych
- łatwa transformacja danych przed estymacją

Oparte na modelach mieszanych

- alternatywne podejście do segmentacji
- zapewniają bardziej zaawansowaną estymację i bardziej dokładną predykcję

Metoda K-średnich

Model zakłada następująca hipotezę:

$$x_{ij} = \mu_{jg_i} + e_{ij},$$

gdzie:

- x_{ij} - zaobserwowana wartość zmiennej *j* dla jednostki *i*
- μ_{jg}- rzeczywista średnia zmiennej j dla wszystkich członków klastra *g*
- e_{ij}- składnik losowy ze średnią E(e_{ij})=0 i wariancją σ^2, który jest wspólny dla wszystkich zmiennych i klastrów; zakładamy, że błąd ma rozkład normalny

Metoda K-średnich - następstwa

Następstwa wspólnej dla wszystkich klastrów wariancji błędu (założenie modelu):

- rozkład punktów należących do jednego klastra jest okrągły lub sferyczny
- wszystkie klastry mają taki sam kształt i dyspersję

Jeśli zmienne nie wpadają w okrągłe obszary równej wielkości, w metodzie K-średnich może wystąpić błąd (np. obszary mają eliptyczny kształt). Podejścia oparte na modelach mieszanych dopuszczają inne kształty i zmienne wielkości klastrów. Innym sposobem na doprowadzenie klastrów do postaci bardziej sferycznej jest transformacja danych przed estymacją modelu. Nie rozwiązuje to jednak kwestii równych wielkości.

Metoda K-średnich a zmienne kategoryzujące

Metoda K-średnich daje lepsze wyniki dla zmiennych ciągłych. Stosuje się dwie metody, które służą wprowadzeniu do analizy K-średnich zmiennych kategoryzujących:

- partycjonowanie – poza modelem – obserwacji z wykorzystaniem ważnych zmiennych kategoryzujących i zastosowanie modelu na każdej z partycji
- utworzenie sztucznych zmiennych zerojedynkowych i zastosowanie metod ważenia zmiennych

Algorytm K-średnich

1. **Inicjalizacja.** Wybierz początkowy zbiór środków klastrów $\hat{\mu}_{jk}^{0}$, gdzie $k = 1,\ldots,K$, a oznaczenie górne 0 wskazuje numer iteracji. Zapoczątkuj licznik pętli $h = 1$.
2. **Przypisanie do klastrów.** Przypisz obserwację *i* do klastra, którego średnia jest najbliższa obserwacji. Odległość Euklidesa pomiędzy punktem x_i a środkiem klastra $\hat{\mu}_{k}^{h-1}$ podana jest wzorem $d_{ik}^{h} = \sqrt{\sum_{j=1}^{p}\left(x_{ij} + \hat{\mu}_{jk}^{h-1}\right)^2}$. Obserwacja i jest przypisana do klastra, który jest jej najbliższy, tzn. $g_i^h = argmin_k d_{ik}^h$. $argmin_k$ wskazuje, że g_i^h jest równe k minimalizując wartość d_{ik}^h.
3. **Wyliczenie średnich dla klastrów.** Przy stałym przypisaniu do klastrów, wylicz ich średnie. Są one nazywane również centrami lub centroidami klastrów $\hat{\mu}_{jk}^{h} = average\{x_{ij}: g_i^h = k\}$. Innymi słowy, nowe oszacowanie klastrowej średniej k jest prostą średnią wszystkich obserwacji przypisanych do danego klastra ($g_i^h = k$).
4. **Obliczenie SSE.** Jest to suma kwadratów błędów (ang. sum of squared errors, SSE) lub całkowita wariancja wewnątrzklastrowa. $SSE = \sum_{i=1}^{n}\sum_{j=1}^{p}\left(x_{ij} - \hat{\mu}_{jg_i^h}^{h}\right)^2 = \sum_{i=1}^{n} d_{ig_i^h}^2$. Błędy są w rzeczywistości odległościami pomiędzy każdą obserwacją a najbliższym jej środkiem klastra.
5. **Pętla.** Niech $h = h + 1$ i powróć do kroku 2, aż spełnione zostanie kryterium zbieżności lub przekroczona zostanie maksymalna dopuszczalna liczba iteracji.

Algorytm K-średnich

Algorytm dąży do minimalizacji wartości SSE, co stanowi kombinatoryczny problem optymalizacyjny i jest bardzo trudne pod względem obliczeniowym. W szczególności, nie ma pewności czy algorytm będzie zbiegał do rozwiązania optymalnego, a różne punkty startowe mogą dawać różne rozwiązania. Ze względu na te kwestie, estymacja K-średnich może być generowana wielokrotnie dla różnych punktów startowych, a z wielu uzyskanych rozwiązań wybiera się to, które zapewnia najmniejszą wartość SSE.

Zadanie 2: Czytelnictwo prasy

Znajdź rozwiązanie 5-klastrowe na danych dotyczących czytelnictwa z wykorzystaniem procedury PROC FASTCLUS. Liczba jednostek w próbie dla każdej możliwej kombinacji wartości cech jest podana w zmiennej zliczeniowej.

Przypomnienie:

Dysponujemy losową próbą n=2 939 mieszkańców Chicago i tabelą krzyżową przedstawiającą rozkład łącznych ich dwóch cech – ilości czasu spędzanego na czytaniu gazet w ciągu tygodnia oraz liczbą czytanych sekcji w prasie.

Wzory statystyk

Wewnątrzklastrowe odchylenie std. dla klastra *k* (RMS Std Deviation) = $\sqrt{\frac{1}{p(n_k-1)}}\sum_{g_i=k}\sum_{j=1}^{p}\left(x_{ij}-\hat{\mu}_{jk}\right)^2$,

gdzie $\hat{\mu}_{jk}$ to końcowa średnia zmiennej *j* dla klastra *k*

Całkowite odchylenie std. dla zmiennej *j* (Total STD) = $\sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}\left(x_{ij}-\hat{\mu}_{j.}\right)^2}$,

gdzie $\hat{\mu}_{j.}$ to ogólna średnia zmiennej *j*

Wewnątrzklastrowe odchylenie std. dla zmiennej ***j*** (Within STD) = $\sqrt{\frac{1}{(n-K)}\sum_{i=1}^{n}\left(x_{ij}-\hat{\mu}_{jg_i}\right)^2}$,

gdzie $\hat{\mu}_{jg_i}$ to ogólna średnia zmiennej *j* dla klastra *k*, do którego przypisana została obserwacja *i*

R-kwadrat (R-square) = $1-\left(\frac{Within\ STD}{Total\ STD}\right)^2$

Zadanie 3: Czytelnictwo prasy

Wcześniej zasugerowano, że metoda K-średnich jest wrażliwa na wartości startowe algorytmu. Oszacuj rozwiązanie 5-klastrowe dla 100 losowych zestawów punktów startowych.

Transformacja danych

Wcześniej zostało wspomniane, że przed budową modelu segmentacyjnego należy dokonać transformacji danych. W poprzednim zadaniu zmienne wykorzystywane do budowy klastrów przyjmowały wartości z tego samego zakresu, więc ich transformacja nie była konieczna. W innym przypadku dane przed ich modelowaniem należy poddać wstępnemu przetworzeniu.

Zadanie 4: Nastawienie do teatru

Grupa teatrów chce przyciągnąć osoby, które obecnie nie uczęszczają na sztuki. W tym celu przeprowadzono badanie marketingowe na grupie ok. 3 000 osób. Ankietowanym zostało zadanych 14 pytań mierzących ich nastawienie i przekonania odnośnie teatru z wykorzystaniem skali semantycznej. W wyniku analizy czynnikowej uzyskano 5 wymiarów:

1. **attitude** – średnia z 10 pytań (skala od 0 do 7)
2. **planning** – 1 = pod wpływem chwili; 7 = wymaga planowania
3. **parents** – 1 = rodzice nie lubili teatru; 7 = rodzice lubili teatr
4. **goodval** – 1 = zbyt drogi; 7 = wart sw ceny
5. **getto** – 1 = ciężko dotrzeć; 7 = łatwo dotrzeć

Jakie są typy jednostek po względem ich nastawienia do teatru?

Wczytaj dane w SAS, wystandaryzuj zmienne i wygeneruj rozwiązanie 1-, 2- i 3-klastrowe. Zinterpretuj rozwiązanie trzy klastrowe.

Transformacja zmiennych

Transformacja pojedynczej zmiennej oznacza zastosowanie pewnej funkcji na zmiennej, takiej jak na przykład logarytm, formuła Z-score czy obie. Przetransformowane wartości są później wykorzystywane do analizy klastrów. Metoda K-średnich jest bardzo wrażliwa na skalowanie zmiennych i dlatego ważne jest, aby używać zmiennych o współmiernych jednostkach. Metoda ta daje również nieprawidłowe wyniki w przypadku asymetrycznego rozkładu zmiennej.

Wzór na standaryzację wartości (Z-score):

$$z = \frac{X - \mu}{\sigma},$$

gdzie: X to zaobserwowana wartość, μ to średnia w populacji, σ to odchylenie standardowe.

W sytuacji, gdy zmienne mają współmierne jednostki, a ich rozkłady nie są asymetryczne, nie ma potrzeby, by standaryzować zmienne w jakikolwiek sposób, a do analizy klastrów mogą zostać użyte dane surowe.

PCA – principal components analysis

Innym sposobem na zwizualizowanie rozwiązania klastrowego jest użycie analizy głównych składowych (ang. principal components analysis, PCA). Polega ona na znalezieniu dla danych wielowymiarowych ich reprezentacji w przestrzeni o mniejszej liczbie wymiarów. W zadaniu 4 (Nastawienie do teatru) mamy obserwacje pięciowymiarowe. PCA identyfikuje najlepszą dwuwymiarową reprezentację danych (n=2 w PROC PRINCOMP) i rzutuje punkty pięciowymiarowe na tą płaszczyznę. Pojęcie najlepszą ma dwa znaczenia: (1) płaszczyzna pokazuje tak dużo wariancji oryginalnych danych ile jest możliwe (tzn. wariancja nowych dwuwymiarowych obserwacji jest maksymalizowana) i (2) płaszczyzna minimalizuje ilość utraconej po redukcji wymiarów informacji (tzn. suma kwadratów odchyleń ortogonalnych projekcji na płaszczyźnie od punktów oryginalnych jest minimalizowana). W zadaniu 4 PCA można porównać do rzucania cienia punktów pięciowymiarowych na dwuwymiarową płaszczyznę.

Podsumowanie

Kroki w procesie budowy modelu segmentacyjnego:

1. **Wybór zmiennych**
2. **Transformacja zmiennych** (jeśli potrzebna)
3. **Znalezienie różnych rozwiązań klastrowych i porównanie ich** (np. z wykorzystaniem statystyki pseudo F)
4. **Wybór najlepszego rozwiązania segmentacyjnego i jego podsumowanie**