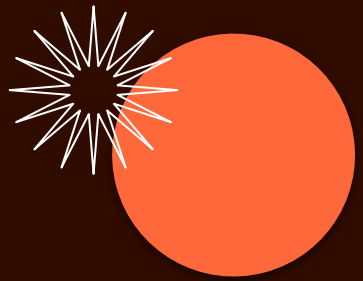
POLITECHNIKA WROCŁAWSKA
WYDZIAŁ INFORMATYKI I TELEKOMUNIKACJI

METODY ANALIZY I EKSPLORACJI DANYCH

Wykład 7 - Klasteryzacja danych

DR INŻ. AGATA MIGALSKA



00

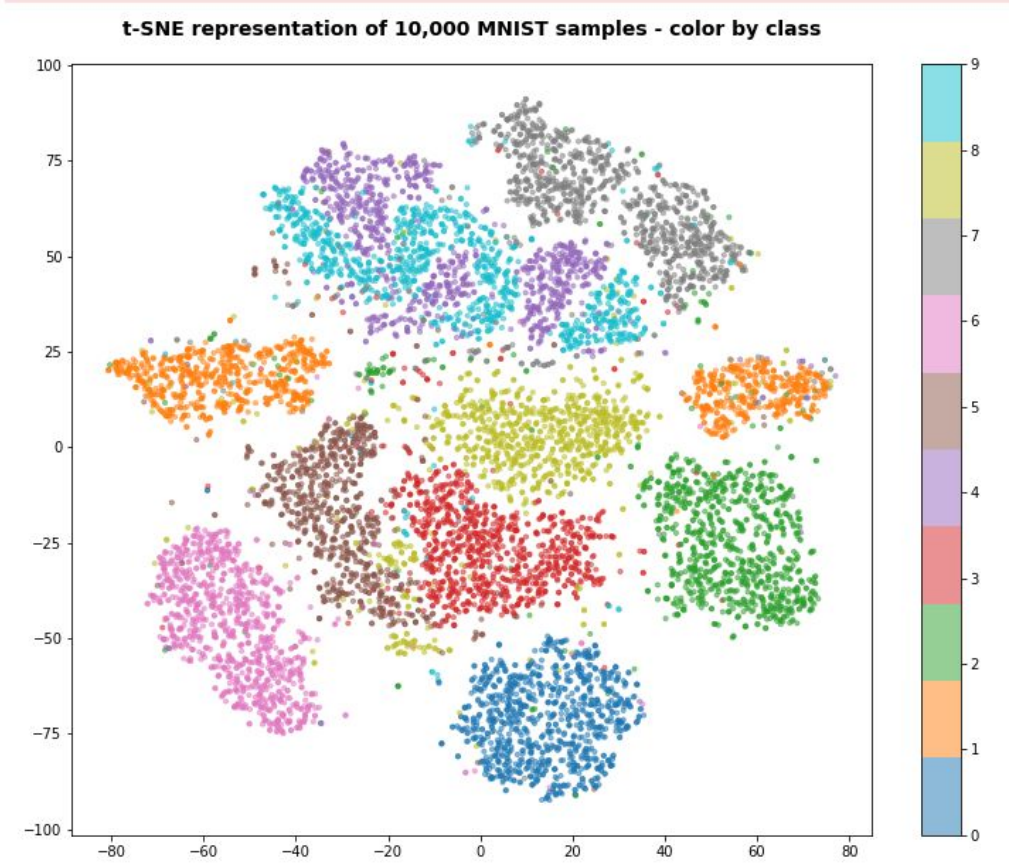
CEL I MOTYWACJA

PROBLEM KLASYFIKACJI

- Grupowanie obiektów jest wymagane do różnych celów w różnych dziedzinach inżynierii, nauki i techniki, nauk humanistycznych, nauk medycznych i naszego codziennego życia.
- Głównym celem badania klasyfikacji jest opracowanie narzędzia lub algorytmu, który można wykorzystać do **przewidywania klasy nieznanego obiektu, który nie jest oznaczony.**

KLASYFIKACJA NADZOROWANA

- Przykład:
  - Weźmy osoby cierpiące na określoną chorobę, które mają pewne wspólne objawy i są umieszczane w grupie oznaczonej etykietą, zwykle nazwą choroby.
  - Osoby nie posiadające tych objawów (a tym samym choroby) nie zostaną umieszczone w tej grupie.
  - Pacjenci zakwalifikowani do tej grupy będą odpowiednio leczeni, podczas gdy pacjenci nienależący do tej grupy powinni być traktowani inaczej.
- W klasyfikacji nadzorowanej, model uczy się rozpoznawać do której grupy dany pacjent należy na podstawie historycznych danych.
- Jednak w wielu przypadkach takie informacje na etykietach nie są podawane z wyprzedzeniem i grupujemy obiekty na podstawie pewnego podobieństwa.


t-SNE representation of 10,000 MNIST samples - color by class
100
75
50
25
0
−25
−50
−75
−100
−80
−60
−40
−20
0
20
40
60
80
9
8
7
6
5
4
3
2
1
0

KLASYFIKACJA NIENADZOROWANA

- W klasyfikacji nienadzorowanej nie ma etykiety przypisanej do żadnego wzorca.
- Klasyfikacja nienadzorowana jest powszechnie znana jako klastrowanie.
- Klastrowanie dzieli wzorce danych na podzbiory w taki sposób, że podobne wzorce są grupowane razem.
- Formalnie i konwencjonalnie klastry można przedstawić jako zbiór S podzbiorów $S_1,\ S_2,\ \ldots,\ S_k$ taki, że: $S_1 \cap S_2 \cap \cdots \cap S_k = \emptyset$.
- Grupowanie jest uważane za trudniejsze niż klasyfikacja nadzorowana, ponieważ nie ma etykiety dołączonej do wzorców w klastrowaniu.
- W przypadku klastrowania trudno jest zdecydować, do której grupy będzie należeć wzór w przypadku braku etykiety.
- Klastrowanie = Analiza skupień = Grupowanie

01 METODY HIERARCHICZNE

02 METODY PODZIAŁOWE

03 METODY GĘSTOŚCIOWE

04 JAK DOBRE SĄ MOJE KLASTRY?

01

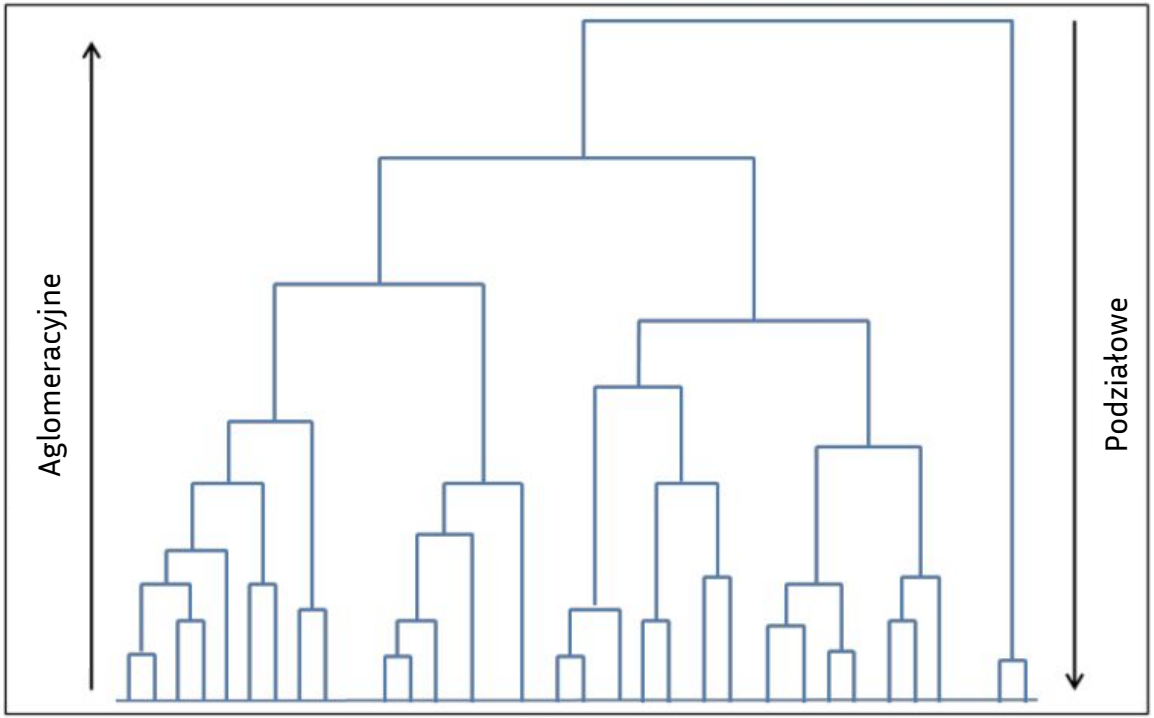
METODY HIERARCHICZNE

KLASTROWANIE HIERARCHICZNE

- **Algorytmy aglomeracyjne** - zaczynamy od zdefiniowania każdego punktu danych jako klastra. W każdym kroku dwa najbliższe klastry są łączone w jeden klaster.
- **Algorytmy podziałowe** - zaczynamy od umieszczenia wszystkich punktów danych w jednym klastrze. W każdym kroku dzielimy istniejący klaster na dwa klastry.

Uwaga: Metody aglomeracyjne są stosowane znacznie częściej niż metody podziałowe.

DENDROGRAM

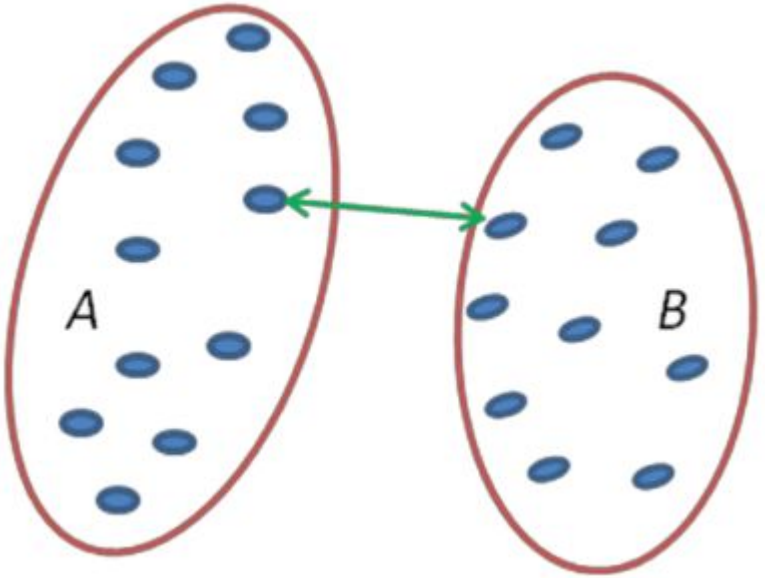


METODA NAJBLIŻSZEGO SĄSIEDZTWA

Odległość między grupami jest ustalona jako odległość między najmniej oddalonymi od siebie obiektami z dwóch grup.

$$\min(d(a,b)) : \ a \in A,\ b \in B$$

Nazywana również metodą pojedynczego wiązania.

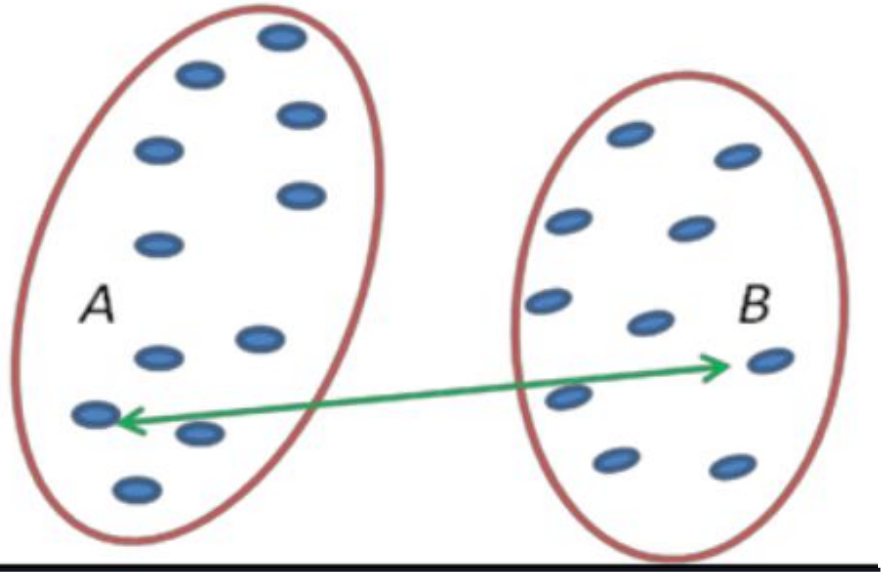


METODA NAJDALSZEGO SĄSIEDZTWA

Odległość między grupami jest ustalona jako odległość między najbardziej oddalonymi od siebie obiektami z dwóch grup.

$$\max(d(a,b)) : \ a \in A, b \in B$$

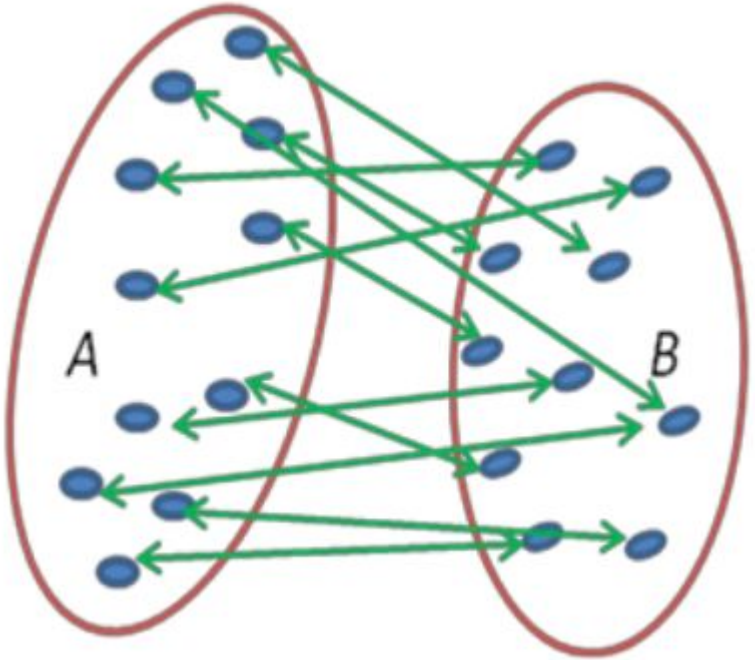
Nazywana również metodą pełnego wiązania.



METODA ŚREDNIEGO WIĄZANIA

Odległość między grupami jest ustalona jako średnia wszystkich odległości między obiektami różnych grup.

$$\frac{1}{|A||B|}\sum_{a\in A}\sum_{b\in B} d(a,b)$$



METODA WARDA

- Metoda Warda jest podejściem opartym na ANOVA i nie definiuje bezpośrednio miary odległości między dwoma punktami lub skupiskami.
- Kryterium minimalnej wariancji Warda minimalizuje całkowitą wariancję wewnątrz klastra.
- Na każdym etapie łączą się dwa klastry, które zapewniają najmniejszy wzrost całkowitej wariancji (wzrost łącznej sumy kwadratów błędów) wewnątrz klastra po połączeniu.
- Kiedy suma kwadratów błędów jest mała, sugeruje to, że nasze dane są zbliżone do wartości średnich klastrów, co sugeruje, że mamy skupisko podobnych jednostek.

ZALETY I WADY METOD HIERARCHICZNYCH

- Elastyczny pod kątem wyboru miary odległości lub podobieństwa obiektów.
- Wszechstronna: może być zastosowana do grupowania dokumentów, sekwencji, zbiorów liczbowych, etc.
- Pozwala na elastyczny wybór zbioru klastrów o określonej ziarnistości.
- Generuje klastry o dowolnym kształcie.

- Kosztowna obliczeniowo i pamięciowo $O(n^2)$
- Czuła na punkty odstające i zaszumione dane
- Może prowadzić do generowania zbiorów klastrów po niskiej jakości, ponieważ po połączeniu klastrów algorytm nie cofa się do stanu sprzed połączenia w celu znalezienia lepszego podziału obiektów do klastrów.

INNE METODY HIERARCHICZNE

- BIRCH
  - połączenie aglomeracyjnego grupowania hierarchicznego z innymi technikami grupowania,
  - charakteryzuje się wysoką skalowalnością, efektywnością i dobrą jakością grupowania
- CURE
  - połączenie aglomeracyjnego grupowania hierarchicznego, próbkowania losowego i partycjonowania danych

02

METODY PODZIAŁOWE

METODY PODZIAŁOWE (PARTYCJONUJĄCE)

- W metodach podziałowych dane są początkowo dzielone na zestaw K klastrów.
- Może to być podział losowy lub podział oparty na pierwszym „dobrym” przypuszczeniu w punktach zarodkowych, które tworzą początkowe centra klastrów.
- Następnie punkty danych są iteracyjnie przenoszone do różnych klastrów, aż nie będzie możliwe rozsądne ponowne przypisanie.
- Początkową liczbę klastrów (K) może określić użytkownik lub algorytm oprogramowania.

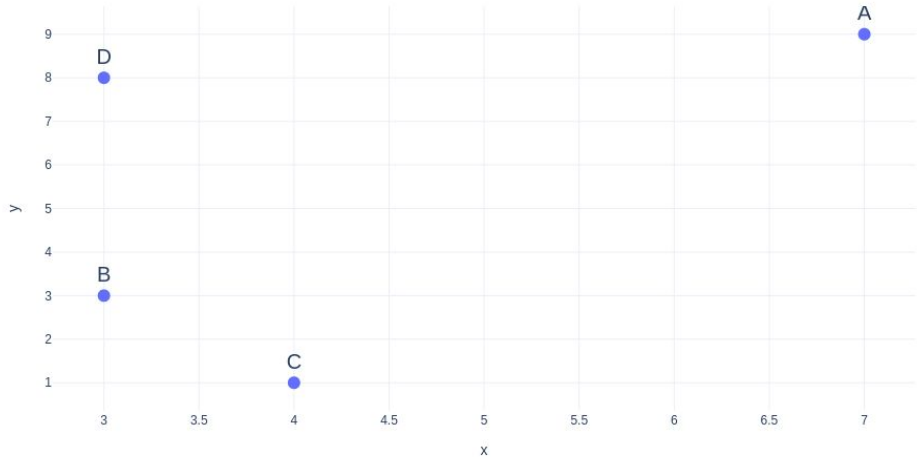
METODA K-ŚREDNICH

- Należy wstępnie określić, ile klastrów należy wziąć pod uwagę. Klastry w tej procedurze nie tworzą drzewa.
- Istnieją dwa podejścia do rozpoczęcia procedury K -średnich:
  - rozpoczęcie od losowego podziału badanych na grupy
  - rozpoczęcie z zestawem punktów początkowych w celu utworzenia centrów klastrów.

- Losowy charakter pierwszego podejścia pozwala uniknąć stronniczości.
- Wyniki algorytmu inicjalizowanego losowo mogą się różnić w każdym przebiegu.

PRZYKŁAD

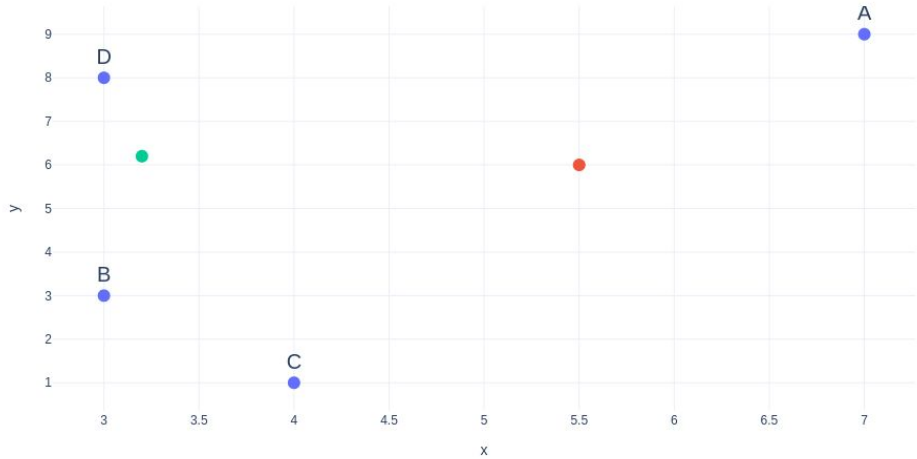
Przedmiot	X_1	X_2
A	7	9
B	3	3
C	4	1
D	3	8



INICJALIZACJA KLASTRÓW

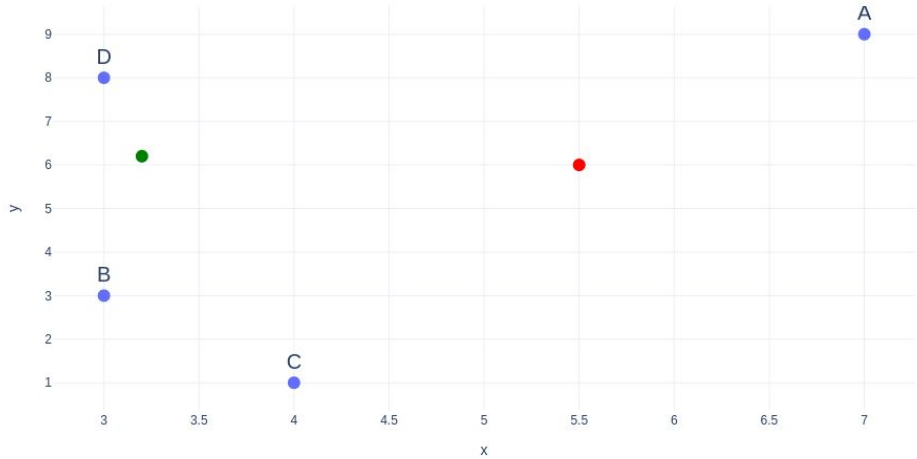
Początkowe punkty (losowe lub wybrane ręcznie):

Początkowe centroidy	X_1	X_2
Czerwony	5.5	6
Zielony	3.2	6.2



ODLEGŁOŚCI DO CENTROIDÓW

	Zielony	Czerwony
A	4.72	**3.35**
B	**3.21**	3.91
C	5.26	**5.22**
D	**1.81**	3.20



NOWE CENTROIDY KLASTRÓW

Środek ciężkości	$\bar{X}_1$	$\bar{X}_2$
A, C	5.5	5
B, D	3	5.5

ALGORYTM

- Krok 1 : Przejrzyj listę n elementów, przypisując każdy element do klastra, którego środek ciężkości (średnia) jest najbliższy.
- Krok 2: Wyznacz nowe środki ciężkości klastrów na podstawie przypisanych punktów.
- Krok 3 : Powtarzaj kroki 1 i 2, aż nie zostaną dokonane żadne ponowne przypisania.

JAK WYBRAĆ LICZBĘ SĄSIADÓW?

- Metoda łokcia (elbow method) - heurystyka stosowana do określania liczby skupień w zbiorze danych.
- Metoda polega na wykreśleniu wyjaśnionej zmienności w danych w zależności od liczby klastrów, a następnie wybraniu punktu, w którym malejące zyski nie są już warte dodatkowych kosztów. W klastrowaniu oznacza to, że należy wybrać taką liczbę klastrów, aby dodanie kolejnego klastra nie dało dużo lepszego modelowania danych.

Dwie pozostałe metody w części 5 “Jak dobre są moje klastry?”

ZALETY I WADY GRUPOWANIA K-ŚREDNICH

- Prosty
- Relatywnie efektywny w porównaniu z metodami klastrowania hierarchicznego
- Złożoność O(knm) gdzie m - liczba iteracji, k - liczba klastrów, n - liczba punktów.
- Elastyczny pod kątem wyboru miary odległości.
- Wynik algorytmu nie zależy od kolejności, w jakiej są analizowane grupowane obiekty.

- Bardzo czuły na dane zaszumione lub dane zawierające punkty osobliwe (odstające).
- Wynik działania algorytmu silnie zależy od początkowego podziału obiektów.
- Najczęściej znajduje lokalne optimum, a nie globalne (k-średnich jest algorytmem zachłannym).
- Ma zastosowanie jedynie do zmiennych liczbowych.
- Nie pozwala na odkrywanie klastrów wklęsłych.

METODA K-MEDOIDS (K-MEDOIDÓW)

- Algorytm bardziej odporny na punkty odstające niż metoda k-średnich.

Medoid klastra definiuje się jako obiekt w klastrze, którego średnia odmienność od wszystkich obiektów w klastrze jest minimalna, to znaczy jest to najbardziej centralnie położony punkt w klastrze.

Główne założenia algorytmu:

- Każdy klaster jest reprezentowany przez jeden ze swoich obiektów.
- Celem metody jest znalezienie k obiektów reprezentujących k klastrów i minimalizujących przyjętą funkcję kryterialną.

METODA K-MEDOIDS (K-CENTROIDÓW)

Krok 1: Wybierz k punktów ze zbioru jako początkowe centroidy klastrów.
Krok 2: Przypisz pozostałe punkty do tego klastra, do którego odległość (lub podobieństwo) obiektów od centroidu klastra jest najmniejsza (lub największe).
Krok 3: Aktualizacja medoidów.

Dla każdego medoidu m i dla każdego punktu danych innego niż medoid o:

Rozważ zamianę m i o i oblicz zmianę kosztu

Jeśli zmiana kosztów jest obecnie najlepsza, zapamiętaj tę kombinację m i o

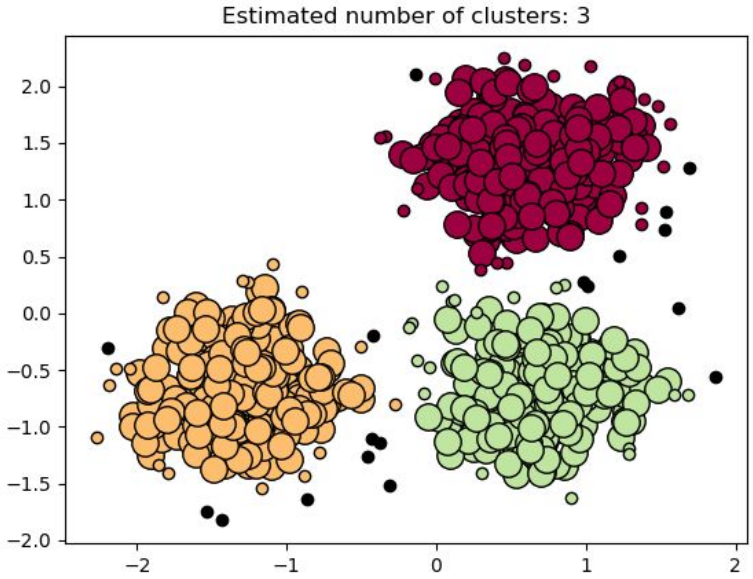
Wykonaj najlepszą zamianę m_best i o_best, jeśli zmniejsza ona funkcję kosztu. W przeciwnym razie algorytm się kończy.

Implementacja w scikit-learn:
https://scikit-learn-extra.readthedocs.io/en/stable/auto_examples/plot_kmedoids.html

03

METODY GRUPOWANIA GĘSTOŚCIOWEGO

DBSCAN

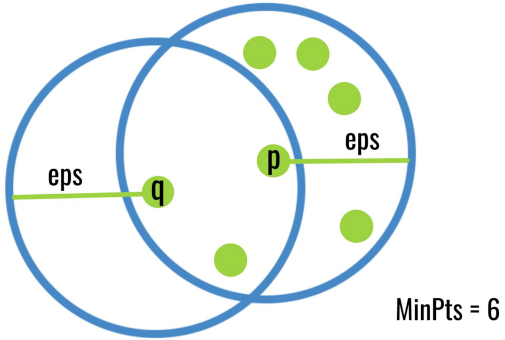


DBSCAN - POJĘCIA

ε-sąsiedztwo obiektu p - zbiór obiektów, których odległość od p jest nie większa niż ε

Obiekt p nazywamy **obiektem centralnym** (lub jądrem) (ang. core object), jeżeli jego ε-sąsiedztwo zawiera co najmniej MinPts obiektów.

Obiekt q jest **bezpośrednio gęstościowo osiągalny** (ang. directly density reachable) z obiektu p, jeżeli q należy do ε-sąsiedztwa obiektu p i obiekt p jest centralny.

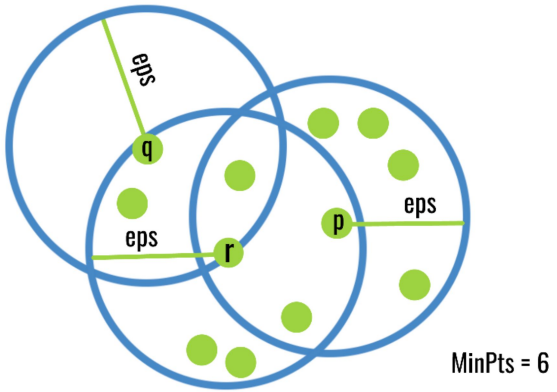


Directly density reachable

Źródło grafik: https://www.geeksforgeeks.org/ml-dbscan-reachability-and-connectivity/

GĘSTOŚCIOWA OSIĄGALNOŚĆ

Obiekt q jest **gęstościowo osiągalny** z obiektu p, jeżeli istnieje łańcuch obiektów $p_1, p_2, \ldots, p_n$, gdzie p_1=p i p_n=q, i obiekt p_{i+1} jest bezpośrednio gęstościowo osiągalny z p_i.

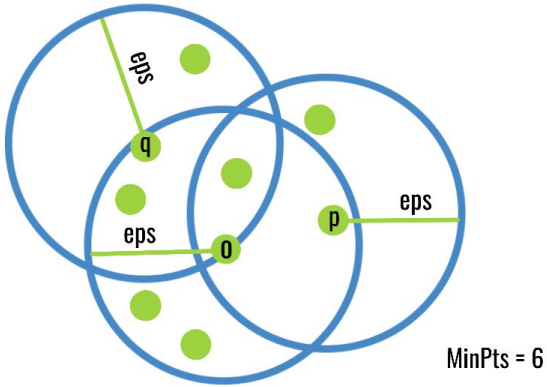


Density reachable

Źródło grafik: https://www.geeksforgeeks.org/ml-dbscan-reachability-and-connectivity/

GĘSTOŚCIOWA POŁĄCZENIOWOŚĆ

Obiekt p jest **gęstościowo połączony** (ang. density connected) z obiektem q, jeżeli istnieje obiekt o taki, że oba obiekty p i q są gęstościowo osiągalne z obiektu o.

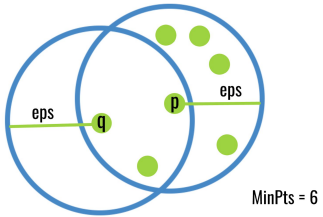


Density connectivity

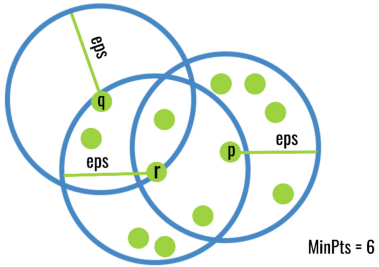
Źródło grafik: https://www.geeksforgeeks.org/ml-dbscan-reachability-and-connectivity/

KLASTER

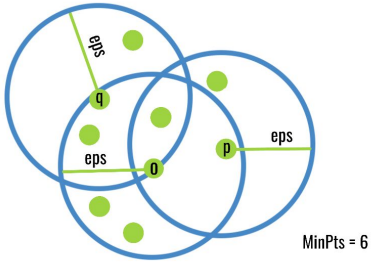
Jeżeli punkt $p \in C$, C - klaster, i punkt q jest gęstościowo osiągalny z punkt p, to $q \in C$.
Dowolne dwa punkty w klastrze są gęstościowo połączone.



Directly density reachable



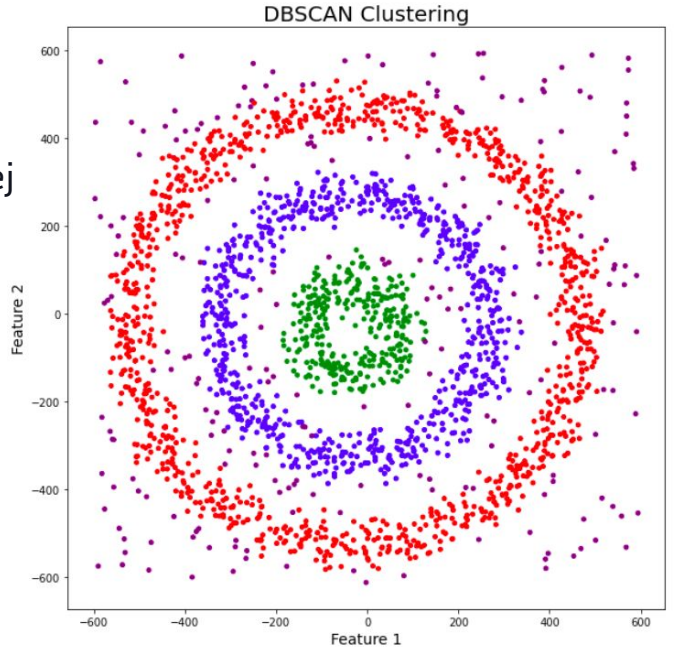
Density reachable



Density connectivity

DBSCAN – ALGORYTM

1. Rozpoczynamy od dowolnego obiektu p ze zbioru D.
2. Sprawdzamy czy p jest punktem centralnym:
   a. Jeżeli ε-sąsiedztwo obiektu p spełnia warunek minimalnej gęstości tzn. jest w nim co najmniej MinPts obiektów, to tworzony jest klaster C i wszystkie obiekty gęstościowo osiągalne z obiektu p są dołączane do klastra C.
   b. W przeciwnym razie wracamy do punktu 1 i wybieramy następny obiekt ze zbioru.
3. Proces grupowania jest kontynuowany tak długo, aż zostaną przetworzone wszystkie obiekty zbioru D.



Obiekty, które nie zostały zaklasyfikowane do żadnego z klastrów, tworzą zbiór punktów osobliwych.

ZALETY I WADY GRUPOWANIA DBSCAN

- Nie wymaga wcześniejszego określenia liczby klastrów.
- Dobrze radzi sobie z dowolnymi kształtami klastrów.
- Jest odporny na wartości odstające i jest w stanie wykryć wartości odstające.

- W niektórych przypadkach określenie odpowiedniej odległości sąsiedztwa (eps) nie jest łatwe i wymaga wiedzy dziedzinowej.
- Jeśli klastry są bardzo różne pod względem gęstości w klastrze, DBSCAN nie nadaje się dobrze do definiowania klastrów.

INNE METODY GĘSTOŚCIOWE

- OPTICS
  - Idea: klastry o większej gęstości zawierają się w klastrach o mniejszej gęstości.
  - Wynikiem działania algorytmu jest kolejność przetwarzania obiektów.
- HDBSCAN
  -
  - https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html

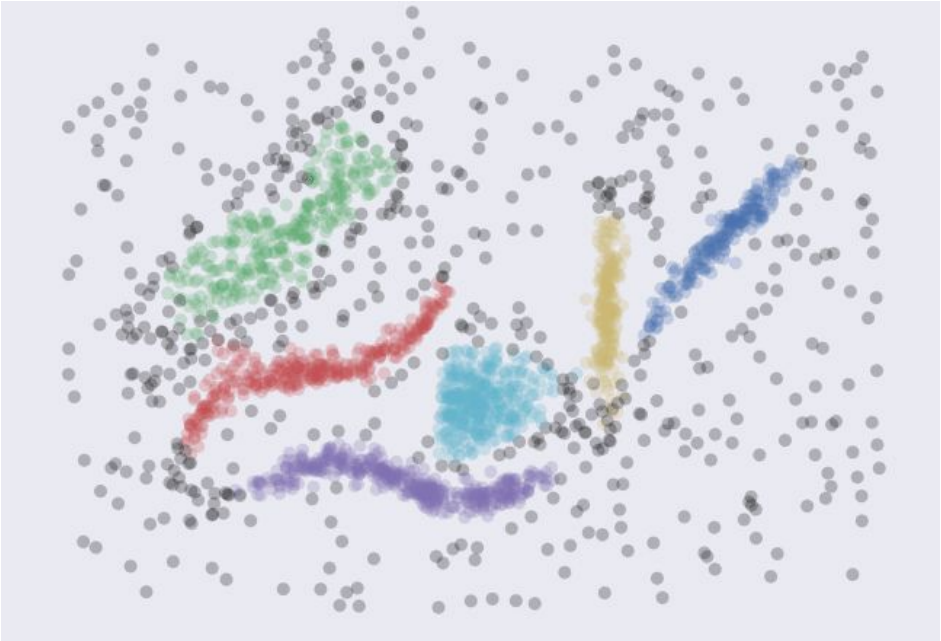
PRZYKŁAD ZASTOSOWANIA

Zbiór danych: informacje o seriach produkcyjnych

Zmienne:

- typ produktu
- chipset
- fabryka
- klient
- doświadczenie w produkcji (w tyg.)
- doświadczenie we współpracy z klientem (w tyg.)

Zastosowany algorytm: HDBSCAN



Rysunek poglądowy. Źródło: https://www.researchgate.net/figure/Example-of-HDBSCAN-clustering_fig2_328530481

INNE PODEJŚCIA DO KLASTROWANIA

- Metody oparte na modelu
  - Zakładają, że dane są generowane przez pewien proces statystyczny.
  - Celem procesu grupowania jest znalezienie modelu statystycznego, który najlepiej opisuje zbiór grupowanych danych.
  - Najbardziej popularny algorytm: Expectation Maximization (algorytm maksymalizacji wartości oczekiwanej)
- Metody grafowe
  - Affinity Propagation
    - oparty na koncepcji “przekazywaniu wiadomości” pomiędzy obiektami zbioru.
- ...

04

JAK DOBRE SĄ MOJE KLASTRY?

INDEKS CH (CALIŃSKIEGO-HARABASZA)

- Indeks CH jest miarą tego, jak obiekt jest podobny do własnego skupienia (spójność) w porównaniu z innymi skupieniami (separacja).
- Spójność jest szacowana na podstawie odległości od punktów danych w klastrze do jego środka ciężkości klastra.
- Separacja jest oparta na odległości między środkami ciężkości klastra od globalnego środka ciężkości.

$$CH = \frac{Separacja}{Spójność}$$

Maksymalizując indeks CH można:

- wybrać końcową liczbę skupień,
- porównywać algorytmy między sobą.

Wyższa wartość wskaźnika CH oznacza, że klastry są gęste i dobrze rozdzielone, chociaż nie ma „akceptowalnej” wartości odcięcia.

ZARYS

- Dla każdej obserwacji i definiujemy Zarys (ang. silhouette) jako

$$Zarys = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

  - a(i) - średnią miarą niepodobieństwa pomiędzy nią, a wszystkimi obserwacjami z tej samej grupy (Spójność)
  - b(i) -średnia odległość do najbliższego skupienia, do którego i nie należy (Separacja).
- Zarys jest miarą tego, jak obiekt jest podobny do własnego skupienia (spójność) w porównaniu z innymi skupieniami (separacja).
- Średnia wartość zarysu s̄ dla każdego skupienia mówi o tym jak dobrze dane są przydzielone do tego skupienia.
- Wybierając optymalną liczbę klastrów, maksymalizujemy średnią wartość zarysu.
- Wartość zarysu przyjmuje wartości od -1 do +1:
  - wysoka wartość wskazuje, że obiekt jest dobrze dopasowany do własnego klastra i słabo dopasowany do sąsiednich klastrów.

THANKS!

DZIĘKUJĘ ZA UWAGĘ