Podstawy uczenia maszynowego
Restauracja "Pod Złotymi Łukami" – raport

Mateusz Kocot

1 maja 2021

Spis treści

1 Przygotowanie zbioru danych

Wczytano załączony do zadania zbiór danych.

1.1 Niepotrzebne kolumny

Na początku usunięto kolumny tekstowe: „Category" i „Item". Następnie usunięto kolumny redundantne:

- Niektóre wartości wyrażone są osobno jako bezwzględne wartości liczbowe oraz wartości procentowe zalecanego dziennego spożycia (np. „Cholesterol" i „Cholesterol (% Daily Value)". Usunięto kolumny z wartościami procentowymi.
- Kolumny „Calories from Fat" i „Total Fat" także wyrażają tą samą cechę. Usunięto kolumnę „Calories from Fat"

1.2 Formatowanie danych

Po usunięciu niepotrzebnych kolumn, w zbiorze danych pozostały prawie tylko wartości liczbowe. Jedynie kolumna „Serving Size" zawiera napisy, z których można jednakże wyciągnąć informację o wadze. Zmodyfikowano więc tę kolumnę tak, by reprezentowała wagę w gramach. Wiele wartości należało przeliczyć. W dwóch przypadkach, gdy informacja była podana tylko w mililitrach, przyjęto 1 $ml = 1.04\ g$ (przybliżenie dla mleka).

Następnie wszystkie kolumny sformatowano do wartości zmiennopozycyjnych.

1.3 Normalizacja

Metoda K-means wykorzystuje algorytm bazujący na odległości. W związku z tym, by żadna cecha nie przeważała, zbiór warto znormalizować, tj. wycentrować i przetransformować tak, by odchylenie standardowe było równe 1. Tak też zrobiono.

2 K-means

2.1 Miara

Wybrałem indeks Calinskiego-Harabasza. Jest on nazywany kryterium stosunku wariancji, gdyż wyrażony jest przez stosunek B/W, gdzie:

- B jest wariancją centrów klastrów (każdy klaster brany jest z wagą równą jego liczności),
- W jest sumą wariancji wewnątrz-klastrowych.

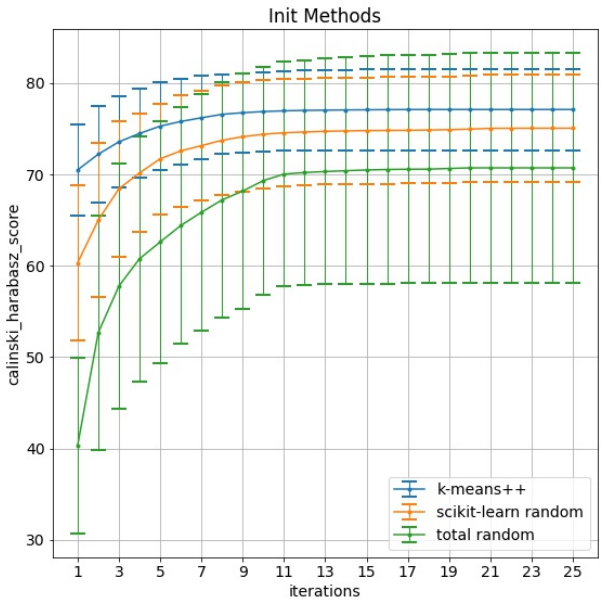
Wartość indeksu Calinskiego-Harabasza jest tym większa, im odległości punktów od centrów swoich klastrów są mniejsze oraz centra klastrów są bardziej rozrzucone i nie są za blisko siebie. W związku z tym, z reguły większa wartość tego indeksu identyfikuje lepsze przypisanie do klastrów.

Miara ta działa dobrze dla klastrów, które są dobrze odseparowane i o podobnej gęstości. Jeżeli jednak klastry są figurami wklęsłymi (np. jeden klaster w kształcie koła w środku drugiego w kształcie okręgu) lub nie mają podobnej liczności, indeks Calinskiego-Harabasza zwróci małą wartość dla dobrej klasteryzacji.

2.2 Metody generowania początkowych centrów klastrów

Na rys. 1 przedstawiono wykres wartości indeksu Calinskiego-Harabasza od iteracji metody k-means dla różnych metod inicjalizacji centrów klastrów. Eksperyment przeprowadzono dla liczby klastrów $k = 6$. Dla każdej liczby iteracji przeprowadzono 25 testów.

Jak widać, najlepiej spisuje się metoda k-means++. Wartość miary jest najlepsza, odchylenia standardowe są najmniejsze oraz po zastosowaniu tej metody, k-means uzyskuje zbieżność po najmniejszej liczbie iteracji. Losowy wybór punktów ze zbioru („scikit-learn random") jest

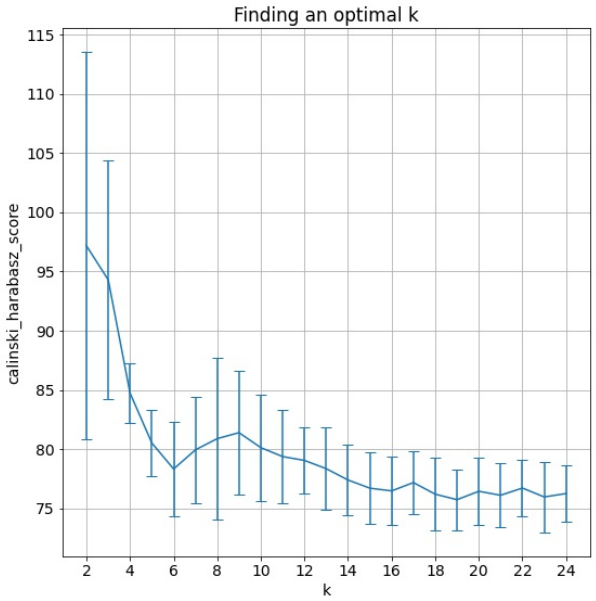


Rys. 1: Porównanie jakości metod inicjalizacji k-means

tylko niewiele gorszy. Najsłabiej wypadł całkowicie losowy wybór początkowych centrów („total random").

Co ciekawe, wybór miary miał duży wpływ na wyniki tego eksperymentu. Wybranie indeksu Daviesa-Bouldina skutkowało uzyskaniem najlepszego rezultatu przez całkowicie losowy wybór.

2.3 Ustalenie odpowiedniej liczby klastrów



Rys. 2: Porównanie jakości klasteryzacji dla różnych wartości k

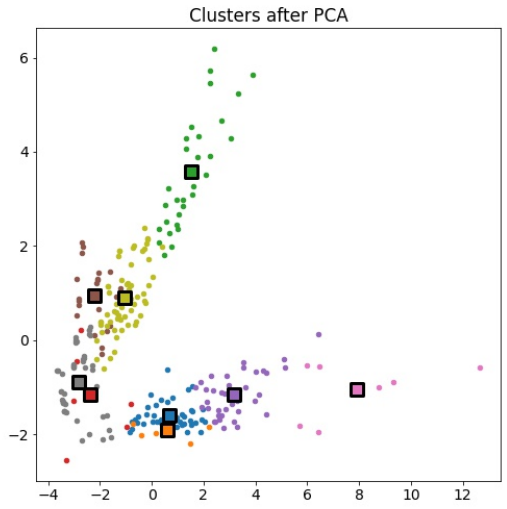
Rys. 2 przedstawia wykres wartości indeksu Calinskiego-Harabasza od liczby klastrów k. Dla każdego k wykonano 50 testów, a w każdym z nich maksymalna liczba iteracji k-means wyniosła 200.

Początkowe duże wartości miary wynikają ze zbyt generalnej klasteryzacji. Warto więc wykorzystać metodę łokcia. „Łokieć" można zaobserwować na wykresie dla $k = 9$ – jest to maksimum lokalne średniej, a wariancja jest stosunkowo nieduża.

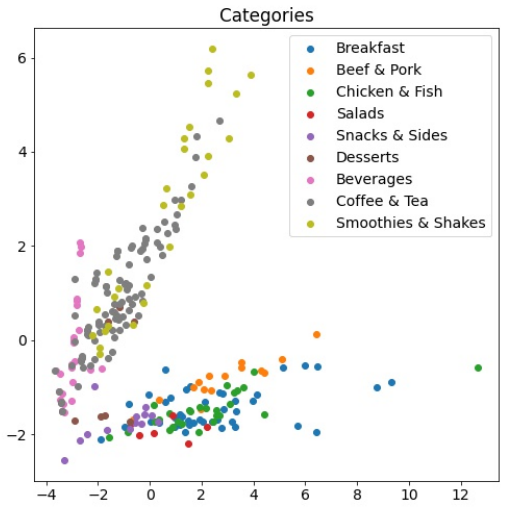
2.4 Próba opisu klasteryzacji

1. Maksimum lokalne na poprzednim wykresie wypadło dla $k = 9$, natomiast wartości dla $k = 8$ oraz $k = 10$ nie są dużo gorsze. Można więc założyć, że liczba klastrów wynosi 8, 9 lub 10.
2. Środki klastrów to uśrednione wartości wszystkich punktów należących do danego klastra. Nie reprezentują one żadnych konkretnych produktów.
3. K-means dąży do tego, by klastry miały podobną liczność albo były mocno odseparowane. Wobec tego, klastry leżące blisko siebie mają podobną liczność, a te bardziej odseparowane – dowolną.
4. Klastry mają sens dla człowieka. W jednym klastrze znajdują się produkty o podobnych cechach, np. rodzaju (płyn, potrawa), kaloryczności, zdrowości, itp. Jednakże, produkty z jednego klastra niekoniecznie należą do tej samej kategorii ze zbioru.

2.5 Wizualizacja przy pomocy PCA



Rys. 3: Rezultat k-means dla $k = 9$



Rys. 4: Domyślnie przypisane kategorie

Na rys. 3 i 4 przedstawiono odpowiednio wizualizację klasteryzacji oraz przypisania produktów do podanych w zbiorze kategorii.

Wyniki są całkiem satysfakcjonujące, jednakże nie oddają dobrze przypisania do kategorii. Produkty z tych samych kategorii często mocno się różnią. Czasami można je przypisać do kilku kategorii. W związku z tym, po zrzutowaniu na dwa wymiary przez PCA, kategorie są mocno przemieszane. Klastry z metody k-means natomiast są bardziej spójne. Reprezentują one po prostu produkty o podobnych cechach.

Początkowe przygotowanie danych ma oczywiście wpływ na uzyskane wyniki. Dzięki wyeliminowaniu cech redundantnych, wyniki klasteryzacji są lepsze. Jednakże, z tak przemieszanymi kategoriami, prawdopodobnie nie da się wybrać cech z tego zbioru w taki sposób, by k-means podzielił produkty na odpowiednie kategorie.