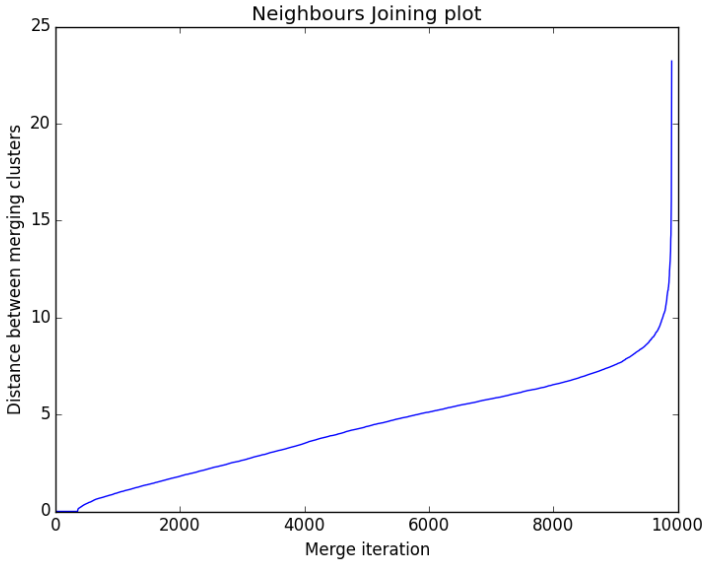
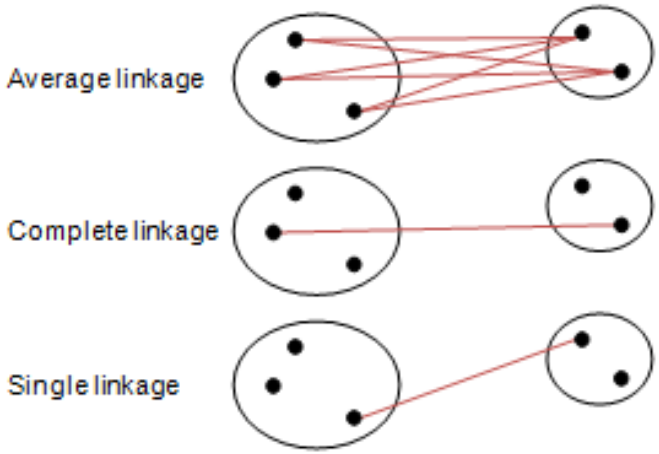




Анализ данных на практике

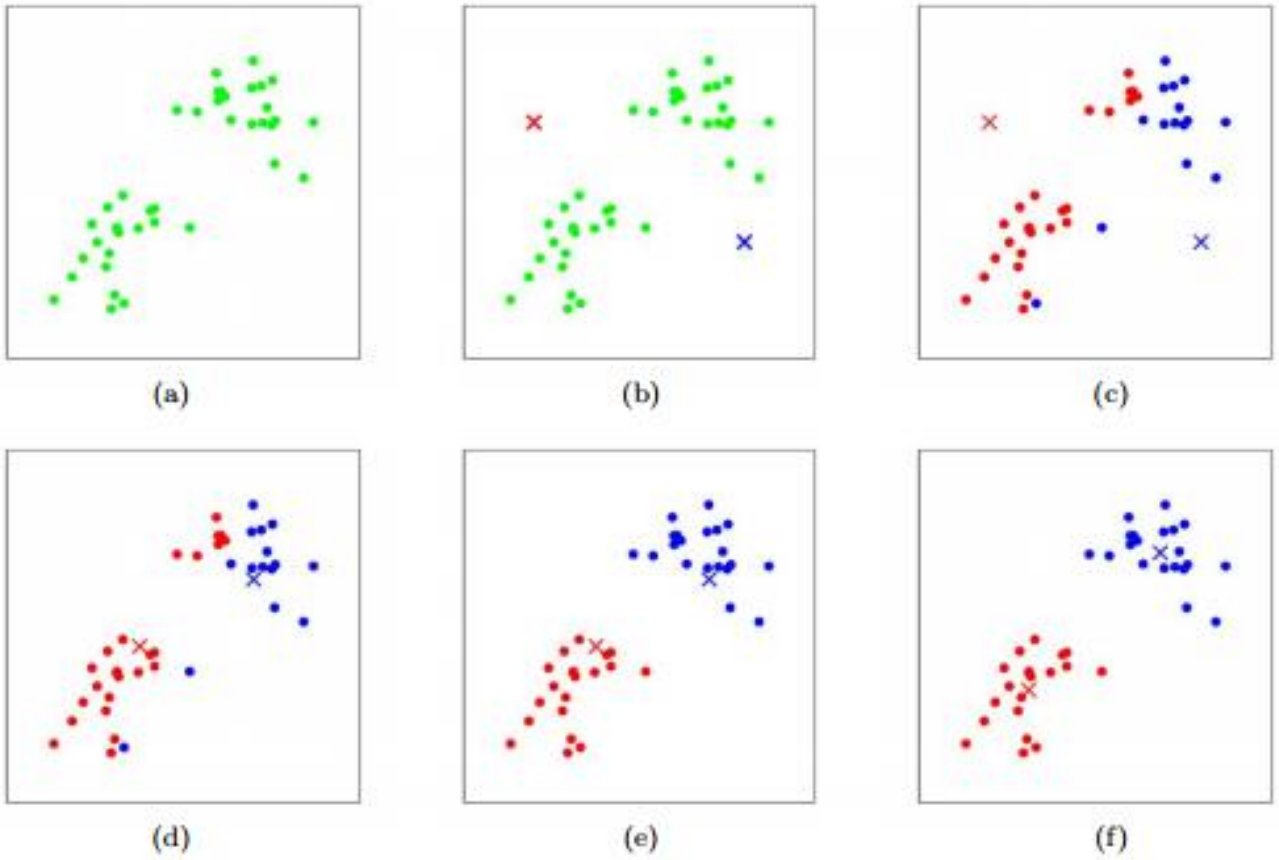
Кластеризация

Виктор Кантор

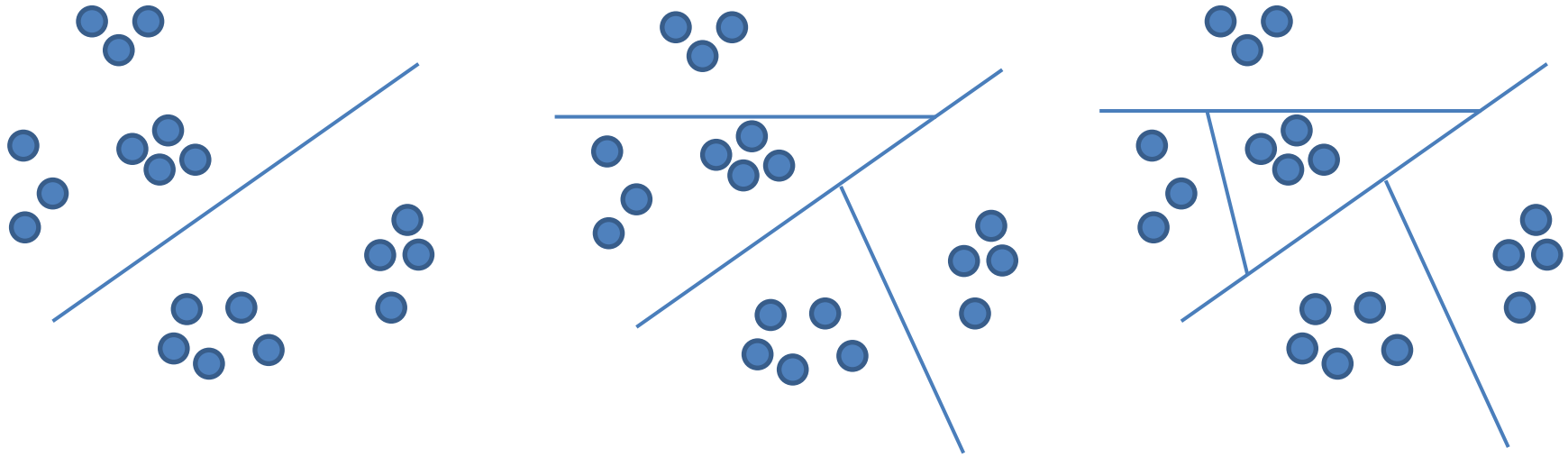
Как могут выглядеть кластеры

Универсального метода кластеризации нет

Простой метод: k-Means



Подбор числа кластеров: BisectKMeans



Более сложный метод: EM

$$p(x) = \sum_{j=1}^{K} w_j p_j(x)\,, \qquad p_j(x) = \varphi(\theta_j; x)$$

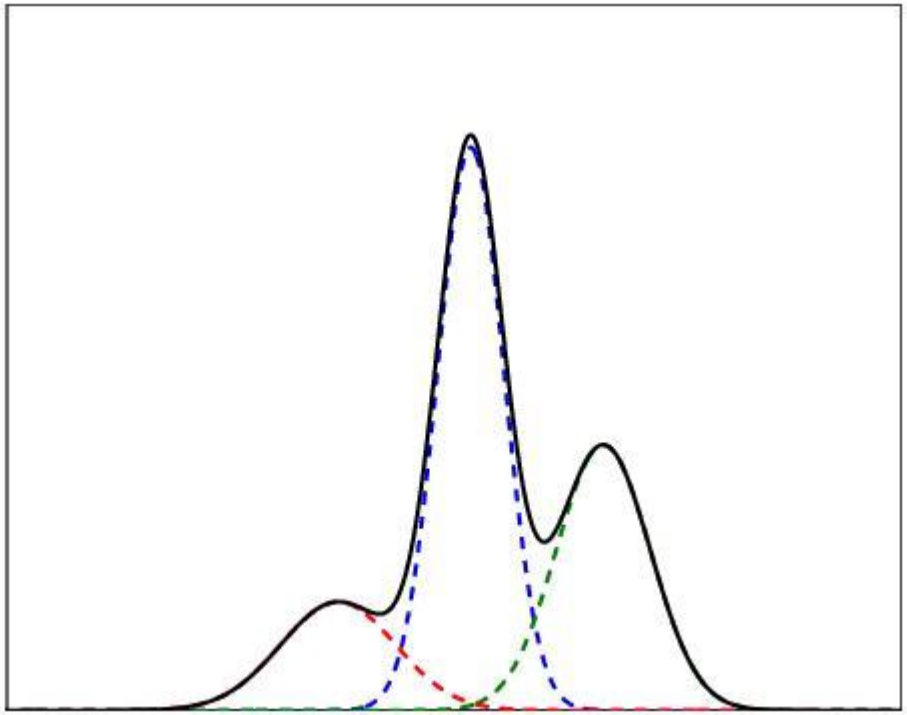
<u>E-шаг:</u>

$$\mathrm{g_{ji}} = p(j|x_i) = \frac{w_j p_j(x_i)}{p(x_i)}$$

<u>M-шаг:</u>

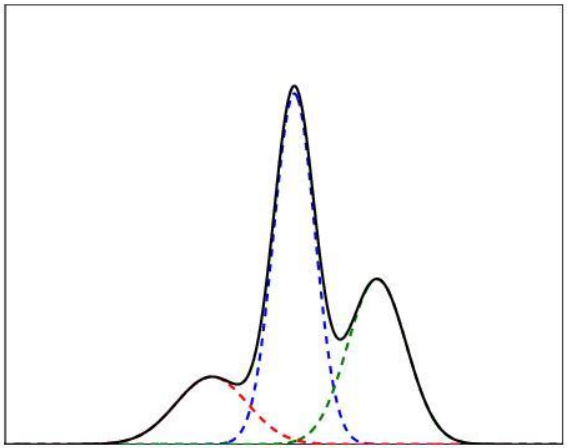
$$w_j = \frac{1}{N}\sum_{i=1}^{N} g_{ji} \qquad \theta_j = argmax_\theta \sum_{i=1}^{N} g_{ji} \ln \varphi(\theta; x)$$

Задача разделения смеси распределений



$$p(x) = \sum_{j=1}^{K} w_j p_j(x), \qquad p_j(x) = \varphi(\theta_j; x)$$

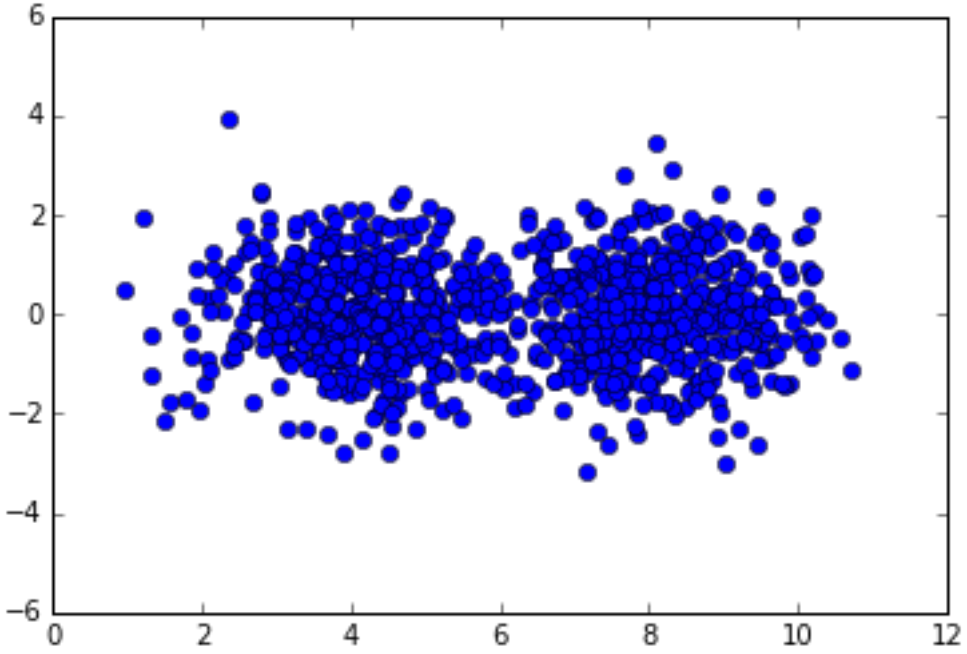
Задача разделения смеси распределений



$$p(x) = \sum_{j=1}^{K} w_j p_j(x), \qquad p_j(x) = \varphi(\theta_j; x)$$

$$w, \theta = argmax_{\theta,w} \sum_{j=1}^{K} \ln p(x_i)$$

Выборка из смеси гауссовских распределений

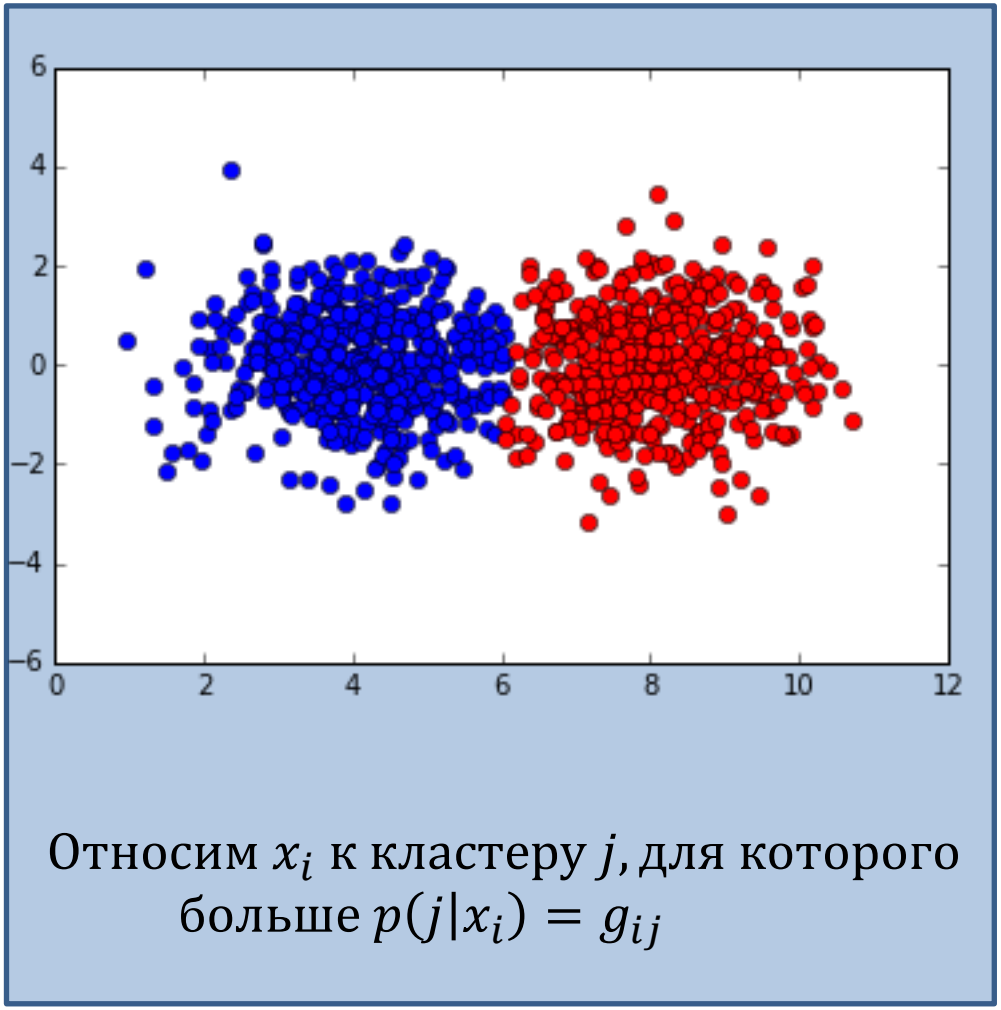


Точки сгенерированы из смеси:

$$p(x) = \frac{1}{2}\mathcal{N}\left(\begin{pmatrix}4\\0\end{pmatrix}, 1\right) + \frac{1}{2}\mathcal{N}\left(\begin{pmatrix}8\\0\end{pmatrix}, 1\right)$$

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

Кластеризация EM-алгоритмом



Относим x_i к кластеру j, для которого больше $p(j|x_i) = g_{ij}$

$$p(x) = w_1 p_1(x) + w_2 p_2(x)$$

<u>E-шаг:</u> $g_{ji} = p(j|x_i) = \frac{w_j p_j(x_i)}{p(x_i)}$

<u>M-шаг:</u>

$$w_j = \frac{1}{N}\sum_{i=1}^{N} g_{ji}$$

$$\mu_j = \frac{1}{N w_j}\sum_{i=1}^{N} g_{ij} x_i$$

$$\Sigma_j = \frac{1}{N w_j - 1}\sum_{i=1}^{N} g_{ij}(x_i - \mu_j)(x_i - \mu_j)^T$$

Идея density-based методов

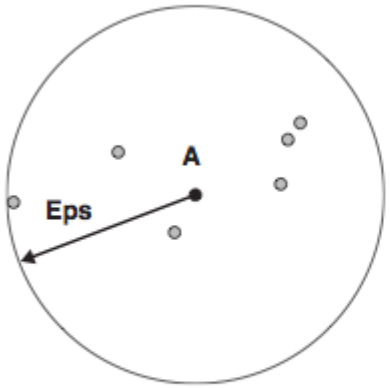


Figure 8.20. Center-based density.

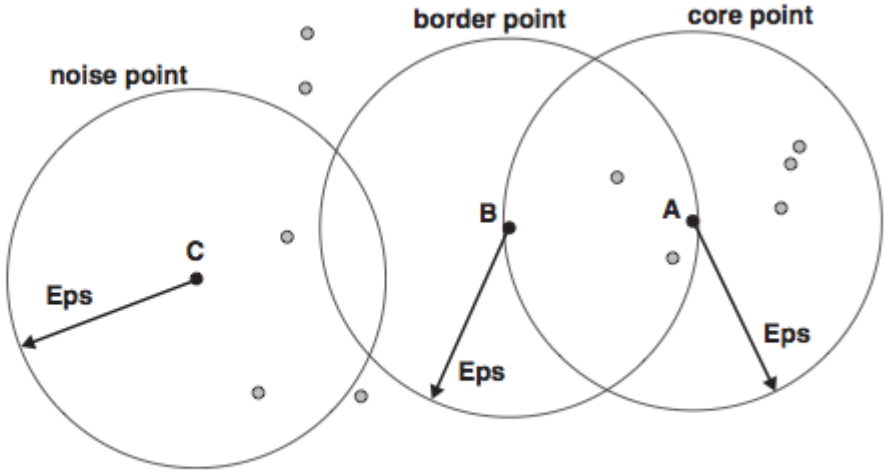
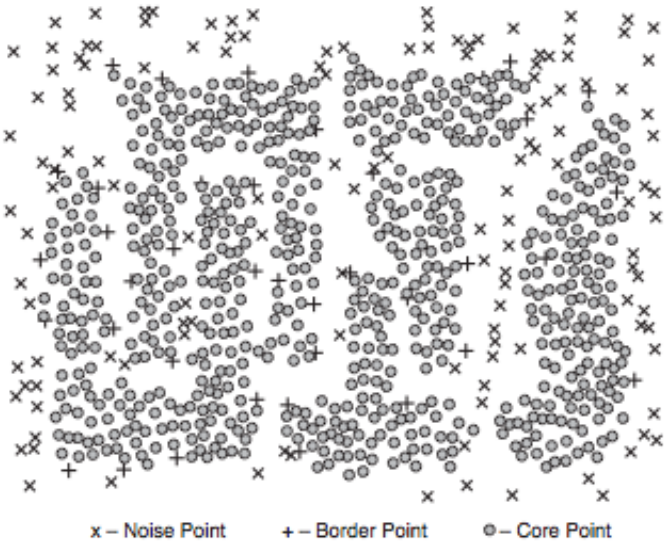


Figure 8.21. Core, border, and noise points.

DBSCAN

(a) Clusters found by DBSCAN.



(b) Core, border, and noise points.

1: Пометить все точки, как основные, пограничные или шумовые.

2: Отбросить точки шума.

3: Соединить все основные точки, находящиеся на расстоянии Eps радиуса одна от другой.

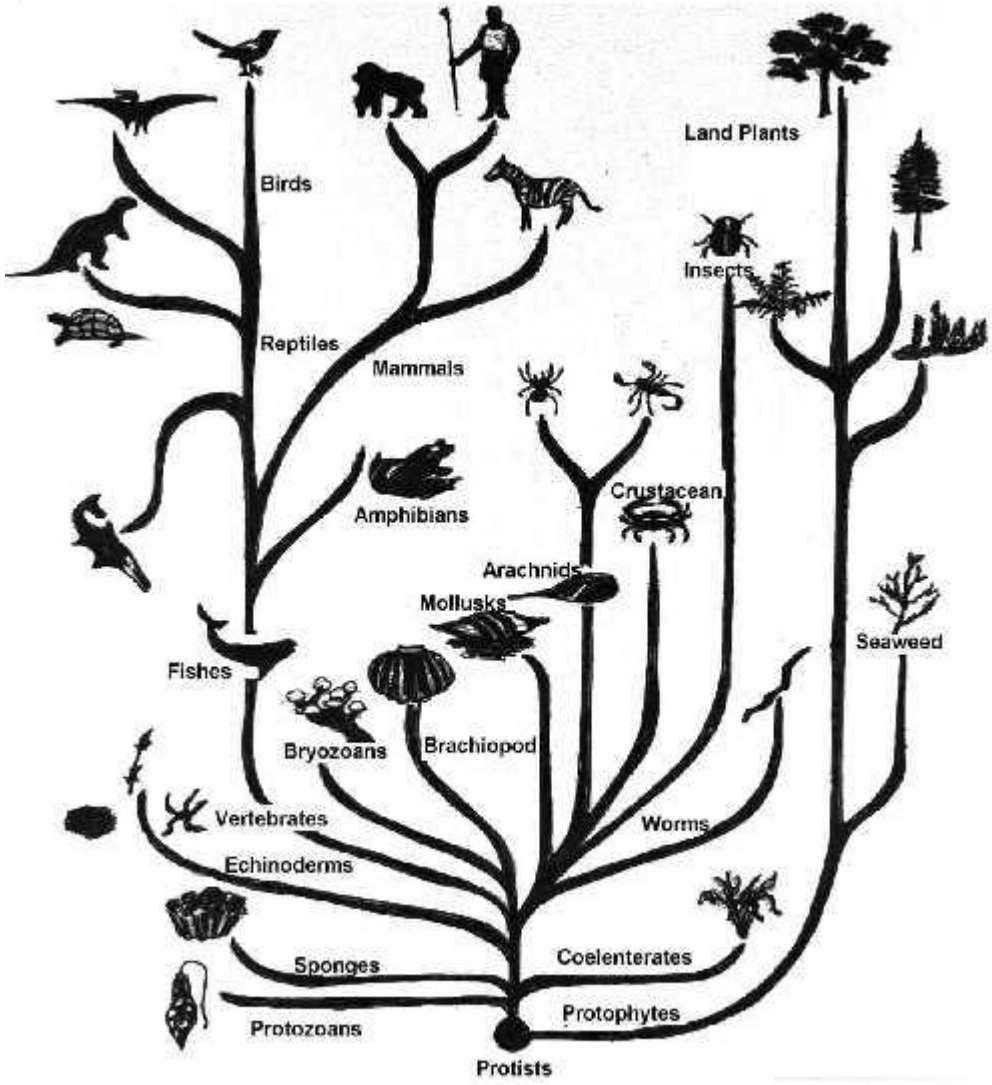
4: Объединить каждую группу соединенных основных точек в отдельный кластер.

5: Назначить каждую пограничную точку одному из кластеров, ассоциированных с ней основных точек.

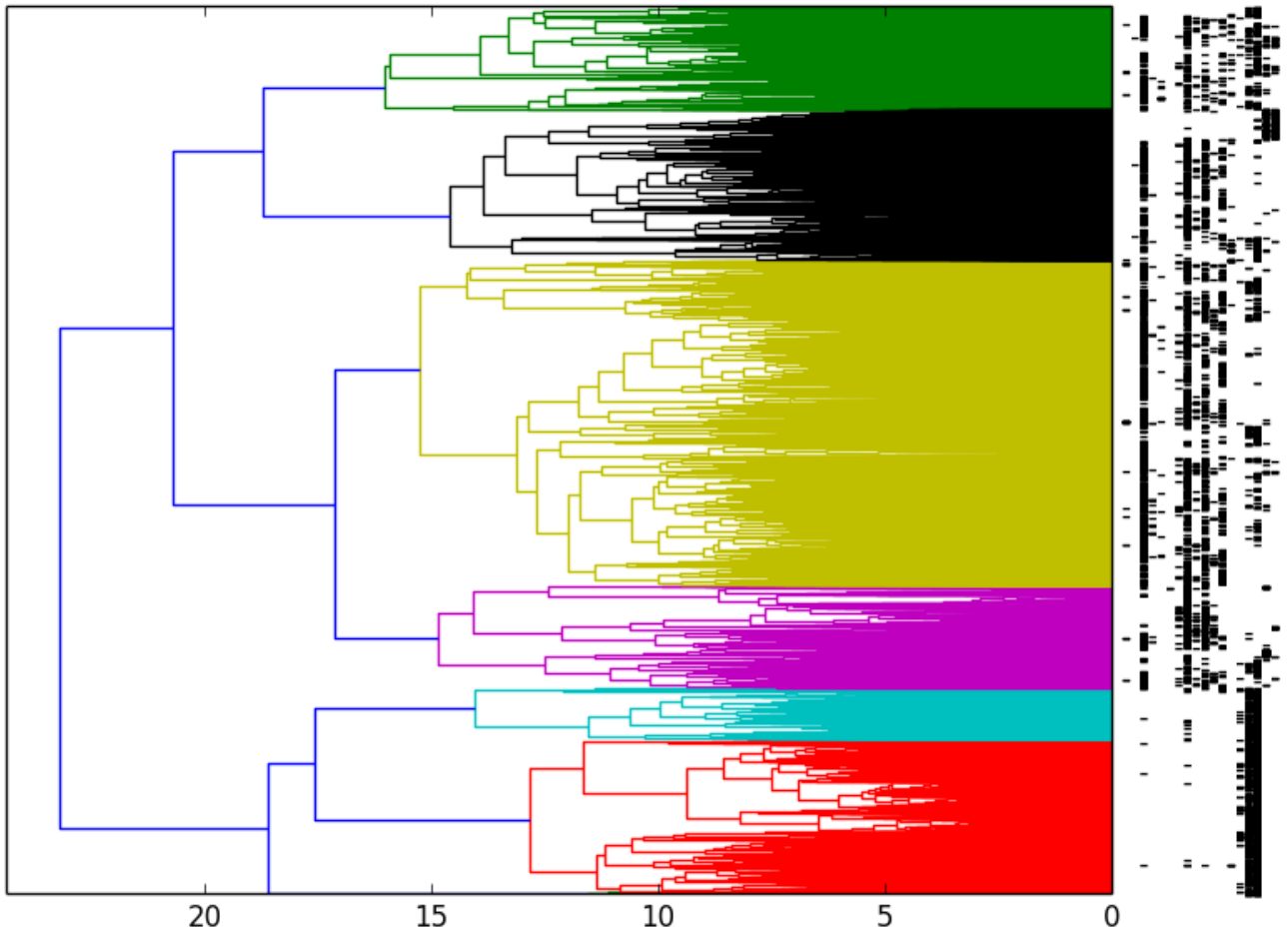
Идея иерархической кластеризации

- Вводим расстояние на объектах
- Пытаемся выстроить «иерархию» вложенных друг в друга кластеров
- Получаем дерево, вершины в котором кластеры
- Дерево можно «обрезать» на какой-то фиксированной глубине и получить нужное число кластеров. Или оставить только достаточно большие кластеры.

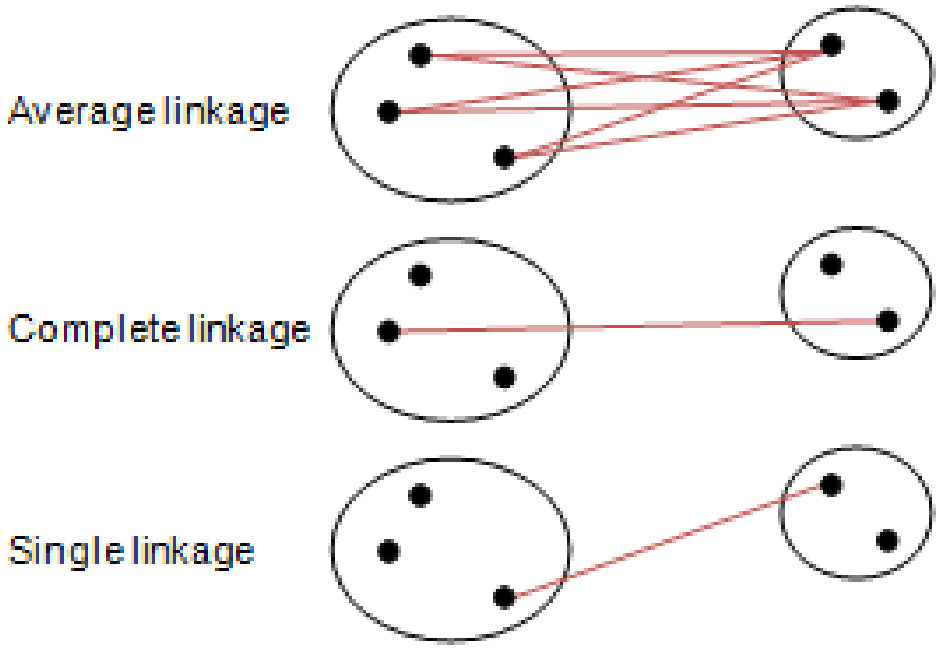
Аналогия из биологии



Дендрограммы



Расстояния между кластерами



Формула Ланса-Вильямса

$$R(U \cup V, S) = \alpha_U R(U, S) + \alpha_V R(V, S) + \beta R(U, V) + \gamma |R(U, S) - R(V, S)|$$

Расстояние ближнего соседа:

$$R^{\text{б}}(W, S) = \min_{w \in W, s \in S} \rho(w, s); \qquad \alpha_U = \alpha_V = \tfrac{1}{2},\ \beta = 0,\ \gamma = -\tfrac{1}{2}.$$

Расстояние дальнего соседа:

$$R^{\text{д}}(W, S) = \max_{w \in W, s \in S} \rho(w, s); \qquad \alpha_U = \alpha_V = \tfrac{1}{2},\ \beta = 0,\ \gamma = \tfrac{1}{2}.$$

Среднее расстояние:

$$R^{\text{с}}(W, S) = \frac{1}{|W||S|} \sum_{w \in W} \sum_{s \in S} \rho(w, s); \qquad \alpha_U = \frac{|U|}{|W|},\ \alpha_V = \frac{|V|}{|W|},\ \beta = \gamma = 0.$$

Расстояние между центрами:

$$R^{\text{ц}}(W, S) = \rho^2\Big(\sum_{w \in W} \frac{w}{|W|}, \sum_{s \in S} \frac{s}{|S|}\Big); \qquad \alpha_U = \frac{|U|}{|W|},\ \alpha_V = \frac{|V|}{|W|},\ \beta = -\alpha_U \alpha_V,\ \gamma = 0.$$

Расстояние Уорда:

$$R^{\text{у}}(W, S) = \frac{|S||W|}{|S|+|W|} \rho^2\Big(\sum_{w \in W} \frac{w}{|W|}, \sum_{s \in S} \frac{s}{|S|}\Big); \quad \alpha_U = \frac{|S|+|U|}{|S|+|W|},\ \alpha_V = \frac{|S|+|V|}{|S|+|W|},\ \beta = \frac{-|S|}{|S|+|W|},\ \gamma = 0.$$

Попробуем систематизировать

- По структуре кластеров:
  - Иерархические
    - Агломеративные
    - Дивизионные
  - Плоские

- По форме
  - Кластеры выпуклой формы
  - Кластеры-ленты
  - Сгустки на «фоне»
  - …

- По присвоению объектов к кластерам:
  - Жесткая кластеризация
  - Мягкая кластеризация

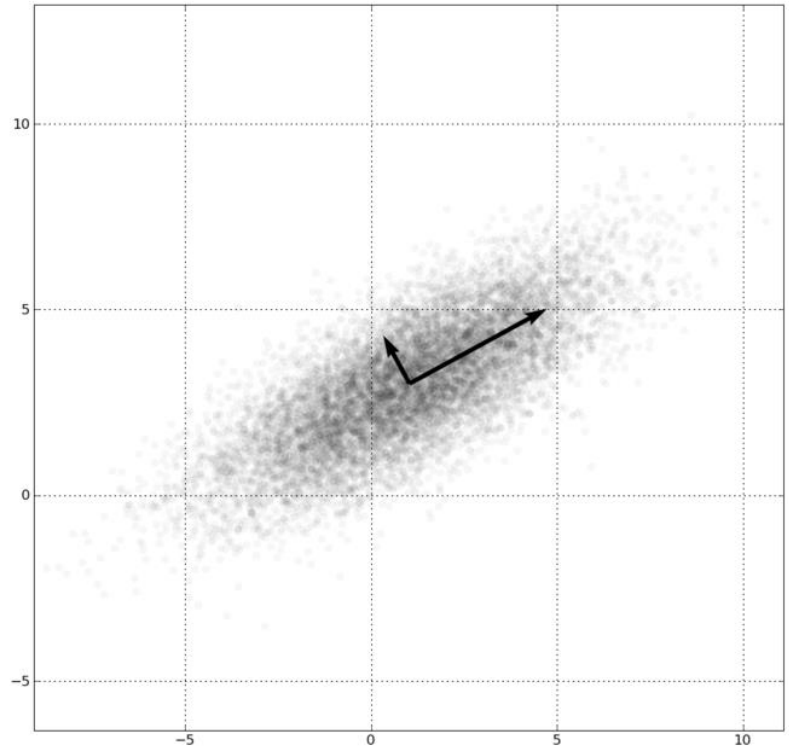
Бонусные слайды

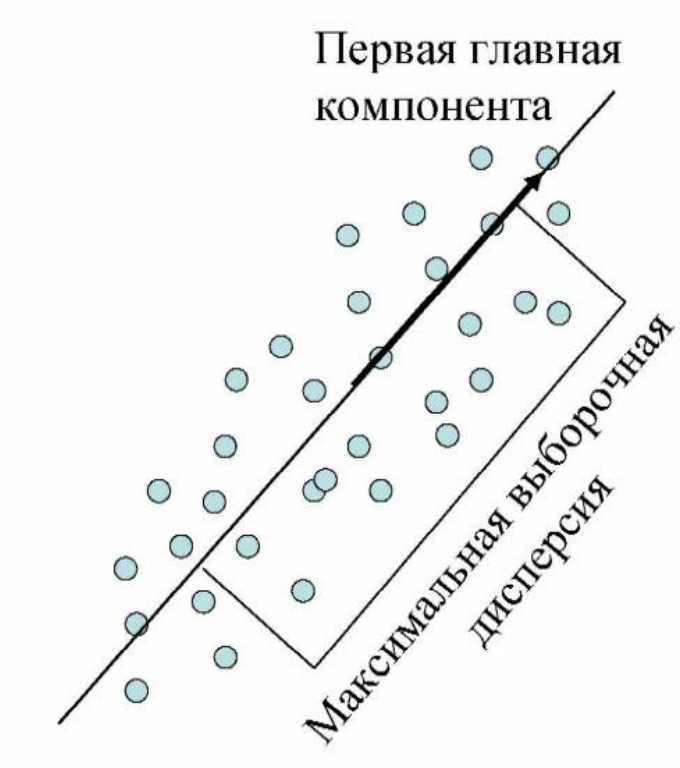
Как проверить наличие кластерной структуры

1. Генерируем p случайных точек из равномерного распределения и p случайных из обучающей выборки
2. Вычисляем величину (статистика Хопкинса):

$$H = \frac{\sum_{i=1}^{p} w_i}{\sum_{i=1}^{p} u_i + \sum_{i=1}^{p} w_i}$$

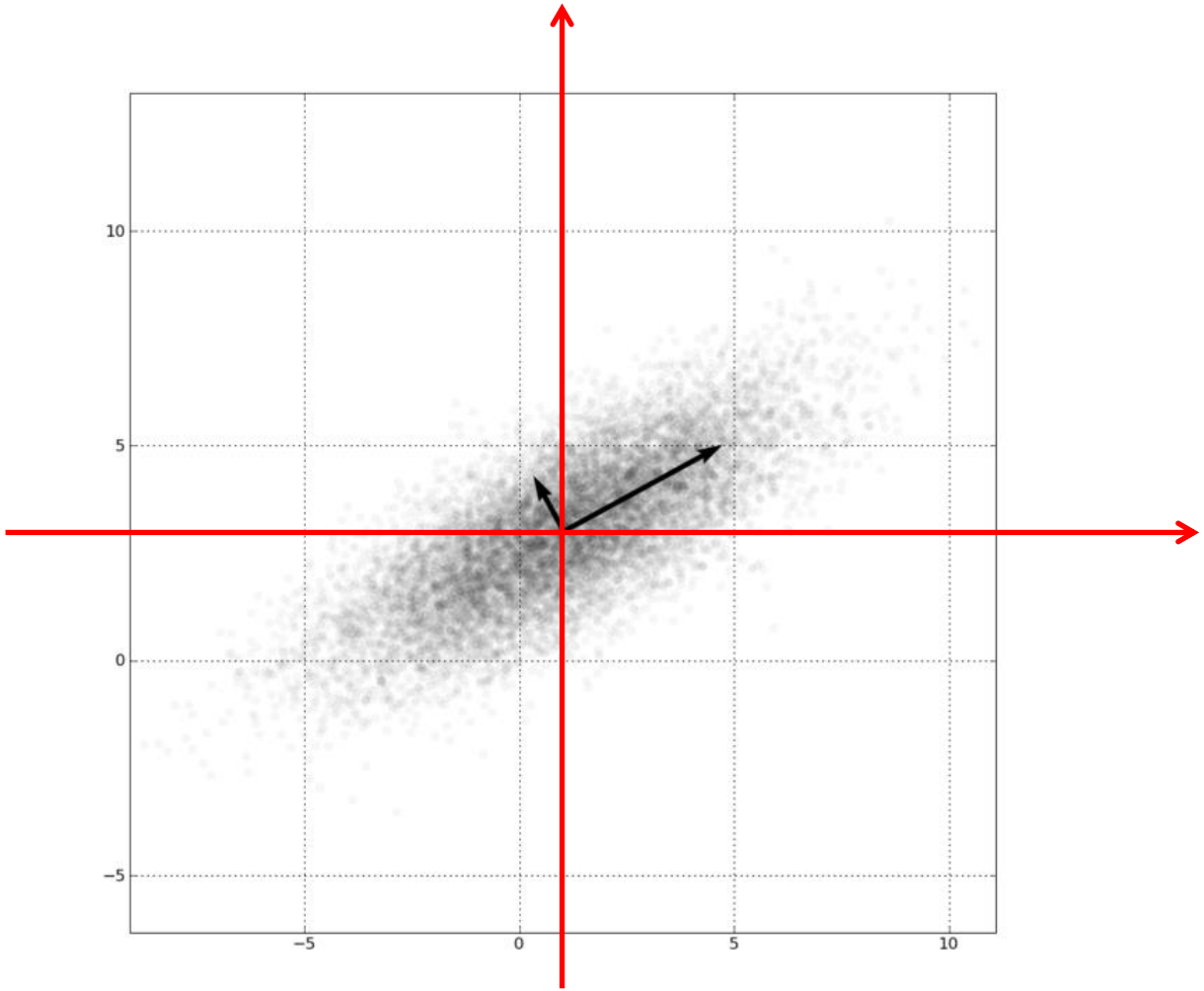
Анонс одной из следующих лекций: PCA



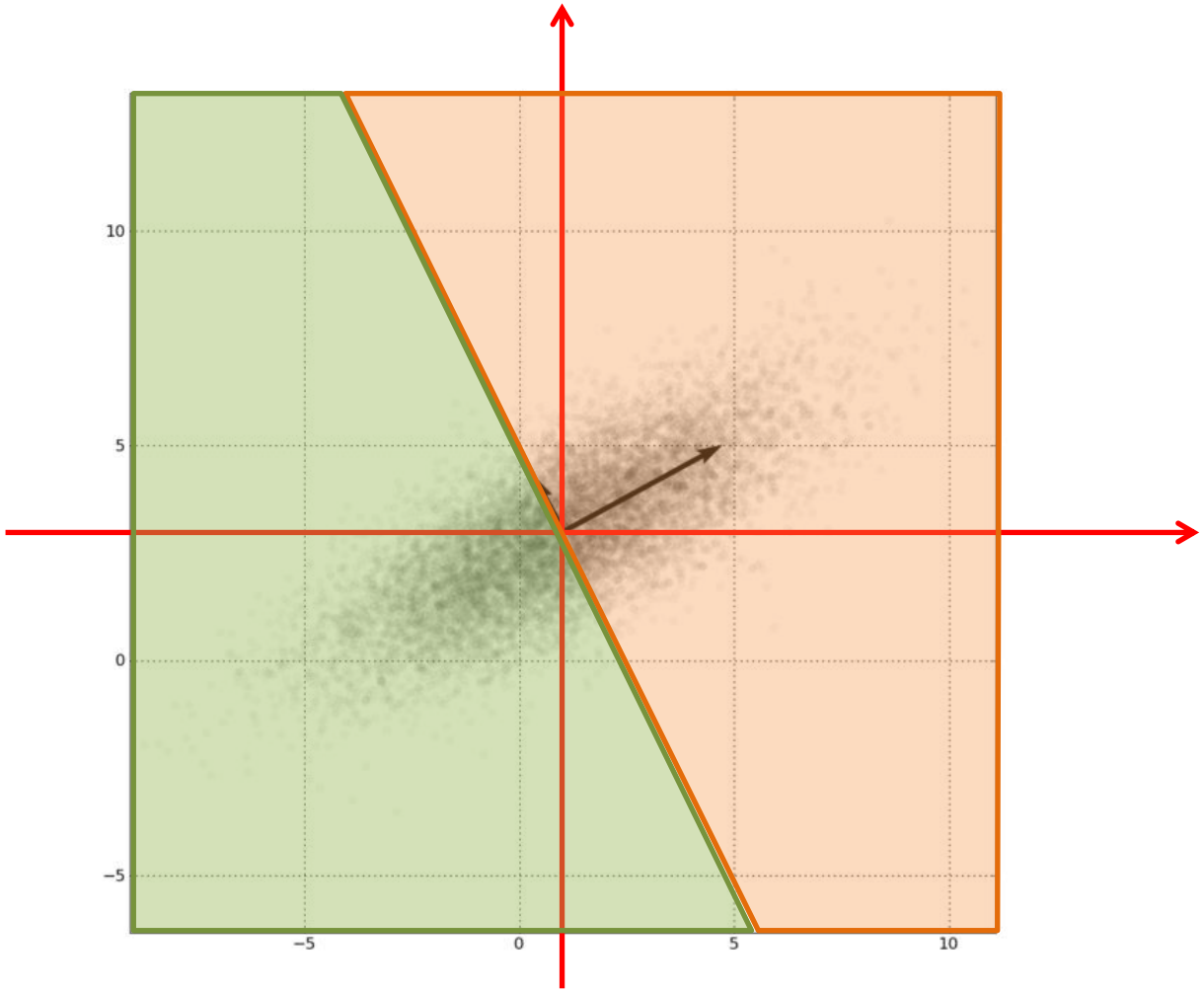


PDDP - Principal Direction Divisive Partition

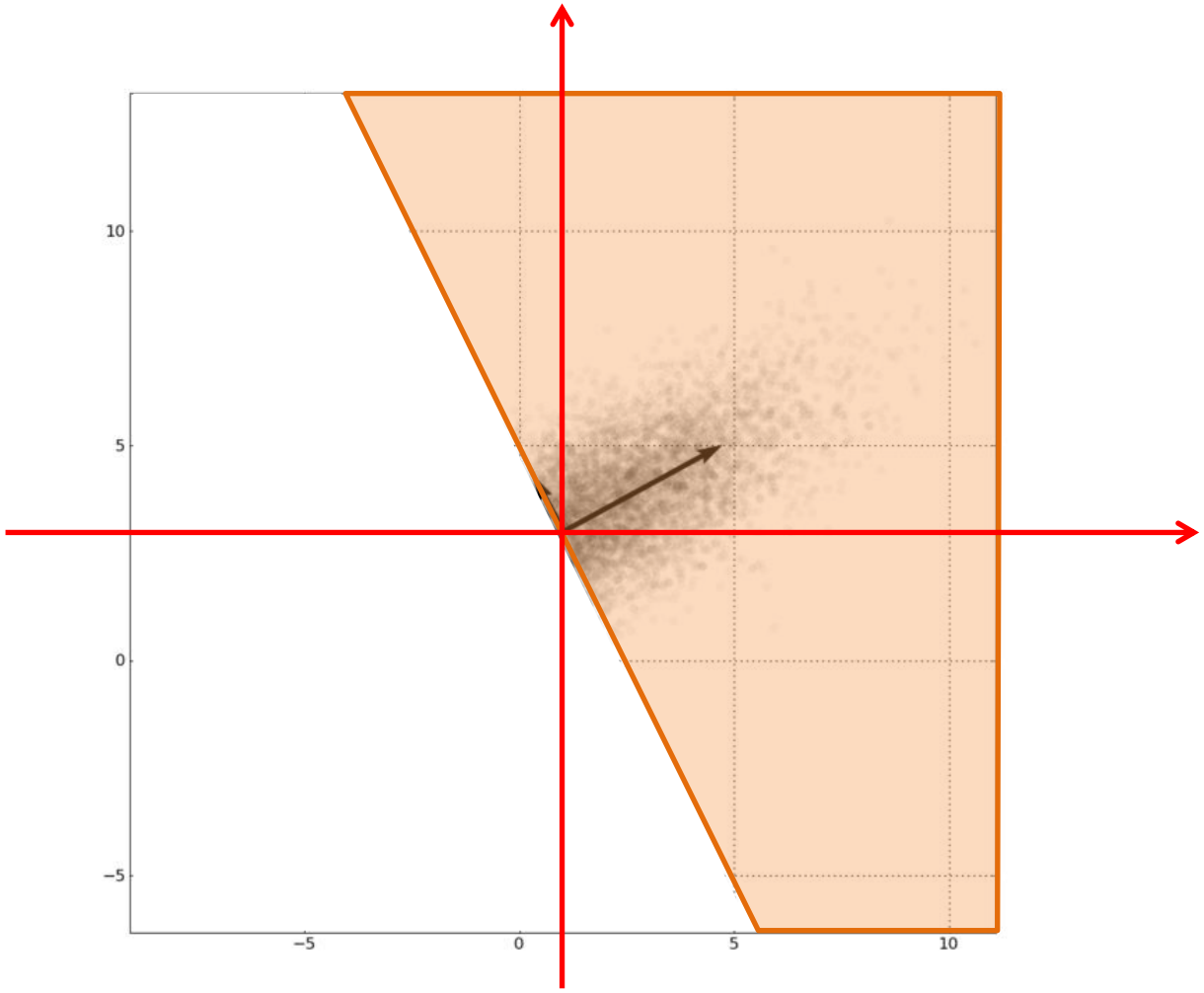
Шаг 1: центрирование данных



Шаг 2: делим по знаку проекции на главную компоненту



Шаг 3: повторяем для кластера с наибольшим разбросом



Шаг N

Повторяем разделения, пока качество кластеризации увеличивается

DBSCAN: выбор параметров