Математические основы информационной безопасности

Груздев Дмитрий Николаевич

Обучение без учителя

Кластеризация

$x_1, x_2, \ldots, x_m$ – обучающая выборка
$\rho(x_i, x_j)$ – расстояние между объектами

Построить:

- Y – множество кластеров
- $A: X \rightarrow Y$ – алгоритм кластеризации объектов

Требования к кластерам:

- кластер состоит из близких объектов
- объекты разных кластеров существенно различны

Кластеризация

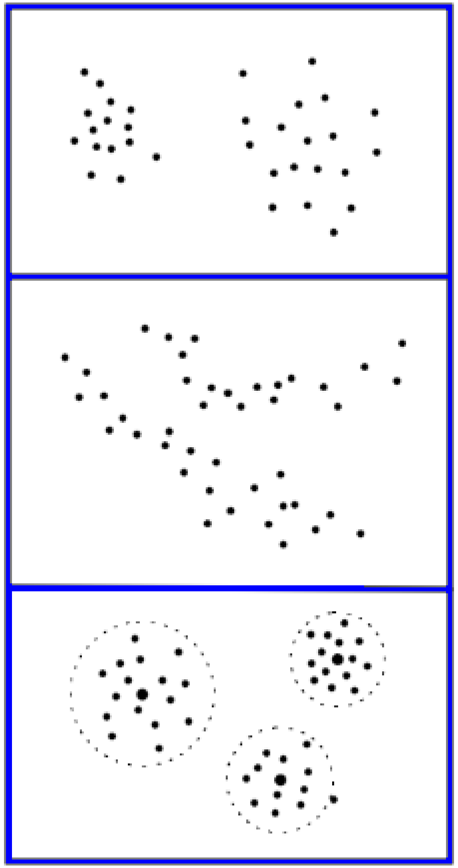
Трудности при формализации задачи

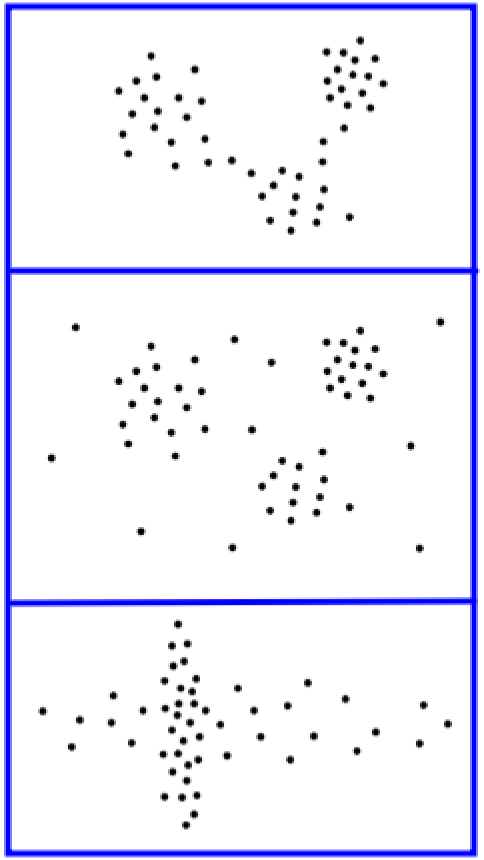
- существует много критериев качества кластеризации
- число кластеров обычно неизвестно заранее

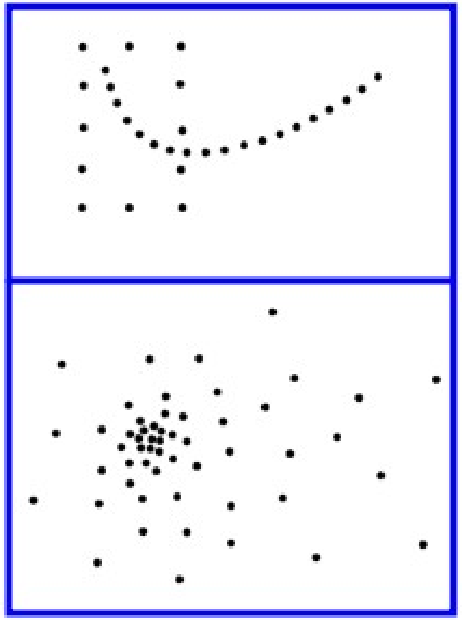
Цели кластеризации:

- упростить обработку данных
- сократить объем хранимых данных
- выделить нетипичные признаки
- построить иерархию множества объектов

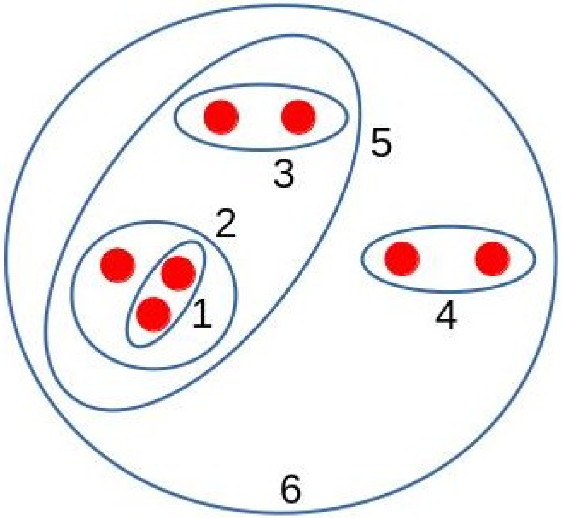
Виды кластеров

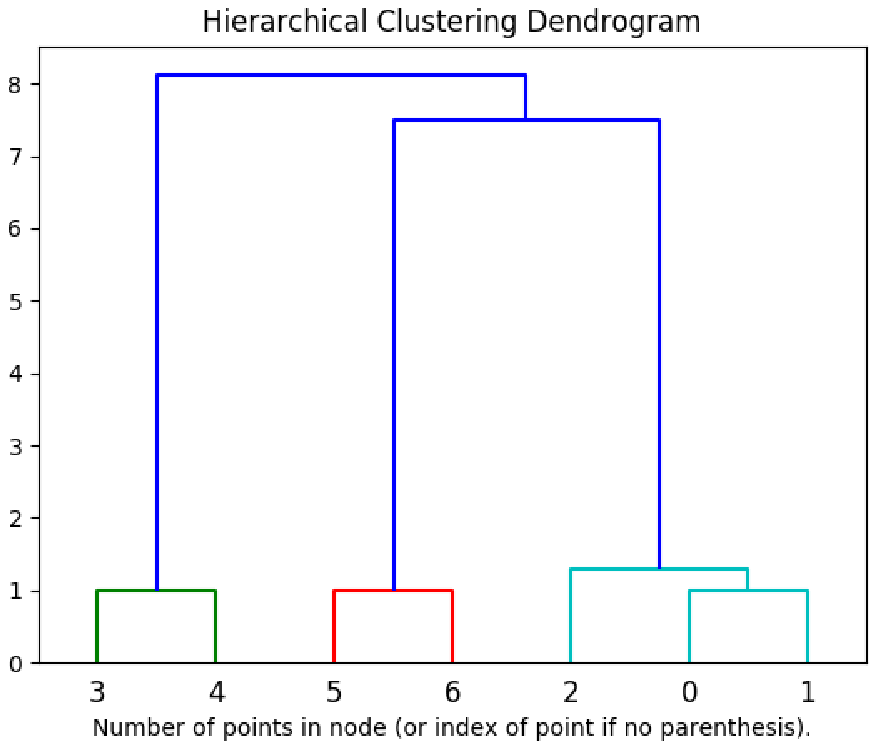






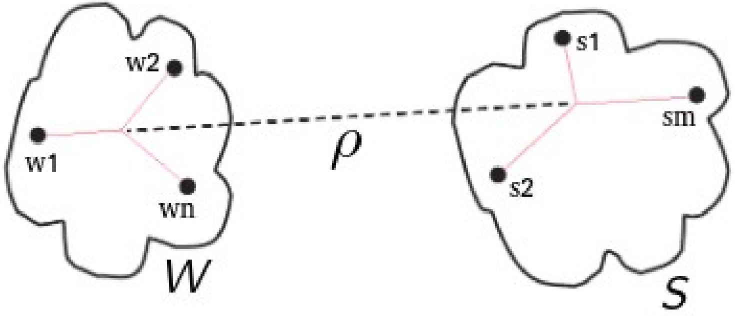
Иерархическая кластеризация





Последовательно объединяем два ближайших друг к другу кластера в новый.

Расстояние между кластерами



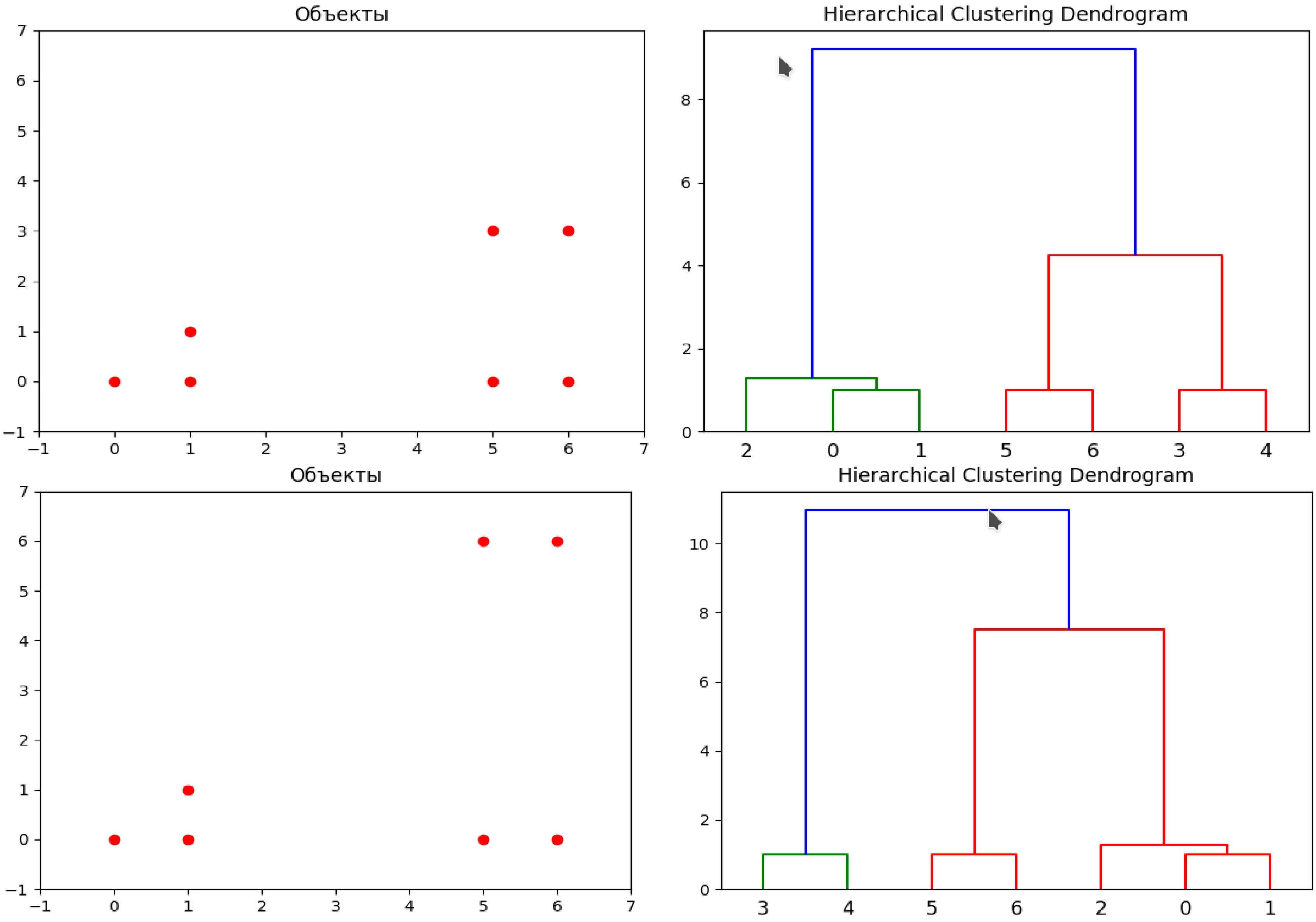
$$\rho(W,S) = \frac{|S||W|}{|S|+|W|}\,\overbrace{\rho^2\left(\sum_{w\in W}\frac{w}{|W|}, \sum_{s\in S}\frac{s}{|S|}\right)}^{\text{расстояние между центрами}}$$ - расстояние Уорда

Индукционное вычисление расстояния Уорда:

W = A⋃B

ρ(A⋃B, S) = α * ρ(A, S) + β * ρ(B, S) + γ * ρ(A, B), где

α = (|A|+|S|)/(|S|+|W|), β = (|B|+|S|)/(|S|+|W|), γ = -|S|/(|S|+|W|)


Объекты
Hierarchical Clustering Dendrogram
Объекты
Hierarchical Clustering Dendrogram

Метод k-средних

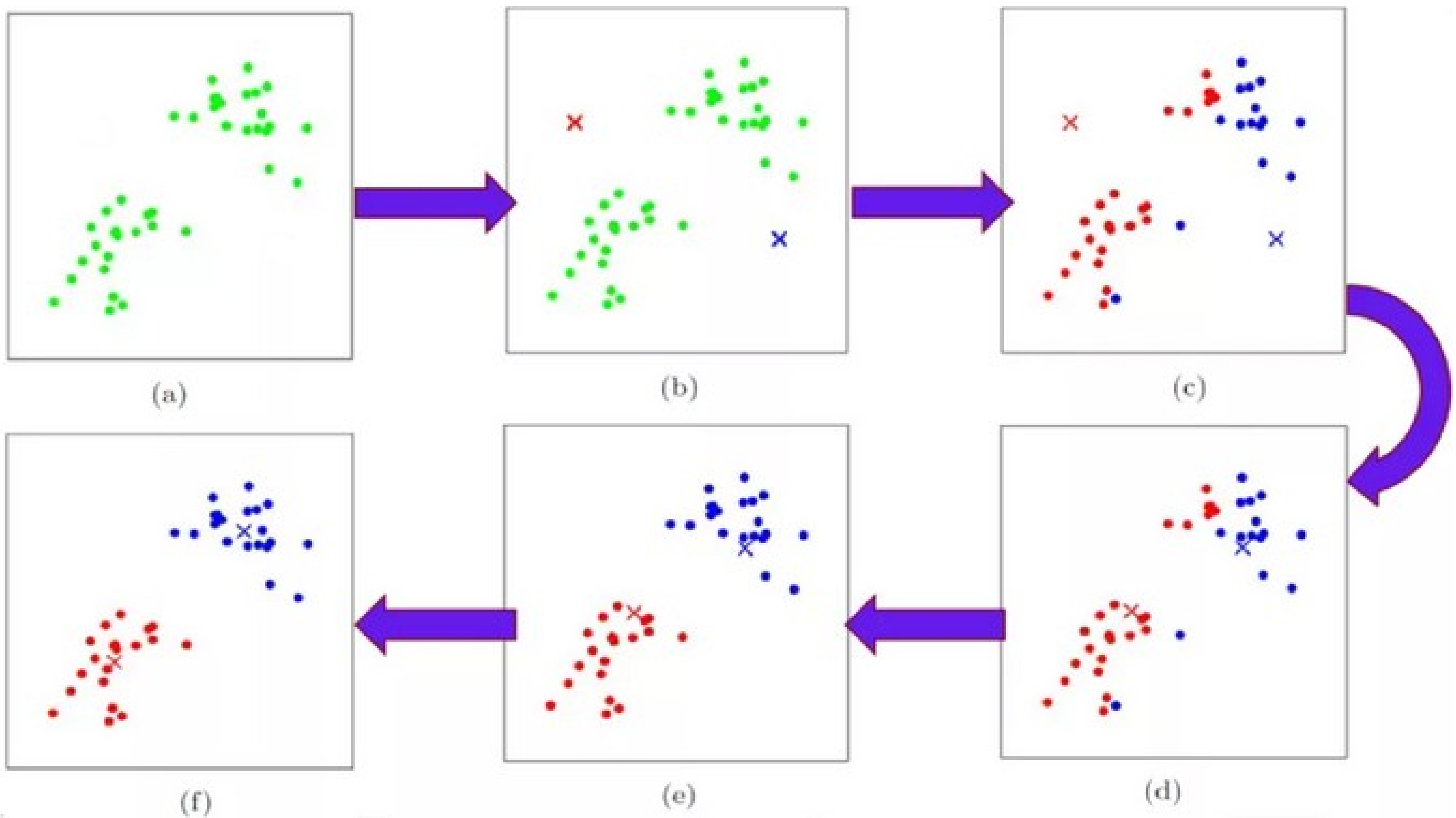
$x_1, x_2, \ldots, x_m$ – обучающая выборка

$x_i = (x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(n)})$

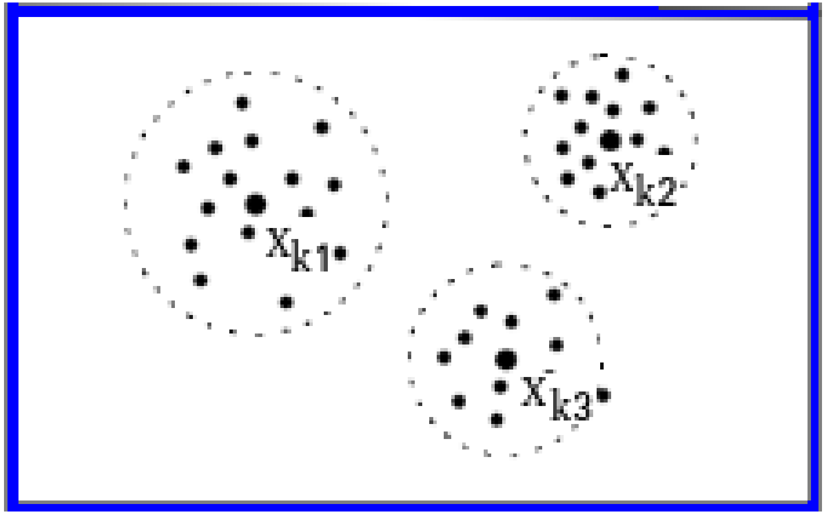
Алгоритм:

1. Задать количество кластеров и их центры $x_{K1}, \ldots x_{Km}$.
2. Отнести каждый объект к ближайшему центру.
3. Перенести координаты центров кластеров в центры тяжести соответствующих групп объектов.
4. Повторять 2 и 3, пока происходят переходы объектов между кластерами.

Метод k-средних



Оптимальное число кластеров

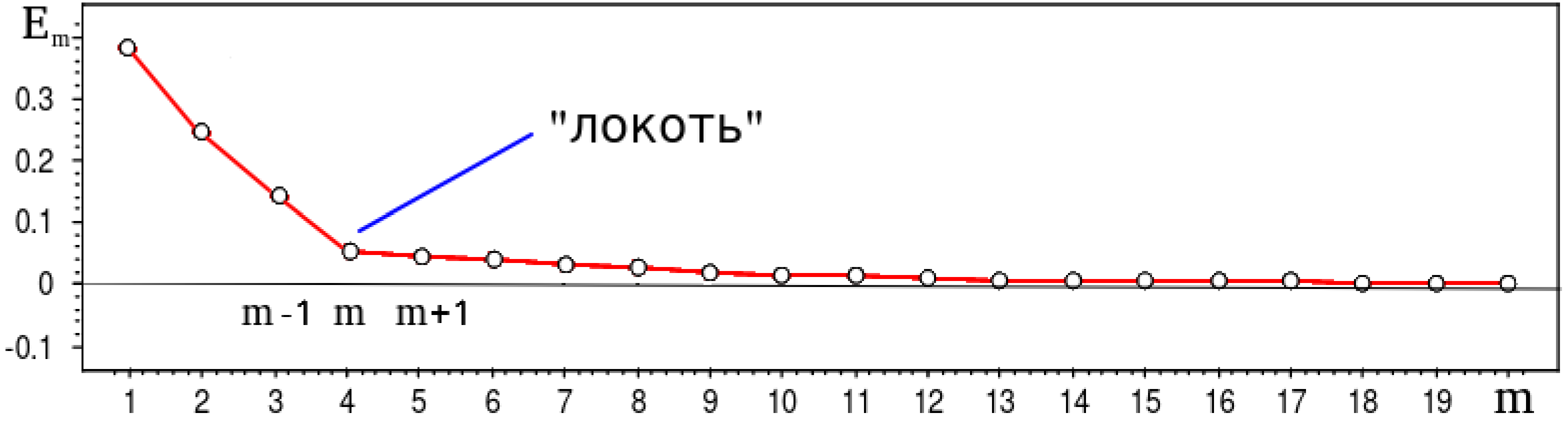


Ошибка кластеризации:

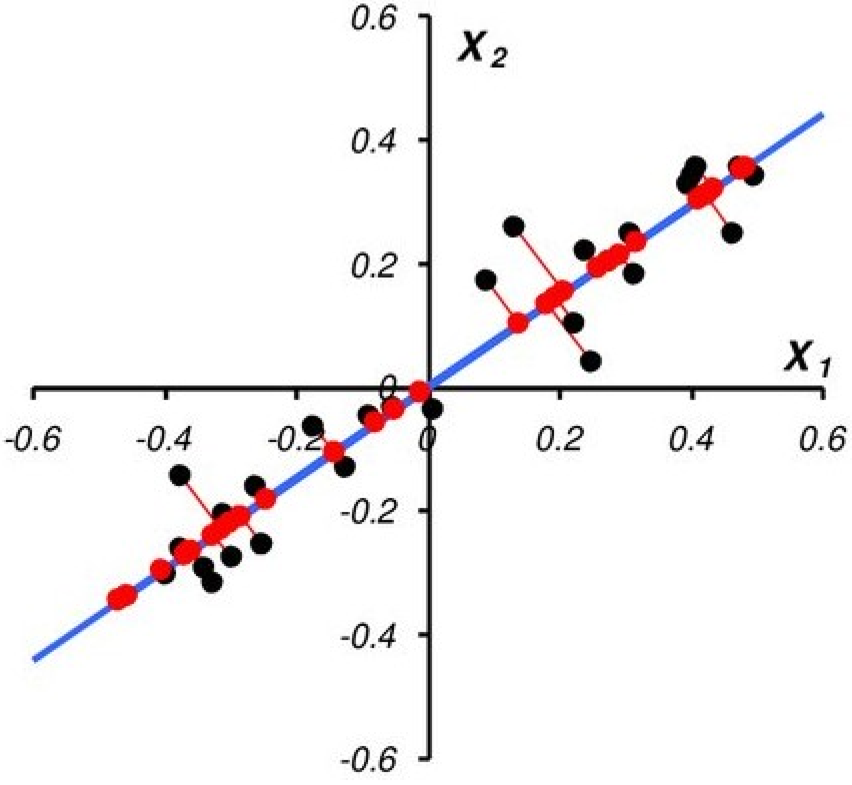
$$E_m = \Sigma_{1\le i\le m}\Sigma_{1\le j\le |Ki|}\|x_{Ki}^{(j)} - x_{Ki}\|^2$$

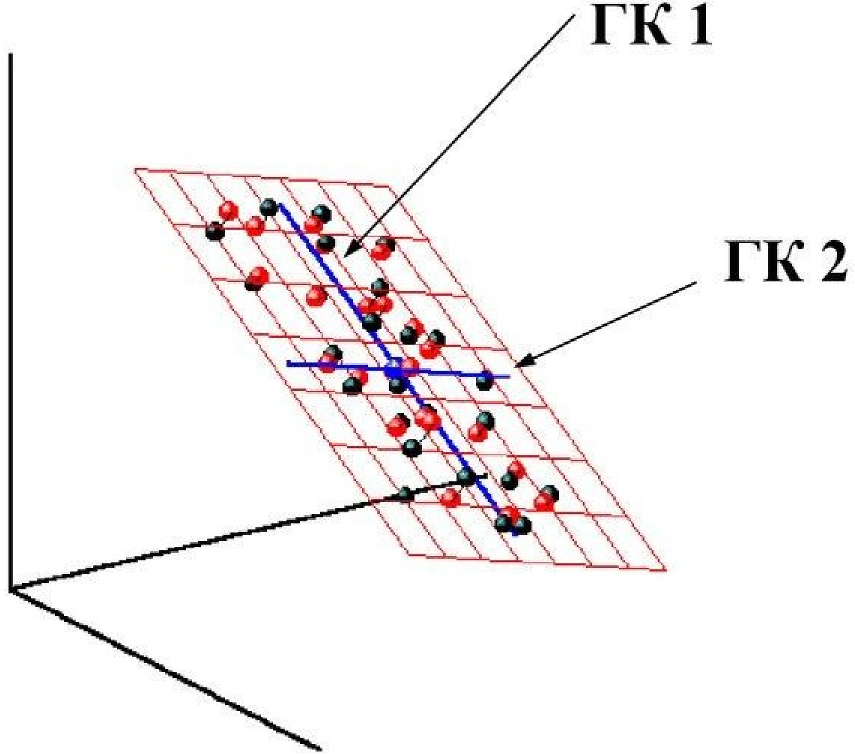
Выбор m – числа кластеров:

$$E_m - E_{m+1} \ggg E_{m-1} - E_m$$



Метод главных компонент





Метод главных компонент

$x_1, x_2, \ldots, x_m$ – обучающая выборка

$x_i = (x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(n)})$

$x_i = (g_i^{(1)}, g_i^{(2)}, \ldots, g_i^{(k)}),\ \ k \le n$

$x_i^{(j)} \approx \Sigma g_i^{(l)} u_l = h_i^{(j)}$

$\Sigma_{1\le i\le m}\Sigma_{1\le j\le n}(h_i^{(j)} – x_i^{(j)})^2 \to \min_{(u,g)}$

Преобразование признаков

$$X = \begin{pmatrix} x_1^{(1)} & \dots & x_1^{(n)} \\ & & \\ x_m^{(1)} & \dots & x_m^{(n)} \end{pmatrix} \quad G = \begin{pmatrix} g_1^{(1)} & \dots & g_1^{(k)} \\ & & \\ g_m^{(1)} & \dots & g_m^{(k)} \end{pmatrix} \quad U = \begin{pmatrix} u_1^{(1)} & \dots & u_1^{(k)} \\ & & \\ u_n^{(1)} & \dots & u_n^{(k)} \end{pmatrix}$$

$X \approx GU^T$; $\|GU^T - X\|^2 \to \min_{G,U}$

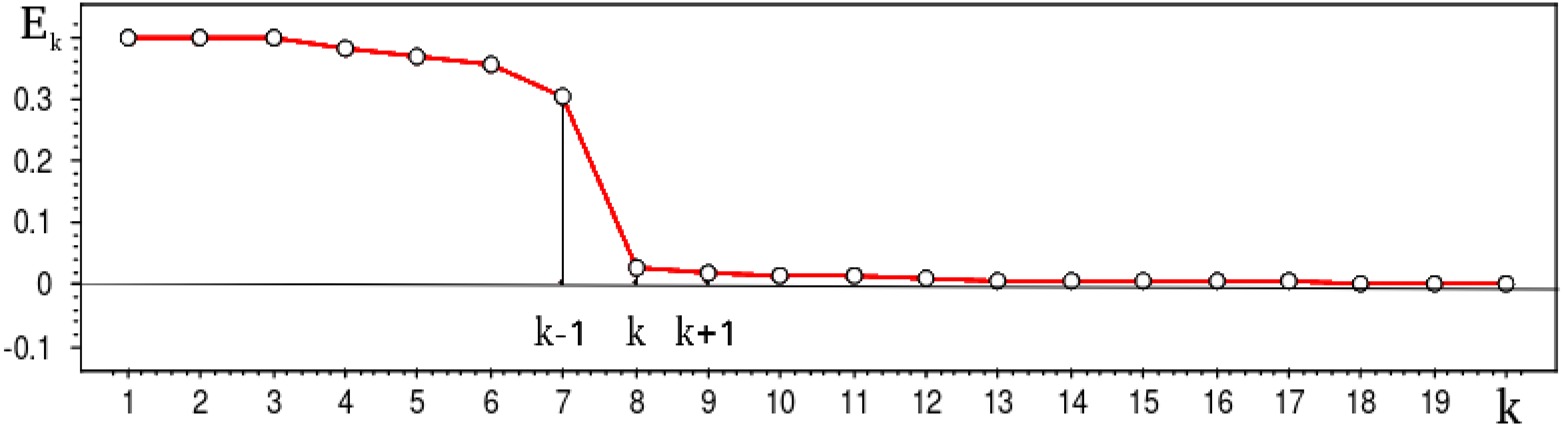
Решение:

$G = \mathrm{diag}(\lambda_1, \lambda_2, \dots, \lambda_k)$, $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_k \geq \dots \geq \lambda_n$

- $UU^T = I_k$ – ортонормирована, и $G \approx XU$, $X \approx GU^T$

$\|GU^T - X\|^2 = \lambda_{k+1} + \lambda_{k+2} + \dots + \lambda_n$

Эффективная размерность



$E_k = \|GU^T - F\|^2 / \|F\|^2 = (\lambda_{k+1} + \ldots + \lambda_n) / (\lambda_1 + \ldots + \lambda_n)$

Установить допустимую погрешность $E_k \leq \varepsilon$ или

если $E_{k-1} \ggg E_k$, то стоит выбрать размерность k.

Диагностика аномалий

$x_1, x_2, \ldots, x_m$ – обучающая выборка

$x_i = (x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(n)})$

Особенности:

- большое количество нормальных результатов,
- малое число (отсутствие) результатов с аномалиями,
- типы аномалий трудно систематизировать,
- возможны неизвестные аномалии.

Случайные величины

Теория вероятности

X, Y – случайные величины

M_X – математическое ожидание X

D_X – дисперсия X

$\sigma_X = D_X^{1/2}$ – стандартное отклонение X

$cov_{XY} = M((X - M_X)(Y - M_Y))$ – ковариация X и Y

$$r_{XY} = \frac{cov_{XY}}{\sigma_X \sigma_Y}$$ -

коэффициент линейной корреляции (коэффициент корреляции Пирсона)

Математическая статистика

$X = \{x_1, \dots, x_N\}$, $Y = \{y_1, \dots, y_N\}$ – выборки

$\overline{X} = \frac{1}{n} \sum x_i$ - выборочное среднее

$S_X^2 = \frac{1}{n} \sum (x_i - \overline{X})^2$ - выборочная дисперсия

$cov_{XY} = \frac{1}{n} \sum (x_i - \overline{X})(y_i - \overline{Y})$ - выборочная ковариация

$$r_{XY} = \frac{cov_{XY}}{S_X S_Y} = \frac{\sum (x_i - \overline{X})(y_i - \overline{Y})}{\sqrt{\sum (x_i - \overline{X})^2 \sum (y_i - \overline{Y})^2}}$$

коэффициент корреляции Пирсона

Вероятность появления объекта

$x(i) \sim N(\mu_i, \sigma_i^2)$

$\mu_i = \Sigma_{1 \le j \le m} x_j(i) / m$ – выборочное среднее

$\sigma_i^2 = \Sigma_{1 \le j \le m} (x_j(i) - \mu_i)^2 / m$ – выборочная дисперсия

$p(x^{*(i)}) = p(x^{*(i)}; \mu_i, \sigma_i^2) = f(x^{*(i)}; \mu_i, \sigma_i^2) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{(x^{*(i)} - \mu_i)^2}{2\sigma_i^2}\right]$

$x^* = (x^{*(1)}, \ldots, x^{*(n)})$

Если $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$ – независимы, то

$p(x^*) = p(x^{*(1)}; \mu_1, \sigma_1^2) * p(x^{*(2)}; \mu_2, \sigma_2^2) * \ldots * p(x^{*(n)}; \mu_n, \sigma_n^2)$

Диагностика аномалий

Алгоритм диагностики:

- выбрать $\varepsilon > 0$ пороговое:
- если $p(x^*) < \varepsilon$, то объект относится к аномальным.

Настройка ε:

- алгоритм должен определять известные аномалии,
- алгоритм не должен относить к аномалиям нормальные объекты
- если не удается подобрать пороговый параметр, то нужны другие признаки объектов

scikit-learn

https://sesc-infosec.github.io/