Machine Learning

lab 4

29.04.2021 Mikołaj Zatorski

Normalizacja danych

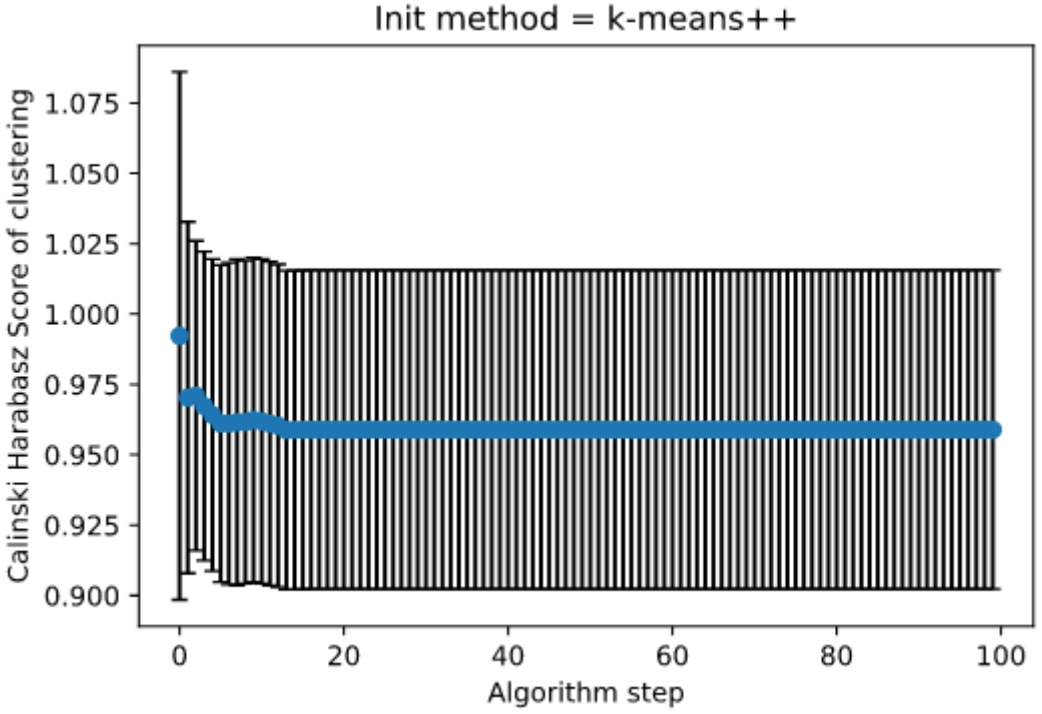
Ze zbioru danych usunąłem nie-liczbowe oraz zbędne wg mnie kolumny. Do dalszej analizy wykorzystałem kolumny *Serving Size, Calories, Total Fat, Carbohydrates, Sugars, Protein*. Cały zbiór danych wycentrowałem oraz znormalizowałem tak, aby *mean = 0 i max - min = 1.0* w obrębie jednej kolumny.

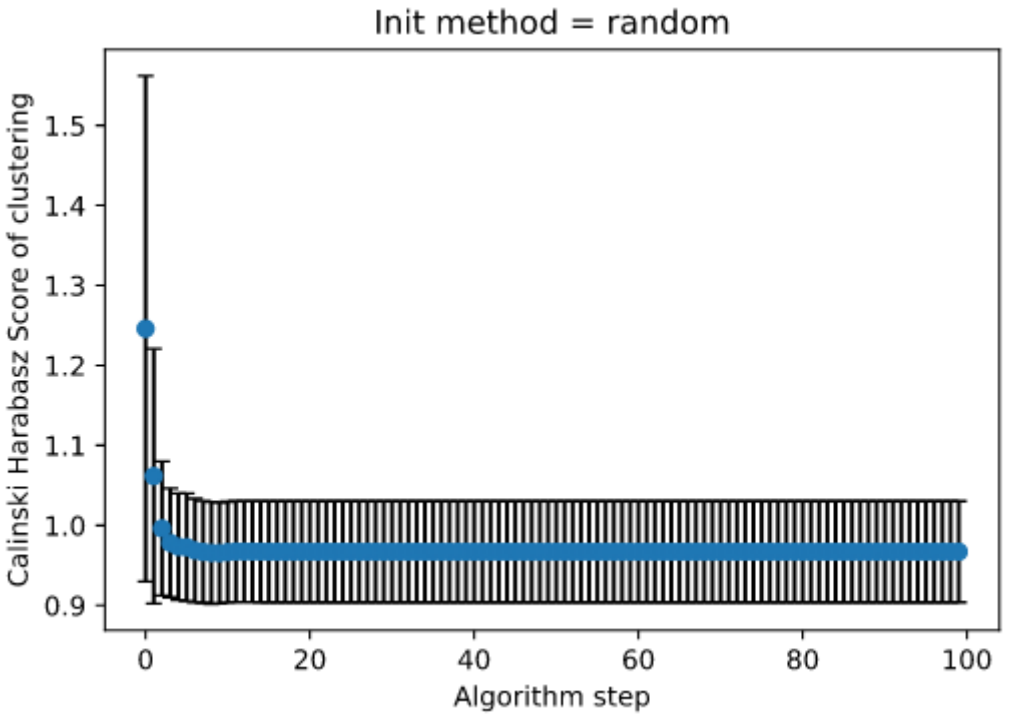
Porównanie technik początkowej generacji środków klastrów

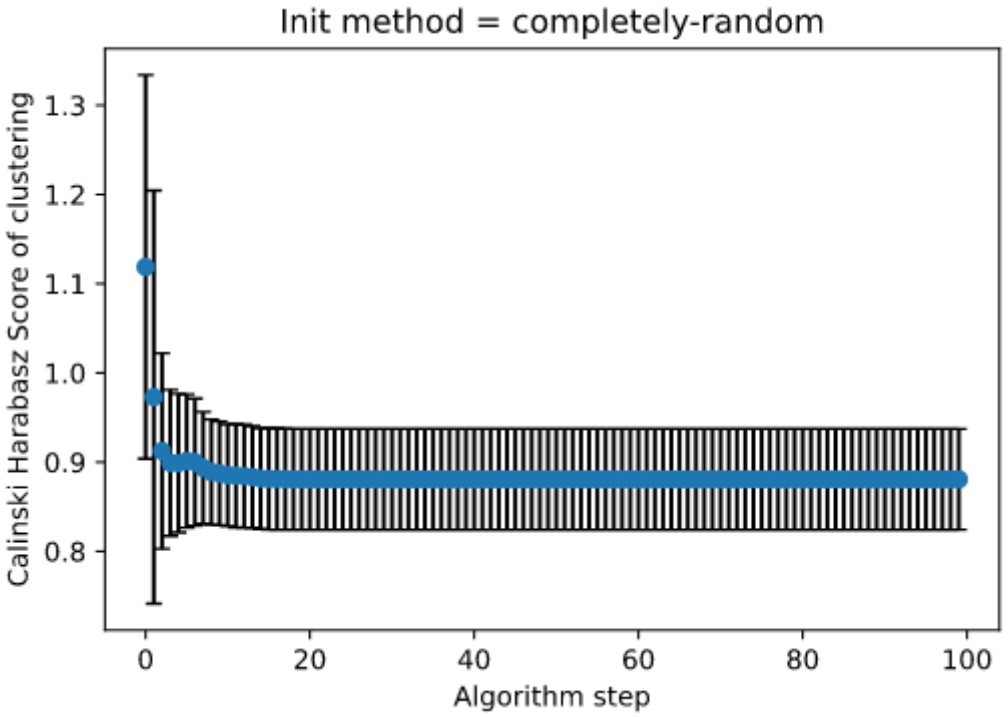
Dla ustalonego *k=5* porównałem techniki k-means++, random oraz własną k_total_random.

Użyta metryka

W celu zmierzenia jakości klasteryzacji użyłem indeksu Daviesa Bouldina. Jest to średnia skala podobieństwa każdego klastra w stosunku do innego, najbardziej podobnego do niego klastra. Im mniejsza jest to wartość, tym lepsza jest klasteryzacja - zatem będziemy dążyli do zminimalizowania tej metryki.



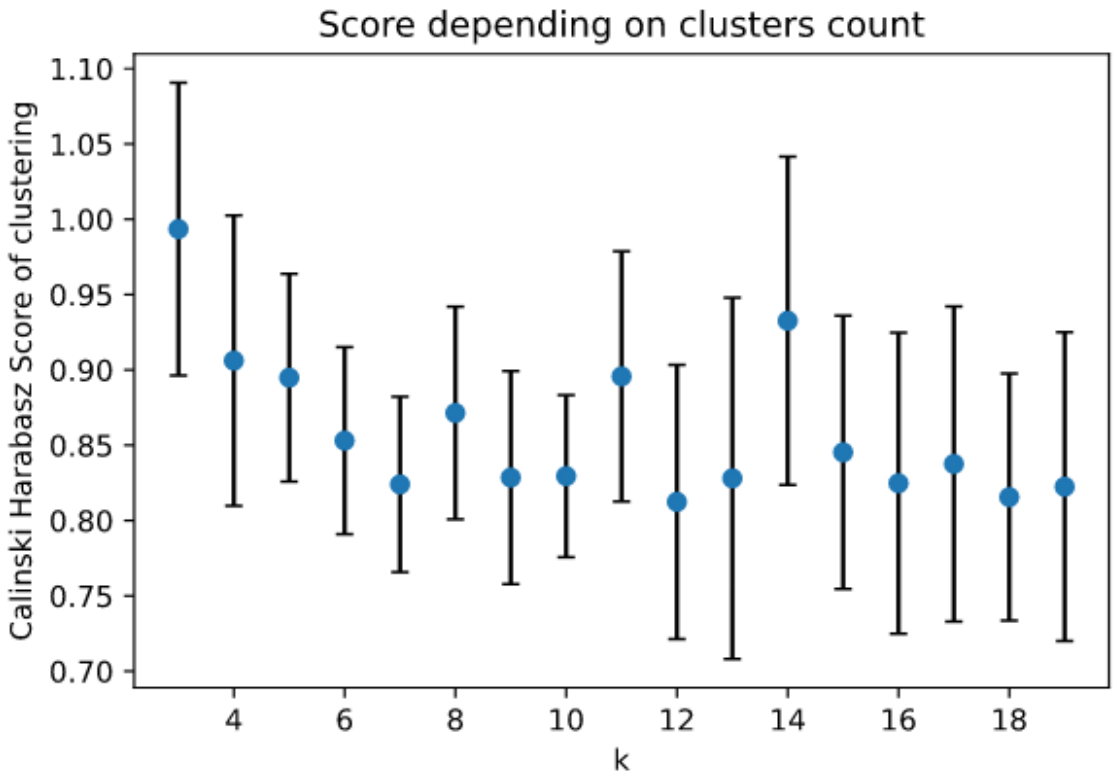




Komentarz:
Najlepszy wynik udało się uzyskać za pomocą metody **completely-random**. W każdym przypadku widać, że już po około 10 krokach algorytmu środki klastra nie zmieniają się. Co więcej odchylenie standardowe dla kolejnych kroków również jest stałe, co sugerowałoby, że w każdym przebiegu algorytmu po szybkim ustabilizowaniu się środków, nie dochodzi już do żadnych zmian.

Ustalanie liczby klastrów

Zgodnie z instrukcją przeprowadziłem badania na temat wpływu liczby klastrów na wynik końcowy metryki:



Komentarz:

Najlepszy wynik udało się uzyskać dla k=12. Dla kilku innych k również występowały bliskie wyniki (np. k=9 - w końcu tyle mamy kategorii dań), jednak doszedłem do wnioski, że może nawet w obrębie jednej kategorii jesteśmy w stanie wyróżnić kilka klastrów “podkategorii”, stąd ostateczny wybór padł na k=12.

Ostateczny podział na klastry

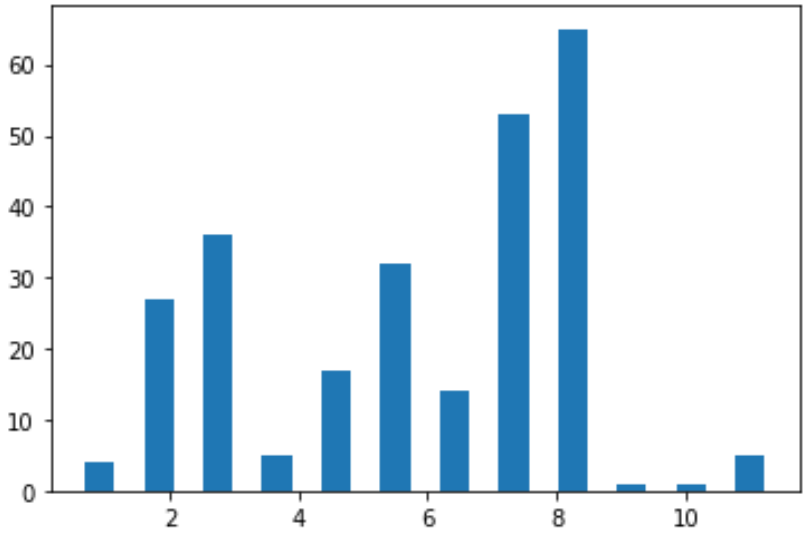
Zgodnie z poleceniem dokonałem ostatecznego podziału zbioru na 12 klastrów.

1. Ile ich jest?
    - 12
2. Gdzie leżą ich środki?

```
Cluster centers: [[ 0.57086849 -0.17461129 -0.10097784 -0.30387343 -0.22596154 -0.13607427]
 [ 0.09105965 -0.15431972 -0.09885919 -0.26197693 -0.17286547 -0.11202148]
 [-0.28854012 -0.07236543 -0.03329893 -0.15316421 -0.18190772 -0.04348168]
 [ 0.05022333  0.35943126  0.3392764   0.38903437 -0.12205529  0.27427056]
 [ 0.06631441  0.13608958  0.06290551  0.34381117  0.41213589 -0.01065169]
 [ 0.25736042 -0.00755677 -0.0685361   0.13207515  0.24693885 -0.03585875]
 [ 0.38653669 -0.07658698 -0.06375023 -0.07590991  0.02682864 -0.09091828]
 [-0.00928976 -0.0534752  -0.06895833 -0.00580037  0.08941287 -0.07090403]
 [-0.16620347  0.08005319  0.10394394 -0.01554828 -0.17001202  0.1382847 ]
 [ 0.32248139  0.80411211  0.87995437  0.50109111 -0.22205529  0.84668435]
 [ 0.10957816  0.29879296  0.15961538  0.65002728  0.77013221  0.07656941]
 [ 0.29667494  0.24028232  0.09012386  0.60321877  0.64200721  0.04208665]]
```

Z powyższych współrzędnych punktów ciężko cokolwiek wywnioskować, postaram się to zwizualizować w następnych punktach.

3. Ile jest obserwacji w każdym klastrze?



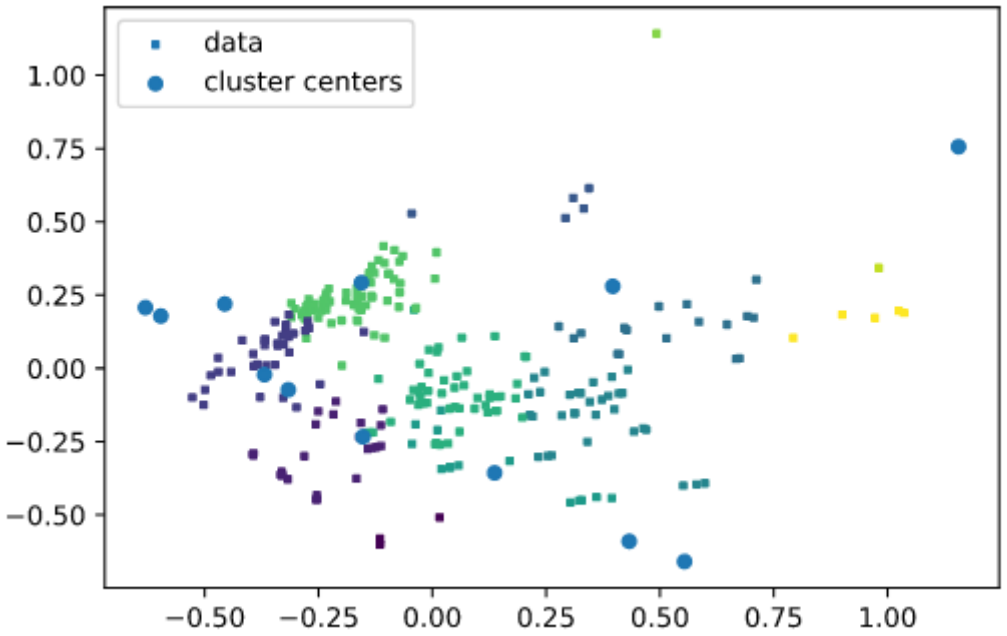
Jak widać liczby te rozkładają się bardzo nierównomiernie.

4. Czy mają sens dla człowieka?

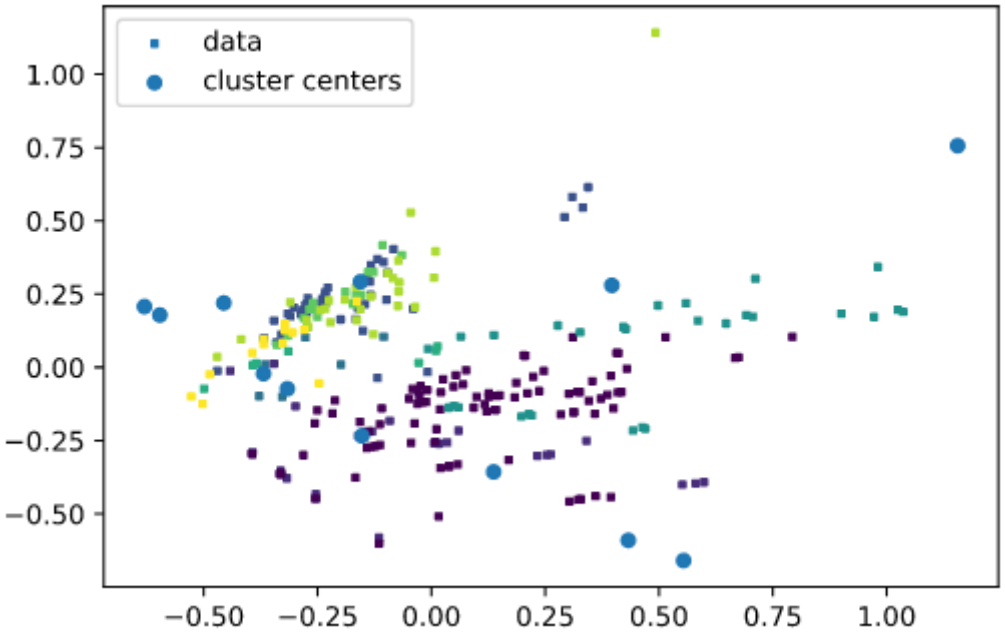
- Patrząc na kategorie przedmiotów, nazwy oraz klastry, można dojść do wniosku, że klastry mniej więcej rozróżniają kategorie przedmiotów. Co więcej, patrząc na np. ostatnie wiersze, desery oznaczone (Small) oraz (Medium) zostały zaklasyfikowane do innego klastra niż oznaczone (Snack). To sugerowałoby, że udało nam się wyróżnić podkategorię w obrębie kategorii deserów.

Wizualizacja rezultatów

Z użyciem metody PCA dokonałem wizualizacji rezultatów zgodnie z wytycznymi. Każdy klaster jest oznaczony innym kolorem, to samo w przypadku kategorii, jednakże kolor klastra != kolor odpowiadającej mu kategorii. Środki klastrów oznaczyłem większymi kołami.



Podział na klastry



Podział na kategorie

Podsumowanie

Widać, że podział na klastry miejscami pokrywa się z podziałem na kategorię. W wielu miejscach jednak się różni - czasem ma to związek z podziałem na podkategorię (przypadki opisane wyżej), czasem jest to po prostu błędna klasyfikacja. Generalnie wyniki są niezłe, biorąc pod uwagę dane na których

operowaliśmy. Ich pre-processing na pewno pomógł w uzyskaniu lepszej klasyfikacji. Wyobrażam sobie, że algorytmy pozwalające dodatkowo na analizę np. składników lub choćby nazw dań dałyby znacząco lepsze efekty.