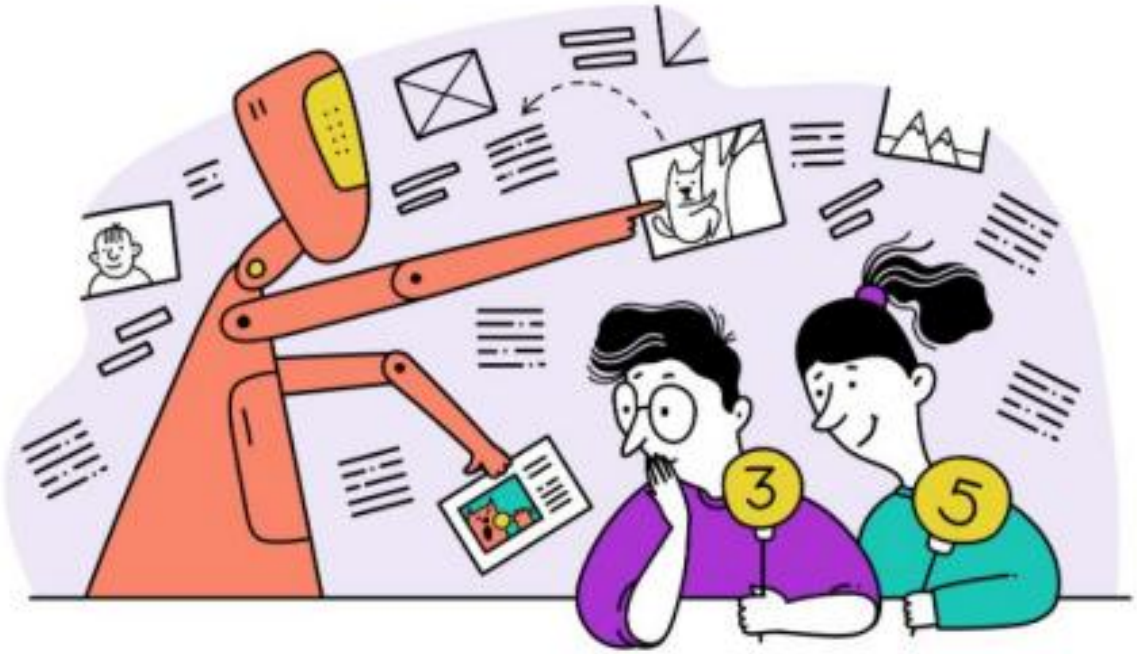
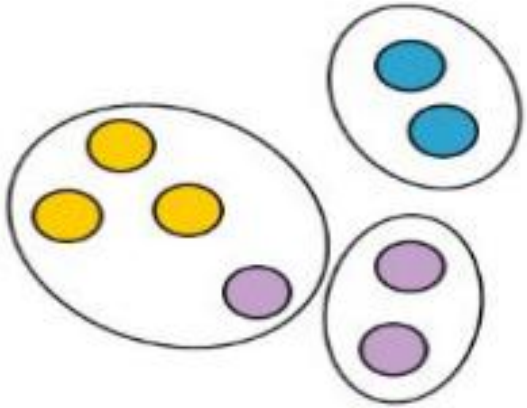
Розпізнавання образів.

Методи кластеризації

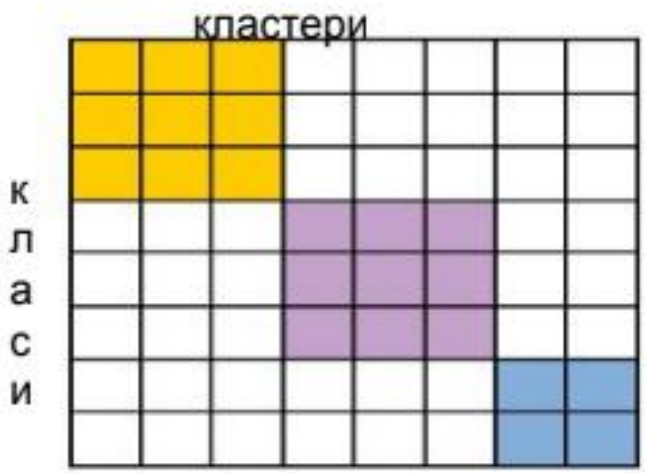


Метрики в sklearn

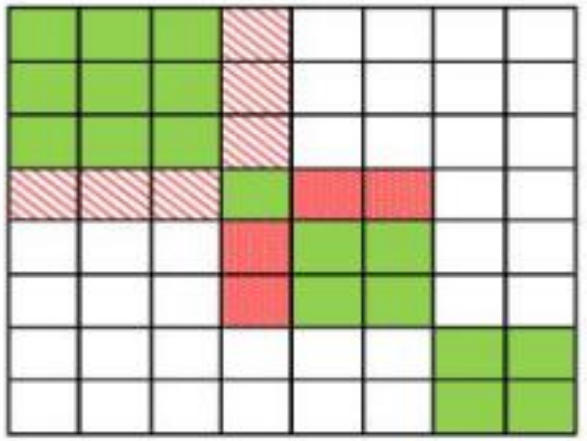
Зовнішні	Внутрішні
Adjusted Rand index Fowlkes-Mallows scores Mutual Information based scores Homogeneity Completeness V-measure	Silhouette Coefficient Davies-Bouldin Index Calinski-Harabasz Index * Dunn Index



Об'єкти кластера



Еталонна матриця відношень



Фактична матриця відношень

$TP = 18$ $TN = 36$

$FP = 6$ $FN = 4$

Точність і повнота

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F_\beta = \frac{(\beta^2 + 1) \cdot P \cdot R}{\beta^2 \cdot P + R}$$

- Recall - здатність алгоритму виявляти даний клас взагалі, а Precision - здатність відрізняти цей клас від інших класів.
- F-міра

Adjusted Rand index

$$Rand = \frac{TP + FN}{TP + TN + FP + FN}$$

$$Jaccard = \frac{TP}{TP + TN + FP}$$

$$\text{Adjusted Index } \widehat{ARI} = \frac{\overbrace{\sum_{ij}\binom{n_{ij}}{2}}^{\text{Index}} - \overbrace{[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}^{\text{Expected Index}}}{\underbrace{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}]}_{\text{Max Index}} - \underbrace{[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}_{\text{Expected Index}}}$$

$X \setminus Y$	Y_1	Y_2	$\dots$	Y_s	Sums
X_1	n_{11}	n_{12}	$\dots$	n_{1s}	a_1
X_2	n_{21}	n_{22}	$\dots$	n_{2s}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_r	n_{r1}	n_{r2}	$\dots$	n_{rs}	a_r
Sums	b_1	b_2	$\dots$	b_s	n

$$p_{ij} = \frac{n_{ij}}{n}, p_i = \frac{a_i}{n}, p_j = \frac{b_j}{n}.$$

Homogeneity, Completeness, V-measure

де $H(C|K)$є **умовною ентропією класів з урахуванням**

присвоєння кластеру і задається як:

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

$$H(C|K) = -\sum_{c=1}^{|C|}\sum_{k=1}^{|K|}\frac{n_{c,k}}{n}\cdot\log\left(\frac{n_{c,k}}{n_k}\right)$$

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

і $H(C)$є **ентропією класів** і задається:

$$H(C) = -\sum_{c=1}^{|C|}\frac{n_c}{n}\cdot\log\left(\frac{n_c}{n}\right)$$

$$v = \frac{(1+\beta)\times\text{homogeneity}\times\text{completeness}}{(\beta\times\text{homogeneity}+\text{completeness})}$$

n загальна кількість елементів, n_c і n_k кількість елементів що відповідно належать класу c і кластеру k, і, нарешті $n_{c,k}$ кількість зразків з класу c призначених кластеру k.

Mutual Information based scores

- Оцінює узгодженість між встановленими мітками класів та спрогнозованими. ↑ [1]

Використовує наступні поняття:

$$H(U) = -\sum_{i=1}^{|U|} P(i)\log(P(i))$$

де $P(i) = |U_i|/N$ - це ймовірність того, що об'єкт вибраний навмання U потрапляє в клас U_i.

Так само для V

$$H(V) = -\sum_{j=1}^{|V|} P'(j)\log(P'(j))$$

$$\mathrm{MI}(U,V) = \sum_{i=1}^{|U|}\sum_{i=1}^{|V|} P(i,j)\log\left(\frac{P(i,j)}{P(i)P'(j)}\right)$$

де $P(i,j) = |U_i \cap V_j|/N$ це ймовірність того, що об'єкт, вибраний навмання, потрапляє в обидва класи U_i і V_j.

Fowlkes-Mallows Index

- Індекс Фоулкса – Малловса

$$\mathrm{FMI} = \frac{\mathrm{TP}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})}}$$

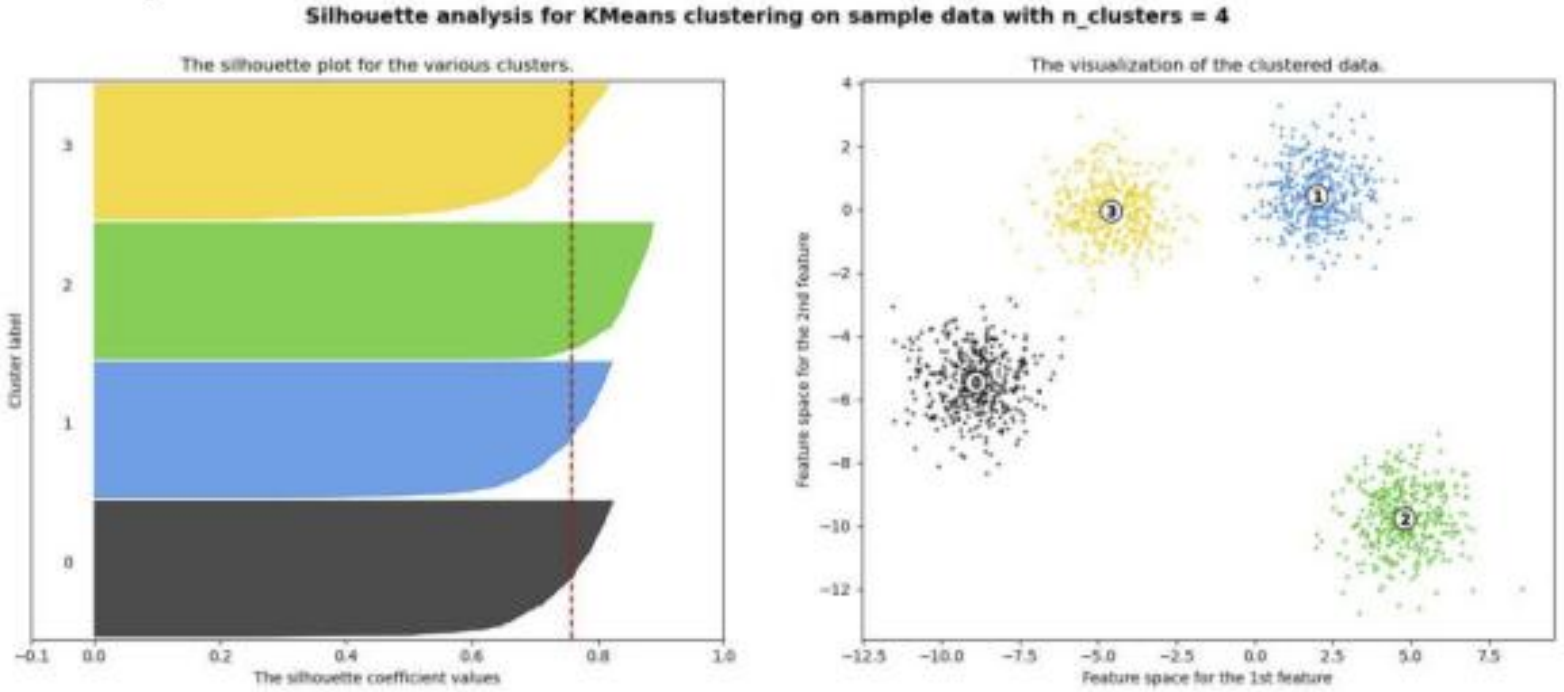
- FMI індекс - це середнє геометричне значення точності (precision) та повноти (recall)

Dunn Index

$$D = \frac{\min_{1 \le i < j \le n} d(i,j)}{\max_{1 \le k \le n} d'(k)}$$

Silhouette Coefficient

- Значення силуету показує, наскільки об'єкт схожий на свій кластер в порівнянні з іншими кластерами.



Silhouette Coefficient

a : Середня відстань між точкою та усіма іншими точками того самого класу.

b : Середня відстань між точкою та усіма іншими точками в *наступному найближчому кластері*.

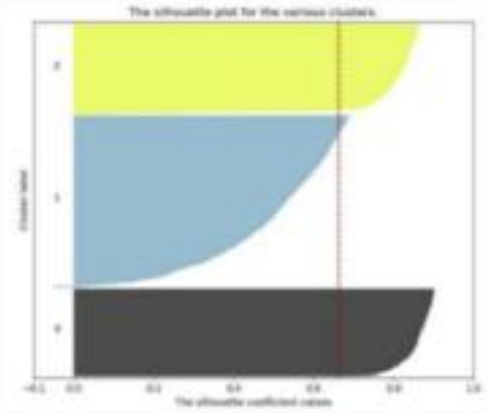
$$s = \frac{b - a}{max(a, b)}$$

- Коефіцієнт силуету для множини елементів рахується як середнє значення коефіцієнта силуету для кожного елементу.

When poll is active, respond at **pollev.com/nataliashovgun288**

Text **NATALIASHOVGUN288** to **37607** once to join

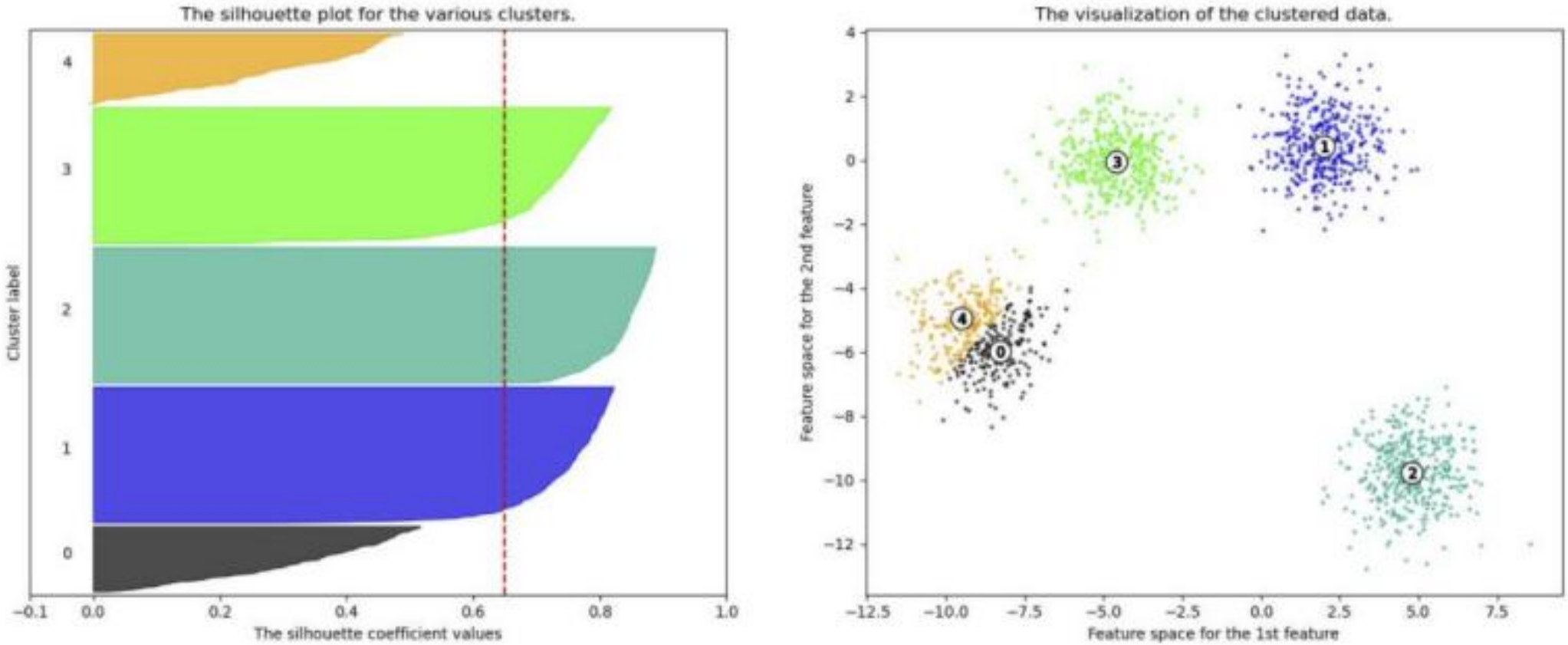
Який висновок ви можете зробити виходячи з такого значення силуету



замале число кластерів

красива картинка

велике число кластерів

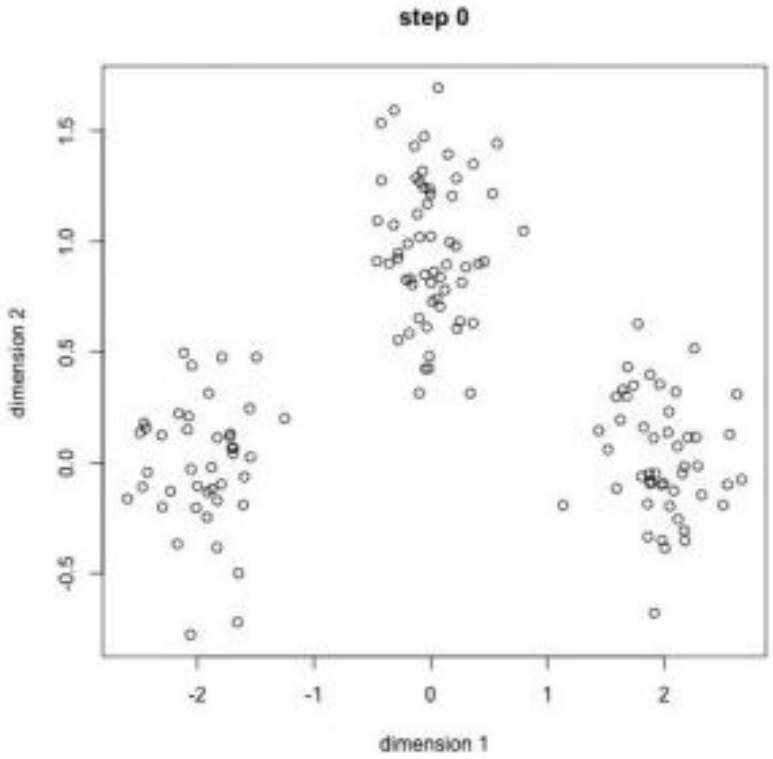

Silhouette analysis for KMeans clustering on sample data with n_clusters = 5
The silhouette plot for the various clusters.
Cluster label
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
Feature space for the 1st feature

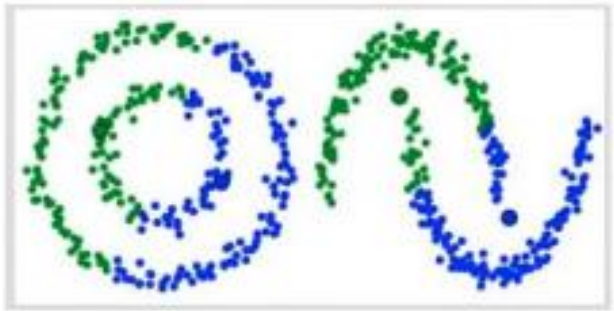
Алгоритм к-середніх

Нехай m – фіксоване число кластерів. Знайти таку функцію кластеризації $f: \Theta \to Y$, $|Y| = m$, щоб $Q^{(3)}(f) \to \min$.

1. Виділяються деякі зразки з навчальної вибірки – початкові центри кластерів $c^{(0)}1 \ldots c^{(0)}m$, к=0
2. Вся навчальна вибірка розбивається на m кластерів за методом найбл[illegible]сіда. Отримуємо деякі кластери $x^{(к)}1 \ldots x^{(к)}m$
3. Розраховуємо нові центри
4. Перевіряємо умову зупинки – центри кластерів не змінюються, інакше на крок 2

$$\mathbf{c}_i^{(k+1)} = \frac{1}{|X_i^{(k)}|} \sum_{\mathbf{x} \in X_i^{(k)}} \mathbf{x},$$



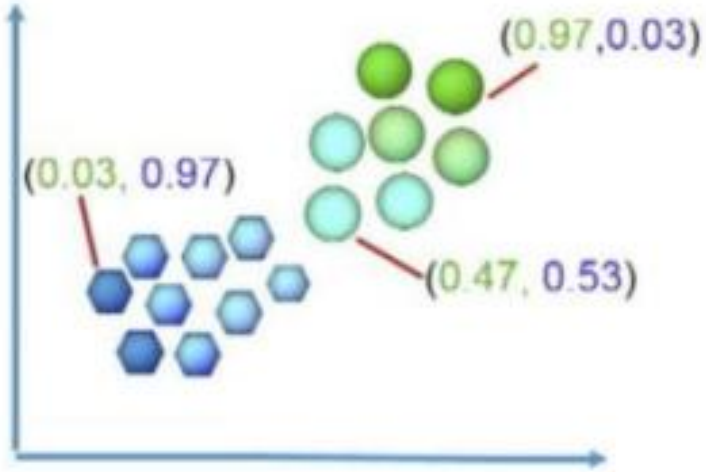


k-means++

Fuzzy c means

Hard Clustering

Soft Clustering



$0<u_{ij}<1$, $\sum u_{ij}=1$

$$E = \sum_{i=1}^{K}\sum_{j=1}^{p} u_{ij}^{m} \left\| c_i - x_j \right\|^2,$$

$$LE = \sum_{i=1}^{K}\sum_{j=1}^{N} u_{ij}^{m} \left\| c_i - x_j \right\|^2 + \sum_{j=1}^{p} \lambda_j \left(\sum_{i=1}^{K} u_{ij} - 1 \right)$$

$$c_i = \frac{\sum_{j=1}^{p} u_{ij}^{m} x_j}{\sum_{j=1}^{p} u_{ij}^{m}} \quad (1)$$

$$u_{ij} = \frac{1}{\sum_{k=1}^{K} \left(\frac{d_{ij}^2}{d_{kj}^2} \right)^{\frac{1}{m-1}}} \quad (2)$$

1. Ініціалізувати значення u_{ij}є (0,1), з виконанням вимоги $\sum u_{ij}=1$
2. Визначити К центрів згідно формули 1.
3. Вирахувати помилку Е. Якщо значення виявиться менше встановленої межі чи зміна цього значення в порівнянні з попередньою ітерацією мала, то зупинити алгоритм. Останні значення центрів і є шуканими.
4. Вирахувати нові значення u_{ij} згідно формули 2, та перейти до пункту 2.

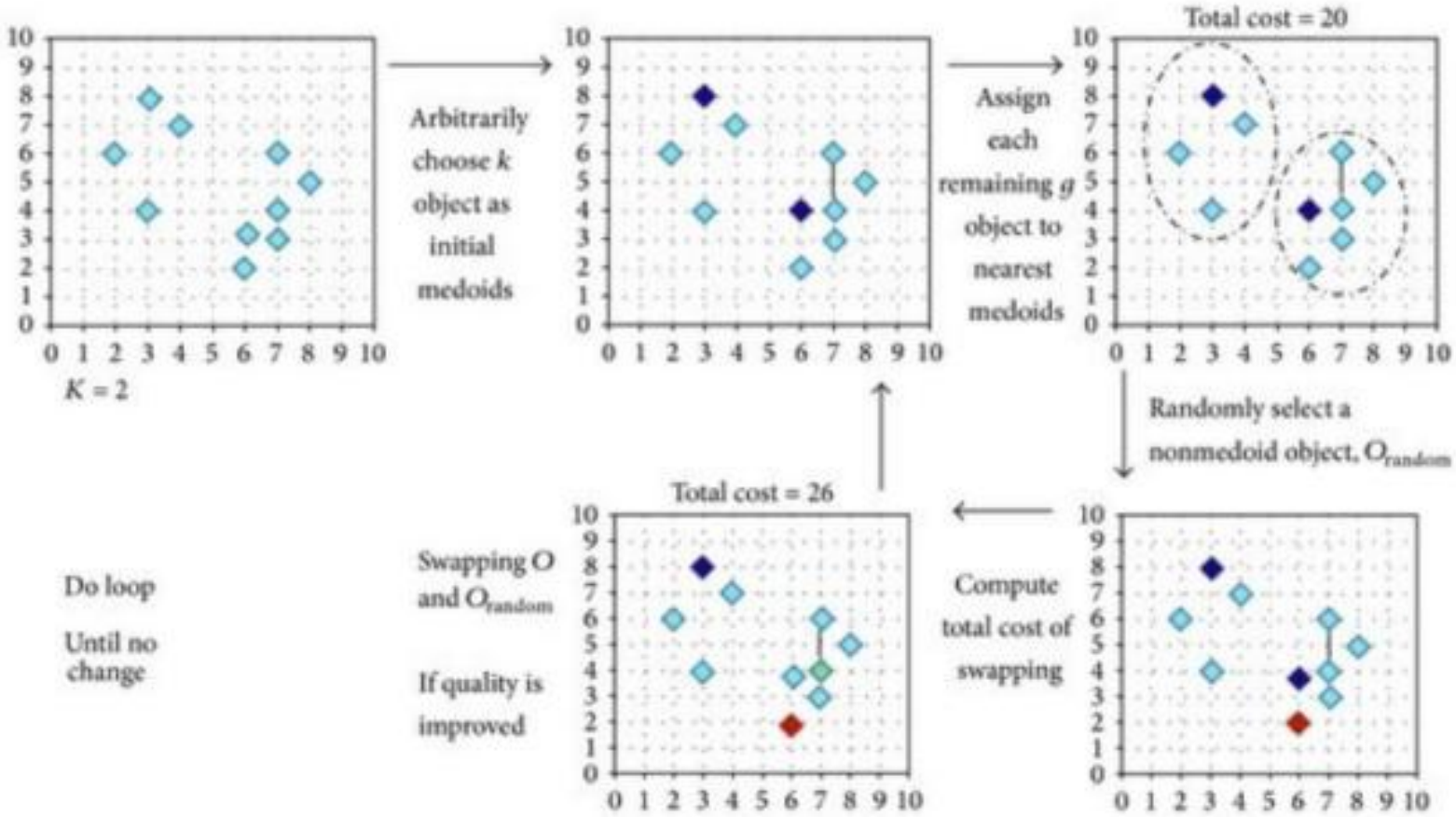
k-medoids

Алгоритм PAM (**Partitioning Around Medoids**), заснований на виборі k об'єктів, які є характерними точками відповідного кластера. Таким чином, кластерам зіставляються належні їм об'єкти, звані медоїдами, на підставі яких розподіляються інші об'єкти за принципом найбільшої схожості.

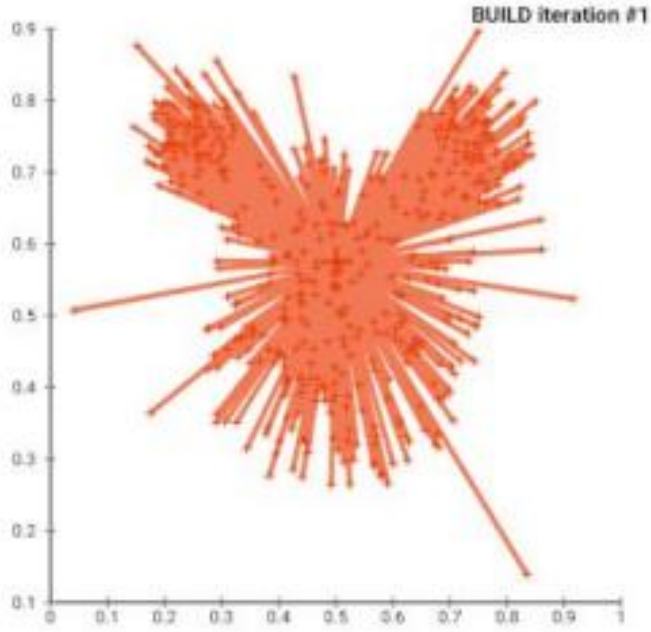
PAM складається з двох фаз: BUILD і SWAP.

$$E = \sum_{j=1}^{n} \min_{1 \le l \le k} \rho(x_{m_l}, x_{o_j})$$

Процедура зміни множини медоїдів повторюється, поки є можливість поліпшення значення цільової функції.


K = 2
Arbitrarily choose k object as initial medoids
Assign each remaining g object to nearest medoids
Total cost = 20
Randomly select a nonmedoid object, O_{random}
Compute total cost of swapping
Total cost = 26
Swapping O and O_{random}
If quality is improved
Do loop
Until no change

- Якість можно оцінити методом силуету
- Він більш стійкий до шуму та викидів у порівнянні з к-середніх , оскільки мінімізує суму попарних відмінностей.
- Медоїд можна визначити як об'єкт кластера, середня відмінність якого від усіх об'єктів кластера мінімальна, тобто це найбільш центральна точка кластера.



sklearn_extra.cluster.**KMedoids**

When poll is active, respond at **pollev.com/nataliashovgun288**

В якості центру кластеру в к-медоїд використовують

False

Алгоритм виділення зв'язних компонент

Вибірка представляється у вигляді графа:

- вершини графа об'єкти xi;
- ребра пари об'єктів з відстанню ρij = ρ (xi, xj) ≤ R.

Алгоритм:

повторювати

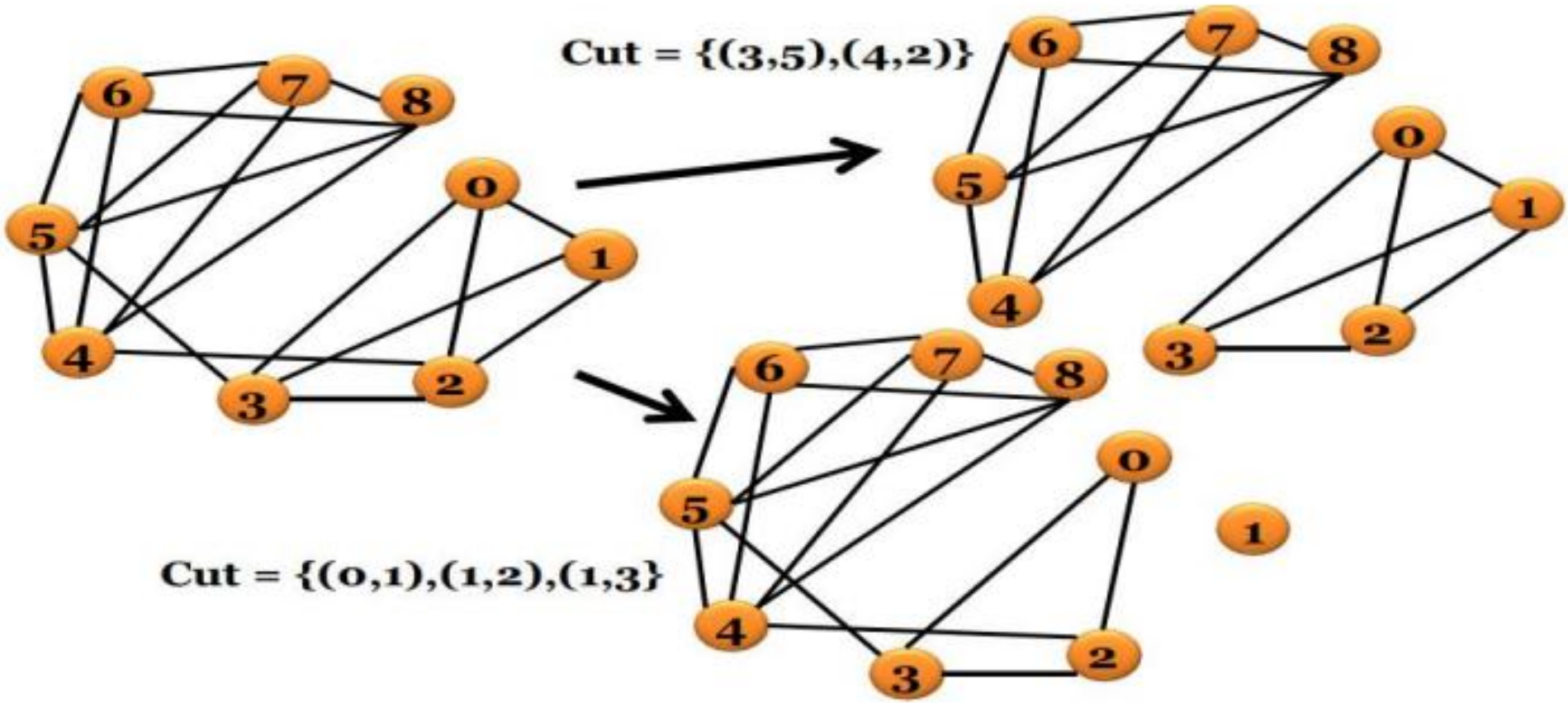
1: видалити всі ребра (i, j), для яких ρij> R;

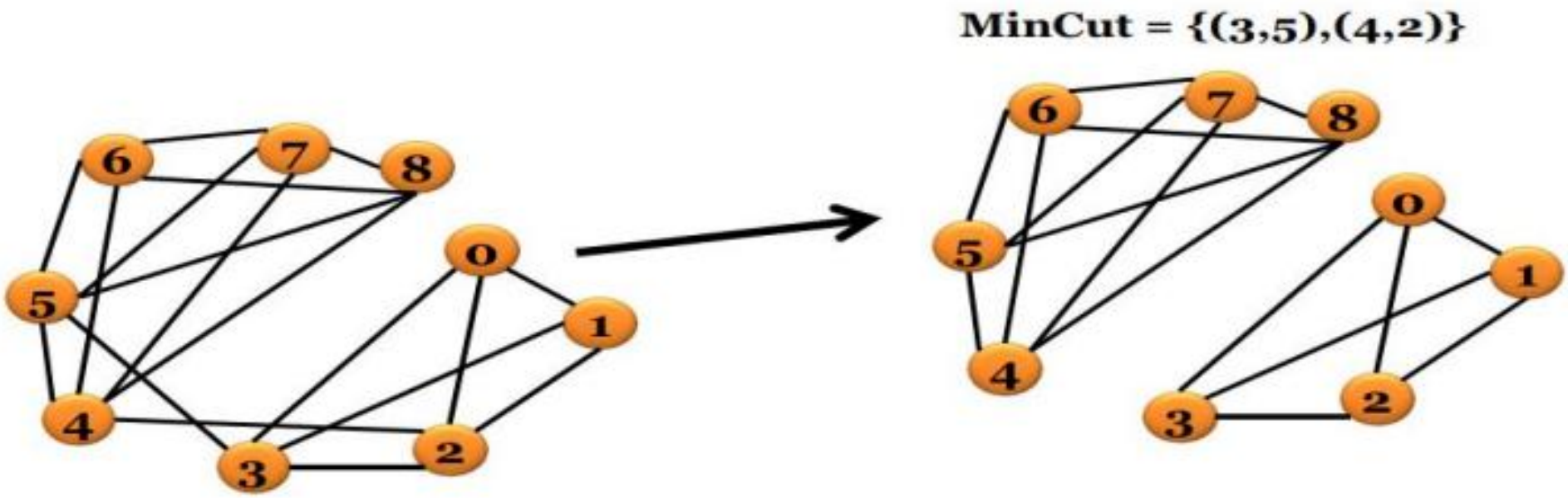
2: K: = число зв'язних компонент;

3: якщо K <K1 то зменшити R;

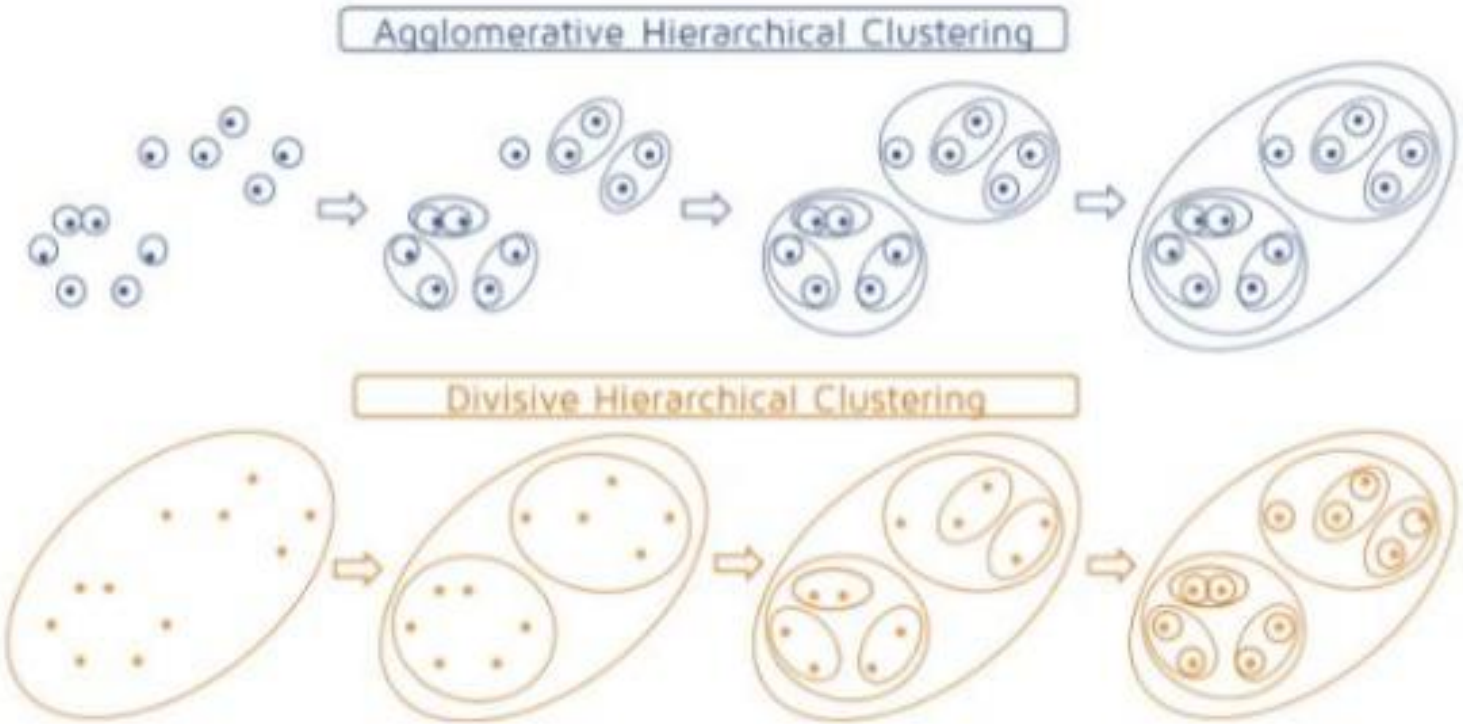
4: якщо K> K2 то збільшити R;

поки K ₵ [K1, K2]


Cut = {(3,5),(4,2)}
Cut = {(0,1),(1,2),(1,3}
6
7
8
5
0
1
4
3
2


MinCut = {(3,5),(4,2)}
6
7
8
0
5
1
4
3
2
6
7
8
0
5
1
4
3
2

Ієрархічна кластеризація



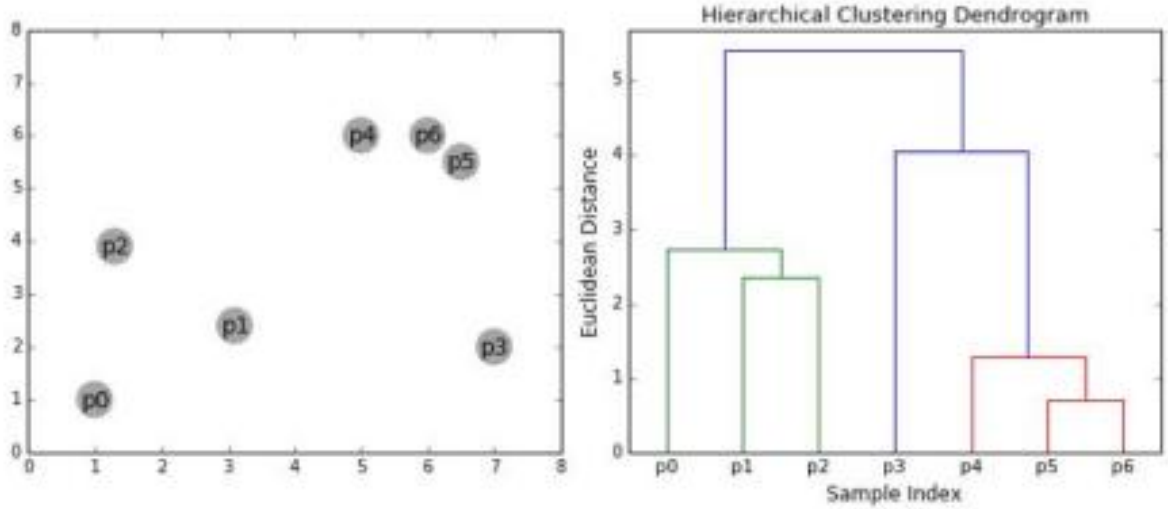
Ієрархічна кластеризація

1: спочатку всі кластери одноелементні:
t := 1; Ct = {x1}, . . . , {xℓ};
R({xi}, {xj }) := ρ(xi , xj);
2: для всіх t = 2, . . . , ℓ (t номер ітерації):
3: знайти в C_{t-1} два найближчі кластери:

$$(U, V) := \arg\min_{U \neq V} R(U, V);$$

$$R_t := R(U, V);$$

4: об'єднати їх в один кластер:
W := U ∪ V ;
$C_t := C_{t-1} \cup \{W\} \setminus \{U,V\}$;
5: для всіх S ∈ Ct
6: вирахувати R(W, S) за формулою Ланса-Вільямса;



sklearn.cluster.**AgglomerativeClustering**

When poll is active, respond at **pollev.com/nataliashovgun288**

Text **NATALIASHOVGUN288** to **37607** once to join

Який параметр на вашу думку найсильніше впливатиме на ієрархічну кластеризацію

кількість кластерів

метрика відстані

етрика оцінки якості кластеризації

Відстань R(W, S) між кластерами W = U ∪ V і S, через відстані R(U, S), R(V , S), R(U,V):

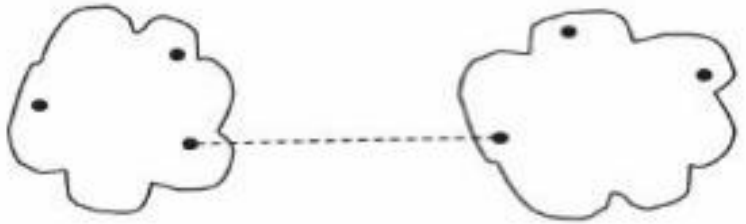
$$
\begin{aligned}
R(U \cup V, S) = \alpha_U \cdot R(U, S) + \\
+ \alpha_V \cdot R(V, S) + \\
+ \beta \cdot R(U, V) + \\
+ \gamma \cdot \left| R(U, S) - R(V, S) \right|,
\end{aligned}
$$

де α_U, α_V, β, γ — числові параметри

1. Відстань найближчого сусіда :

$$R^{б}(W,S) = \min_{w\in W, s\in S} \rho(w,s);$$

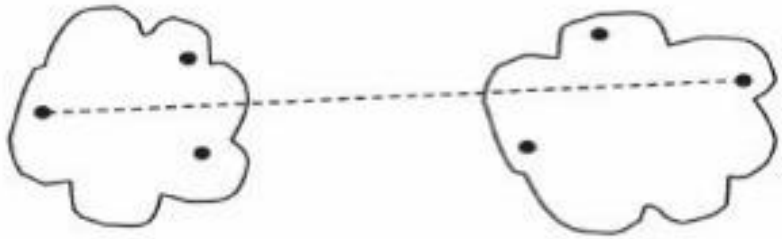
$$\alpha_U = \alpha_V = \tfrac{1}{2},\ \ \beta = 0,\ \ \gamma = -\tfrac{1}{2}.$$



2. Відстань найдальшого сусіда :

$$R^{д}(W,S) = \max_{w\in W, s\in S} \rho(w,s);$$

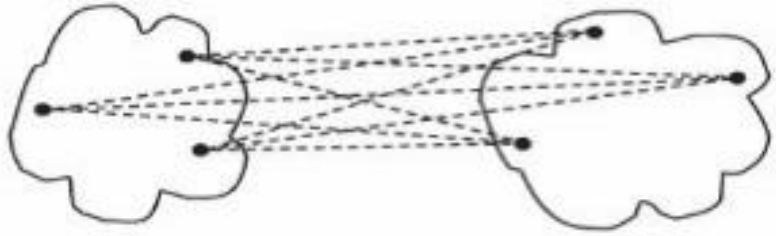
$$\alpha_U = \alpha_V = \tfrac{1}{2},\ \ \beta = 0,\ \ \gamma = \tfrac{1}{2}.$$



3. Групова середня відстань :

$$R^{г}(W,S) = \frac{1}{|W||S|} \sum_{w\in W} \sum_{s\in S} \rho(w,s);$$

$$\alpha_U = \frac{|U|}{|W|},\ \ \alpha_V = \frac{|V|}{|W|},\ \ \beta = \gamma = 0.$$

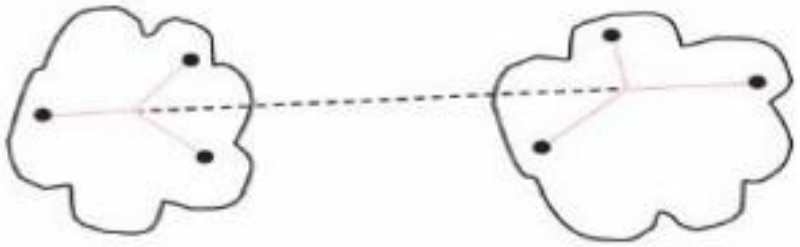


4. Відстань між центрами :

$$R^{ч}(W,S) = \rho^2\Big(\sum_{w\in W}\frac{w}{|W|}, \sum_{s\in S}\frac{s}{|S|}\Big);$$

$$\alpha_U = \frac{|U|}{|W|},\ \ \alpha_V = \frac{|V|}{|W|},$$

$$\beta = -\alpha_U\alpha_V,\ \ \gamma = 0.$$



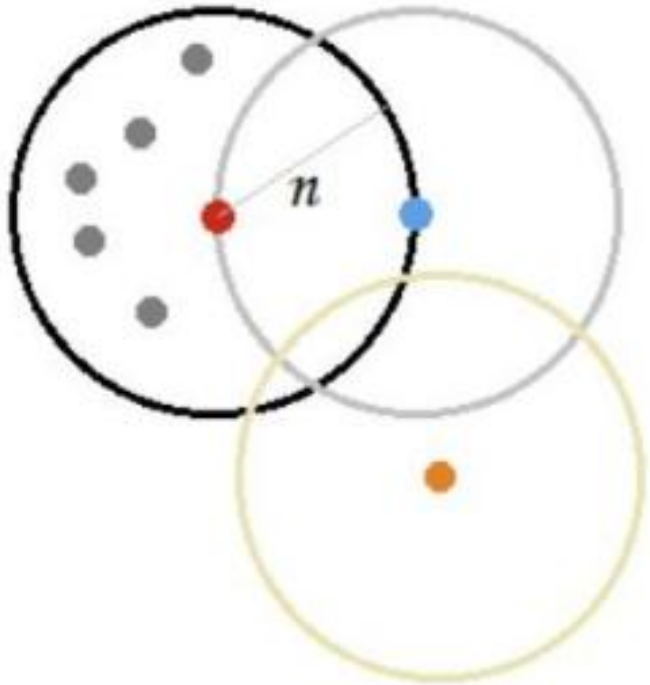
5. Відстань Уорда :

$$R^{у}(W,S) = \frac{|S||W|}{|S|+|W|}\,\rho^2\Big(\sum_{w\in W}\frac{w}{|W|}, \sum_{s\in S}\frac{s}{|S|}\Big);$$

$$\alpha_U = \frac{|S|+|U|}{|S|+|W|},\ \ \alpha_V = \frac{|S|+|V|}{|S|+|W|},\ \ \beta = \frac{-|S|}{|S|+|W|},\ \ \gamma = 0.$$

Алгоритм кластеризації DBSCAN

Об'єкт x ∈ U, його ε-околиця Uε (x) = u ∈ U: ρ (x, u) ≤ ε

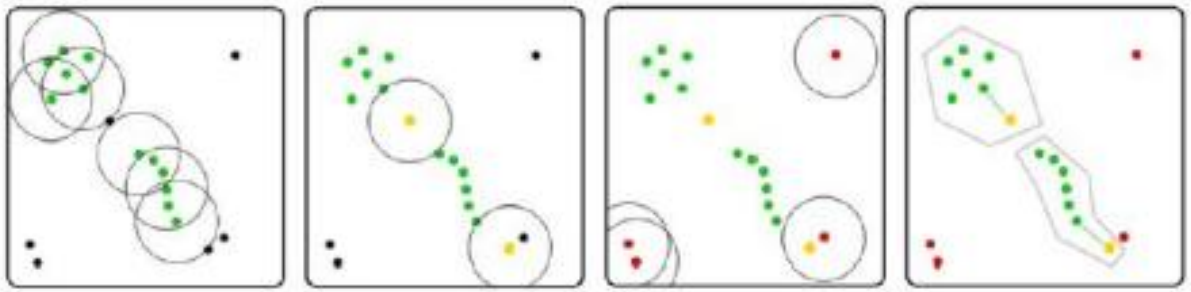


- Core Point | Uε (x) | > M.
- Border Point
- Noise Point

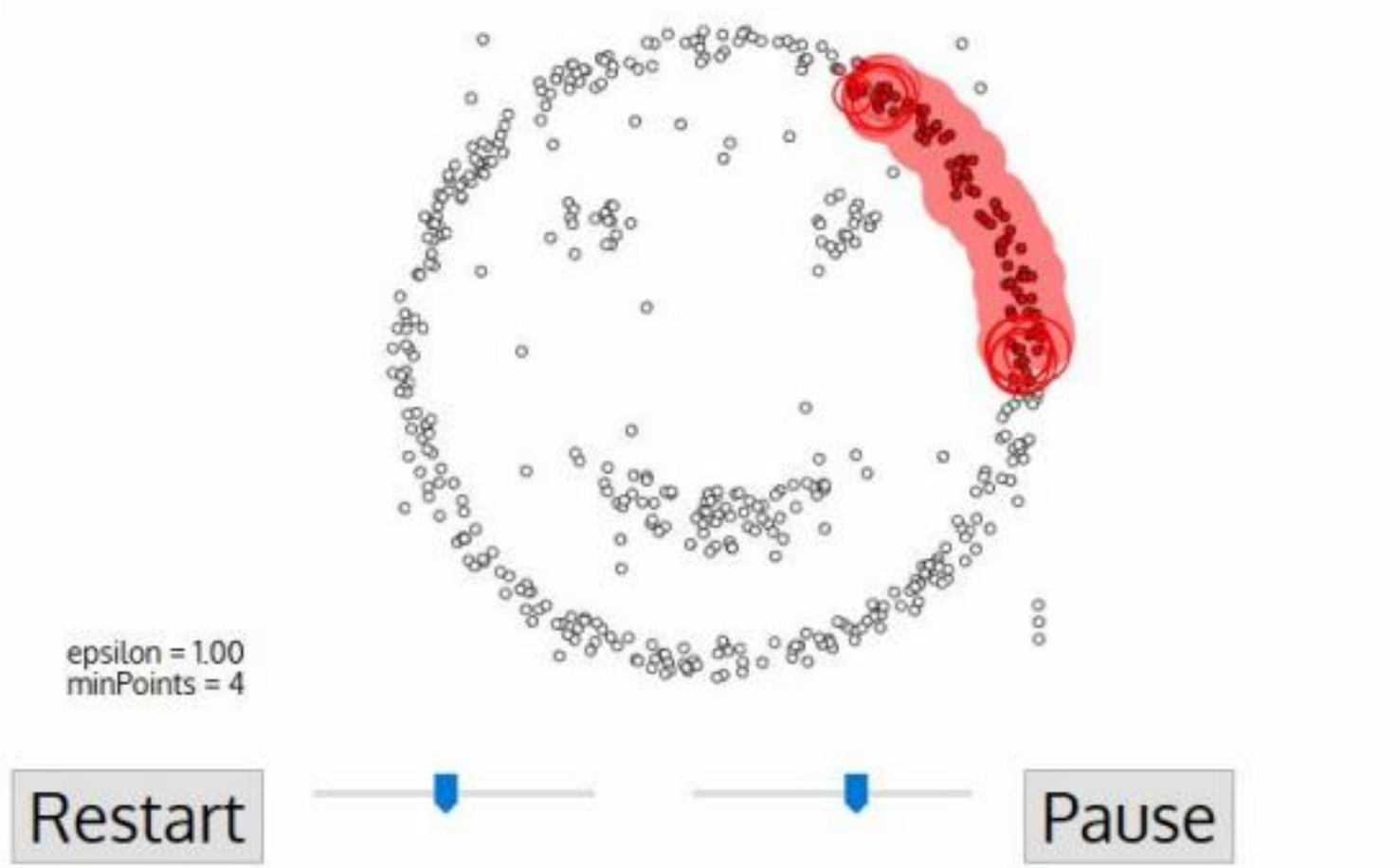
n = Neighbourhood

m = 4

DBSCAN



sklearn.cluster.**DBSCAN**


epsilon = 1.00
minPoints = 4
Restart
Pause

Affinity Propagation

Метрика "схожості", визначається тим, що s (xi, xj)> s (xi, xk) якщо спостереження xi більше схоже на спостереження xj ніж на xk.

Матриця r(i, k) буде описувати, наскільки добре k-те спостереження підходить для того, щоб бути "прикладом для наслідування" для i-того спостереження щодо всіх інших потенційних "прикладів« (responsibility).

Матриця a(i, k) буде описувати, наскільки правильним було б для i-того спостереження вибрати k-те в якості такого "прикладу« (availability).

Affinity Propagation

Поки не досягнуте максимальне значення ітерацій, повторювати послідовне коригування матриць S, R, A. На початку R=0, A=0.

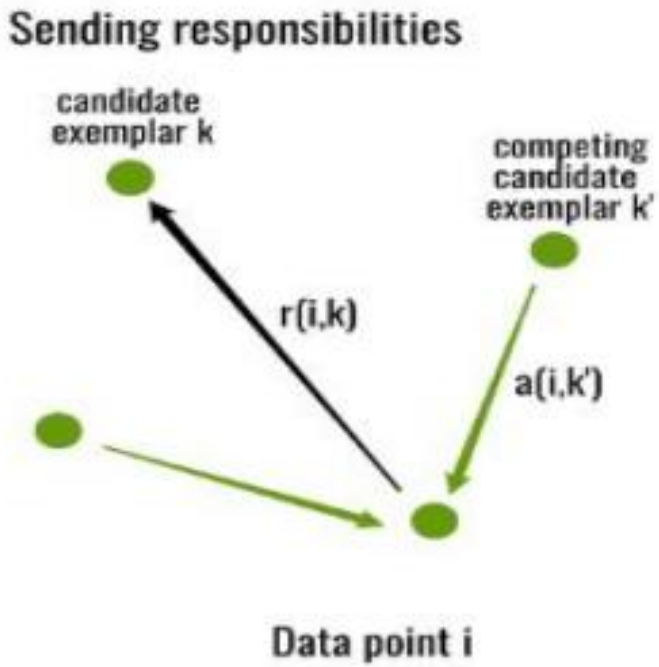
$$r_{i,k} \leftarrow s_(x_i, x_k) - \max_{k' \neq k} \{a_{i,k'} + s(x_i, x'_k)\}$$

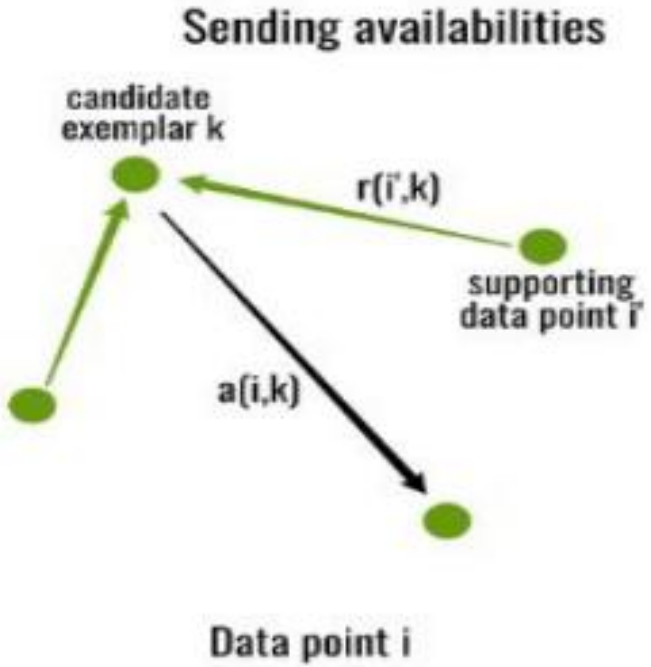
$$a_{i,k} \leftarrow \min \left(0, r_{k,k} + \sum_{i' \notin \{i,k\}} \max(0, r_{i',k})\right), \quad i \neq k$$

$$a_{k,k} \leftarrow \sum_{i' \neq k} \max(0, r_{i',k})$$

$$r_{t+1}(i,k) = \lambda \cdot r_t(i,k) + (1-\lambda) \cdot r_{t+1}(i,k)$$

$$a_{t+1}(i,k) = \lambda \cdot a_t(i,k) + (1-\lambda) \cdot a_{t+1}(i,k)$$





sklearn.cluster.**AffinityPropagation**

EM-алгоритм

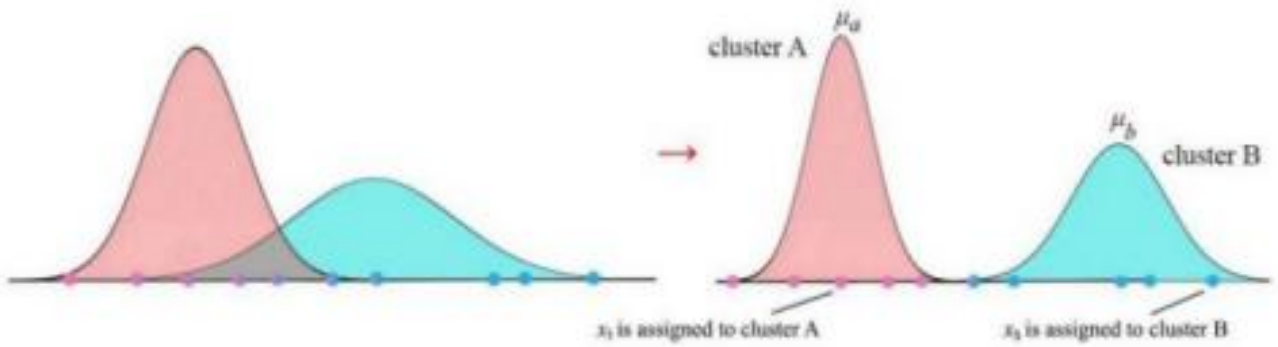
Це загальний метод знаходження оцінок функції правдоподібності в моделях з прихованими змінними, який з суміші розподілів дозволяє будувати (наближати) складні імовірнісні розподіли.

E-Step. Оцінюються відсутні змінні у наборі даних.

M-Крок. Максимізуються параметри моделі за наявних даних.

EM-алгоритм

$$pdf(x_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-\mu)^2/2\sigma^2}$$



EM-алгоритм

$$P(x_1, x_2, x_3, \ldots, x_{100} \mid \underbrace{\mu_a, \sigma_a^2, \mu_b, \sigma_b^2}_{\theta_1}, \underbrace{\text{cluster assignments for 100 points}}_{\theta_2})$$

$$P(x_{new} \mid \underbrace{\mu_a, \sigma_a^2, \mu_b, \sigma_b^2}_{\theta_1})$$

- На початку ми випадковим чином задаємо параметри θ_1 (μa, σa^2, μb, σb^2). Робимо присвоєння P (θ_2) для кожної точки вибірки даних. Обчислюємо ймовірність *належності* x_i до певного кластера. Потім на основі цього розподілу P (θ_2) ми оптимізуємо очікувану логарифмічну ймовірність спостереження x присвоєння кластеру (θ_1). Отже, ми фіксуємо θ_1 і виводимо θ_2. Тоді ми оптимізуємо θ_1 з фіксованим θ_2 . Повторюємо ітерації.
- **E-Step**. Оцінюємо очікуване значення для кожної прихованої змінної.
- **M-Крок**. Оптимізуємо параметри розподілу, використовуючи максимальну ймовірність.

EM-алгоритм

E-step

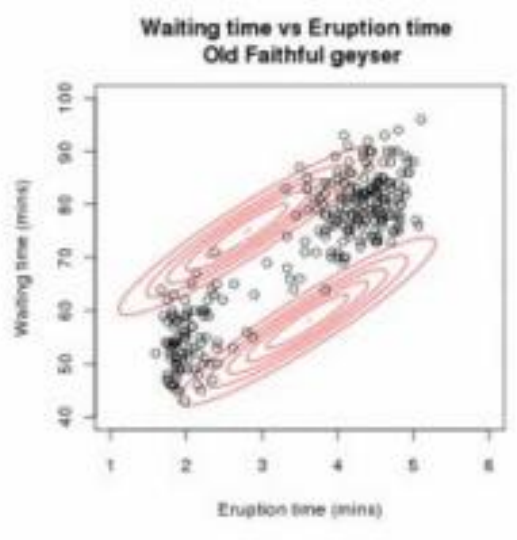
$$P(x_i|b) = \frac{1}{\sigma_b\sqrt{2\pi}} e^{-(x-\mu_b)^2/2\sigma_b^2}$$

$$b_i = P(b|x_i) = \frac{P(x_i|b)P(b)}{P(x_i|b)P(b) + P(x_i|a)P(a)} \quad \text{(for every data point)}$$

$$a_i = P(a|x_i) = 1 - b_i$$

$$P(b) = \frac{b_1 + b_2 + \cdots + b_n}{n}$$

$$P(a) = 1 - P(b)$$



M-step

$$\mu_b = \frac{b_1x_1 + b_2x_2 + \cdots + b_nx_n}{b_1 + b_n + \cdots + b_n}$$

$$\sigma_b^2 = \frac{b_1(x_1 - \mu_b)^2 + \cdots + b_1(x_n - \mu_b)^2}{b_1 + b_2 + \cdots + b_n}$$

$$\mu_a = \frac{a_1x_1 + a_2x_2 + \cdots + a_nx_n}{b_1 + a_n + \cdots + a_n}$$

$$\sigma_a^2 = \frac{a_1(x_1 - \mu_a)^2 + \cdots + a_1(x_n - \mu_a)^2}{a_1 + a_2 + \cdots + a_n}$$

a.k.a.

$$\arg\max_{\theta_1} p(x \mid \theta_1) = \sum_{\theta_2} p(x, \theta_2 \mid \theta_1)$$

$$\theta_1 = \{\mu_a, \sigma_a, \mu_b, \sigma_b\}$$

Кластеризація за допомогою EM-алгоритму

Нехай кластер характеризується гаусівським розподілом, с параметрами w_y –ймовірності кластерів, μ_y –середнє, g – приховані змінні.

1: обрати початкове наближення для всіх кластерів $y \in Y$:

$w_y := 1/|Y|$;

μ_y $\sigma_{yj}^2 := \frac{1}{\ell|Y|}\sum_{i=1}^{\ell}(f_j(x_i) - \mu_{yj})^2,\ \ j = 1,\ldots,n;$ ки;

2: повторювати

3: E-крок (expectation):

$$g_{iy} := \frac{w_y p_y(x_i)}{\sum_{z \in Y} w_z p_z(x_i)},\ \ y \in Y,\ \ i = 1,\ldots,\ell;$$

4: М-крок (maximization):

$$w_y := \frac{1}{\ell}\sum_{i=1}^{\ell} g_{iy}, \quad y \in Y;$$

$$\mu_{yj} := \frac{1}{\ell w_y}\sum_{i=1}^{\ell} g_{iy} f_j(x_i), \quad y \in Y, \quad j = 1, \ldots, n;$$

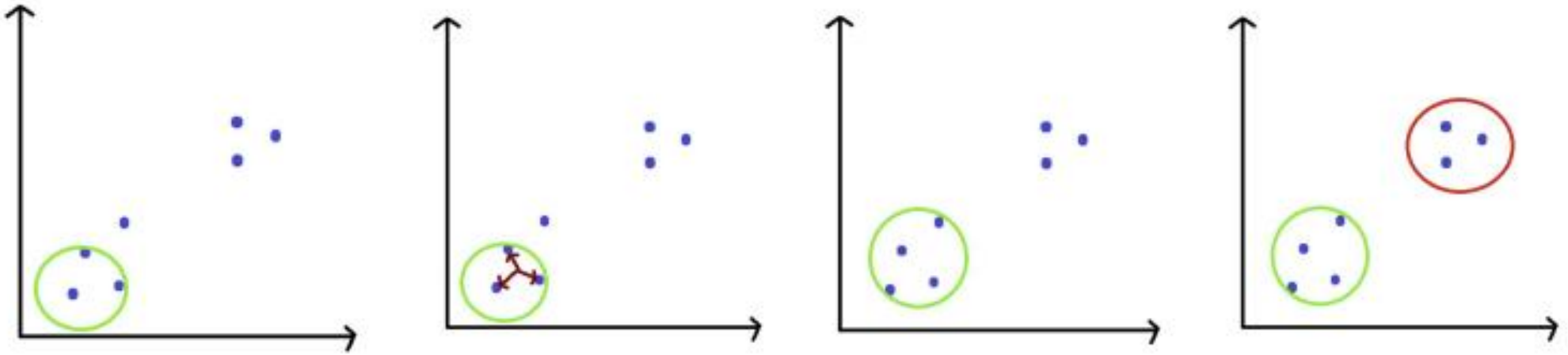
$$\sigma^2_{yj} := \frac{1}{\ell w_y}\sum_{i=1}^{\ell} g_{iy}(f_j(x_i) - \mu_{yj})^2, \quad y \in Y, \quad j = 1, \ldots, n;$$

5: Співставити об’єкти і кластери за байєсівським правилом:

$$y_i := \arg\max_{y \in Y} g_{iy}, \quad i = 1, \ldots, \ell;$$

6: поки уі не припинять змінюватися

Mean Shift

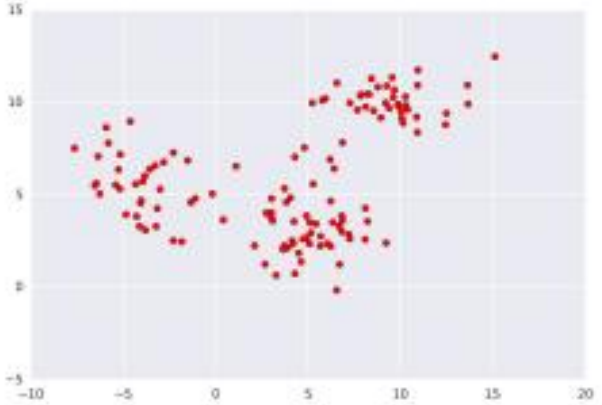


$$K(x_i - x) = e^{-c||x_i - x||^2}$$

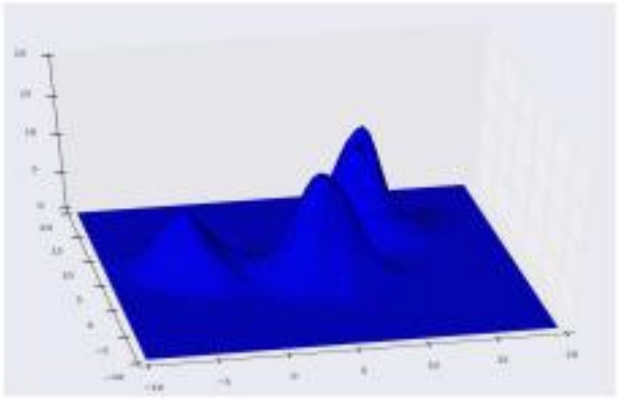
$$m(x) = \frac{\sum_{x_i \in N(x)} K(x_i - x) x_i}{\sum_{x_i \in N(x)} K(x_i - x)}$$

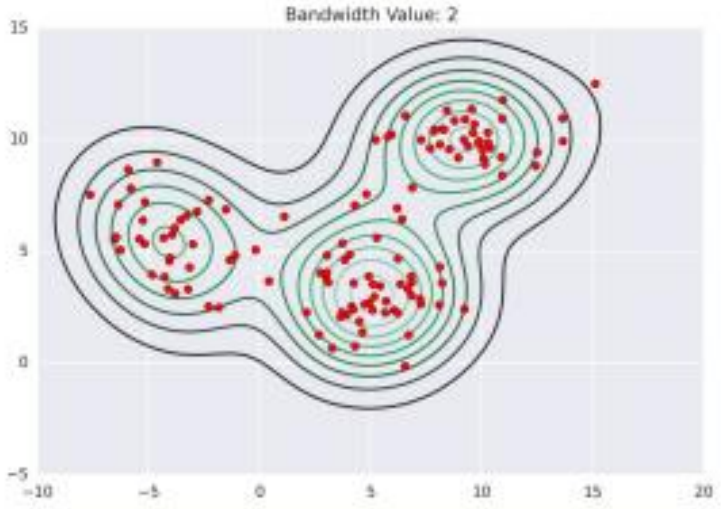
$$x_i^{t+1} = m(x_i^t)$$

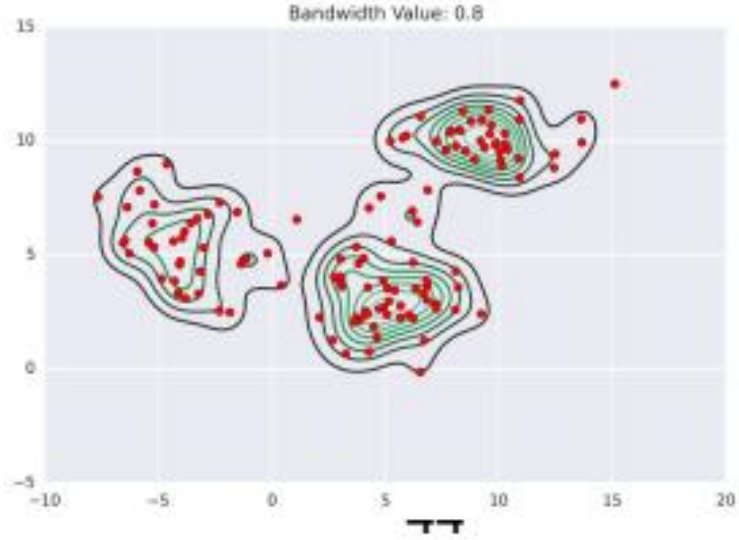
Mean Shift



$$\sum_i K(x - x_i) = \sum_i k\left(\frac{\|x - x_i\|^2}{h^2}\right)$$







When poll is active, respond at **pollev.com/nataliashovgun288**

Чи відомий вам SVM?

Так

Ні

Powered by Poll Everywhere

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**