Laboratorium Metody Sztucznej Inteligencji
6. Metody klasteryzacji

Klasteryzacja
Jej celem jest pogrupowanie wzorców (obserwacji) na możliwie jak najbardziej jednorodne zbiory (klastry). Jest to proces nienadzorowany, więc klastry, ani nawet ich liczba, nie są znane z góry – klasteryzacja bazuje jedynie na podobieństwie/odległości pomiędzy poszczególnymi obserwacjami.
Jest jedną z podstawowych technik analizy eksploracyjnej danych – umożliwia poznanie struktury danych, identyfikację zbliżonych wzorców, może także posłużyć do wykrywania anomalii lub efektu paczki.

Klasteryzacja hierarchiczna
Polega na utworzeniu hierarchii klastrów w oparciu o odległości między nimi. Wyróżnia się dwa podstawowe podejścia:

- Aglomeracyjne (*agglomerative)* – każda z obserwacji traktowana jest na początku jako osobny klaster. W kolejnych krokach najbardziej zbliżone klastry są łączone.
- Deglomeracyjne (*divisive)* – na początku wszystkie obserwacje znajdują się w jednym klastrze. W kolejnych krokach wyodrębniane są coraz bardziej jednorodne klastry.

Wynik klasteryzacji hierarchicznej jest silnie zależny od przyjętej miary odległości (np. odległość euklidesowa, Manhattan, Jaccarda) oraz od połączenia, a dokładniej definicji odległości pomiędzy klastrami (*linkage*). Połączenie może być na przykład:

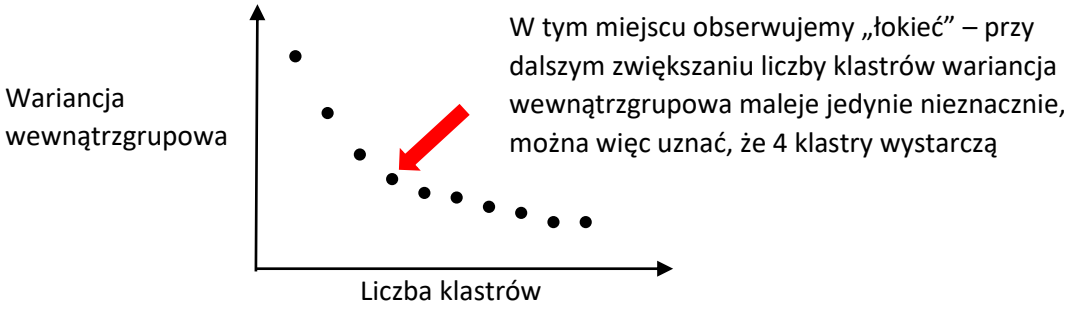
- Kompletne (*complete)* – odległość między klastrami to największa odległość pomiędzy obserwacjami jakie do nich należą
- Pojedyncze (*single)* – najmniejsza odległość pomiędzy obserwacjami w klastrach
- Średnie (UPGMA) – średnia odległość
- Centroidowe (UPGMC) – odległość pomiędzy centroidami klastrów

Hierarchię utworzoną w procesie klasteryzacji można zwizualizować w postaci dendrogramu.

Metoda k-średnich (*k-means*)
Polega na podziale zbioru obserwacji na k klastrów definiowanych przez ich centroidy tak, aby zminimalizować wariancję wewnątrzgrupową. Przebieg algorytmu można opisać następująco:

1. Inicjalizacja algorytmu – wybierz początkowe położenie centroidów.
2. Przypisz każdą obserwację do klastra, z którego centroidem dzieli ją najmniejsza odległość.
3. Oblicz nowe centroidy (średnie obserwacji dla każdego klastra).
4. Jeżeli zmiana położenia centroidów jest większa niż założony margines tolerancji, wróć do punktu 2.

W metodzie k-średnich z góry zakłada się, jaka będzie liczba klastrów. Zwykle testuje się zatem różne liczby, a następnie wyznacza optymalną przy pomocy na przykład tzw. „metody łokcia”.



Zadania

Zbiór danych zawiera listy składników potrzebnych do przygotowania 100 przepisów kulinarnych. Brak wartości w kolumnie odpowiadającej danemu składnikowi oznacza, że nie jest on używany. Jednostki, w jakich podawane są składniki, są takie same w obrębie poszczególnych kolumn, ale mogą różnić się pomiędzy kolumnami. Z podanych przepisów powstają różne liczby porcji.

1. Przygotuj trzy macierze zawierające listy składników – oryginalną, zakodowaną binarnie (składnik występuje/nie występuje) oraz ograniczoną do pięciu wybranych składników.
2. Przedstaw dane przy pomocy heatmap (bez skalowania).
3. Wykonaj klasteryzację hierarchiczną dla oryginalnych danych (wyniki przedstaw w formie dendrogramów):
    a. Zbadaj wpływ miary odległości – dla połączenia „complete” przetestuj odległość euklidesową oraz Manhattan. Na czym polegają te miary?
    b. Zbadaj wpływ połączenia – dla odległości euklidesowej przetestuj połączenia „complete”, UPGMA, „single” oraz UPGMC.
    c. Zbadaj wpływ przetworzenia danych – dla odległości euklidesowej i połączenia „complete” wykonaj klasteryzację dla dwóch pozostałych macierzy.
    d. Czy uzyskane wyniki klasteryzacji pokrywają się z faktycznymi kategoriami potraw? Podaj przykład obserwacji, która znalazła się w „niewłaściwym” klastrze i zastanów się dlaczego tak się stało. Jakie informacje można dodać do zbioru aby uzyskać wyniki bliższe intuicji?
4. Dla przygotowanych macierzy wykonaj klasteryzację metodą k-średnich. Wyznacz najlepszą liczbę klastrów przy pomocy „metody łokcia”. Zwizualizuj dane przy pomocy PCA, kolorami zaznaczając przynależność do klastrów. Jak oceniasz uzyskane grupowanie?

Zadania można wykonać w dowolnie wybranym języku programowania. Sprawozdanie powinno zawierać wyniki symulacji i wnioski (w tym odpowiedzi na pytania znajdujące się w instrukcji). Czas na wykonanie sprawozdania to dwa tygodnie od daty laboratorium.