Eksploracja danych
7. Grupowanie

Piotr Szwed
Katedra Informatyki Stosowanej AGH
2021

Grupowanie

Grupa obiektów (ang. cluster) to zbiór obiektów wyodrębniony z danych

- Obiekty wewnątrz grupy powinny cechować się podobieństwem
- Obiekty różnych grup powinny być niepodobne do siebie

Grupowanie, klasteryzacja, analiza skupień (ang. clustering, cluster analysis) ma na celu **odkrywanie wiedzy**: znalezienie interesujących wzorców wewnątrz danych, np. często występujących kombinacji atrybutów lub grup podobnych obiektów.

Wynikiem grupowania jest przydział etykiet (numerów grup) do obiektów.

Porównanie

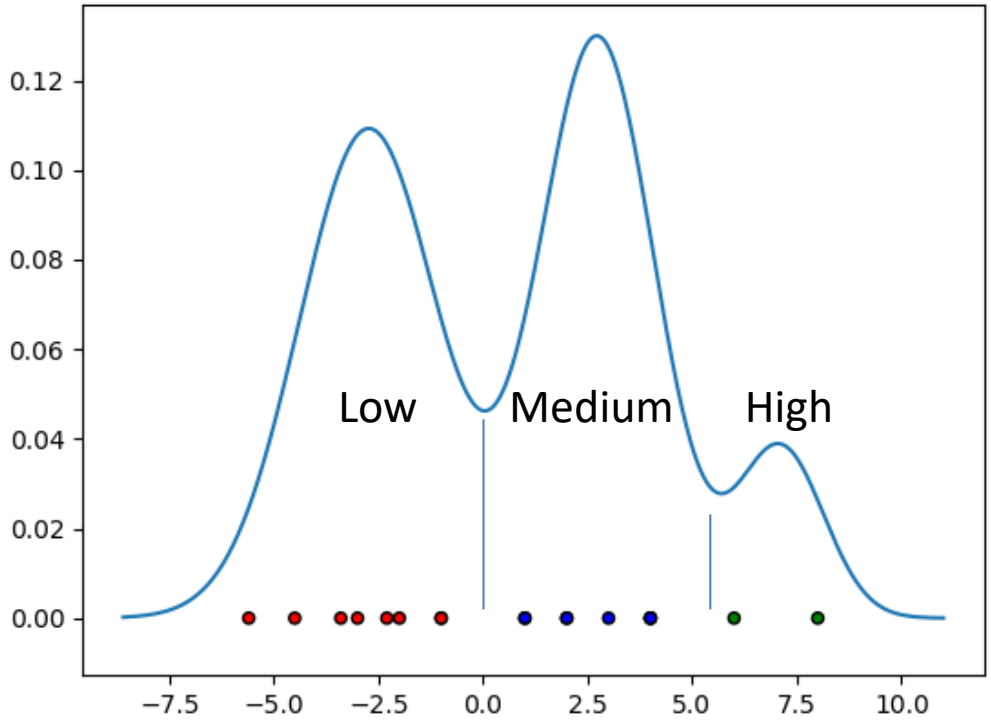
Zagadnienie	Dane uczące	Cel	Metoda oceny
Regresja	$D = \{(x_i, y_i)\}$ $x_i \in X$ $y_i \in R$	Znalezienie funkcji $f: X \to R$ odwzorowującej całą przestrzeń X	Zbiór testowy Sprawdzian krzyżowy (kroswalidacja)
Klasyfikacja	$D = \{(x_i, y_i)\}$ $x_i \in X$ $C = \{c_1, \dots, c_k\}$ – etykiety klas $y_i \in C$	Znalezienie funkcji $f: X \to C$ odwzorowującej całą przestrzeń X	Zbiór testowy Sprawdzian krzyżowy (kroswalidacja)
Grupowanie	$D = \{x_i\}$ $x_i \in X$	Znalezienie funkcji $f: D \to L$ która przypisuje etykiety ze zbioru L instancjom w D	Wartość założonej funkcji celu lub wskaźnika jakości, na podstawie dostępnych etykiet $Y = y_i$, użyteczność w dalszym przetwarzaniu

Uczenie nienadzorowane

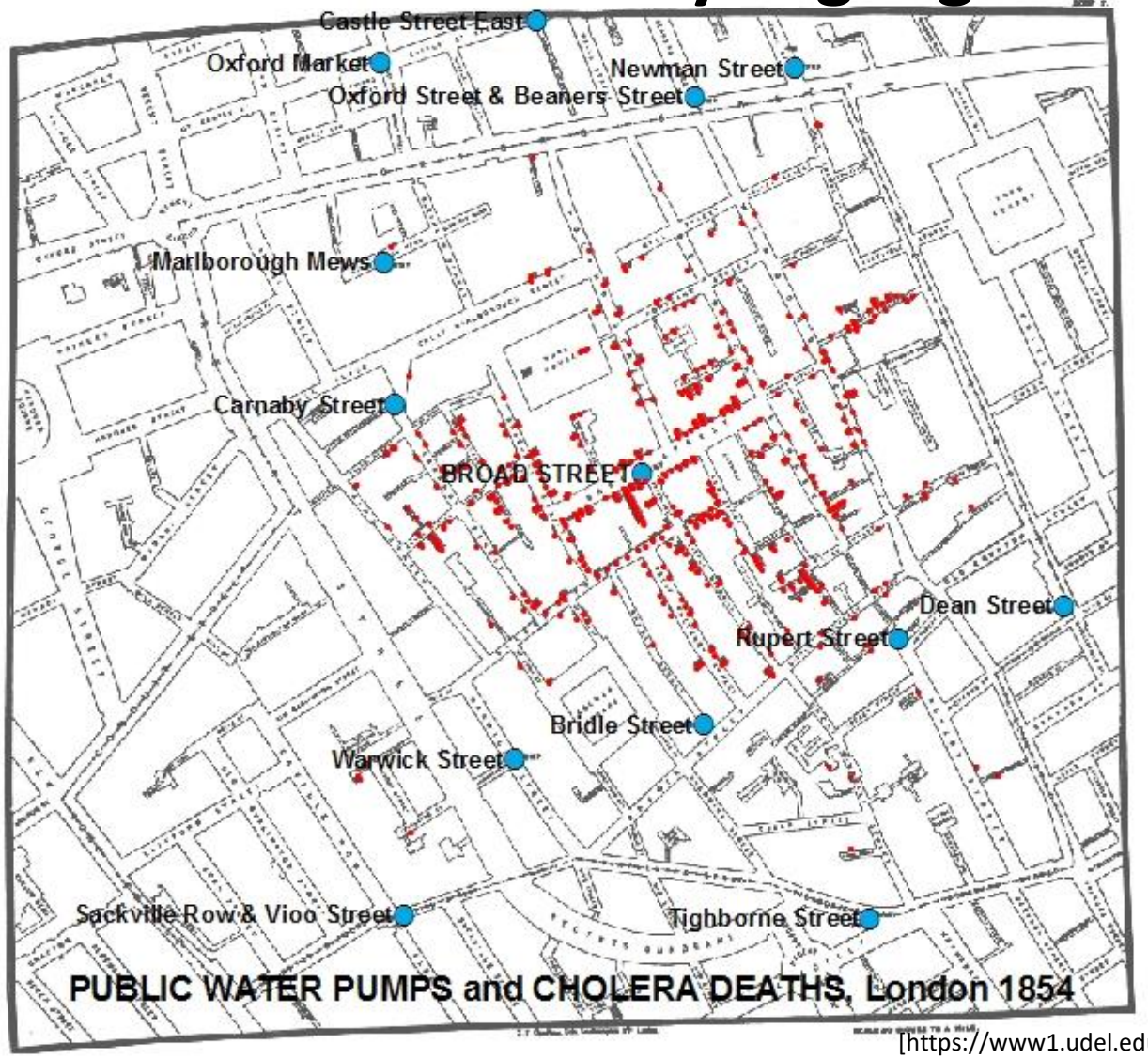
- Jeżeli w procesie uczenia i budowy modeli wykorzystuje się znane wartości wyjściowe y_i dla obserwacji (x_i, y_i), porównuje się je z wyjściami modelu i koryguje jego parametry – jest to **uczenie nadzorowane**.
- **Klasyfikacja** i **regresja** to przykłady zagadnień uczenia nadzorowanego
- Jeżeli wartości wyjściowe y_i nie są znane/wykorzystywane, podczas budowy modelu lub jego działania jest to **uczenie nienadzorowane**.
- Podczas **grupowania** nie są znane wartości wyjściowe y_i (ciągłe lub etykiety klas). Jest to więc przykład uczenia nienadzorowanego.
- Czy kNN (metodę klasyfikacji) należy zaliczyć do grupy algorytmów uczenia nadzorowanego?

Zastosowania

- Wykorzystywane jako samodzielne narzędzie analizy danych np. do znalezienia interesujących wzorców i rozkładów statystycznych
- Jako wstępny etap przetwarzania danych:
  - Grupowanie danych numerycznych, ustalanie przedziałów i zamiana na nominalne
  - Identyfikacja skupisk w celu automatycznego nadania obserwacjom etykiet klas



Przykłady i zastosowania: analiza danych geograficznych



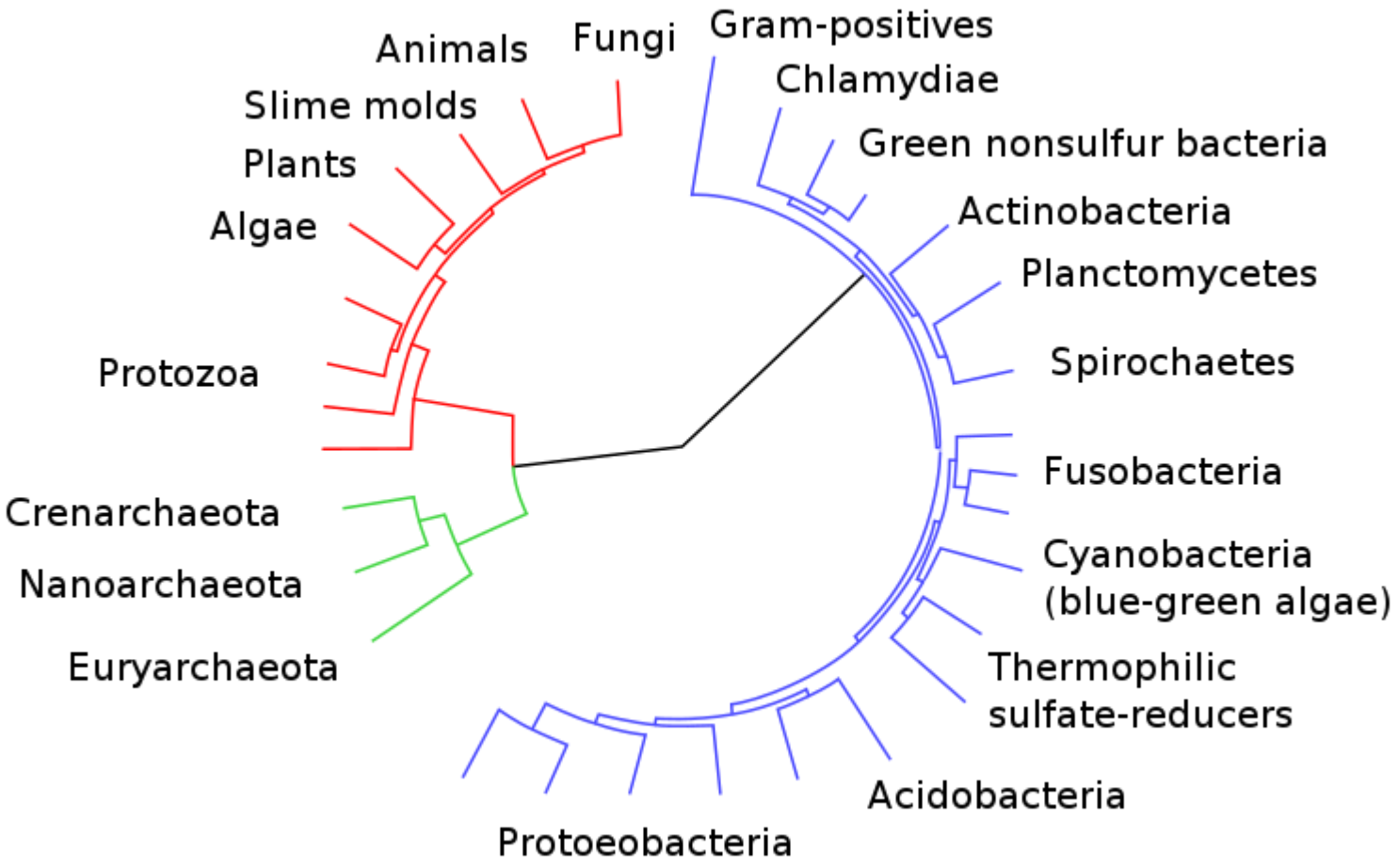
Skupiska cholery podczas epidemii w 1854 roku

[http://tasteforhealth.com/1854-broad-street-cholera-outbreak/]

[https://www1.udel.edu/johnmack/frec682/cholera/cholera2.html]

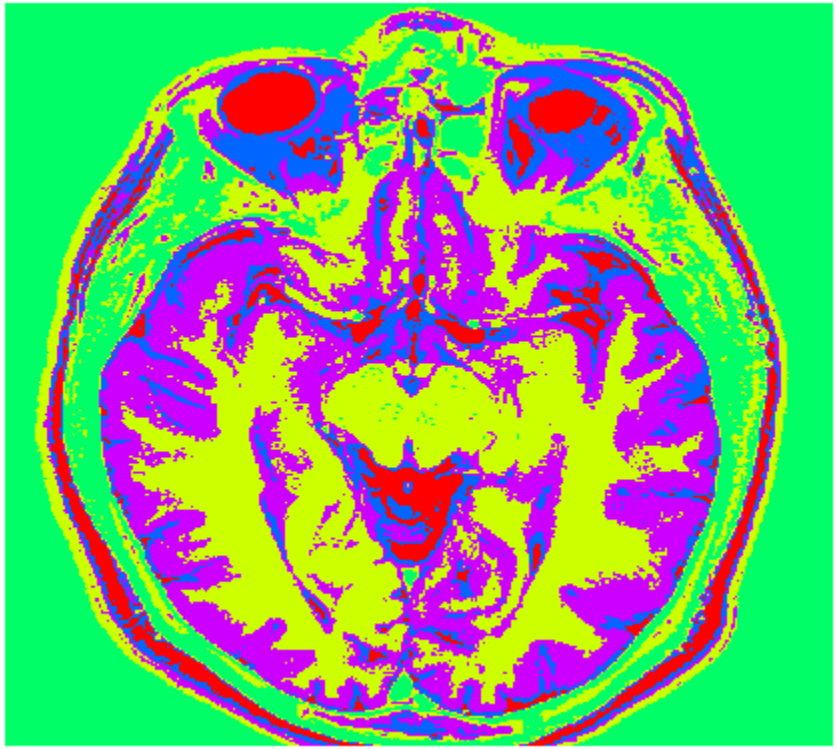
Przykłady i zastosowania: wyszukiwanie dokumentów tekstowych

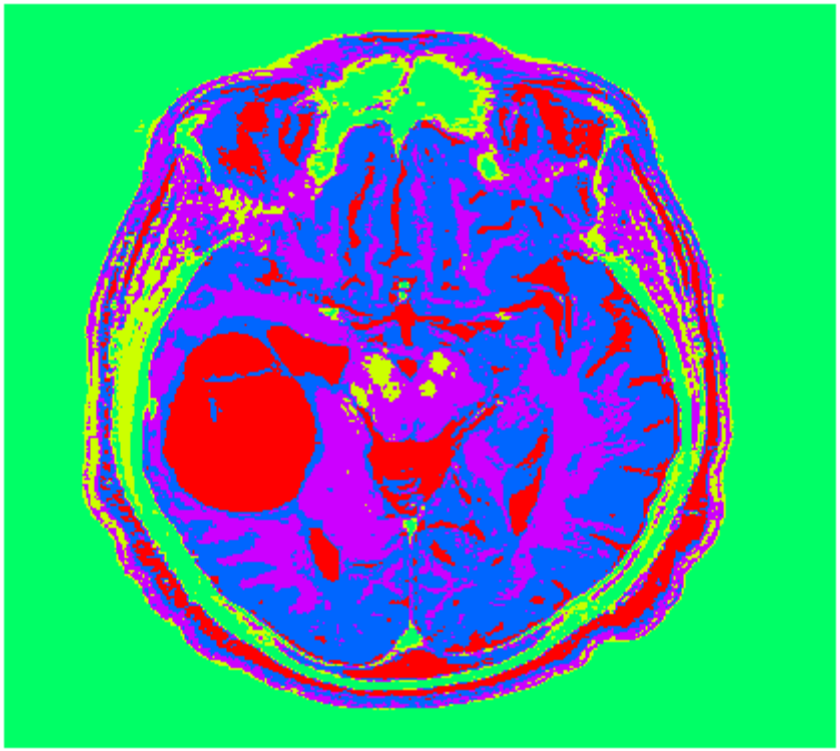
Przykłady i zastosowania: tworzenie taksonomii



[https://en.wikipedia.org/wiki/Taxonomy_(biology)]

Przykłady i zastosowania: segmentacja obrazów





[http://mpmendespt.github.io/MRI-image-segmentation.html]

Zastosowania

- Marketing – poszukiwanie grup podobnych klientów w celu stworzenia dedykowanych programów marketingowych
- Ubezpieczenia – identyfikacja grup klientów o podobnych profilach ryzyka
- Służba zdrowia – analiza kosztów dla różnych wariantów terapii
- Planowanie transportu - analiza ruchu pasażerów
- Rynek nieruchomości – grupowanie ofert o podobnych parametrach

Metody grupowania

- Hierarchiczne
- Rozłączne
- Oparte na modelach statystycznych
- Wykorzystujące informacje o gęstości instancji
- Rozmyte

Grupowanie hierarchiczne

Grupowanie hierarchiczne (aglomeratywne)

Algorytm

Dane:

- $D = \{x_1, \dots, x_m\} \subset X$
- $d(A, B)$, gdzie $A, B \subset X$, odległość zbiorów A i B

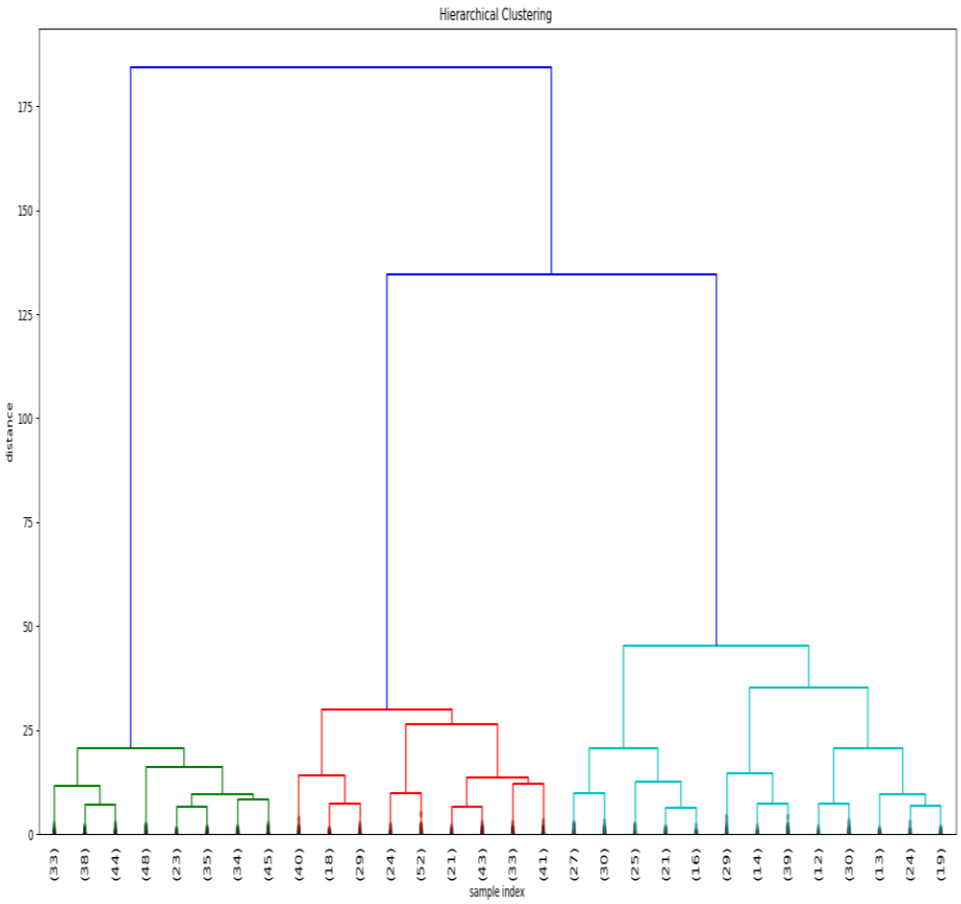
1. Utwórz tyle grup, ile jest obserwacji $\boldsymbol{G} \leftarrow \{C_1, \dots, C_m\}$, gdzie $C_i = \{x_i\}$
2. Dopóki $|\boldsymbol{G}| > 1$:
   3. Znajdź takie dwie grupy C_r i C_s należące do $\boldsymbol{G}$, że odległość między nimi jest najmniejsza

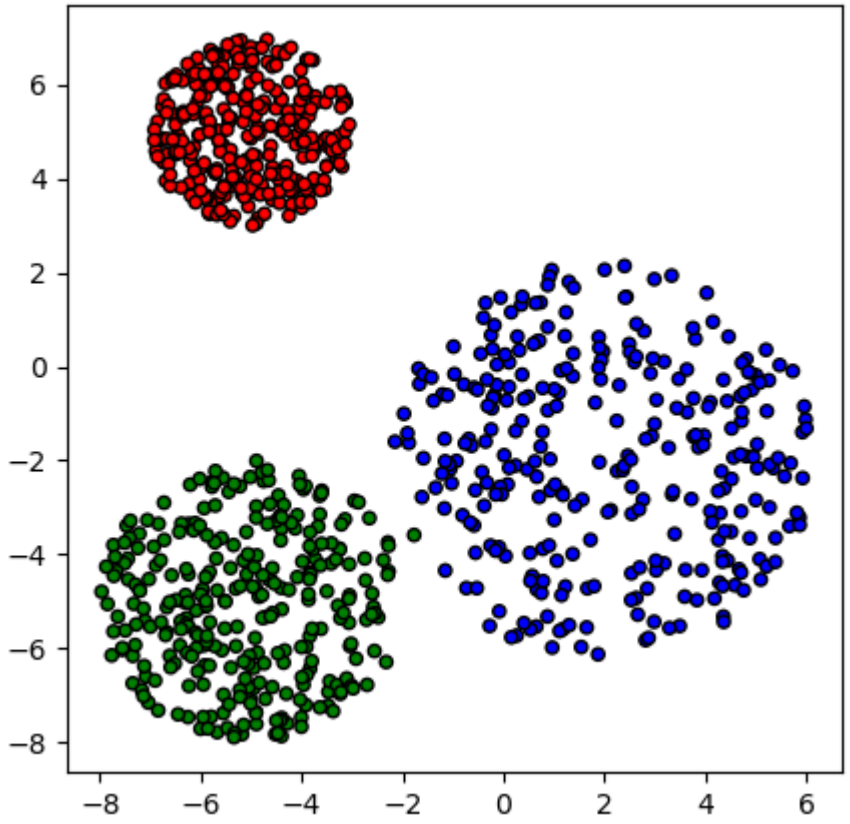
$$(C_r, C_s) = \arg \min_{\substack{(A,B) \in G \times G \\ A \neq B}} d(A, B)$$

   4. Połącz grupy: $\boldsymbol{G} \leftarrow \boldsymbol{G} \backslash \{C_r, C_s\} \cup \{C_{rs}\}$, gdzie $C_{rs} = C_r \cup C_s$

Informacja o połączeniach grup jest zapisywana: grupa C_{rs} jest nadrzędną względem C_r i C_s

Przykład





Rysunek po prawej stronie pokazuje przydział etykiet do obserwacji na trzecim poziomie dendrogramu.

Odległość zbiorów obserwacji

Może zostać zdefiniowana na wiele sposobów:

- Single link: odległość między najbliżej położonymi puntami

$$d(A,B) = \min_{\substack{x_i \in A \\ x_j \in B}} dist(x_i, x_j)$$

- Complete link: odległość między najdalszymi puntami

$$d(A,B) = \max_{\substack{x_i \in A \\ x_j \in B}} dist(x_i, x_j)$$

- Średnia odległość

$$d(A,B) = \sum_{\substack{x_i \in A \\ x_j \in B}} \frac{dist(x_i, x_j)}{|A| \cdot |B|}$$

Odległość zbiorów obserwacji

- Odległość pomiędzy środkami grup (centroidami)

$$d(A,B) = dist(\bar{x}_A, \bar{x}_B)$$

gdzie $\bar{x}_A = \frac{\sum_{x \in A} x}{|A|}$, $\bar{x}_B = \frac{\sum_{x \in B} x}{|B|}$

- Metoda ważona. Zakładając, że grupa A została uformowana przez połączenie dwóch grup A_1 i A_2

$$d(A,B) = \frac{d(A_1,B) + d(A_2,B)}{2}$$

Metoda Warda

- Metoda Warda wybierając grupy do złączenia posługuje się kryterium najmniejszej sumy kwadratów błędów (alternatywnie wariancji -- jak podzielimy RSS przez n).

$$RSS(A) = \sum_{x \in A} (x - \bar{x}_A)^T \cdot (x - \bar{x}_A)$$

- Poszukując grup do połączenia wybierane są te dwie, dla których wzrost kwadratu błędu (wariancji) będzie najmniejszy

$$(C_i, C_j) = \arg \min_{C_i, C_j \in G} RSS(C_i \cup C_j) - RSS(C_i) - RSS(C_j)$$

- Przyrost kwadratu błędu można także obliczać jako:

$$\Delta(C_i, C_j) = RSS(C_i \cup C_j) - RSS(C_i) - RSS(C_j)$$
$$= \frac{|C_i| \cdot |C_j|}{|C_i| + |C_j|} \left\| \bar{x}_{C_i} - \bar{x}_{C_j} \right\|^2$$

Algorytm k-means
k-średnich

Środek grupy - centroid

Załóżmy, że $x \in R^n$. Punkt środkowy (centroid) grupy $C_i = \{x_{i1}, \dots, x_{ir}\}$ obliczamy jako:

$$\bar{x}(i) = \frac{\sum_{x \in C_i} x}{|C_i|}$$

W idealnym przypadku grupa C_i zawiera się wewnątrz kuli, której środkiem jest centroid $\bar{x}(i)$ oraz kule nie przecinają się.

Miarą, która pozwala określić jakość reprezentacji grupy przez centroid jest suma kwadratów błędów (*residual sum of squares*)

$$RSS(i) = \sum_{x \in C_i} \|x - \bar{x}(i)\|^2$$

Miarą jakości podziału na grupy jest:

$$RSS = \sum_{i=1}^{k} RSS(i) = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \bar{x}(i)\|^2$$

Podział na grupy

- Niech $C = (C_1, \dots, C_k)$ będzie podziałem na grupy zbioru $D = \{x_j : j = 1, m\}$ spełniającym:
  $\forall r \neq s : C_r \cap C_s = \emptyset$ oraz
  $\bigcup_{i=1}^{k} C_i = D$
- Najlepszy podział na grupy minimalizuje RSS, stąd:

$$(C_1, \dots, C_k)^* = \arg\min_{C} \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \bar{x}(i)\|^2$$

Zagadnienie można więc potraktować jako problem optymalizacyjny: znaleźć taki podział na rozłączne grupy, które zminimalizuje funkcję celu RSS.

Problem NP zupełny

Inna postać definicji RSS

Wprowadźmy m zmiennych Z_j o wartościach ze zbioru $\{1, \dots, k\}$ opisujących przydział obserwacji do grupy:
$Z_j = i \Leftrightarrow x_j \in C_i$

$$RSS = \sum_{i=1}^{k} \sum_{j=1}^{m} \mathbf{1}(Z_j = i) \left\| x_j - \bar{x}(i) \right\|^2$$

Wartości m zmiennych Z można wybrać na m^k sposobów (i w wielomianowym czasie obliczyć RSS)

- Funkcja $\mathbf{1}(expr)$ przyjmuje wartość 1, gdy $expr$ jest prawdziwe; 0 w przeciwnym przypadku.
- Funkcja celu ma wiele minimów lokalnych, znalezienie rozwiązania dokładnego jest trudne. W zamian używa się algorytmu zachłannego k-means (k-średnich)

Algorytm k-means

Dane

- $D = \{x_i \in R^n\}, i = 1, \dots, m$ – zbiór obserwacji
- k – zadana liczba grup
- $d(x_1, x_2) = \|x_1 - x_2\|^2$ - odległość euklidesowska

Algorytm

1. Wybierz początkowe położenia k centroidów $\bar{x}(1), \dots, \bar{x}(k)$
2. Przypisz obserwacje do najbliższych centroidów:

for $j = 1 \dots m$:

$$Z_j = \arg \min_{i \in \{1,\dots,k\}} d(x_j, \bar{x}(i))$$

3. Uaktualnij położenie centroidów

for $i = 1 \dots k$:

$$\bar{x}(i) = \frac{\sum_{j=1}^{m} 1(Z_j = i) \cdot \mathrm{x}_j}{\sum_{j=1}^{m} 1(Z_j = i)}$$

Suma wektorów należących do grupy

Liczba elementów w grupie

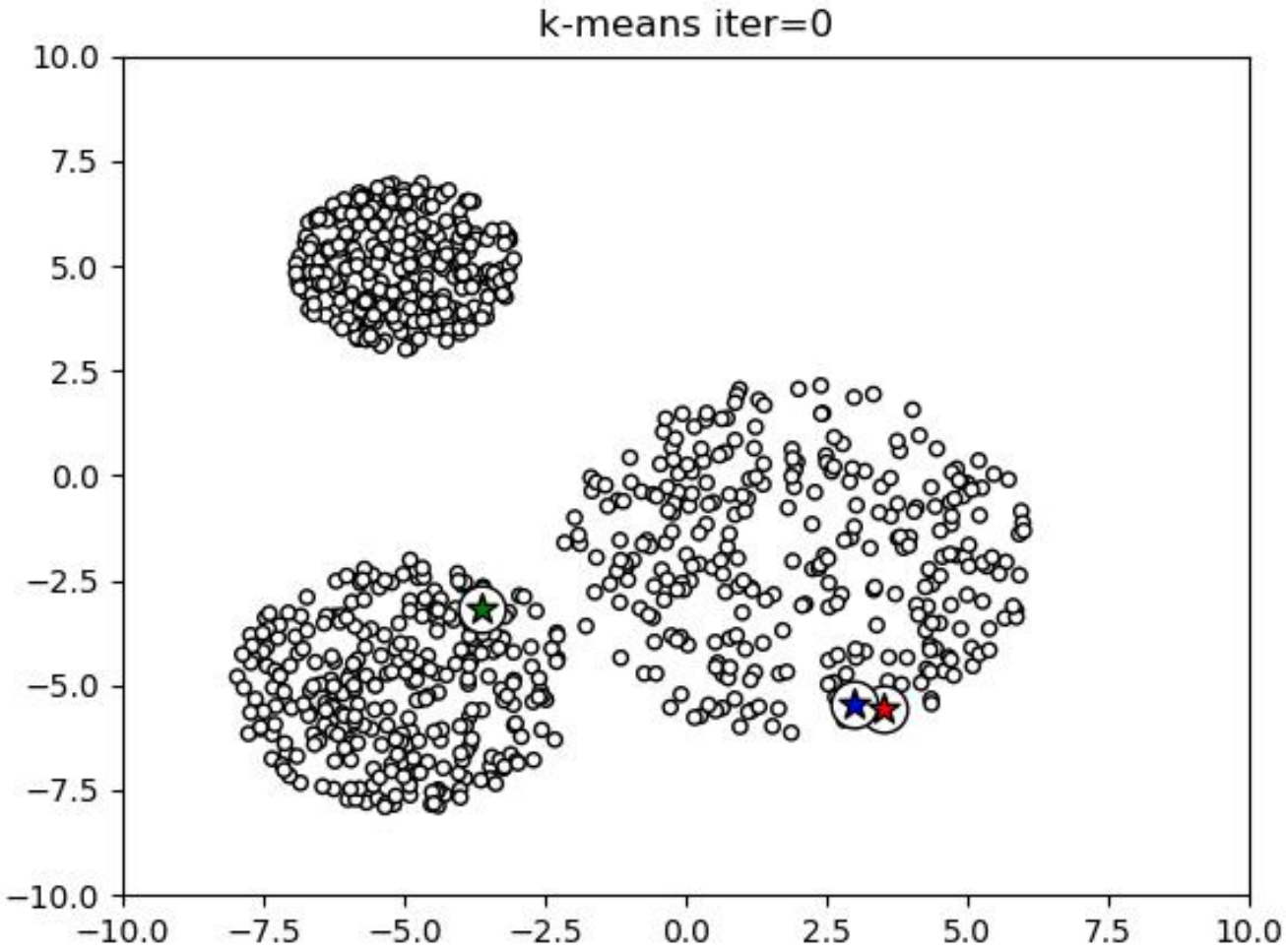
4. Jeżeli nie jest spełnione kryterium stopu – wróć do 2

Algorytm k-means – kryterium stopu

- Brak zmiany przydziału obserwacji do grup (brak zmiany wartości zmiennych Z_j)
- Osiągnięcie maksymalnej liczby iteracji
- Brak poprawy funkcji celu RSS lub względna zmiana mniejsza od zadanego progu:

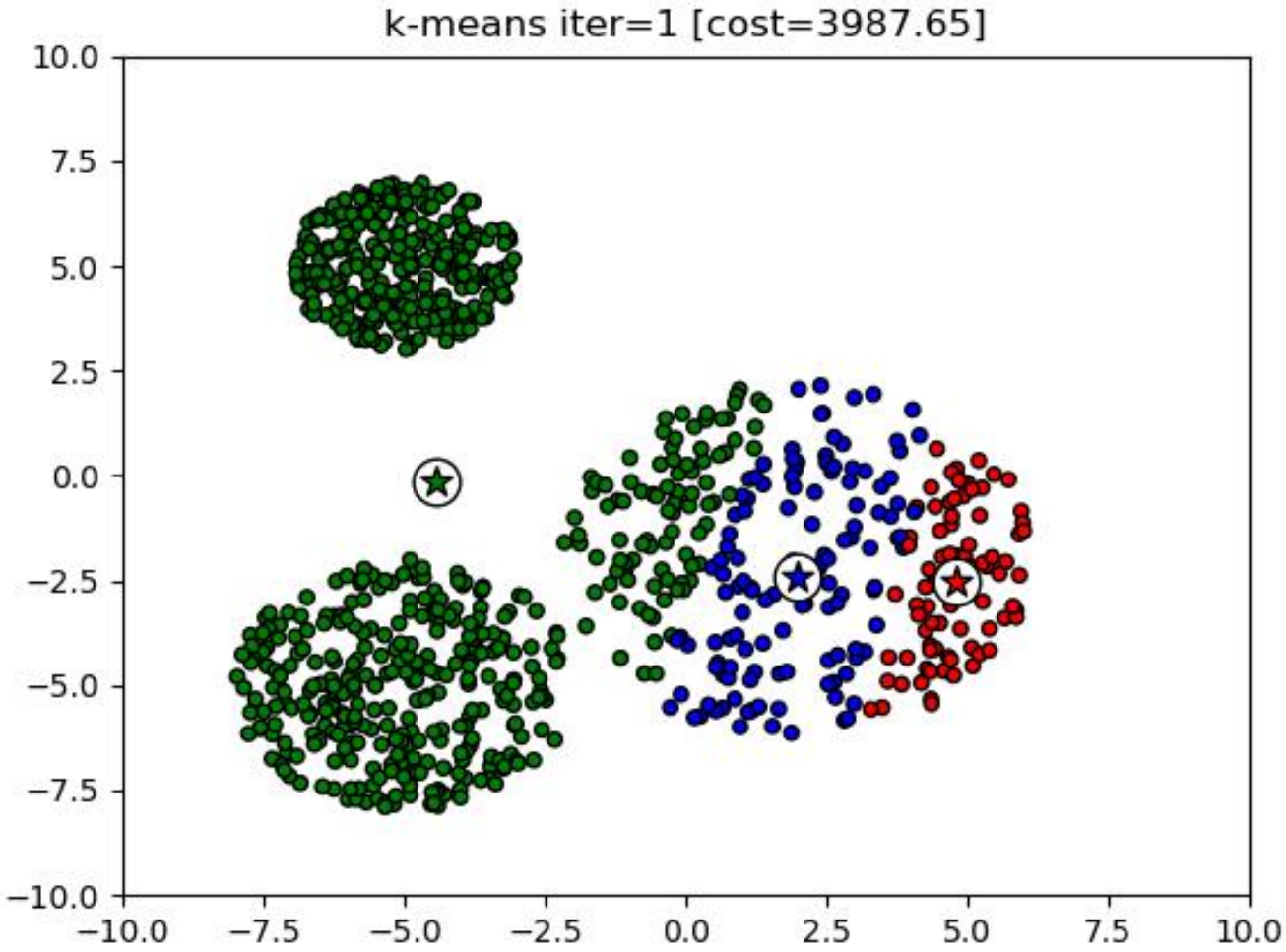
$$\frac{RSS^{(t)} - RSS^{(t-1)}}{RSS^{(t-1)}} < \epsilon$$

Przebieg algorytmu - inicjalizacja



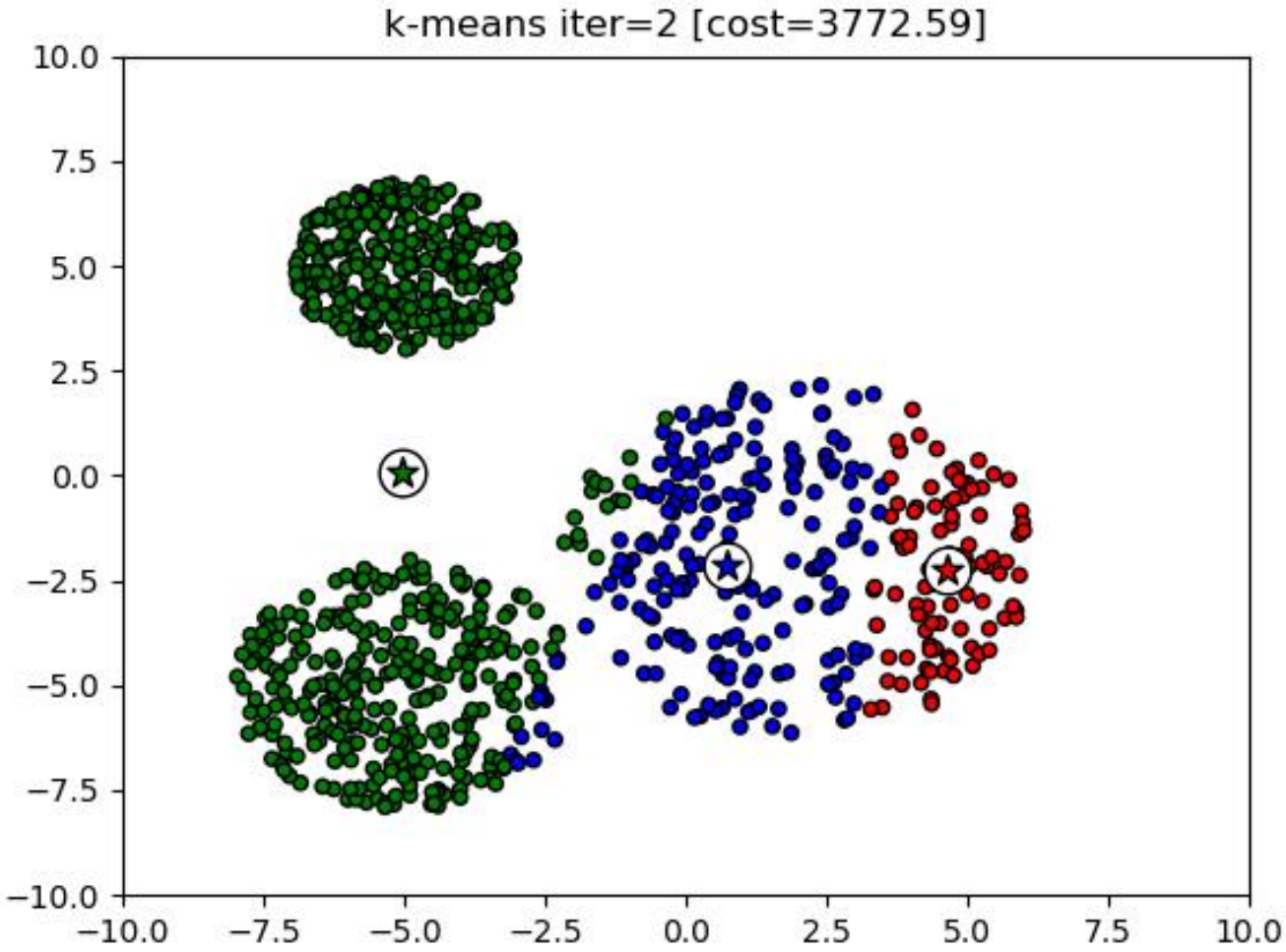
Wylosowano 3 spośród punktów jako środki skupisk

Przebieg algorytmu – po pierwszej iteracji



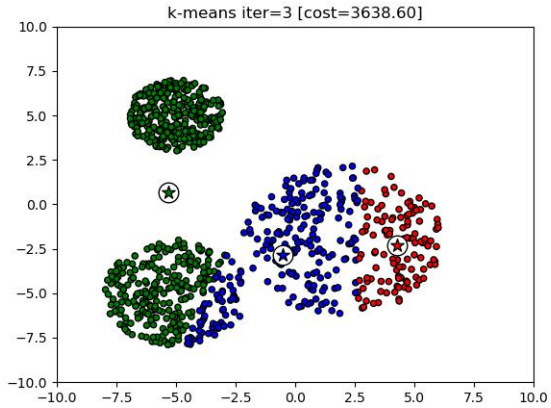
Przesunięcie środków skupisk i uaktualnienie przydziału obserwacji

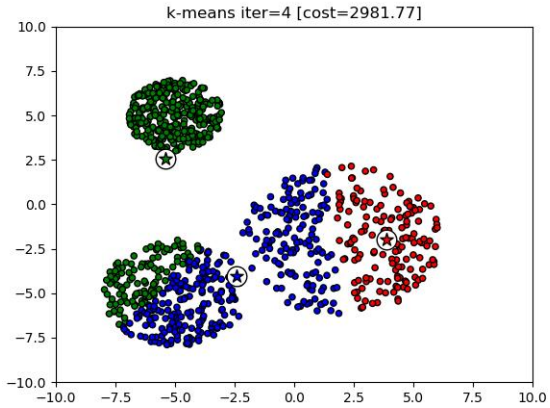
Przebieg algorytmu – po 2 iteracji

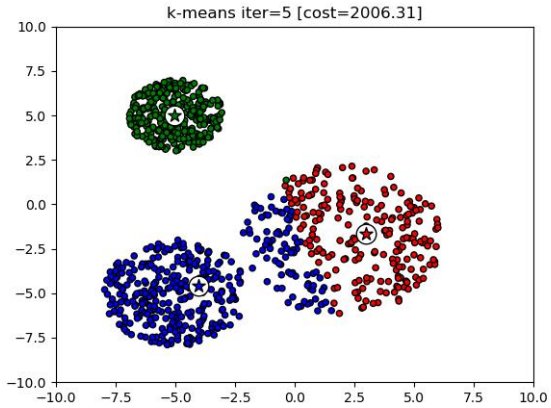


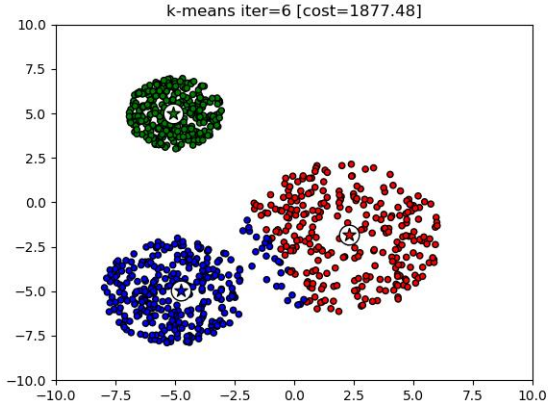
Algorytm jest kontynuowany dopóki następuje zmiana przydziału (etykiet) obserwacji. Zazwyczaj także spada koszt (minumum lokalne)

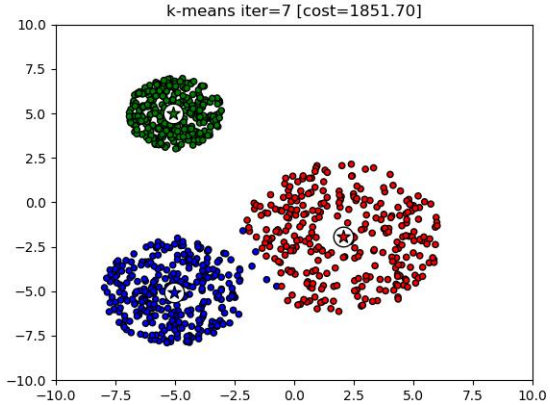
Przebieg algorytmu – kolejne iteracje

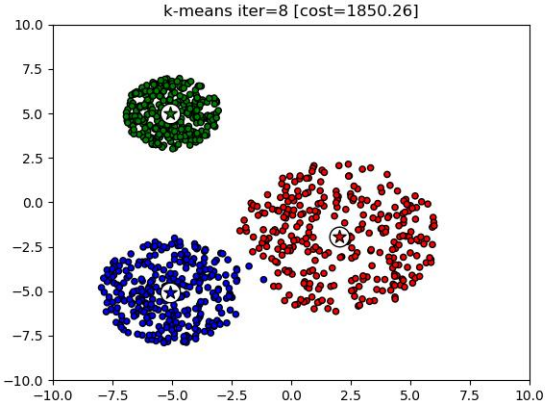




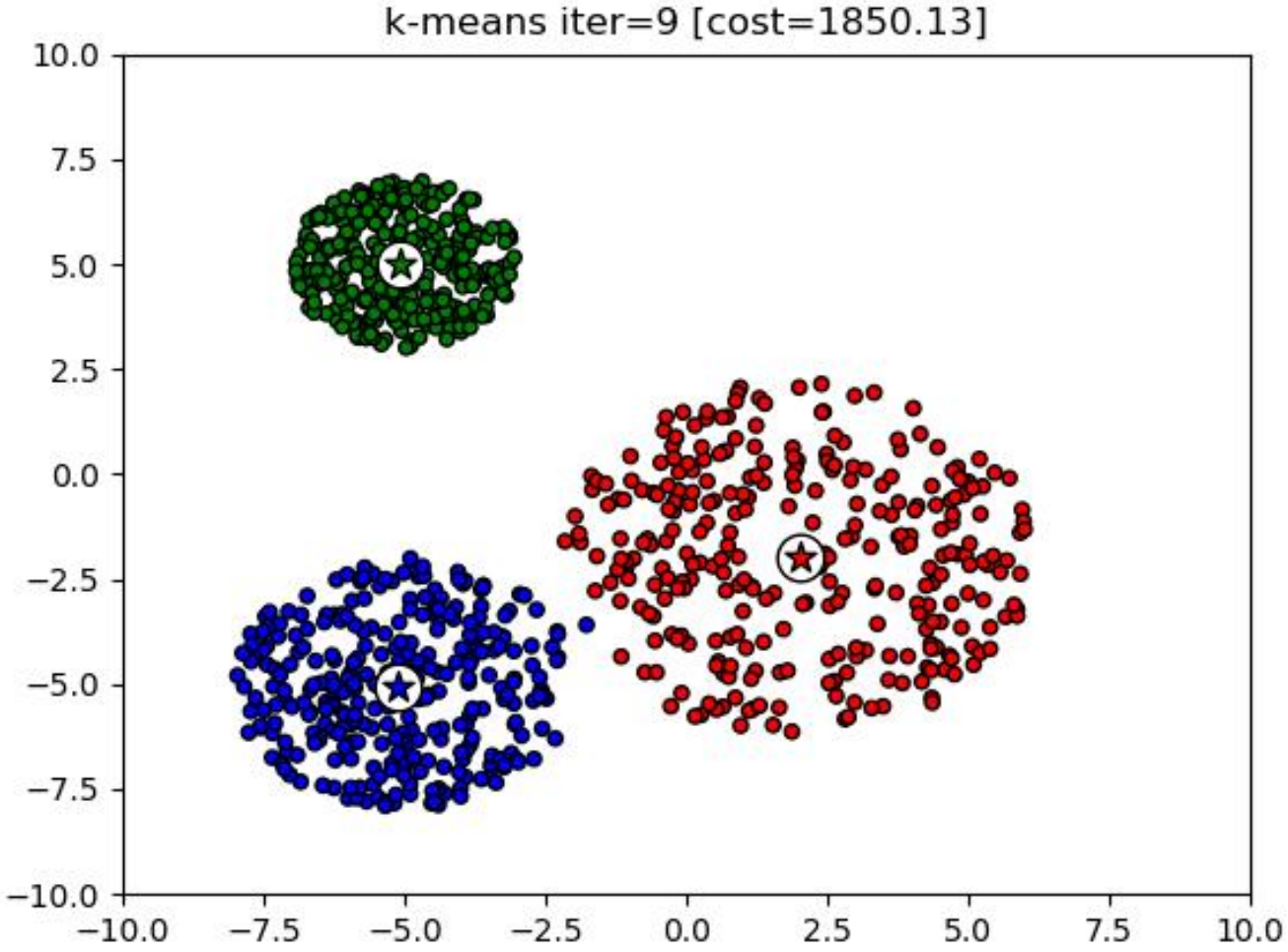






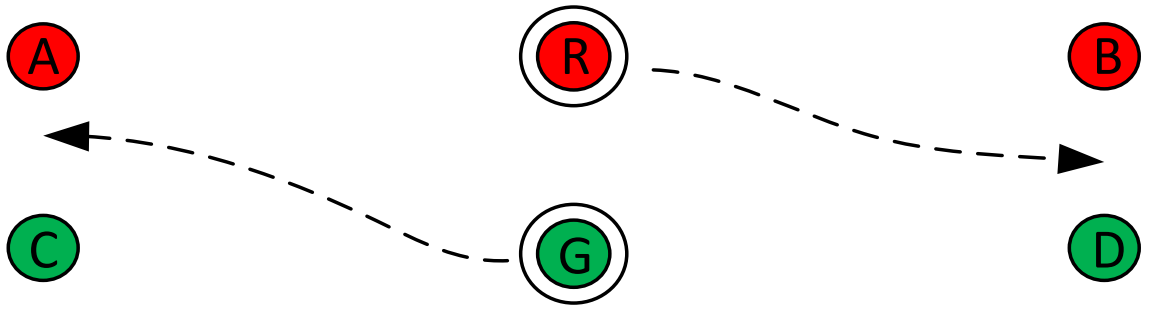


Przebieg algorytmu - koniec



K-means - inicjalizacja

Algorytm jest wrażliwy na wybór początkowych środków skupisk. Niewłaściwy wybór może spowodować utknięcie algorytmu w jednym z minimów lokalnych o bardzo dużej wartości RSS



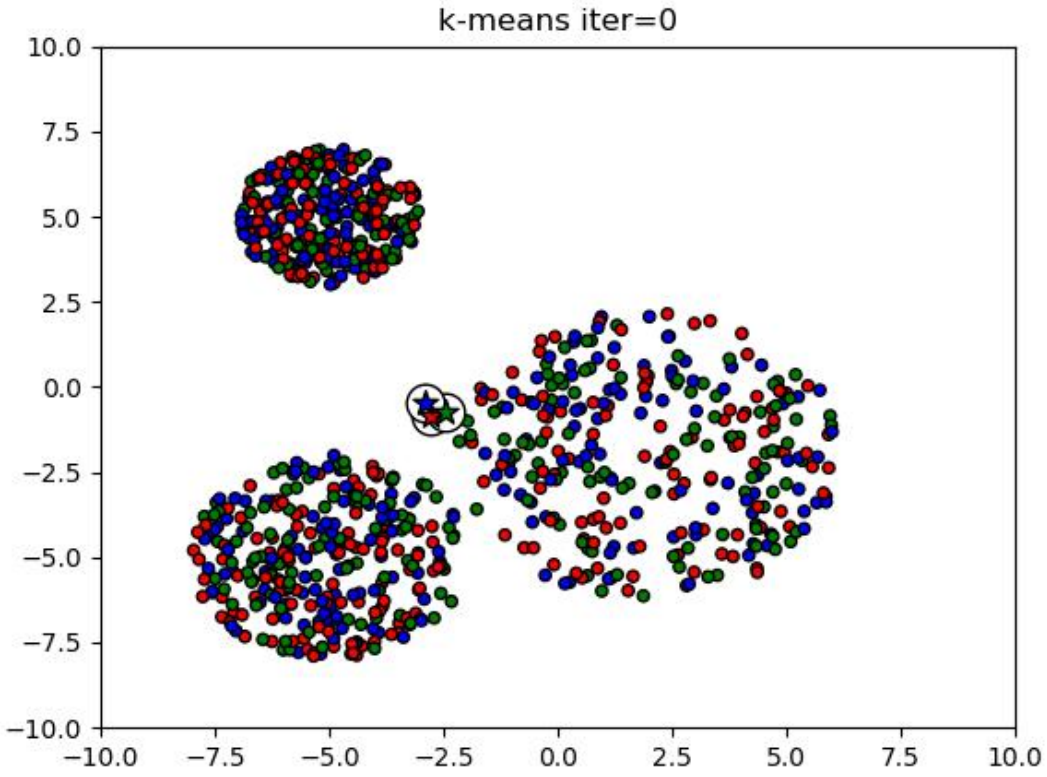
Przykład

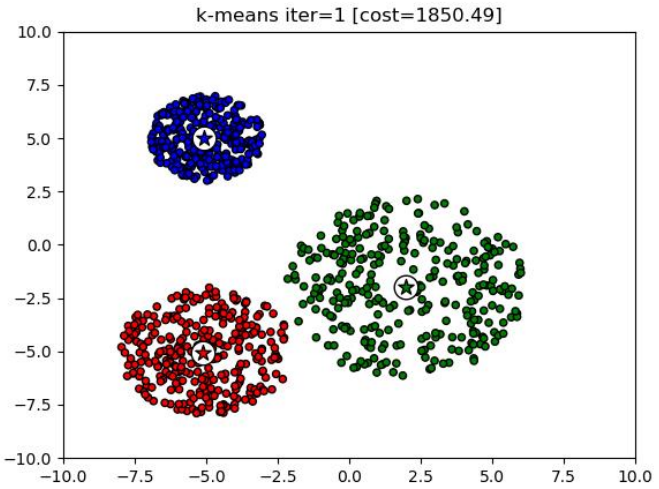
- R i G to centroidy będące środkami grup $\{A, B\}$ i $\{C, D\}$
- Optymalny podział na grupy to $\{A, C\}$ i $\{B, D\}$.
- Jednakże zmiana podziału nie jest możliwa, np. punkt A nigdy nie zmieni grupy, ponieważ $|AG| > |AR|$. Podobnie zachowają się pozostałe punkty B, C i D.

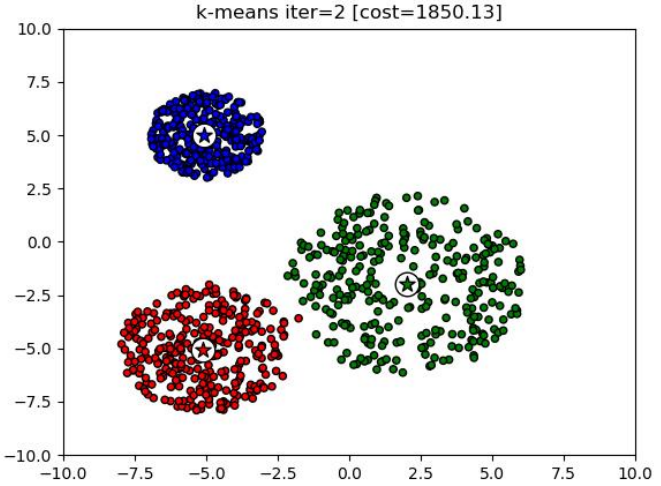
Trzy strategie inicjalizacji

- Losowanie obserwacji (pokazane wcześniej)
- Losowy przydział obserwacji do skupisk
- K-means++
  Losowy wybór kolejnych punktów z prawdopodobieństwem wzrastającym wraz z odległością od wybranych środków

Inicjalizacja – losowy przydział obserwacji







Przy losowaniu przydziału do grup zazwyczaj centroidy są położone w pobliżu środka masy całego zbioru danych

Algorytm inicjalizacji k-means++

1. Przypisz $i \leftarrow 1$. Wylosuj środek $\bar{x}(1)$.
2. Przypisz $i \leftarrow i + 1$
3. Dla każdej obserwacji x_j oblicz najmniejszą odległość od ustalonych środków:

$$d_{min}(x_j) \leftarrow \min\{d(x_j, \bar{x}(r): r = 1, i - 1)\}$$

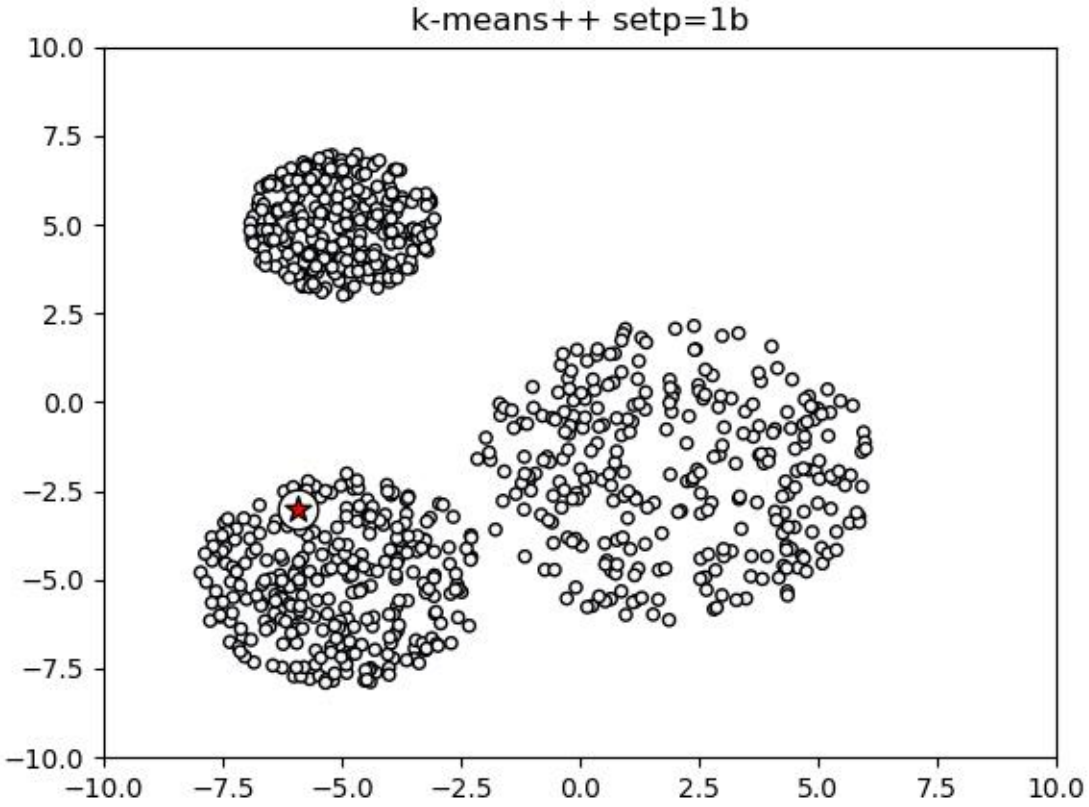
4. Przypisz każdej obserwacji prawdopodobieństwo wyboru

$$P(x_j) \sim d_{min}(x_j)^2$$

5. Wylosuj $\bar{x}(i)$
6. Jeżeli $i = k$, STOP. W przeciwnym przypadku przejdź do 2

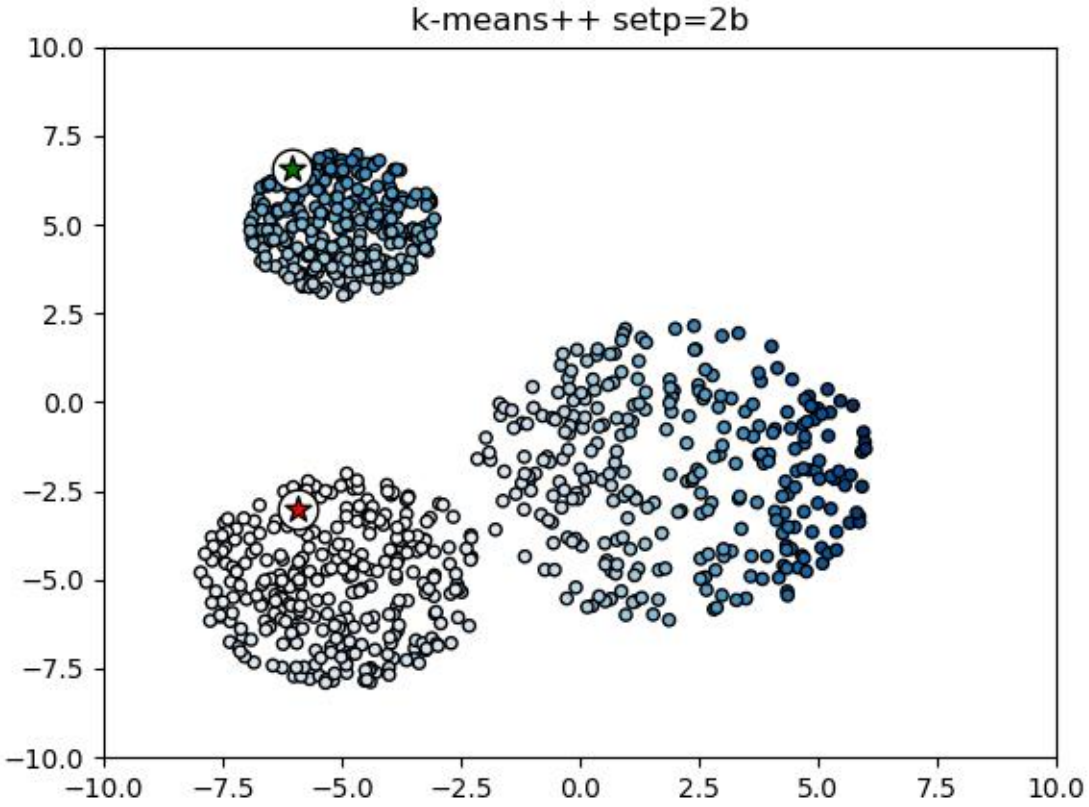
Jest to obecnie najczęściej stosowany algorytm inicjalizacji k-means (z reguły standardowa opcja).

K-means++



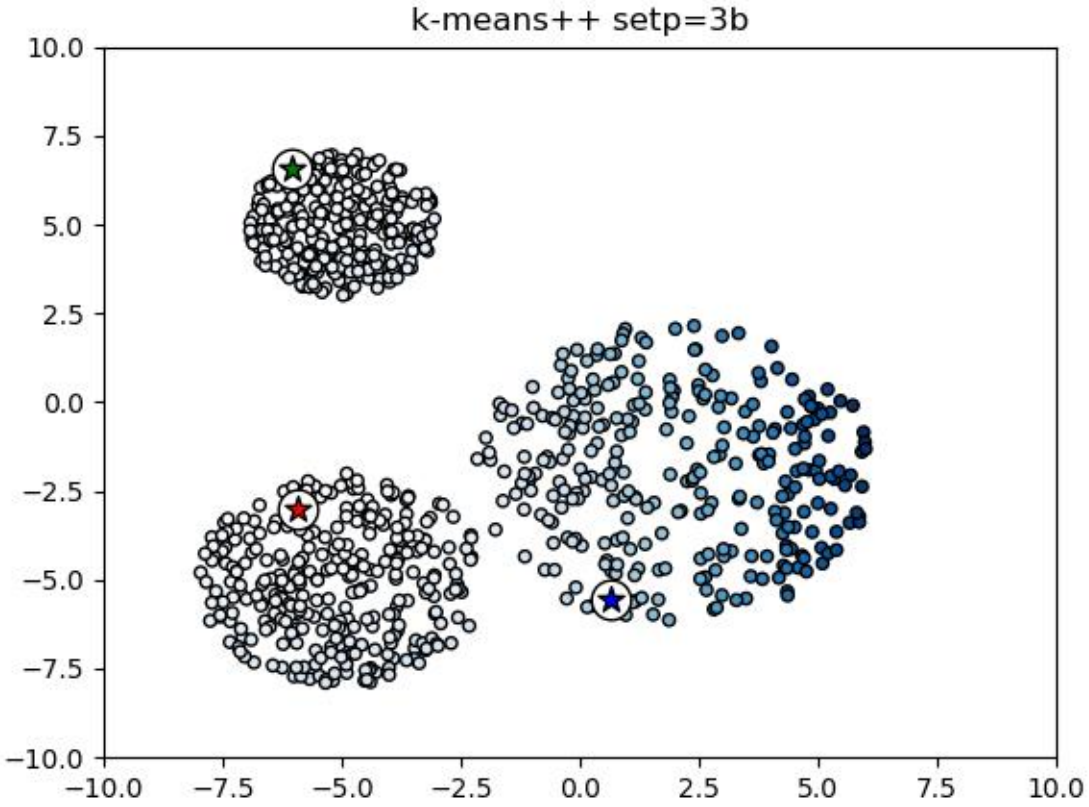
Losowany jest pierwszy punkt centralny…

K-means++



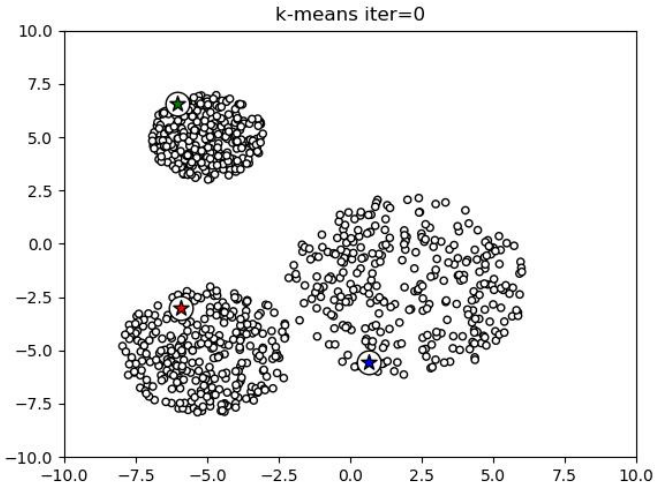
Wylosowany zostaje kolejny środek grupy. Prawdopodobieństwo rośnie wraz z kwadratem odległości od pierwszego punktu centralnego.

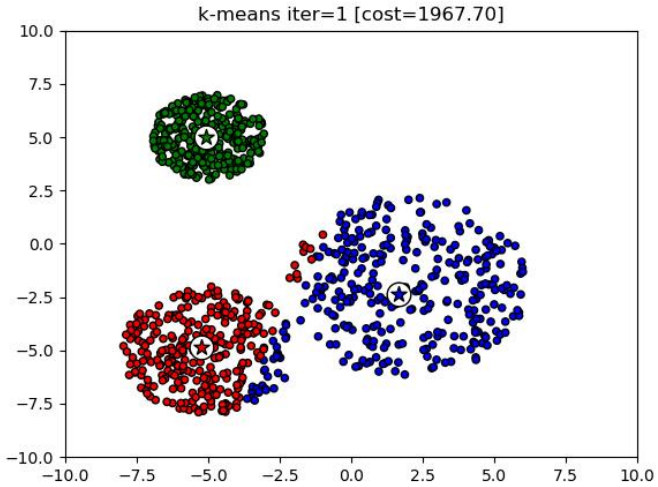
K-means++

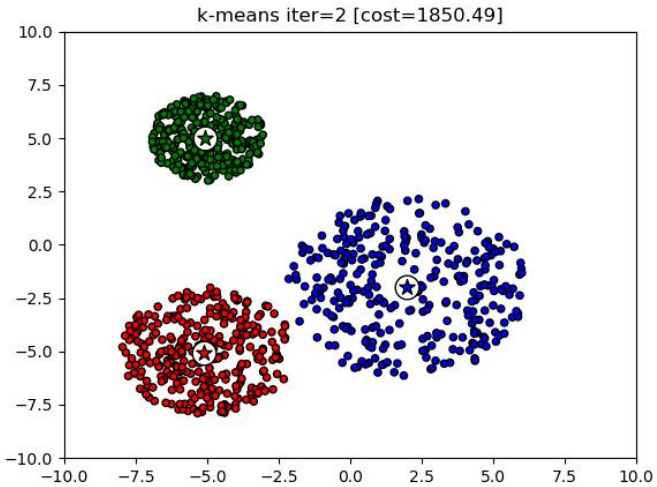


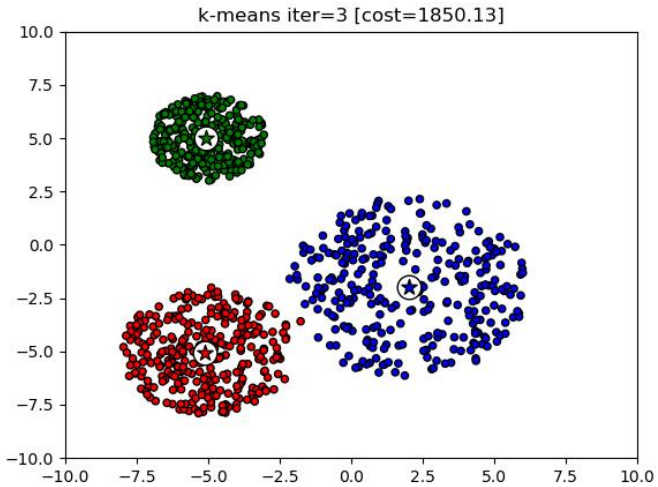
Losowany jest trzeci punkt. Nie musi być to najdalej położona obserwacja (ale prawdopodobieństwo wylosowania dalej położonych punktów jest większe)

Dalszy przebieg algorytmu k-means









EM
Mieszanina rozkładów Gaussa
(ang. Gaussian Mixture Model)

Rozkład Gaussa

- Rozkład Gaussa (normalny) dla $x \in \mathrm{R}$

$$p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

  - μ – wartość oczekiwana (średnia wartość)
  - σ^2 - wariancja

- Rozkład Gaussa dla $x \in \mathrm{R}^n$

$$p(x|\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

  - $\mu \in \mathrm{R}^n$ - wartość oczekiwana dla x, $\mu = E(X)$
  - Σ – macierz kowariancji $\Sigma = E|(X-\mu)(X-\mu)^T|$
  - $|\Sigma|$ - wyznacznik macierzy, musi spełniać $|\Sigma|>0$

Rozkład Gaussa – przypadek 2D

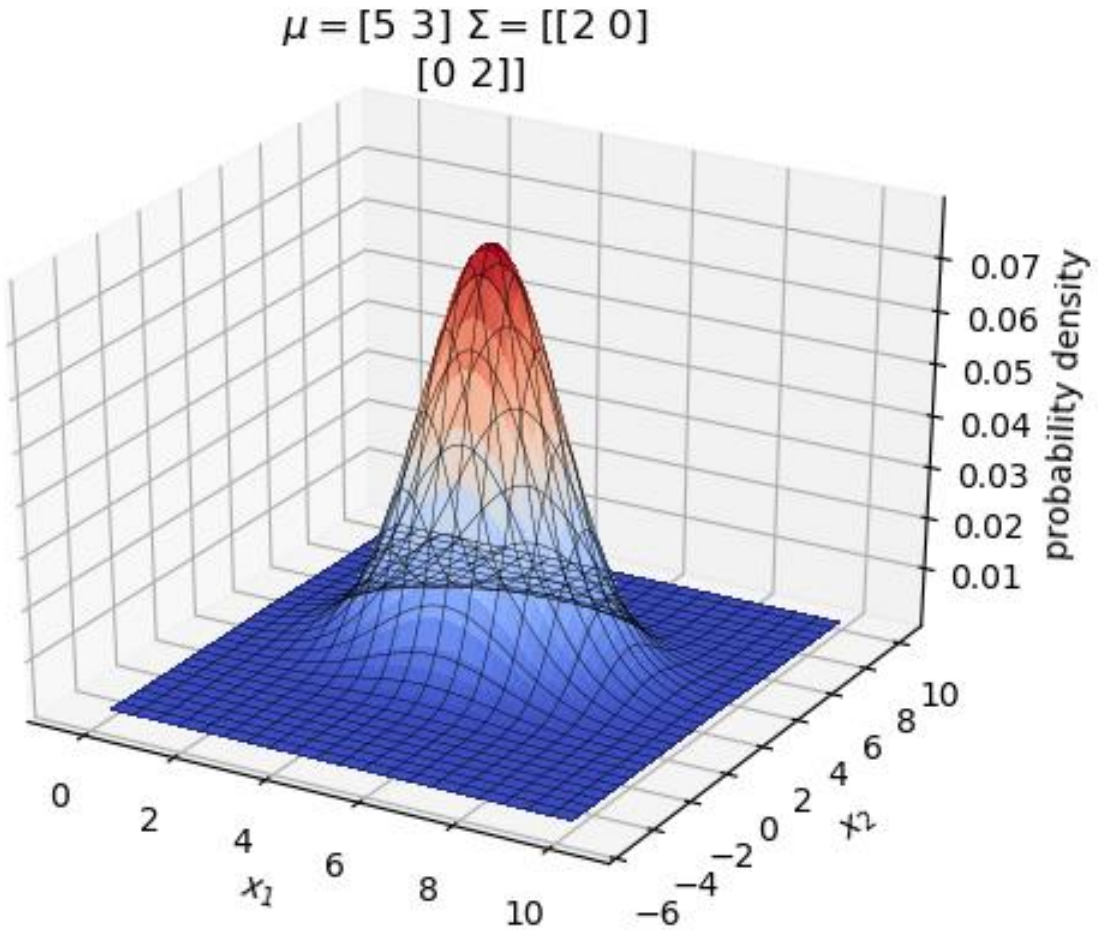
- $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in R^2$
- Środek rozkładu: $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \in R^2$
- Macierz kowariancji $\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & r\sigma_1\sigma_2 \\ r\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$,

  $\sigma_{12} = \sigma_{21}$ - kowariancja X_1 i X_2 (r – to współczynnik Pearsona)
- Punkty, o jednakowej wartości prawdopodobieństwa tworzą elipsę opisaną równaniem:

$$\frac{(x_1-\mu_1)^2}{\sigma_1^2} - 2r\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} = const$$

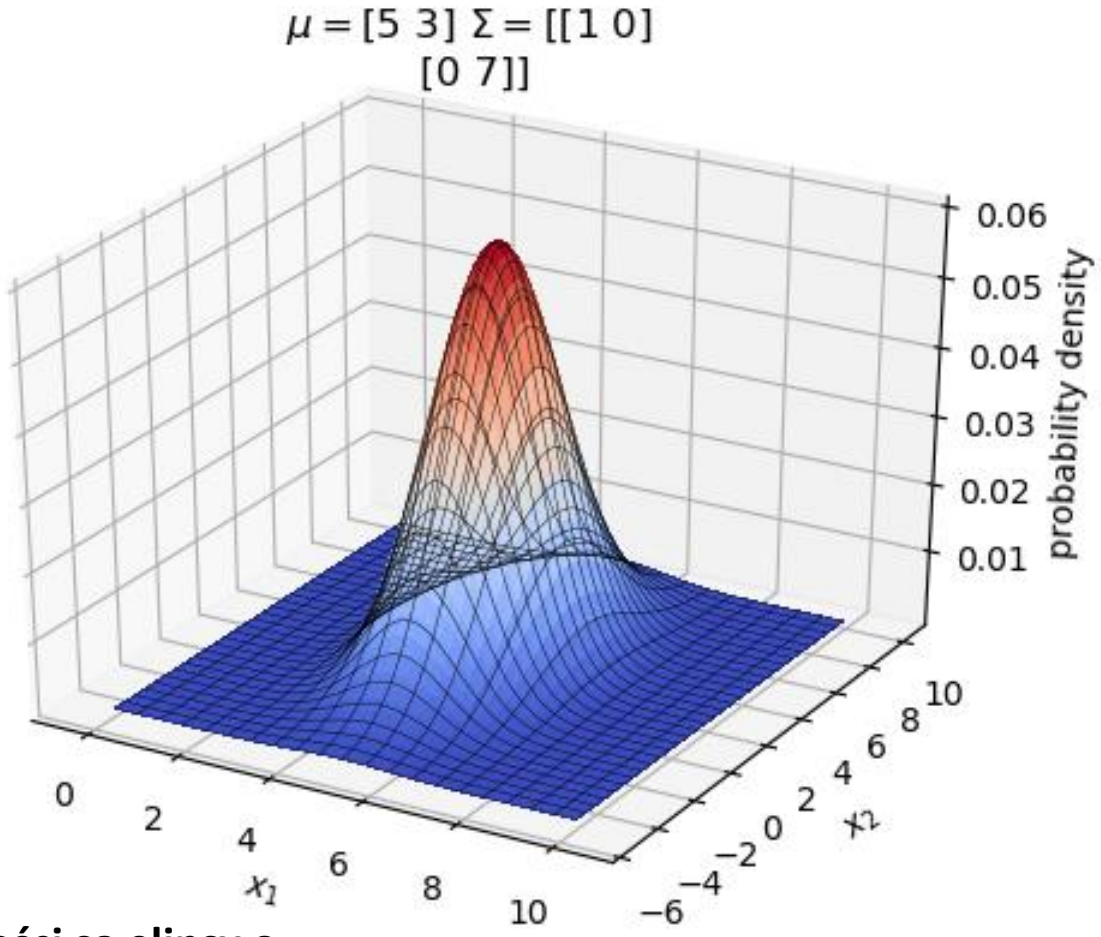
- Długości osi zależą od wartości σ_1 oraz σ_2
- Kąt obrotu zależy od wartości współczynnika Pearsona r (lub korelacji)

Przykład: $\sigma_1^2 = \sigma_2^2$



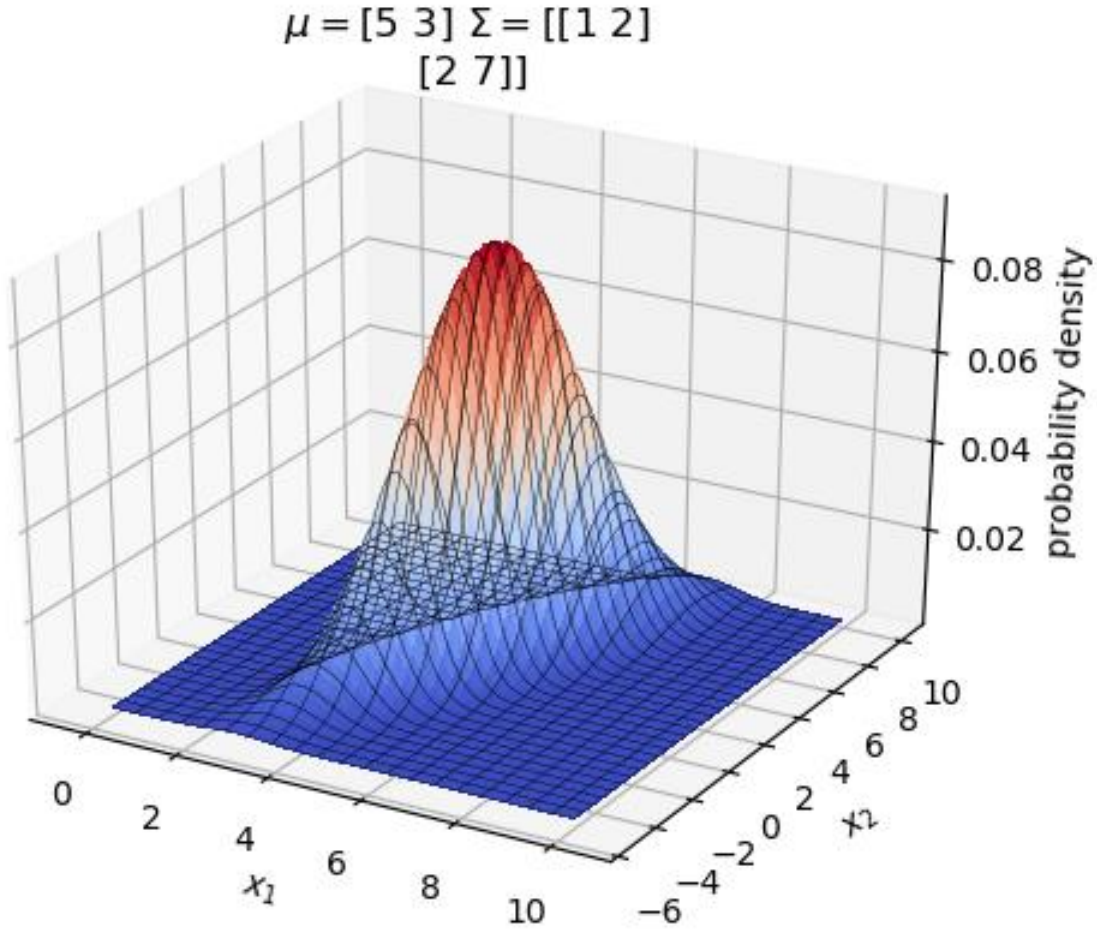
Przekrojami funkcji gęstości są okręgi

Przykład: $\sigma_1^2 \neq \sigma_2^2, r = 0$



Przekrojami funkcji gęstości są elipsy o osiach zgodnych z kierunkami x_1 i x_2

Przykład: $\sigma_1^2 \neq \sigma_2^2, r \neq 0$

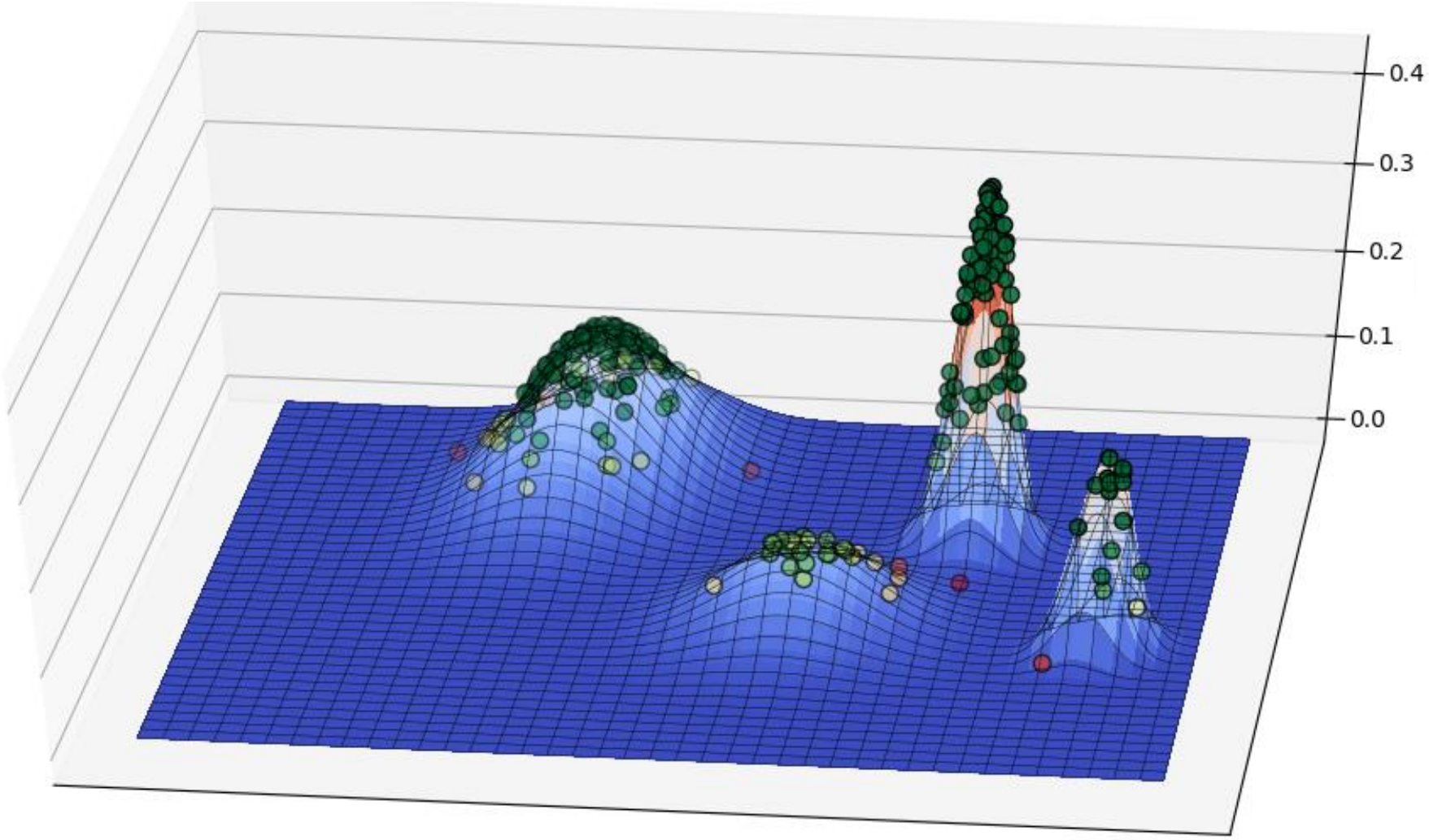


Przekrojami funkcji gęstości są obrócone elipsy

Mieszanina k rozkładów Gaussa

- Załóżmy, że model składa się z k klas $C_1,\ldots,C_k$, które mogą generować dane z prawdopodobieństwem opisanym rozkładem Gaussa $p(x|\mu_i, \Sigma_\mathrm{i})$. Modelem klasy jest więc para $\theta_i = (\mu_i, \Sigma_\mathrm{i})$.
- Proces losowania zaobserwowanych danych $x_j \in \mathrm{R}^n$ przebiega następująco:
  - Wpierw losowana jest i-ta klasa z prawdopodobieństwem π_i
  - Następnie losowany jest punkt x_j zgodnie z rozkładem
$$p(x|\mu_i, \Sigma_\mathrm{i}) = p(x|\theta_i)$$
- Rozkład prawdopodobieństwa $f(x|\Theta)$ jest sumą rozkładów przeskalowanych przez współczynniki π_i (prawdopodobieństwa a priori klas) spełniające: $\forall i: \pi_i \geq 0$ oraz $\sum_{i=1}^{k} \pi_i = 1$
$$p(x|\Theta) = \sum_{i=1}^{k} \pi_i p(x|\theta_i)$$
- Pełne parametry modelu
$$\Theta = (\pi_1, \ldots, \pi_k, \theta_1, \ldots, \theta_k)$$
obejmują zarówno prawdopodobieństwa π_i, jak i parametry rozkładów Gaussa $\theta_i = (\mu_i, \Sigma_\mathrm{i})$.

Przykład 2D: mieszanina k rozkładów Gaussa



Mieszanina k rozkładów Gaussa (podział przestrzeni)

- Jeżeli znany jest pełny model $\Theta = (\pi_1, \dots, \pi_k, \theta_1, \dots, \theta_k)$, wówczas prawdopodobieństwo przynależności dowolnego punktu $x \in R^n$ do grupy C_i można obliczyć ze wzoru Bayesa:

$$P(C_i|x) = \frac{\pi_i \, p(x|\theta_i)}{\sum_{j=1}^{k} \pi_j \, p(x|\theta_j)}$$

  - π_i to prawdopodobieństwo a priori
  - θ_j to model dla C_i
  - mianownik $\sum_{j=1}^{k} \pi_j \, p(x|\theta_j)$ to prawdopodobieństwo zaobserwowania x dla modelu Θ
- Na podstawie tego prawdopodobieństwa -- lub wyłącznie korzystając z licznika $\pi_i \, p(x|\theta_i)$ - można wyznaczyć naturalny podział przestrzeni na grupy

$$C_i = \left\{ x : \pi_i \, p(x|\theta_i) > \max_{j \neq i} \pi_j \, p(x|\theta_j) \right\}$$

Mieszanina k rozkładów Gaussa

- Zakładając, że obserwacje ze zbioru danych $D = \{x_j\},\ j = 1, \dots, m$ są niezależne, prawdopodobieństwo jego zaobserwowania, dla danego zestawu parametrów Θ wynosi:

$$P(D|\Theta) = \prod_{j=1}^{m} p(x_j|\Theta) = \prod_{j=1}^{m} \left(\sum_{i=1}^{k} \pi_i \cdot p(x_j|\theta_i) \right)$$

$$= \prod_{j=1}^{m} \left(\sum_{i=1}^{k} \pi_i \cdot \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)} \right)$$

- Parametry modelu Θ są jednak nieznane, natomiast znane są obserwacje D. Prawdopodobieństwo (ang. likelihood), że znane obserwacje D powstały jako wynik losowania dla badanego modelu Θ ma taką samą wartość

$$L(\Theta|D) = P(D|\Theta)$$

Mieszanina k rozkładów Gaussa

- Po zastosowaniu logarytmu (zamiana iloczynu małych wartości prawdopodobieństwa na sumę ich logarytmów) otrzymujemy log-likehood

$$LL(\Theta|D) = \ln \prod_{j=1}^{m} \left(\sum_{i=1}^{k} \pi_i \cdot \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)} \right) =$$

$$\sum_{j=1}^{m} ln \left(\sum_{i=1}^{k} \pi_i \cdot \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)} \right)$$

oraz przepis na optymalną wartość parametrów

$$\Theta^* = \arg \max_{\Theta} LL(\Theta|D)$$

- Niestety logarytm naturalny sumy jest trudny do optymalizacji…

Algorytm EM 1

- W algorytmie EM zakłada się, że model zawiera m (m to liczba danych) ukrytych zmiennych $Z = \{Z_j\}$.
- Zmienne te przyjmują wartości ze zbioru $\{1, \dots, k\}$. Jeżeli $Z_j = i$, to oznacza, że obserwacja x_j została wygenerowana przez i-ty komponent mieszaniny zgodnie z rozkładem

$$\pi_i \cdot p\left(x_j \middle| \theta_i\right) = \pi_i \cdot \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

(czyli x_j należy też do i-tej grupy).

Algorytm EM 2

- Wówczas zamiast maksymalizować

$$L(\Theta|D) = P(D|\Theta)$$

maksymalizuje się

$$L(\Theta|D, Z) = P(D, Z|\Theta)$$

- Przekształcając:

$$P(D, Z|\Theta) = \frac{P(D,Z,\Theta)}{P(\Theta)} = \frac{P(Z|D,\Theta)\cdot P(D,\Theta)}{P(\Theta)} = P(Z|D,\Theta)\cdot P(D|\Theta)$$

Pozwala to skonstruować iteracyjną procedurę, złożoną z dwóch kroków:

Expectation. Wpierw wyznacza się prawdopodobieństwa $P\left(Z\middle|D,\Theta^{(t-1)}\right)$ poszczególnych wartości zmiennych Z na podstawie obserwacji ze zbioru D i parametrów z modelu poprzedniej iteracji $\Theta^{(t-1)}$

Maximization. Następnie poszukuje się nowego modelu $\Theta^{(t)}$ maksymalizującego prawdopodobieństwo $P(D|\Theta^{(t)})$ przy ustalonym w fazie E rozkładzie zmiennych Z

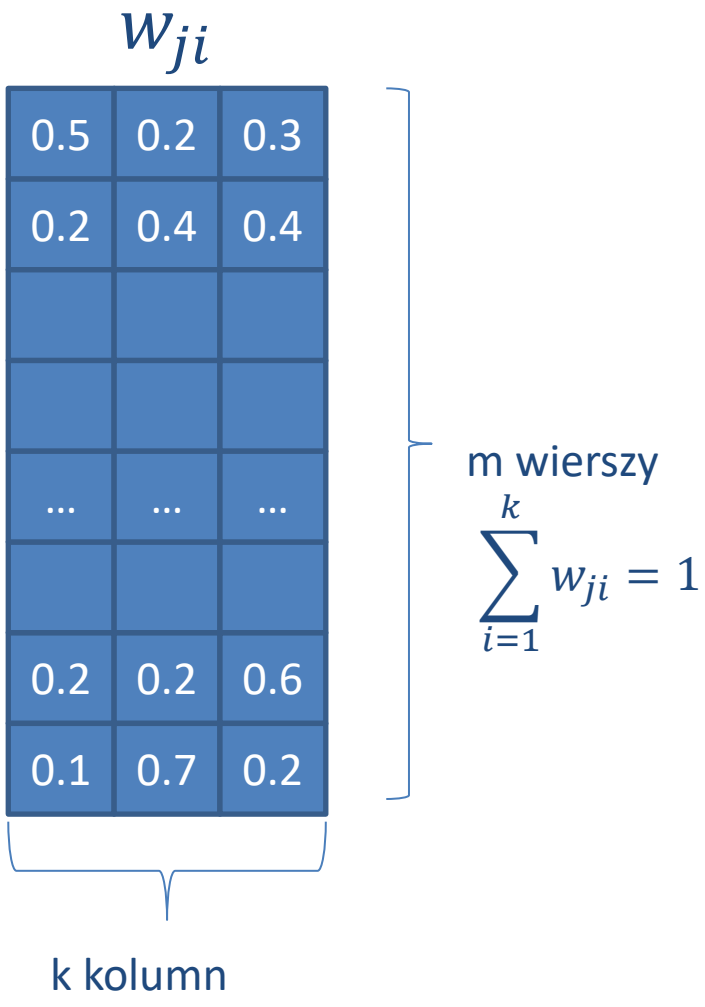
Expectation

- Oznaczmy przez w_{ji} prawdopodobieństwo, że $Z_j = i$ (czyli, że obserwacja x_j została wygenerowana przez i-ty komponent)

$$w_{ji} = \frac{\pi_i \cdot p\left(x_j \middle| \theta_i\right)}{\sum_{l=1}^{k} \pi_l \cdot p\left(x_j \middle| \theta_l\right)}$$

- Wartość w_{ji} może być traktowana jako waga określająca przynależność obserwacji x_j do i-tej grupy. Opisuje ona niepewność, który komponent wygenerował obserwację x_j dla danych parametrów Θ

Maximization



Parametry nowego modelu $\Theta^{(t)}$:

- Wagi (prawdopodobieństwa) komponentów:

$$\pi_i = \frac{\sum_{j=1}^{m} w_{ji}}{m}$$

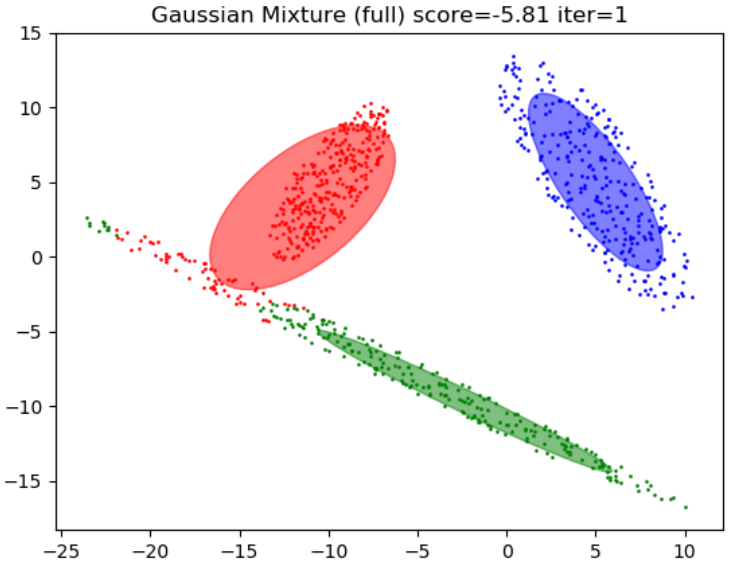
- Środki komponentów

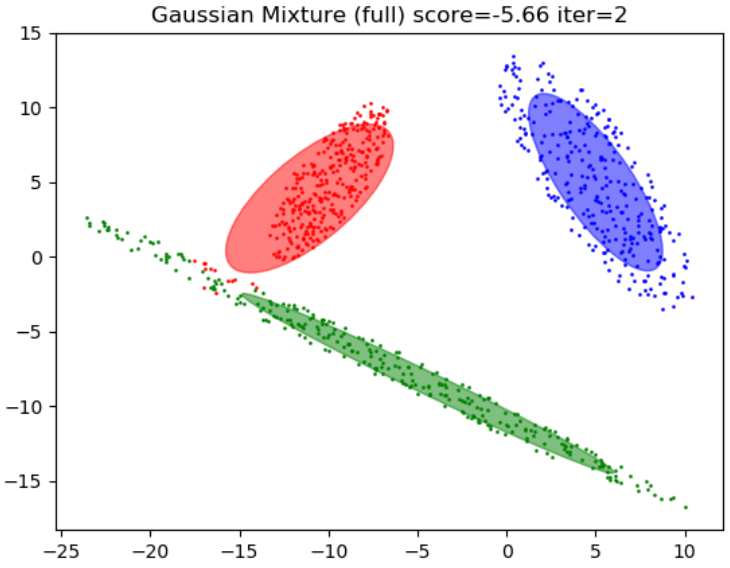
$$\mu_i = \frac{\sum_{j=1}^{m} w_{ji} \cdot x_j}{\sum_{j=1}^{m} w_{ji}}$$

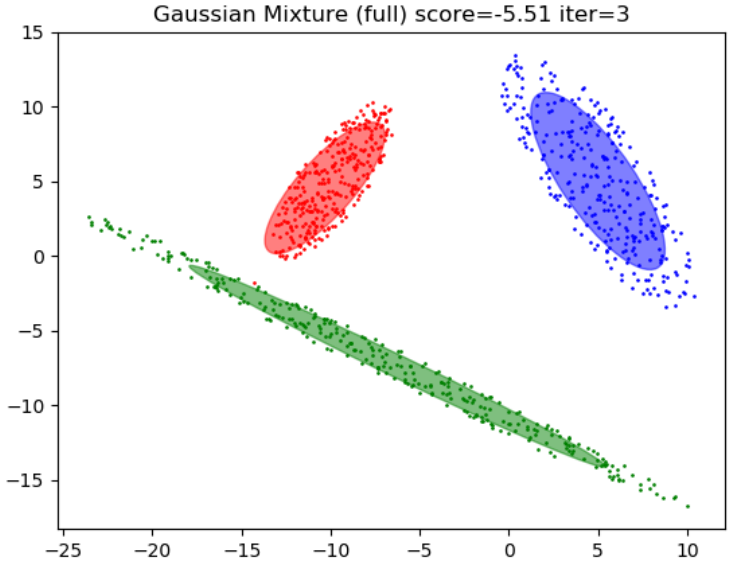
- Macierze kowariancji

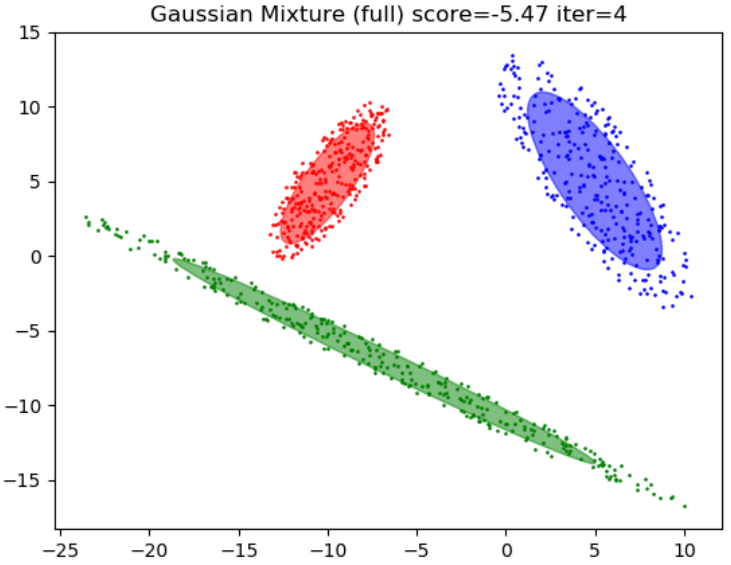
$$\Sigma_i = \frac{\sum_{j=1}^{m} w_{ji} \cdot (x_j - \mu_i)(x_j - \mu_i)^T}{\sum_{j=1}^{m} w_{ji}}$$

Przykładowy przebieg algorytmu









Inicjalizacja i kryterium stopu

- Inicjalizacja możliwa jest na dwa sposoby
  - Wstępny wybór parametrów Θ i przeprowadzenie kroku E
  - Wstępny wybór wag w_{ji} i przeprowadzenie kroku M
- W obu przypadkach dane mogą być wybrane losowo (np. przez wylosowanie punktów środkowych i wartości wariancji) lub też przez zastosowanie jakiegoś algorytmu heurystycznego. Typowe rozwiązanie – użycie k-means.
- Kryteria stopu
  - Brak poprawy funkcji celu

    $$J(\Theta) = \sum_{j=1}^{m} ln\left(\sum_{i=1}^{k} \pi_i \cdot \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}\right)$$

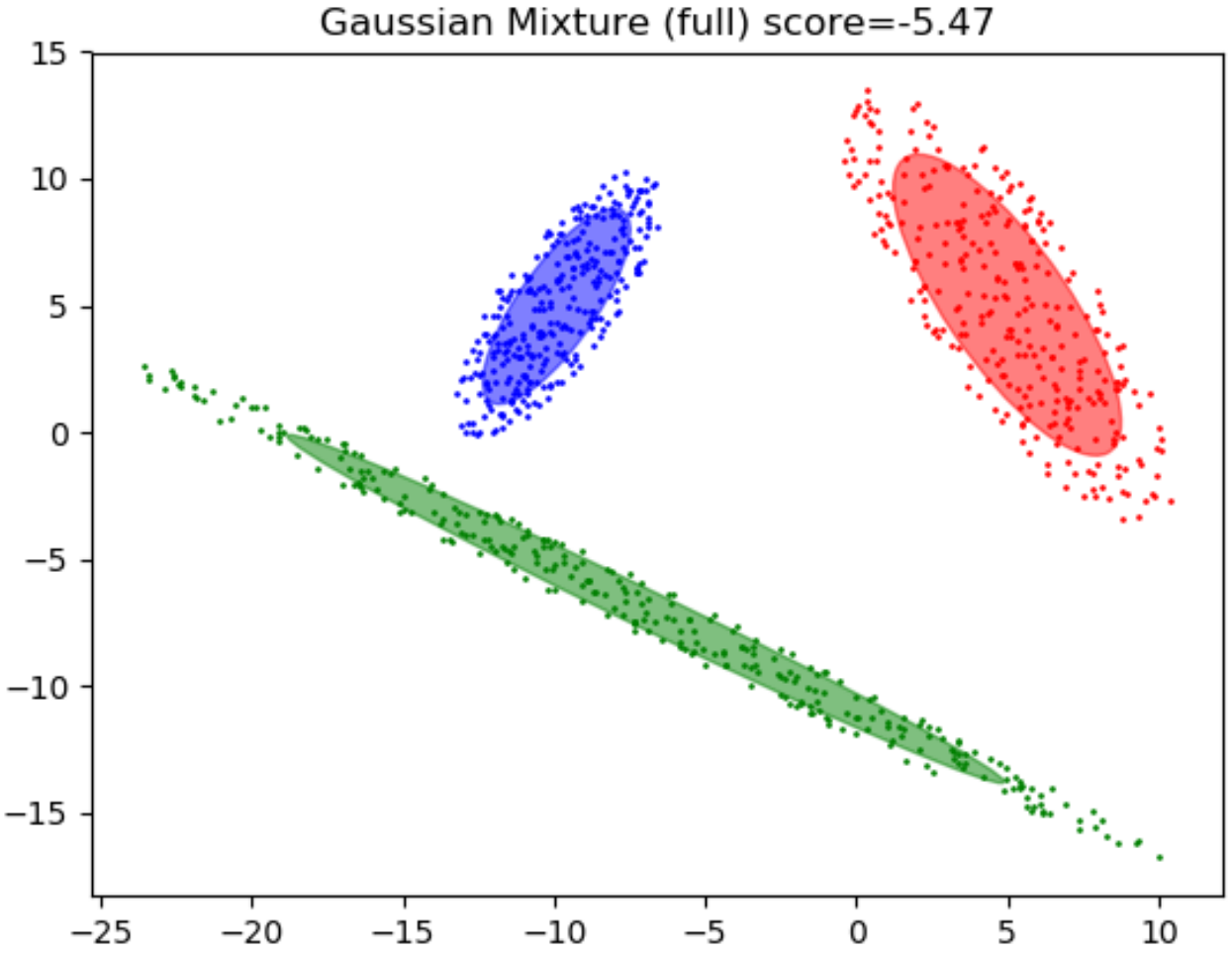
  - Maksymalna liczba iteracji

Postać modelu

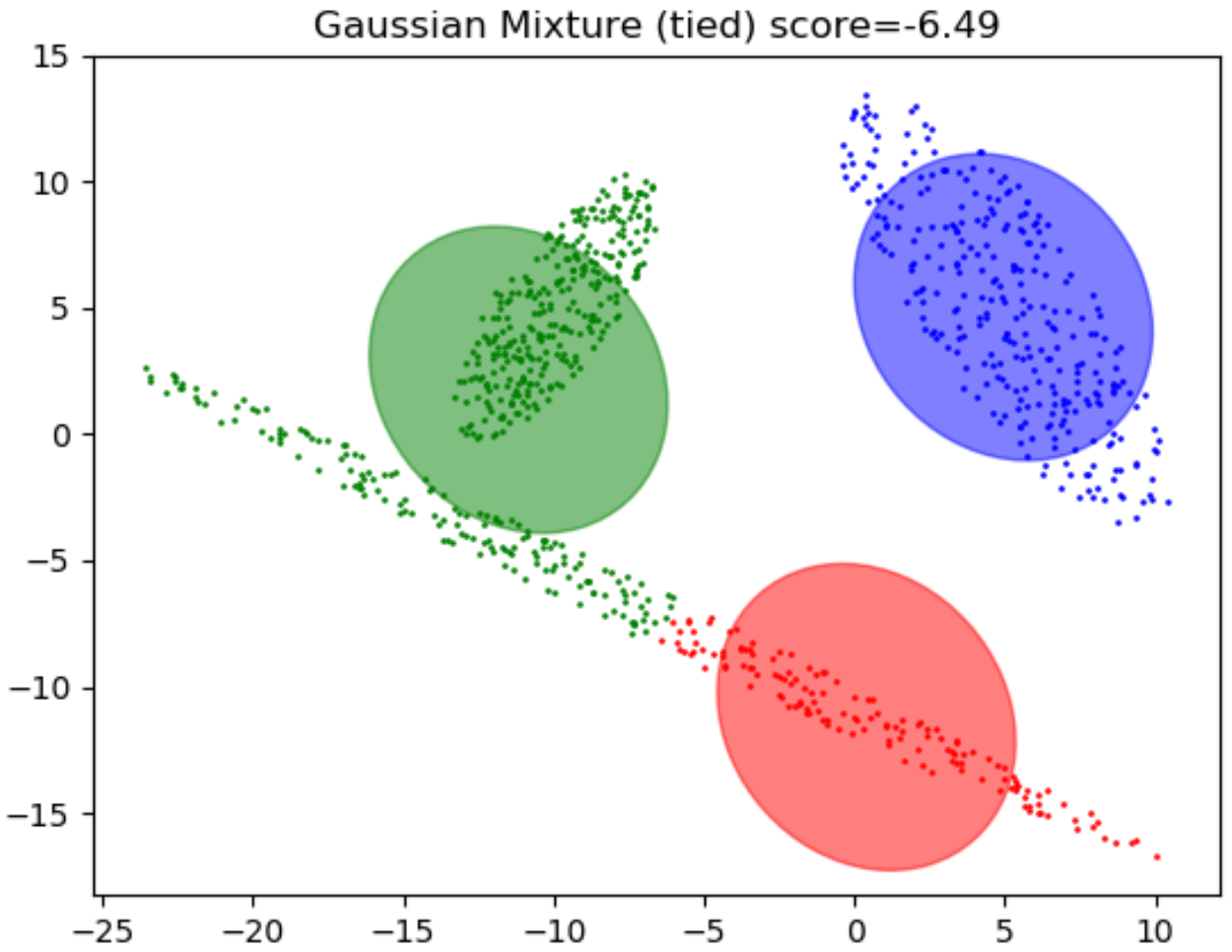
W zależności od implementacji mogą być budowane uproszczone modele.

- Python – wybór pomiędzy
  - full – pełna macierz kowariancji
  - tied – jedna macierz kowariancji dla wszystkich grup
  - diag – diagonalna macierz kowariancji (czyli zawierająca wyłącznie wariancje dla poszczególnych atrybutów)
  - spherical – jednakowa wariancja dla wszystkich wymiarów i grup
- Weka – implementuje wyłącznie model diag

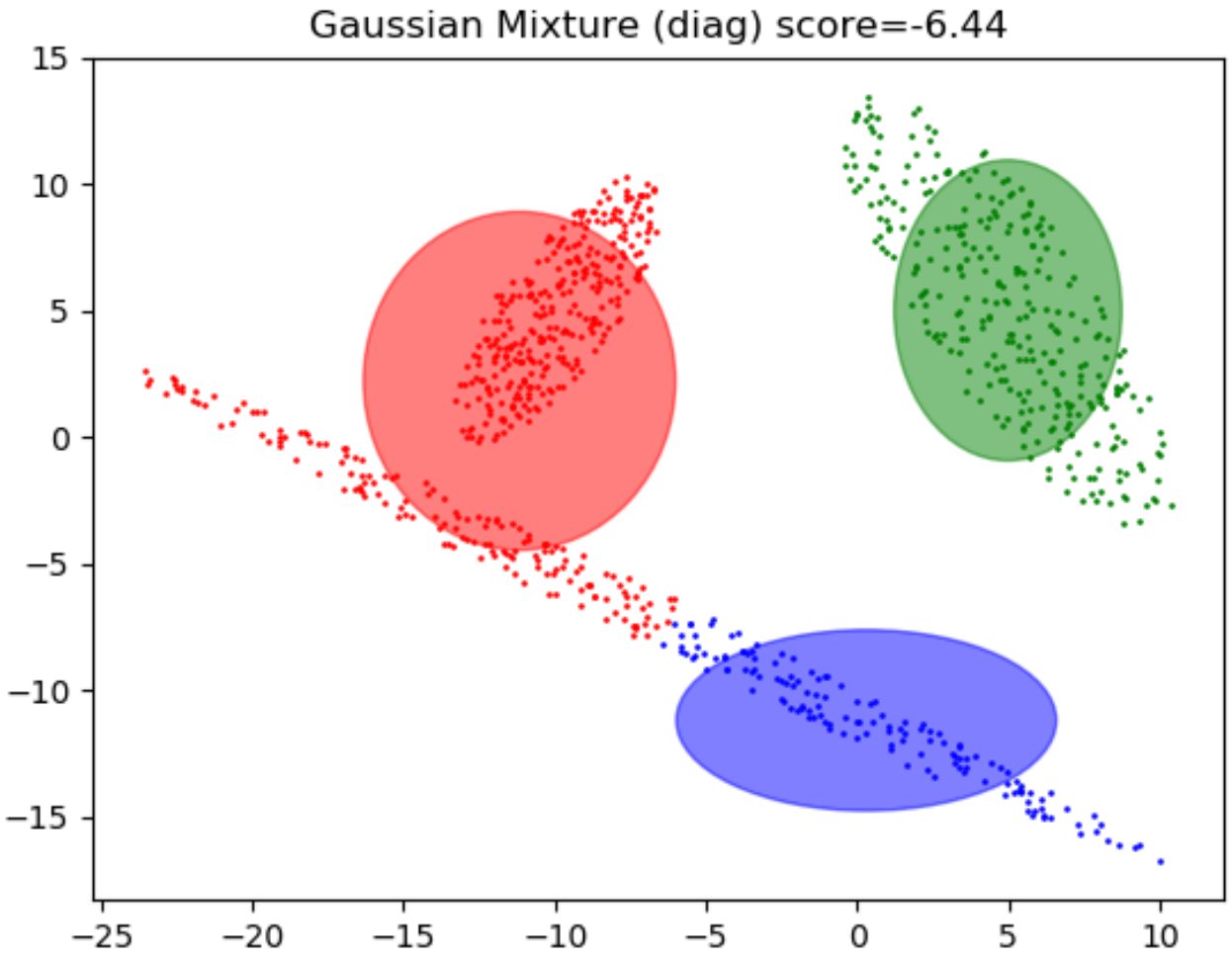
Model full – pełna macierz kowariancji



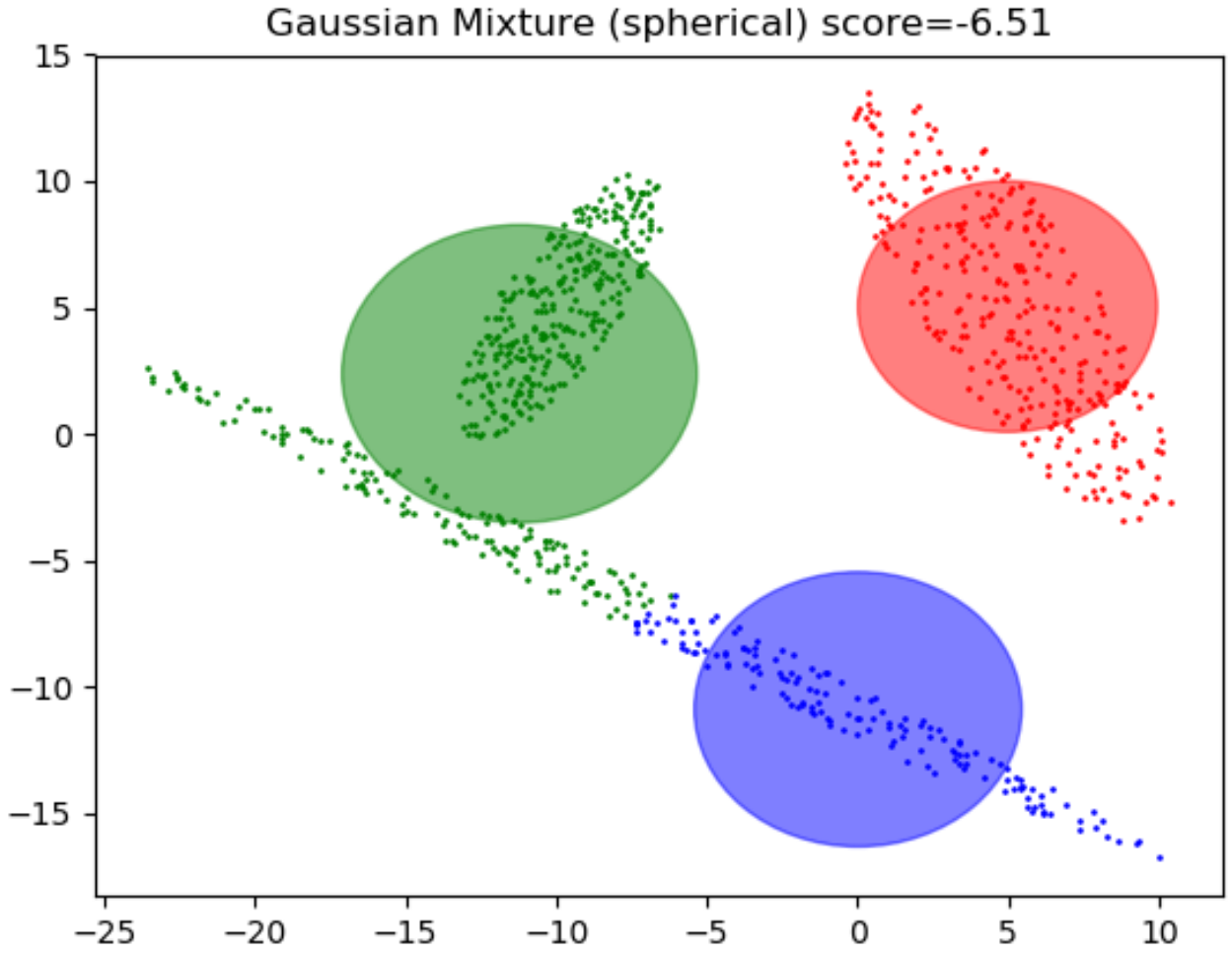
Model tied – jedna macierz kowariancji dla wszystkich grup



Model diag - wyłącznie wariancje dla poszczególnych atrybutów



Model spherical - jednakowa wariancja dla wszystkich wymiarów i grup



K-means vs. EM

- Model spherical mieszaniny Gaussa odpowiada założeniom k-means.
- Dodatkowo , w k-means wariancje muszą być małe (tak, aby kule zmieściły się w komórkach Woronoja)
- W rzeczywistości algorytm k-means może być potraktowany jako uproszczona procedura EM:
  - $\Theta = \{\bar{x}(1), \ldots, \bar{x}(k)\}$
  - **Expectation** – „twardy” przydział obserwacji do grup przy ustalonych położeniach centroidów z poprzedniej iteracji $\Theta^{(t-1)}$
  - **Maximization** – uaktualnienie parametrów modelu $\Theta^{(t)}$ przy założonym przydziale obserwacji do grup

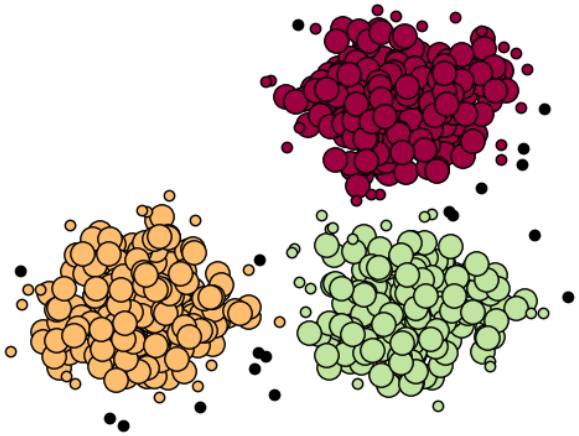
DBSCAN

DBSCAN - założenia

Algorytm DBSCAN (Density-Based Spatial Clustering of Applications with Noise) wykorzystuje informacje o gęstości punktów w przestrzeni do przeprowadzenia grupowania.

Założenia:

- Skupienia mogą zostać zidentyfikowane jako zbiory gęsto (blisko) położonych punktów
- W przestrzeni pomiędzy skupieniami mogą znaleźć się rzadko rozmieszczone punkty, które są uznane za szum.



[http://scikit-learn.org/stable/auto_examples/cluster/plot_dbscan.html#sphx-glr-auto-examples-cluster-plot-dbscan-py]

DBSCAN - definicje

Parametry algorytmu:

- $d(x_1, x_2)$ – odległość pomiędzy obiektami x_1 i x_2
- ε – promień otoczenia obiektu
- $minPts$ – minimalna liczba punktów w otoczeniu

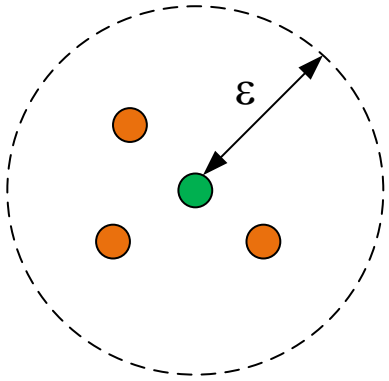
Definicje

- Otoczenie $N(x) = \{z \in \mathrm{D}: d(x, z) < \varepsilon\}$
- Punkt jest punktem **wewnętrznym** (ang. core point), jeżeli jego otoczenie zawiera co najmniej $minPts$ punktów

$$core(x) \Leftrightarrow |N(x)| \geq minPts$$

- Punkt z jest **bezpośrednio osiągalny** (ang. directly reachable) z x, jeżeli x jest punktem wewnętrznym i z należy do sąsiedztwa x

$$x \xrightarrow{\mathrm{dr}} z \Leftrightarrow core(x) \wedge z \in \mathrm{N}(x)$$



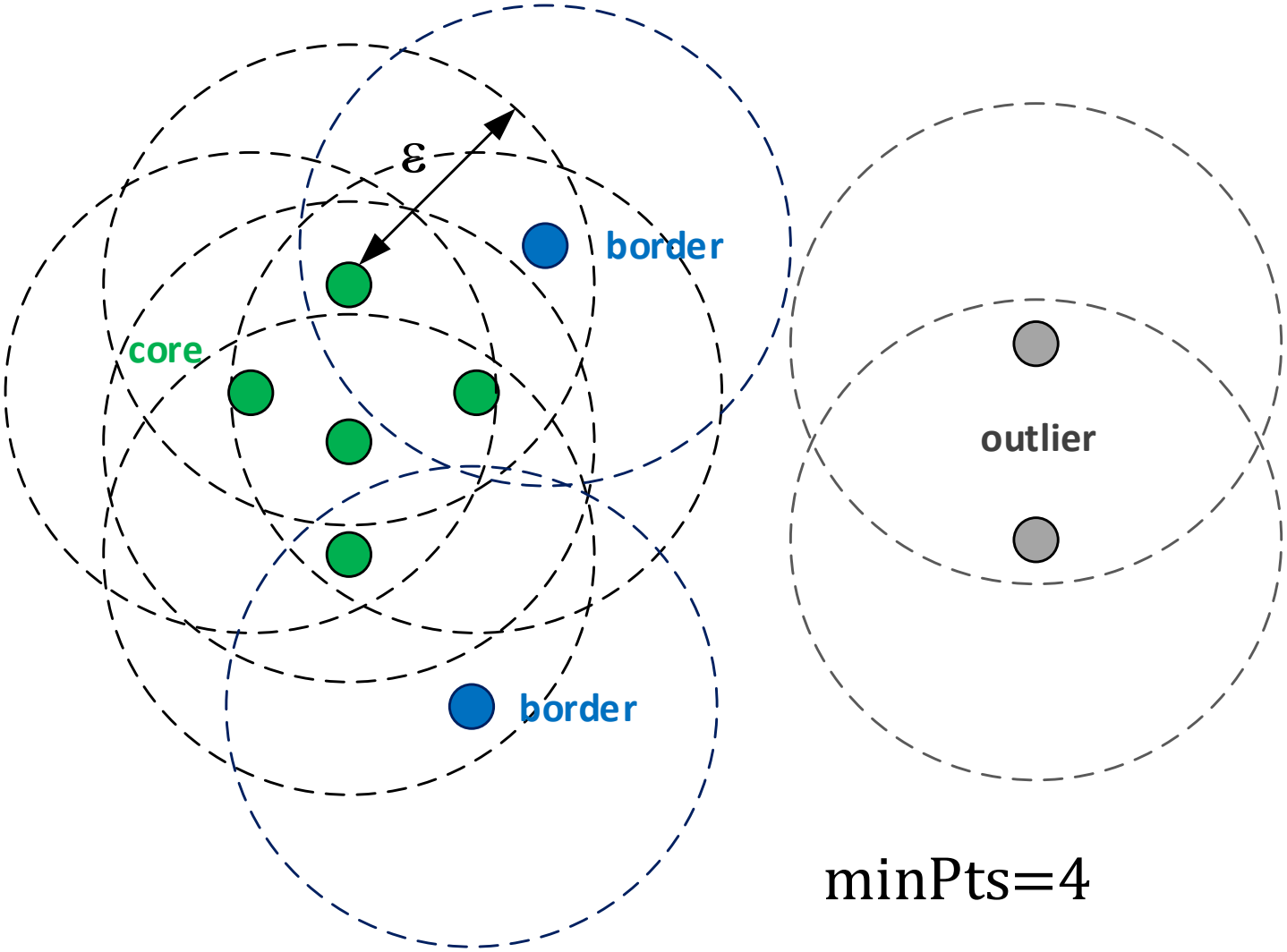
DBSCAN -definicje

- Punkt z jest **osiągalny** z x, jeżeli istnieje ścieżka bezpośrednio osiągalnych punktów prowadząca od x do z

$$x \xrightarrow{\text{dr}} v_1 \xrightarrow{\text{dr}} \ldots \xrightarrow{\text{dr}} \text{v}_\text{k} \xrightarrow{\text{dr}} z$$

  czyli wszystkie punkty $x, v_1, \ldots, v_n$ muszą być punktami wewnętrznymi (mieć co najmniej $minPts$ sąsiadów) i być położone w odległości nie większej niż ε od siebie
- Jeżeli z nie jest punktem wewnętrznym (w jego otoczeniu jest mniej niż $minPts$ punktów), to nazywany jest punktem **brzegowym**.
- Punkty, które nie są osiągalne z żadnego punktu wewnętrznego są punktami poza grupami (outliers)

DBSCAN definicje - ilustracja



DBSCAN - algorytm

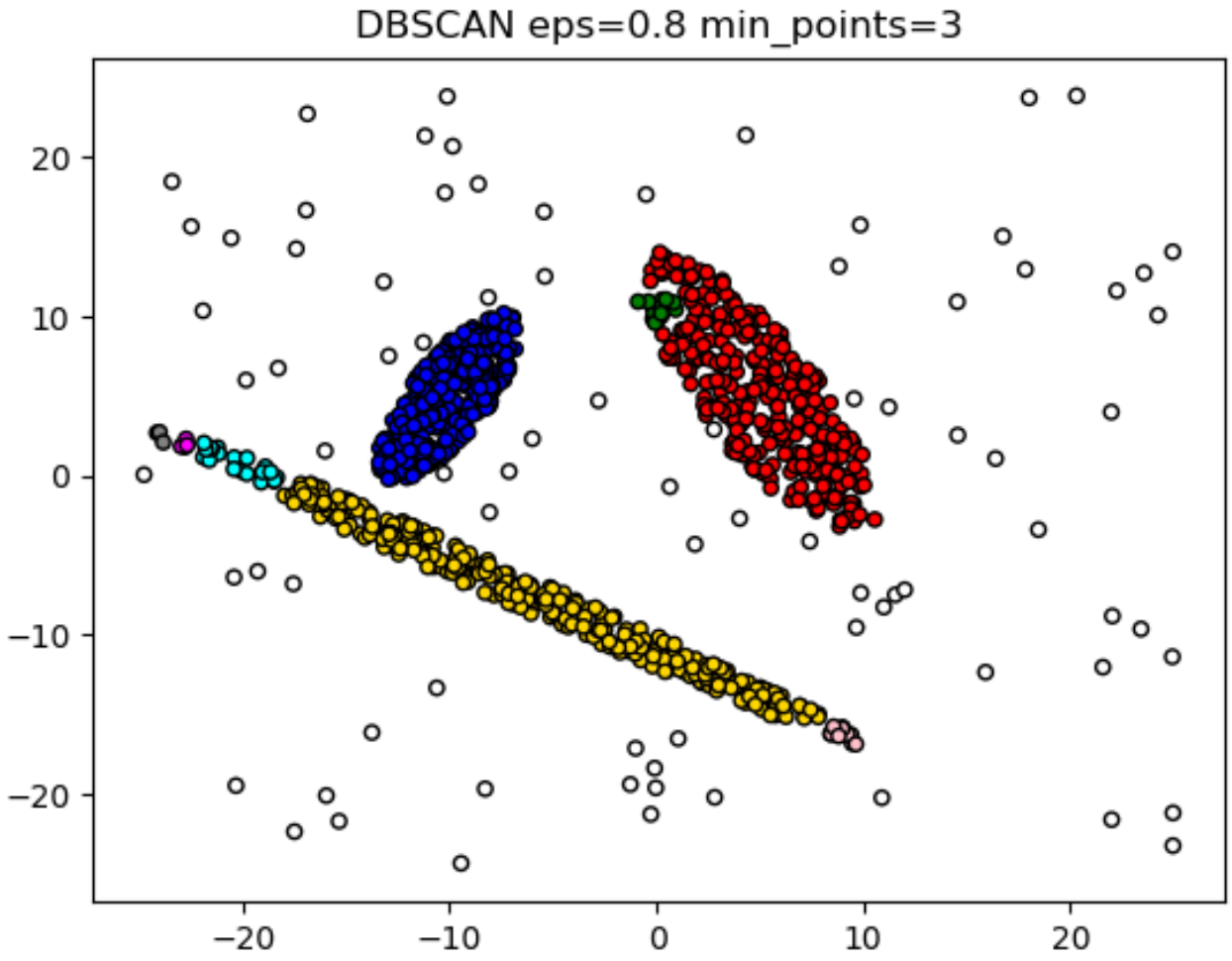
ALGORITHM 1: Pseudocode of Original Sequential DBSCAN Algorithm

```
Input: DB: Database
Input: ε: Radius
Input: minPts: Density threshold
Input: dist: Distance function
Data: label: Point labels, initially undefined
1  foreach point p in database DB do                          // Iterate over every point
2      if label(p) ≠ undefined then continue                  // Skip processed points
3      Neighbors N ← RangeQuery(DB, dist, p, ε)               // Find initial neighbors
4      if |N| < minPts then                                   // Non-core points are noise
5          label(p) ← Noise
6          continue
7      c ← next cluster label                                 // Start a new cluster
8      label(p) ← c
9      Seed set S ← N \ {p}                                   // Expand neighborhood
10     foreach q in S do
11         if label(q) = Noise then label(q) ← c
12         if label(q) ≠ undefined then continue
13         Neighbors N ← RangeQuery(DB, dist, q, ε)
14         label(q) ← c
15         if |N| < minPts then continue                      // Core-point check
16         S ← S ∪ N
```

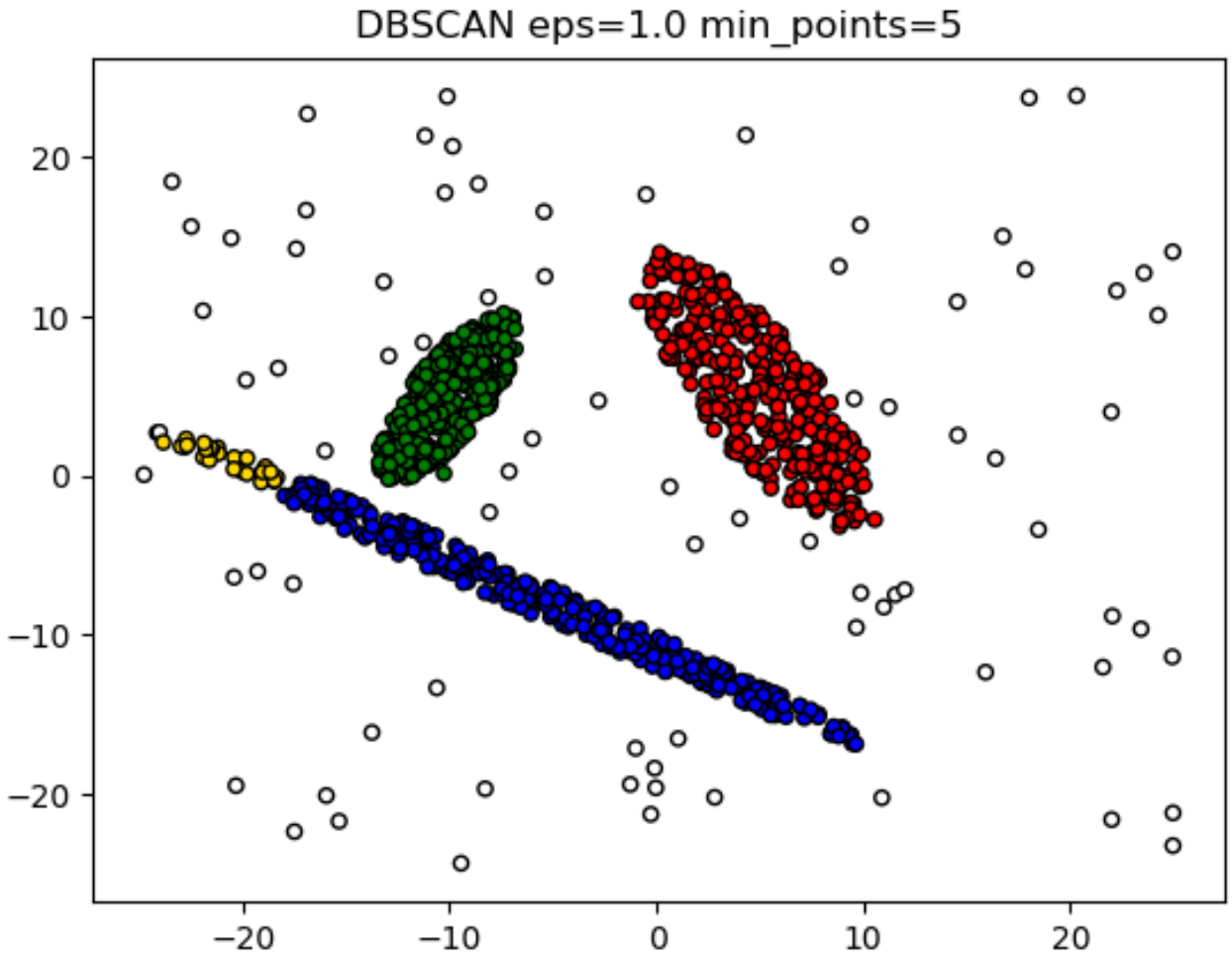
Niejednoznaczność dla punktów brzegowych

Schubert, Erich; Sander, Jörg; Ester, Martin; Kriegel, Hans Peter; Xu, Xiaowei (July 2017). "DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN". ACM Trans. Database Syst. 42 (3): 19:1–19:21.

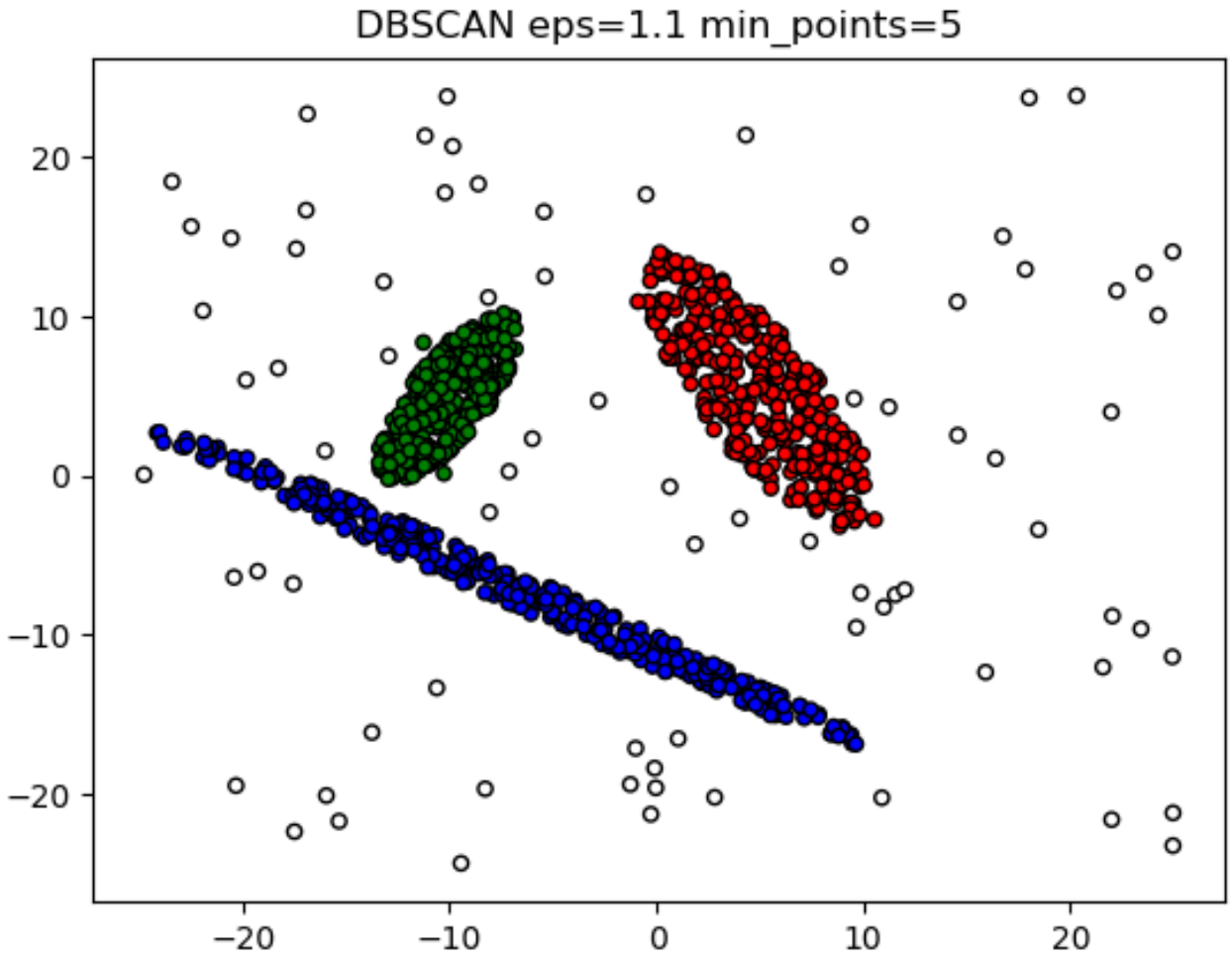
DBSCAN – przykład 1



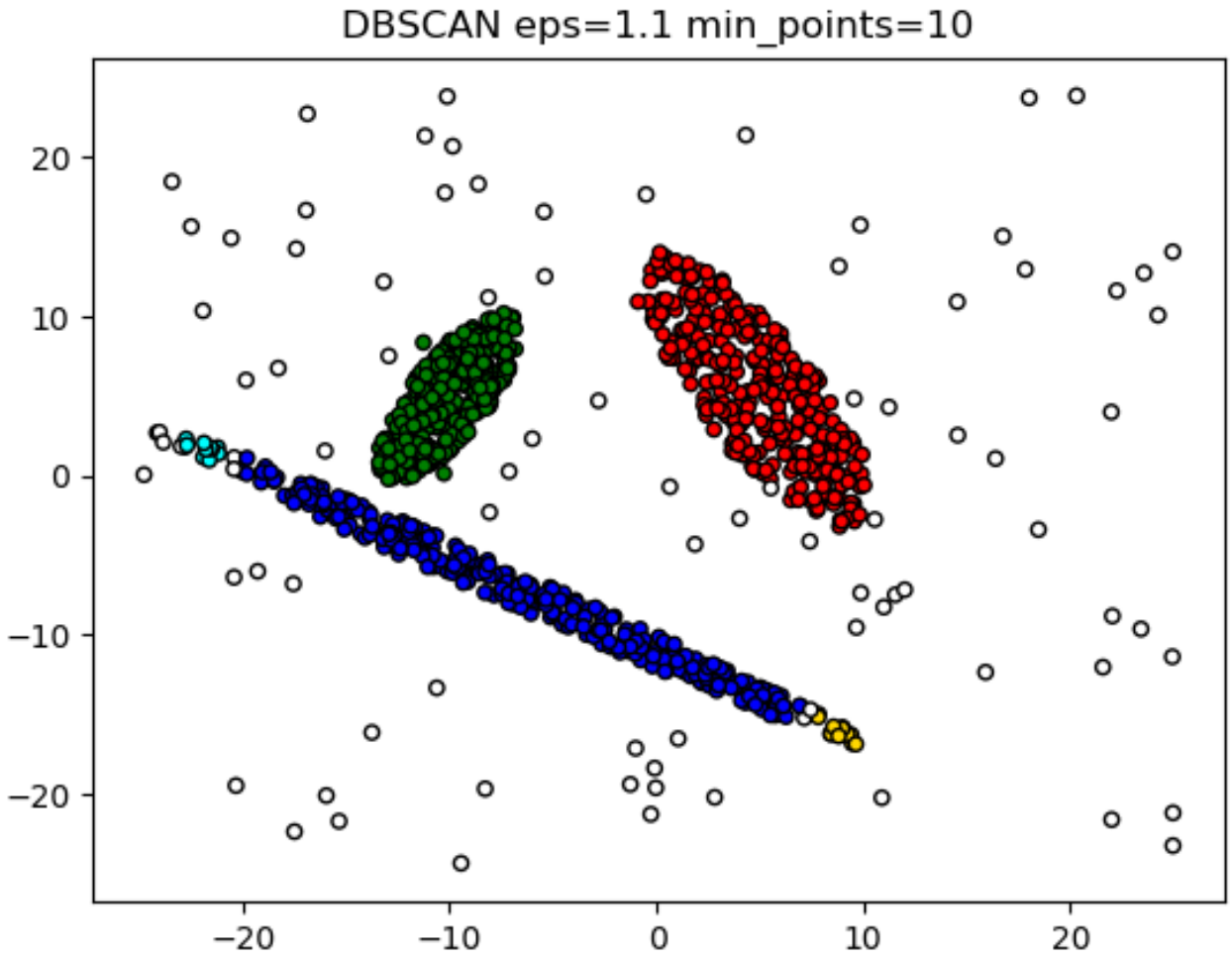
DBSCAN – przykład 2



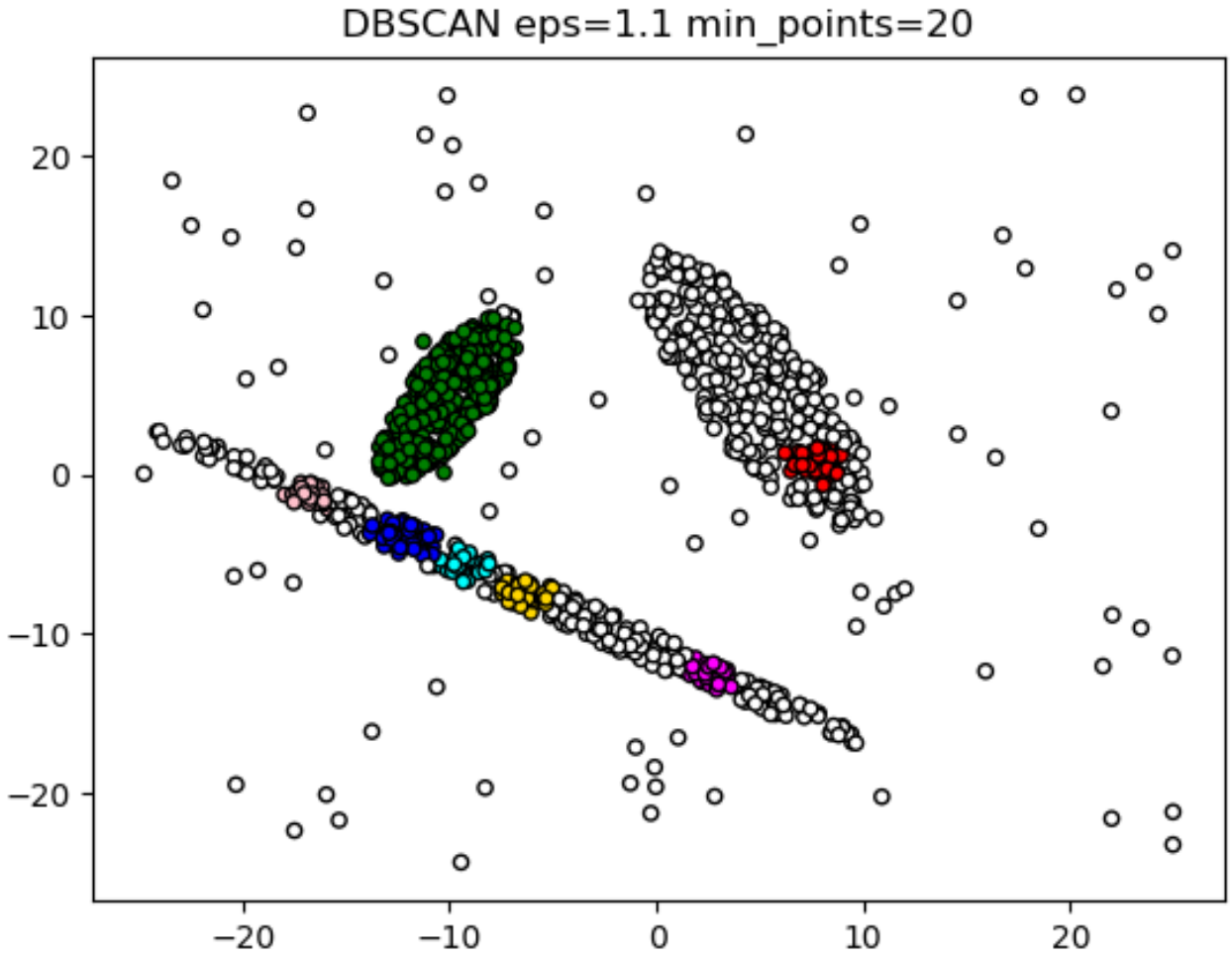
DBSCAN – przykład 3



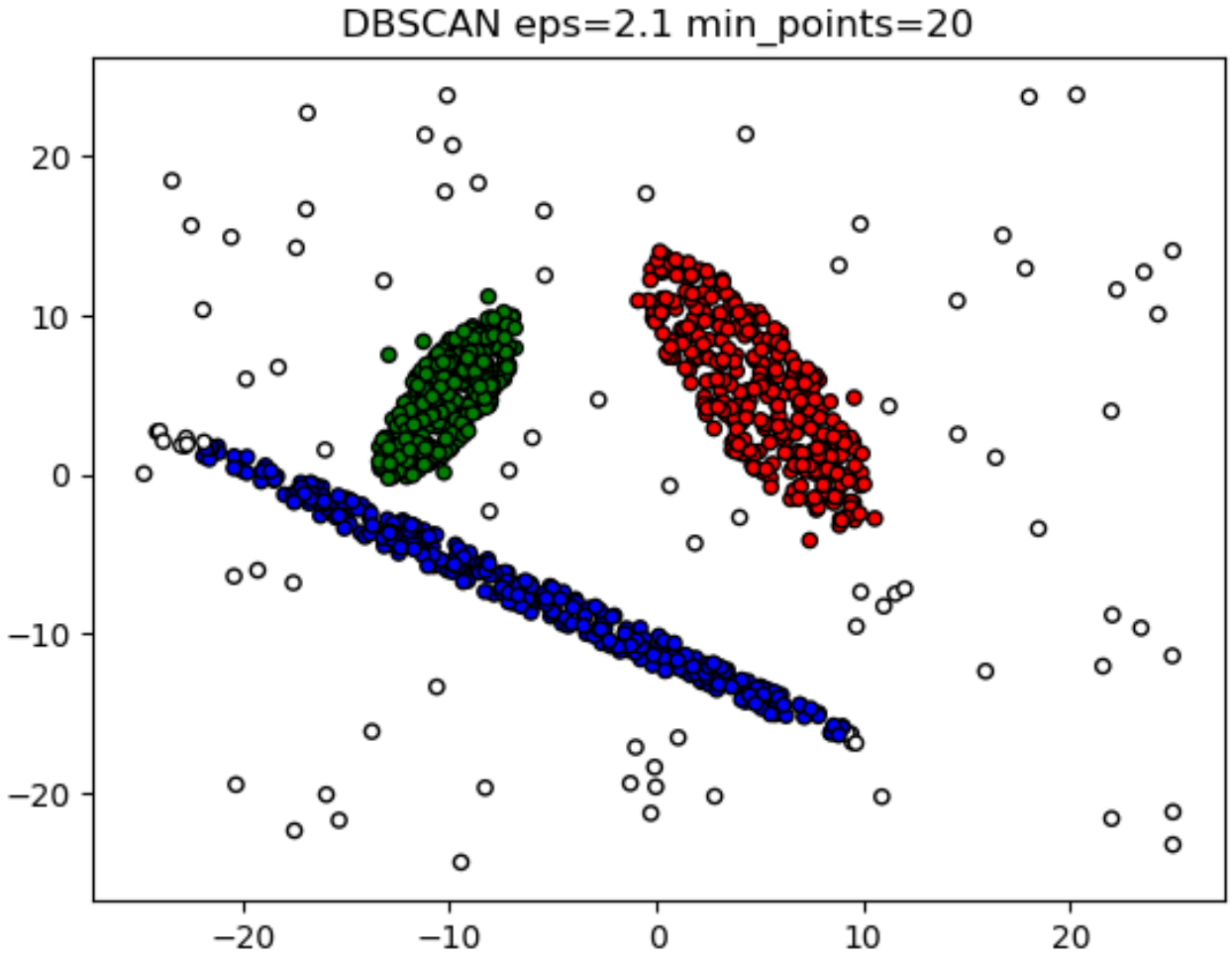
DBSCAN – przykład 4



DBSCAN – przykład 5

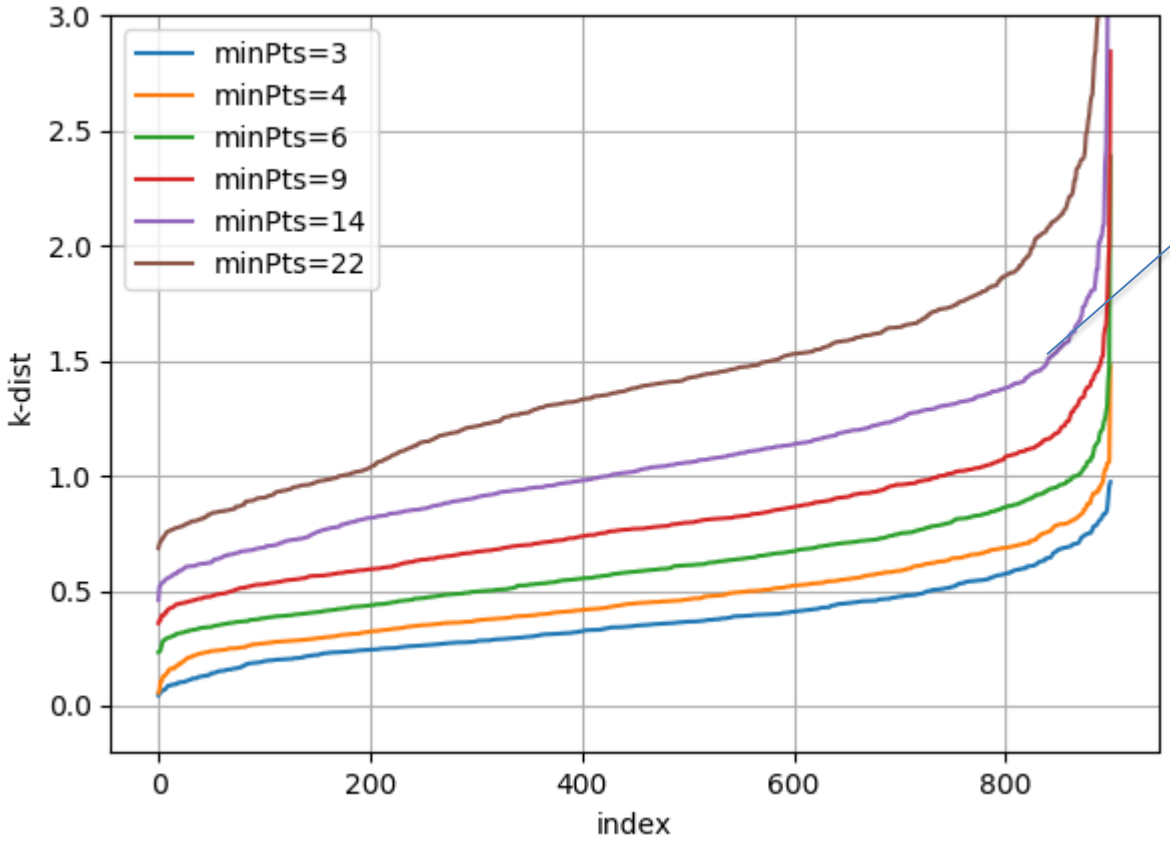


DBSCAN – przykład 6



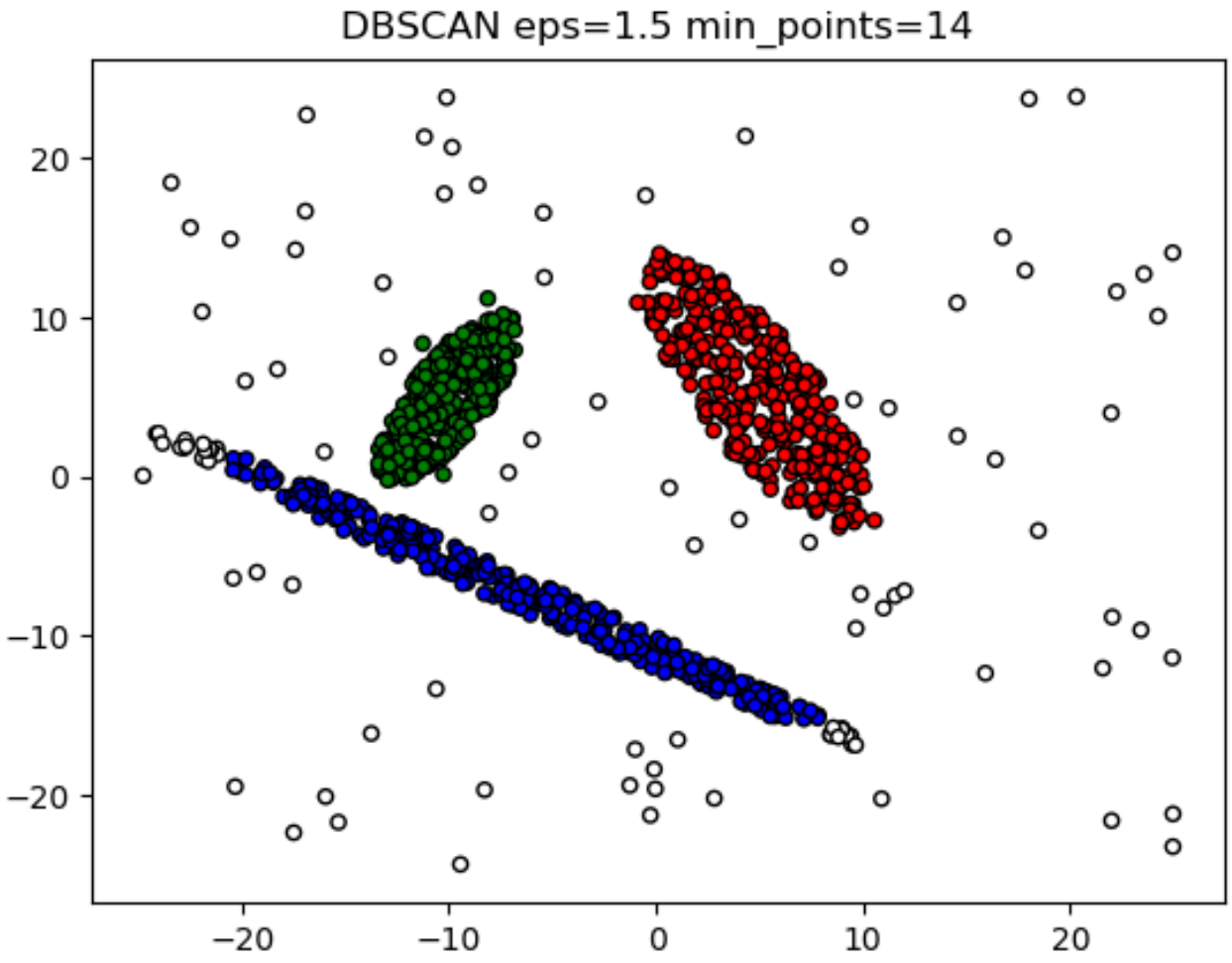
Jak dobrać minPts i ε ?

Wykres k-dist pokazuje posortowany zbiór odległości do najbliższego k-tego sąsiada. Należy go sporządzić dla $k = minPts - 1$



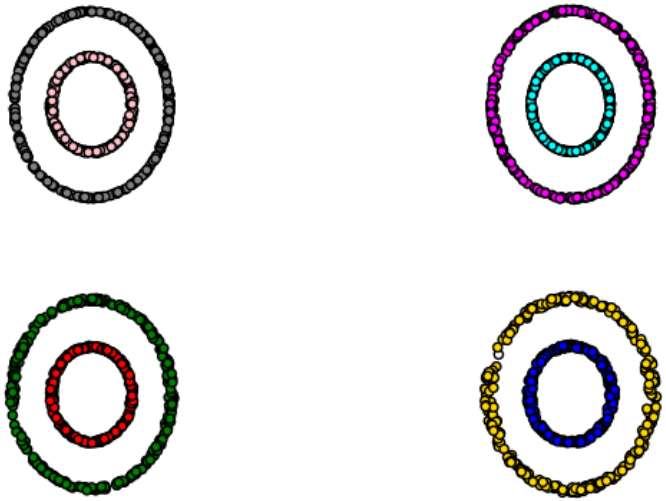
Odczytaj dist dla punktu przegięcia i wybierz go jako ε, np. $minPts = 14$, $\varepsilon = 1.5$

Jak dobrać minPts i ε -wynik ?



DBSCAN - podsumowanie

- Nie wymaga podania liczby grup
- Może znaleźć grupy o dowolnym kształcie
- Może zidentyfikować odstające punkty (outliers)
- Jeżeli punkty są przetwarzane w takiej samej kolejności – zachowuje się deterministycznie
- Może dawać różne przydziały dla punktów brzegowych
- Dobór parametrów nie zawsze jest oczywisty (zwłaszcza dla większych wymiarów)
- Może używać kD-tree lub ball tree do przechowywania danych



Metody oceny grupowania

Podział metod

- Ocena **wewnętrzna** (ang. internal evaluation) – ocena na podstawie danych poddanych grupowaniu z użyciem funkcji oceny. Zwykle powiązanej z algorytmem grupowania.
- Ocena **zewnętrzna** (ang. external evaluation) – ocena na podstawie danych które nie były użyte do grupowania, np. znanych etykiet klas reprezentujących ground truth
- Ocena **ekspercka** (ang. manual evaluation)
- Ocena **pośrednia** (ang. indirect) – ocena użyteczności grupowania

Ocena wewnętrzna - indeks Daviesa-Boudina

- Oznaczmy przez $\bar{d}(i)$ średnią odległość od środka wewnątrz i-tej grupy:

$$\bar{d}(i) = \frac{\sum_{j=1}^{m} \mathbf{1}(Z_j = i) d(x_j, \bar{x}(i))}{\sum_{j=1}^{m} \mathbf{1}(Z_j = i)}$$

- Indeks Daviesa-Boudina jest zdefiniowany jako:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left\{ \frac{\bar{d}(i) + \bar{d}(j)}{d(\bar{x}(i), \bar{x}(j))} \right\}$$

- Indeks odzwierciedla założenie, że grupy powinny skupiać obiekty położone blisko siebie (mała wartość licznika), natomiast grupy powinny być oddalone (duża wartość mianownika). Niższa wartość DB oznacza lepsze grupowanie.

Ocena wewnętrzna – indeks Dunna

- Niech $d(i,j)$ oznacza odległość pomiędzy i-tą i j-tą grupą, natomiast $\bar{d}(i)$ odległość wewnątrz grupy. Obie te metryki mogą być zdefiniowane na różne sposoby (patrz hierarchiczne grupowanie dla $d(i,j)$).
- Indeks Dunna jest zdefiniowany jako:

$$D = \frac{\min\{d(i,j): 1 \leq i,j \leq k\}}{\max\{\bar{d}(i): 1 \leq i \leq k\}}$$

Oczekiwane jest duże podobieństwo/mała odległość wewnątrz grupy (mianownik mały) i duża odległość pomiędzy grupami, stąd większe wartości D są lepsze.

Ocena wewnętrzna – współczynnik kształtu

- Współczynnik kształtu (ang. Silhouette Coefficient) dla pojedynczej próbki jest zdefiniowany jako

$$s = \frac{b - a}{\max\{a, b\}}$$

gdzie:

  - a – to średnia odległość pomiędzy próbką i innymi punktami z tej samej grupy
  - b – to średnia odległość pomiędzy próbką i punktami z **najbliżej położonej** grupy
- Współczynnik przyjmuje wartości pomiędzy -1 (dla złego grupowania) a 1 (dla idealnego grupowania)
- Współczynnik jest wyższy, dla gęstych i wypukłych grup (np. otrzymanych z k-means, natomiast może być zawodny dla niewypukłych grup, które może zwrócić np. DBSCAN.

Ocena wewnętrzna – indeks Calińskiego Harabasza

Oznaczmy przez $\bar{x}(i)$ środek i-tej grupy, a przez c punkt środkowy całego zbioru danych.

Macierze W_k i B_k definiują kowariancję wewnątrz grup (W_k) oraz pomiędzy grupami (B_k)

- $W = \sum_{i=1}^{k} \sum_{x \in C_i} \left(x - \bar{x}(i)\right) \left(x - \bar{x}(i)\right)^T$
- $B = \sum_{i=1}^{k} (\bar{x}(i) - c) \, (\bar{x}(i) - c)^T$

Wskaźnikiem określającym stopień rozproszenia (dyspersji) dla macierzy kowariancji Σ jest jej ślad $tr(\Sigma)$, czyli suma elementów na przekątnych (wariancji).

Indeks CH jest zdefiniowany jako:

$$CH(k) = \frac{tr(B)}{tr(W)} \frac{m-k}{k-1}$$

Indeks rośnie dla gęstych (mała wartość $tr(W)$) i dobrze odseparowanych (duża wartość $tr(B)$) grup.

Ocena zewnętrzna

- W ocenie zewnętrznej używa się etykiet klas (np. wprowadzonych manualnie dla części danych) lub też informacji o rozkładach użytych do generacji syntetycznych zbiorów danych.
- Zdefiniowane wskaźniki mierzą, podobieństwa pomiędzy strukturą klas i grup (stopień zgodności etykiet klas i grup).

Ocena zewnętrzna – indeks Randa

Indeks Randa (William Rand) pozwala ocenić zgodność dwóch podziałów zbioru na rozłączne podzbiory

Niech $O = \{o_1, \dots, o_n\}$ będzie zbiorem obiektów, $X = \{X_1, \dots, X_r\}$ oraz $Y = \{Y, \dots, Y_r\}$ dwoma podziałami.

Rozważmy parę różnych elementów z O: (o_i, o_j). Mamy cztery możliwe przypadki:

A. o_i,o_j należą do jednego ze zbiorów w X oraz do jednego ze zbiorów w Y
B. o_i,o_j należą do dwóch różnych zbiorów w X oraz różnych zbiorów w Y
C. o_i,o_j należą do tego samego zbioru w X i różnych zbiorów w Y
D. o_i,o_j należą do różnych zbiorów w X i tego samego zbioru w Y

Oznaczmy przez a, b, c, d liczby par spełniające warunki A, B, C i D.

$$RI = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}$$

RI zwraca wartości pomiędzy 0 i 1 i może być interpretowany jako prawdopodobieństwo, że dowolnie wybrana para jest w analogiczny sposób sklasyfikowana w obu grupowaniach.

Skorygowany indeks Randa (ang. Adjusted Rand Index) wprowadza poprawkę uwzględniającą prawdopodobieństwo dwa algorytmy grupowania zachowując się losowo równocześnie rozdzielą parę lub dołączą do jednej grupy (rozkład hipergeometryczny).

$$ARI = \frac{RI - \boldsymbol{E}(RI)}{max(RI) - \boldsymbol{E}(RI)}$$

[https://link.springer.com/content/pdf/10.1007%2FBF01908075.pdf]
[https://en.wikipedia.org/wiki/Rand_index]

Entropia (przypomnienie)

- Niech X będzie zmienną losową przyjmującą k dyskretnych wartości, każdą z nich z prawdopodobieństwem $P(X = x_k)$
- Entropia zdefiniowana jest jako:

$$H(X) = -\sum_{i=1}^{k} P(X = x_k) \log_2 P(X = x_k)$$

- Entropia (mierzona w bitach) podaje minimalną liczbę bitów niezbędną do zakodowania losowo wybranej wartości X.
- Jeśli $P = 0$, przyjmuje się, że $0 \cdot \log_2 0 \ = \ 0$

Symbol	Częstość
A	0.17
B	0.33
C	0.50

- $H = -(0.17 \cdot \log_2(0.17) + 0.33 \cdot \log_2(0.33) + 0.5 \cdot \log_2(0.5))$ = 1.46
- Proponowane kodowanie (średnio 1.55 bitu/symbol):
  C: 0
  B: 10
  A: 11

Entropia warunkowa (przypomnienie)

Entropia warunkowa obliczana jest dla dwóch zmiennych Y i X

$$H(Y|X) = -\sum_{i=1}^{k}\sum_{j=1}^{r} P\big(Y = y_i \wedge X = x_j\big) \log_2 \frac{P\big(Y = y_i \wedge X = x_j\big)}{P\big(X = x_j\big)}$$

przekształcając:

$$H(Y|X) = -\sum_{i=1}^{k}\sum_{j=1}^{r} P\big(Y = y_i|X = x_j\big)\, P(X = x_j) \log_2 P(Y = y_i|X = x_j)$$

$$H(Y|X) = -\sum_{j=1}^{r} P(X = x_j) \sum_{i=1}^{k} P\big(Y = y_i|X = x_j\big)\, \log_2 P(Y = y_i|X = x_j)$$

- Wydzielany jest podzbiór $X \times Y$, gdzie $X = x_j$
- Obliczane jest entropia w tym podzbiorze
- Obliczana jest suma z wagami $P(X = x_j)$

Ocena zewnętrzna V-measure

Załóżmy, że mamy podział zbioru obserwacji na klasy (ground truth) $C = C_1, \dots, C_n$ oraz zbiór grup $K = K_1, \dots, K_m$ będących efektem działania algorytmu.

Podziałowi K stawiane są dwa wymagania:

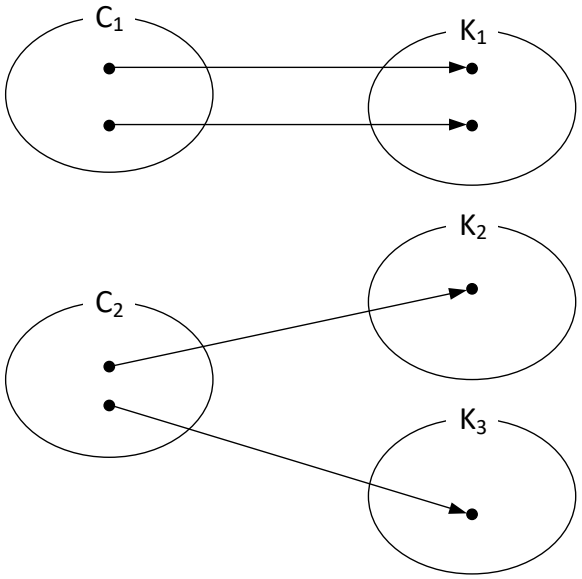
- Jednorodności (ang. homogenity) – grupowanie powinno przydzielić do jednej grupy **wyłącznie** elementy jednej klasy (nie powinny być pomieszane).
- Zupełności (ang. completness) – **wszystkie** elementy jednej klasy powinny zostać przydzielone do **jednej grupy**

Jednorodność

- Formalnie jednorodność jest zdefiniowana jako

$$h = 1 - \frac{H(C|K)}{H(C)}$$

- W idealnym przypadku $H(C|K) = 0$, ponieważ dla każdej klasy C_i i grupy K_j prawdopodobieństwo $P(C_i|K_j)$ wynosi 1 lub 0. Wtedy $h = 1$. Czynnik $H(C)$ jest użyty do skalowania. Entropia $H(C)$ zależy od rozkładu liczby obserwacji w klasach.



Przykład
Grupy zawierają jednorodne obserwacje (pochodzące z tej samej klasy).

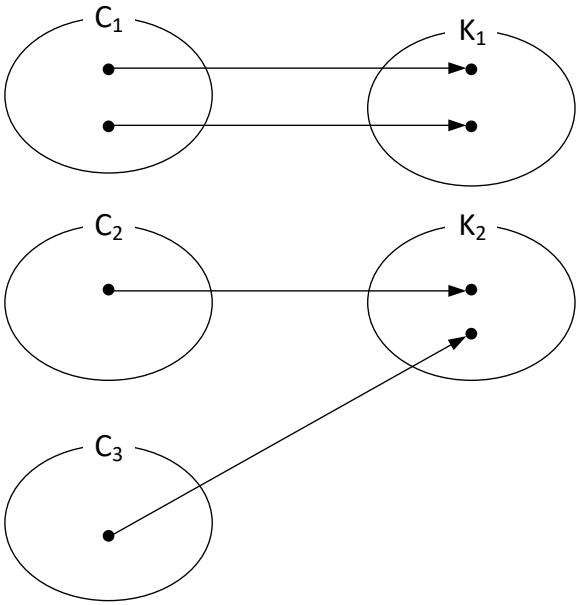
Np. $P(C_1|K_1) = 1$, $P(C_2|K_3) = 1$
Wówczas $h = 1$

Zupełność

- Zupełność jest zdefiniowana jako

$$c = 1 - \frac{H(K|C)}{H(K)}$$

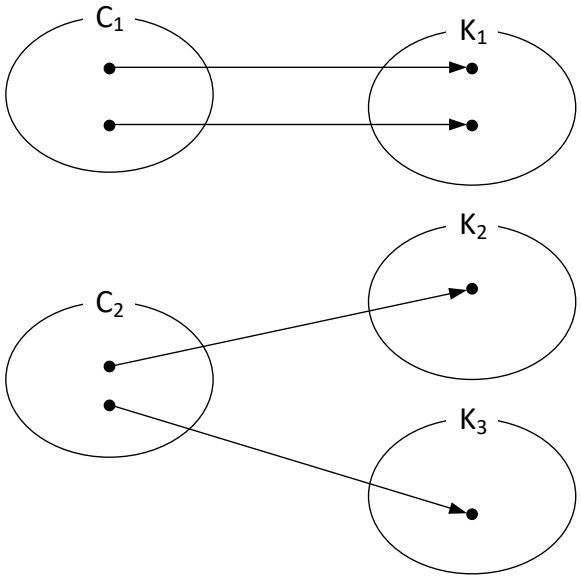
- Jest to miara zdefiniowana symetrycznie, odpowiadająca wymaganiu, aby wszystkie elementy danej klasy trafiły do tej samej grupy



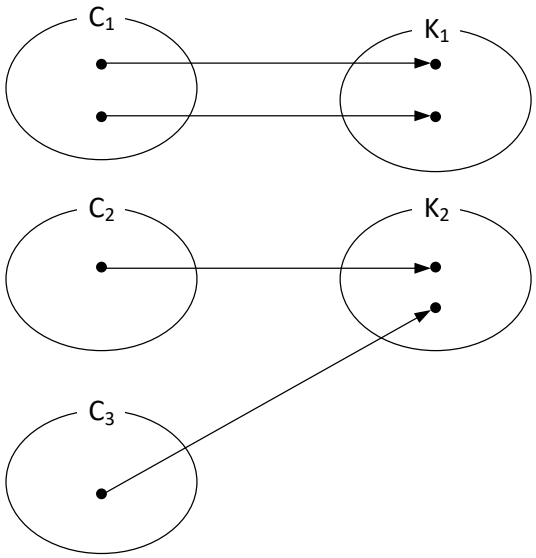
Przykład
Wszystkie elementy klas trafiły do tych samych grup.

Np. $P(K_1|C_1) = 1, P(K_2|C_2) = 1$
Wówczas $c = 1$

Ocena zewnętrzna V-measure



homogenity $h = 1$
completness $c = 2/3$



homogenity $h = 2/3$
completness $c = 1$

- Miara V-measure jest średnią harmoniczną h i c:

$$v = 2\frac{h \cdot c}{h + c}$$

[Andrew Rosenberg and Julia Hirschberg: V-Measure: A conditional entropy-based external cluster evaluation measure http://www.aclweb.org/anthology/D07-1043]

Ocena zewnętrzna – wzajemna informacja

- Dla dwóch zmiennych losowych X i Y wskaźnik wzajemnej informacji (ang. mutual information) mierzy w bitach ile na podstawie jednej zmiennej można uzyskać informacji o drugiej. Formalnie zdefiniowany jest jako

$$I(Y,Z) = \sum_{y \in Y} \sum_{z \in Z} p(y,z) \log_2 \left(\frac{p(y,z)}{p(y)p(z)} \right)$$

- Niech $Y = \{Y_1, \dots Y_i, \dots, Y_r\}$ oraz $Z = \{Z_1, \dots Z_j, \dots, Z_r\}$ będą rozłącznymi podziałami zbioru n obiektów. Wówczas
  - $p(i) = \frac{|Y_i|}{n}$ będzie prawdopodobieństwem, że losowo wybrany obiekt znajdzie się w Y_i,
  - $p(j) = \frac{|Z_j|}{n}$ że znajdzie się w Z_j
  - $p(i,j) = \frac{|Y_i \cap Z_j|}{n}$ – że losowo wybrany obiekt znajdzie się w Y_i oraz Z_j
- Indeks wzajemnej informacji MI dla dwóch podziałów na grupy jest zdefiniowany jako:

$$MI(Y,Z) = \sum_{i=1}^{|Y|} \sum_{j=1}^{|Z|} p(i,j) \log_2 \left(\frac{p(i,j)}{p(i)p(j)} \right)$$

- Indeks przyjmuje wartości z przedziału [0,1].
- Wraz ze wzrostem liczby (małych) grup indeks przyjmuje większe wartości, stąd korekta uwzględniająca przypadkową zgodność przydziału (adjusted mutual information)

$$AMI = \frac{MI - E(MI)}{\max(MI) - E(MI)}$$