Обучение без учителя. Кластеризация.

Обучение без / с учителем

Обучение с учителем

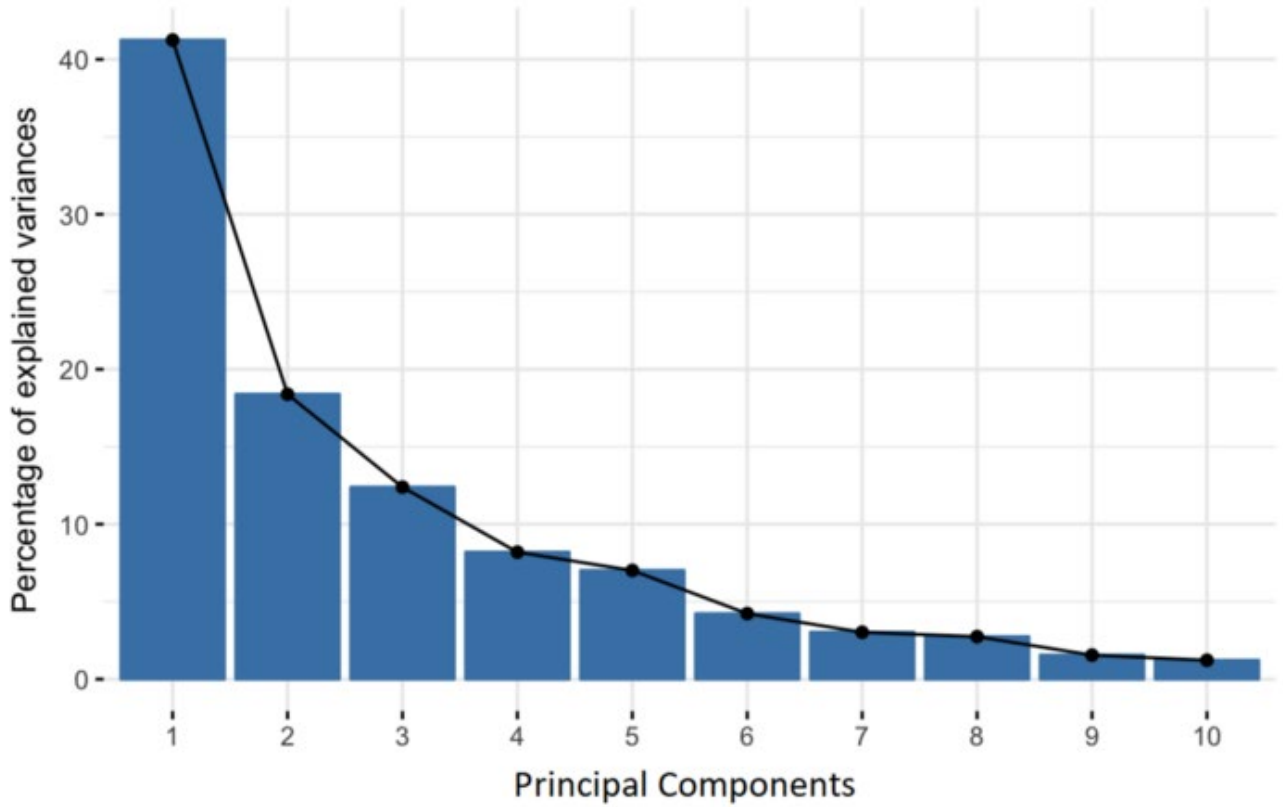
- Есть размеченная выборка
- Классификация
- Регрессия

Обучение без учителя

- Нет размеченной выборки
- Кластеризация
- Снижение размерности
- Векторное кодирование (обучение с частичным привлечением учителя)

Снижение размерности методом выделения главных компонент

Главные компоненты – новые переменные, линейные комбинации старых, построенные таким образом, чтобы они были не коррелирующими и большая часть информации (дисперсии) сжата в первую компоненту.



PCA

- Центрируем данные $z = x - E(x)$
- Вычисляем матрицу ковариации
  $cov(X, Y) = \mathrm{E}[(X - \mathrm{E}X)(Y - \mathrm{E}Y)] = \mathrm{E}(X, Y) - \mathrm{E}(X) \cdot \mathrm{E}(Y)$
- $cov(X, X) = var(X), cov(X, Y) = cov(Y, X)$
- Ищем собственные вектора и значения $Ax = \lambda x$
- Отбираем по собственным значениям топ
- $Result = EigenVectors.T \cdot DataSet.T$
- Объясненная дисперсия $\max(e_{val}) / sum(e_{val})$

Кластеризация

- Задача - упорядочить объекты в сравнительно однородные группы.
- Формально: если задано пространство объектов X, с обучающей выборкой $X^l = \{x_i\}_{i=1}^l$ и расстоянием $\rho: X \times X \to [0, \infty)$, то надо найти Y – множество кластеров и $a: X \to \mathrm{Y}$ – алгоритм кластеризации, такие что:
  - каждый кластер состоит из близких объектов
  - объекты разных кластеров существенно различны.

Постановка задачи некорректна

Решение задачи кластеризации принципиально неоднозначно:

- точной постановки задачи кластеризации нет
- существует много критериев качества кластеризации
- существует много эвристических методов кластеризации
- число кластеров $|Y|$, как правило, неизвестно заранее
- результат кластеризации существенно зависит
- от метрики ρ, которую эксперт задаёт субъективно

Цели кластеризации

- Упростить дальнейшую обработку данных,
- Сократить объём хранимых данных
- Выделить нетипичные объекты
- Построить иерархию множества объектов

Первоначальные проблемы

Кластеры могут быть

- различной формы (размытые кластеры, ленточные кластеры, кластеры с центром)
- подвержены шуму (разреженный фон, перекрытие кластеров, соединения перемычками)

Бонусом – кластеров может не быть

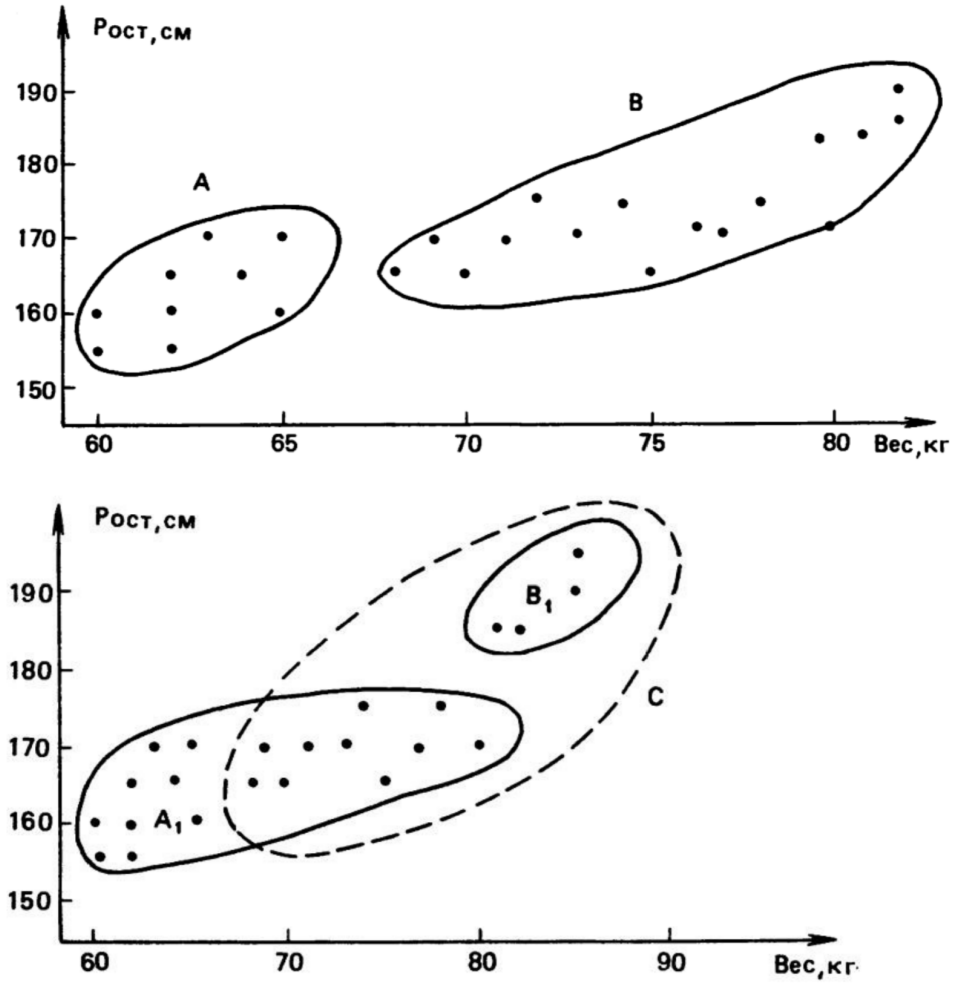
Каждый метод кластеризации имеет свои ограничения и выделяет кластеры лишь некоторых типов.

Чувствительность к нормировке

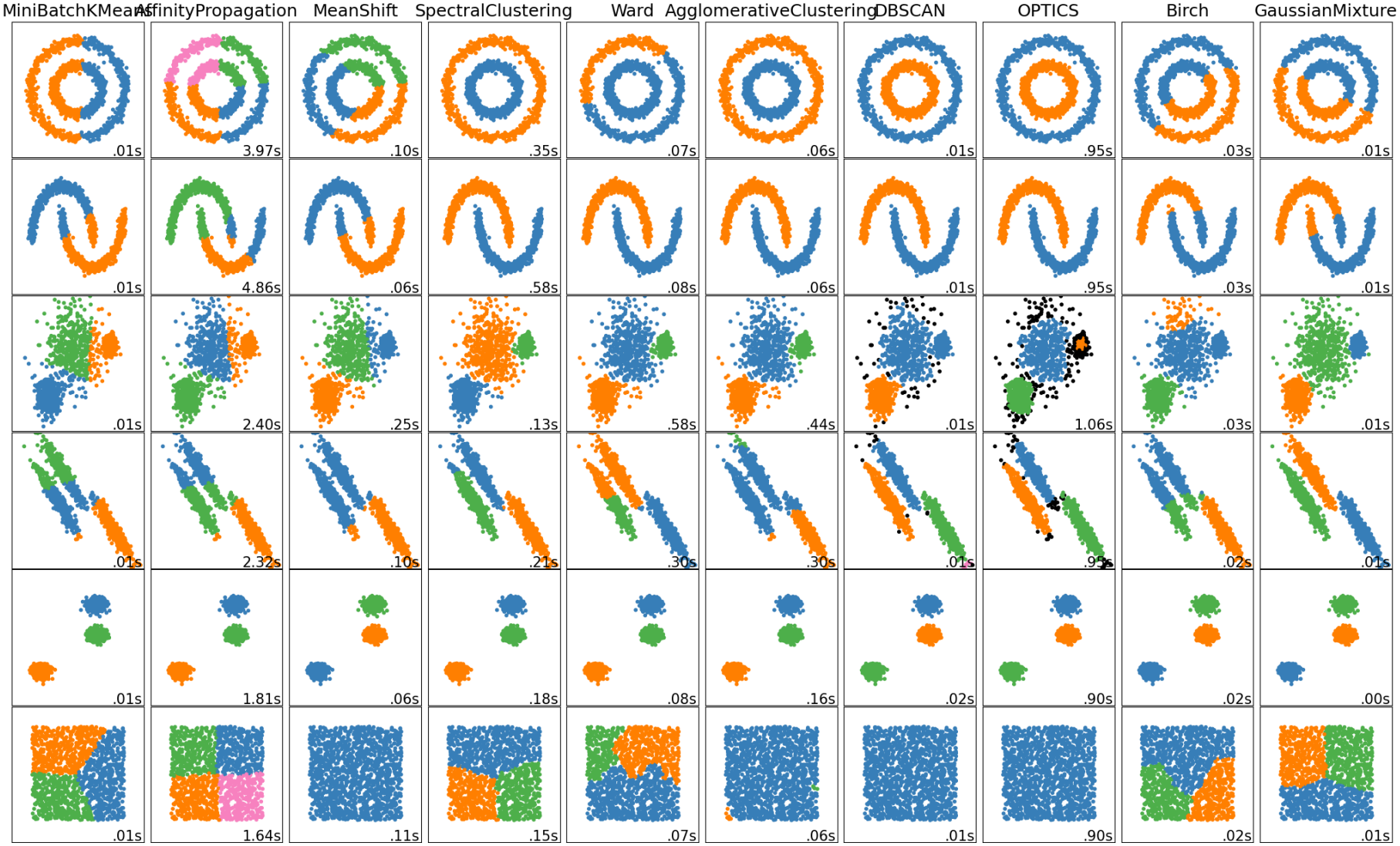
A – студентки

B – студенты

После нормировки сжимаем ось веса вдвое



Алгоритмы кластеризации



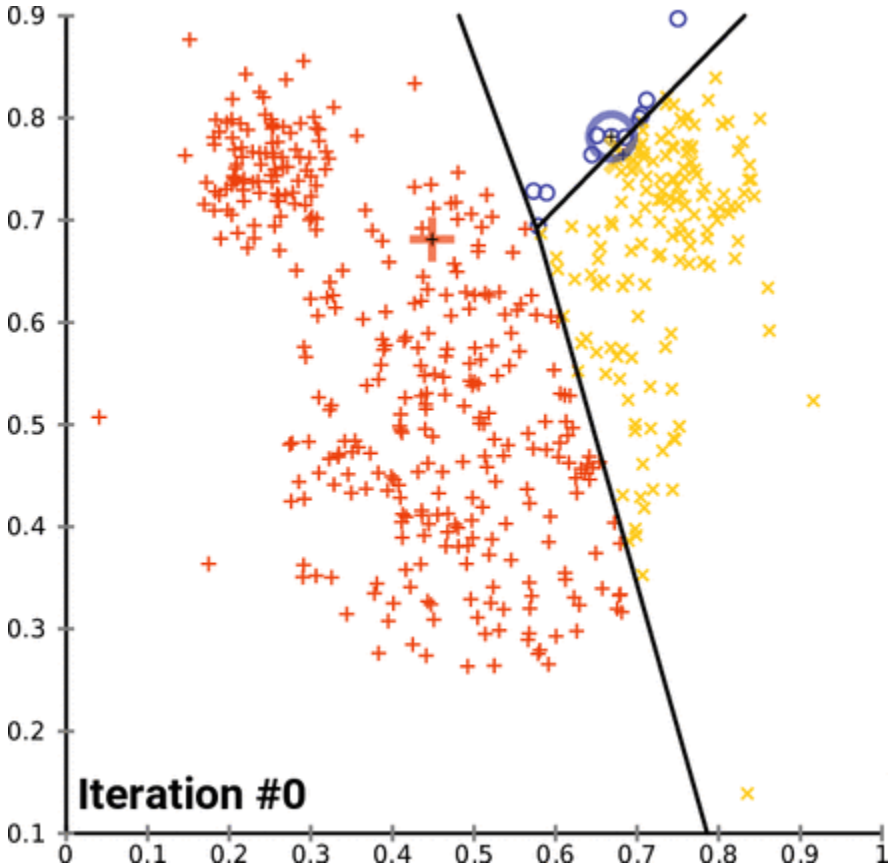
K-means (метод k-средних)

- Алгоритм разделяет выборку на K непересекающихся кластеров, каждый из которых описывается средним μ_i (центройды) всех элементов в кластере.
- Центройды выбираются так, чтобы минимизировать суммарное квадратичное отклонение точек кластеров от центройдов $\sum_{i=0}^{n} \min_{\mu_j \in C} \left(\|x_i - \mu_j\|^2 \right)$

K-means (метод k-средних)

1. Выбрать гиперпараметр k (число кластеров)
2. Инициализировать k центройдов
3. Отнести каждый элемент выборки к ближайшему центройду
4. Обновить центройды
5. Повторять шаги 3-4 n итераций или пока центройды не перестанут меняться

K-means (метод k-средних)



https://en.wikipedia.org/wiki/K-means_clustering#/media/File:K-means_convergence.gif

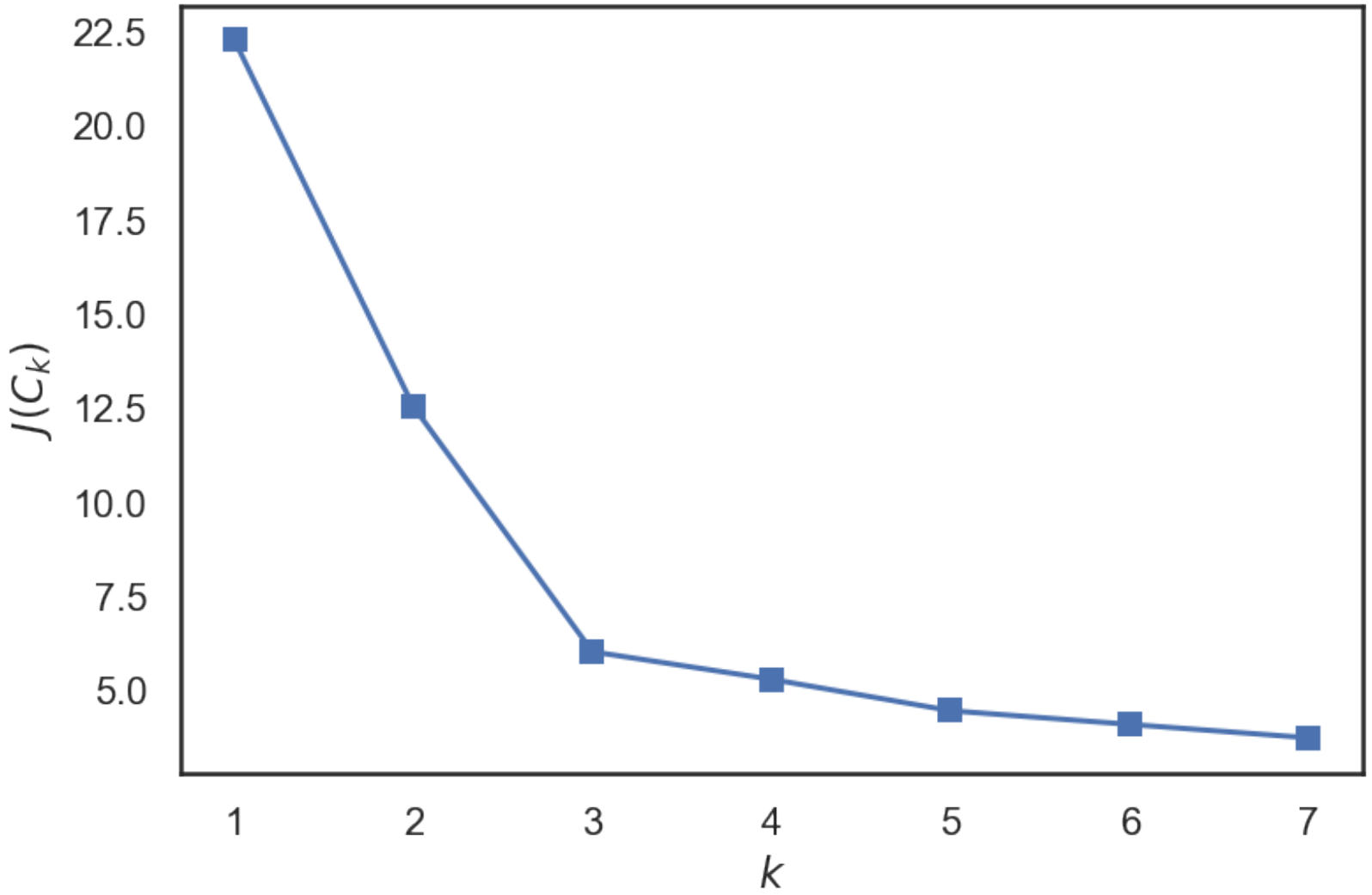
K-means (выбор k)

- Как выбрать число k?
- Хотим, чтобы суммарное квадратичное отклонение было минимальным

K-means (выбор k)

- Как выбрать число k?
- Хотим, чтобы суммарное квадратичное отклонение было минимальным
- $J(C) = \sum_{i=0}^{n} \min_{\mu_j \in C} \left(\left\| x_i - \mu_j \right\|^2 \right)$
- Будем выбирать то k, после которого $J(C)$ уменьшается не сильно (правило локтя)
- $D(k) = \frac{|J(C_k) - J(C_{k+1})|}{|J(C_{k-1}) - J(C_k)|} \to \min_k$

K-means (выбор k)

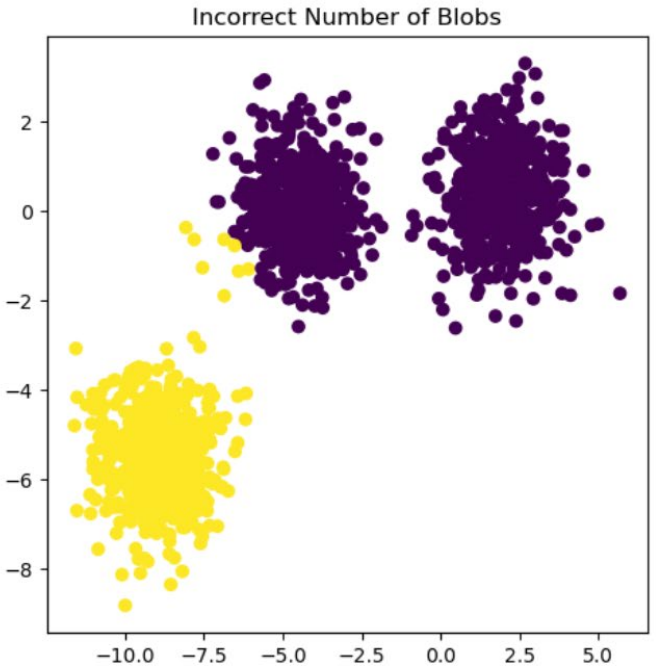


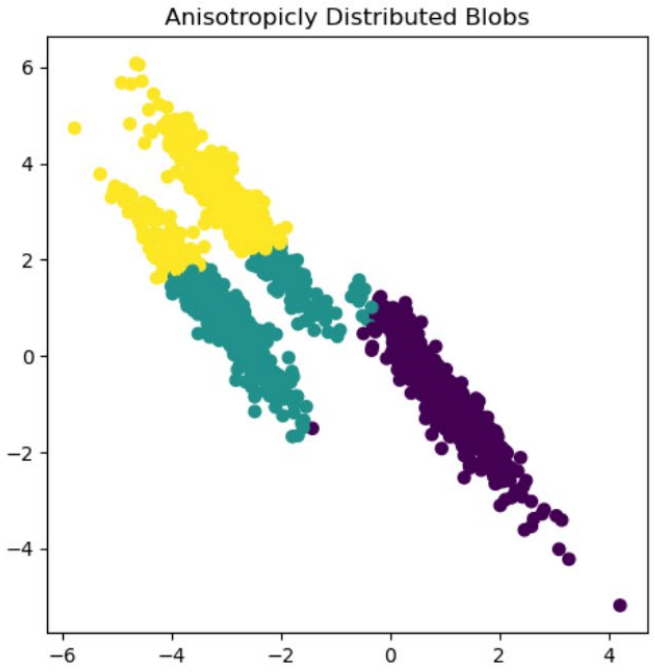
Проблемы K-means

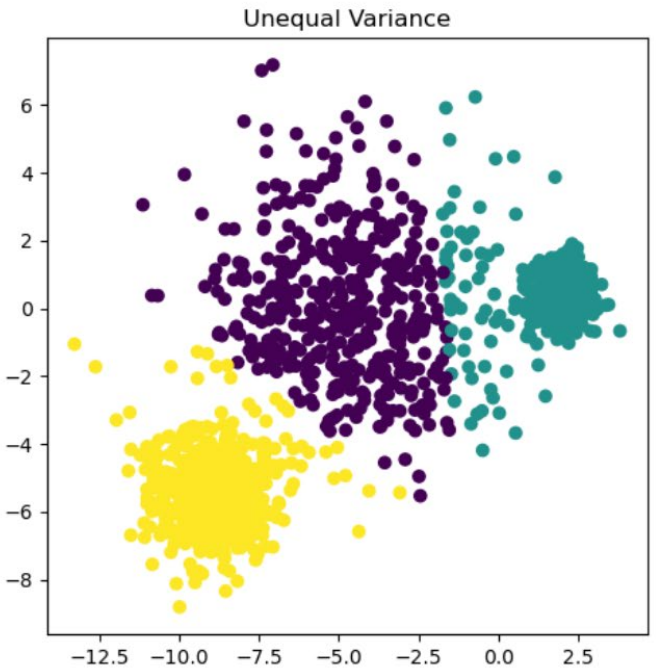
- Метрика суммарного квадратичного отклонения предполагает, что кластеры выпуклые и равномерные
- Не гарантируется глобальный минимум
- Алгоритм нестабилен из-за начальной инициализации
- Долго сходится

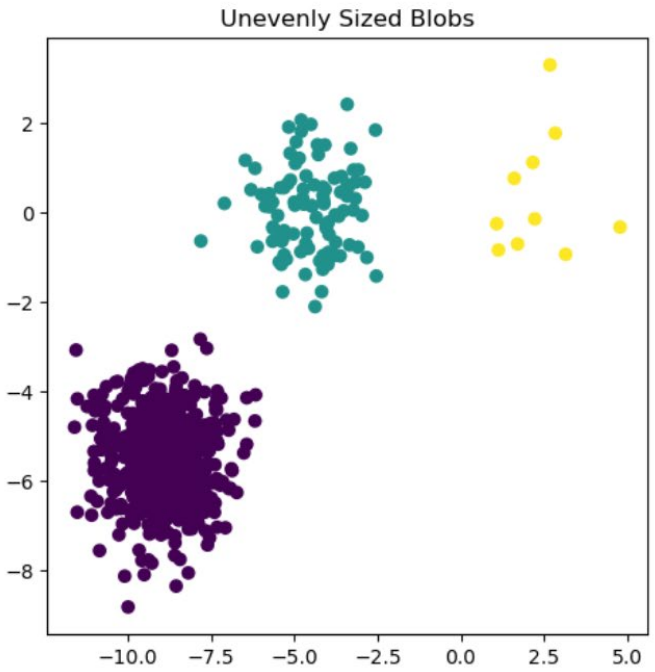
Проблемы K-means

Сложные случаи

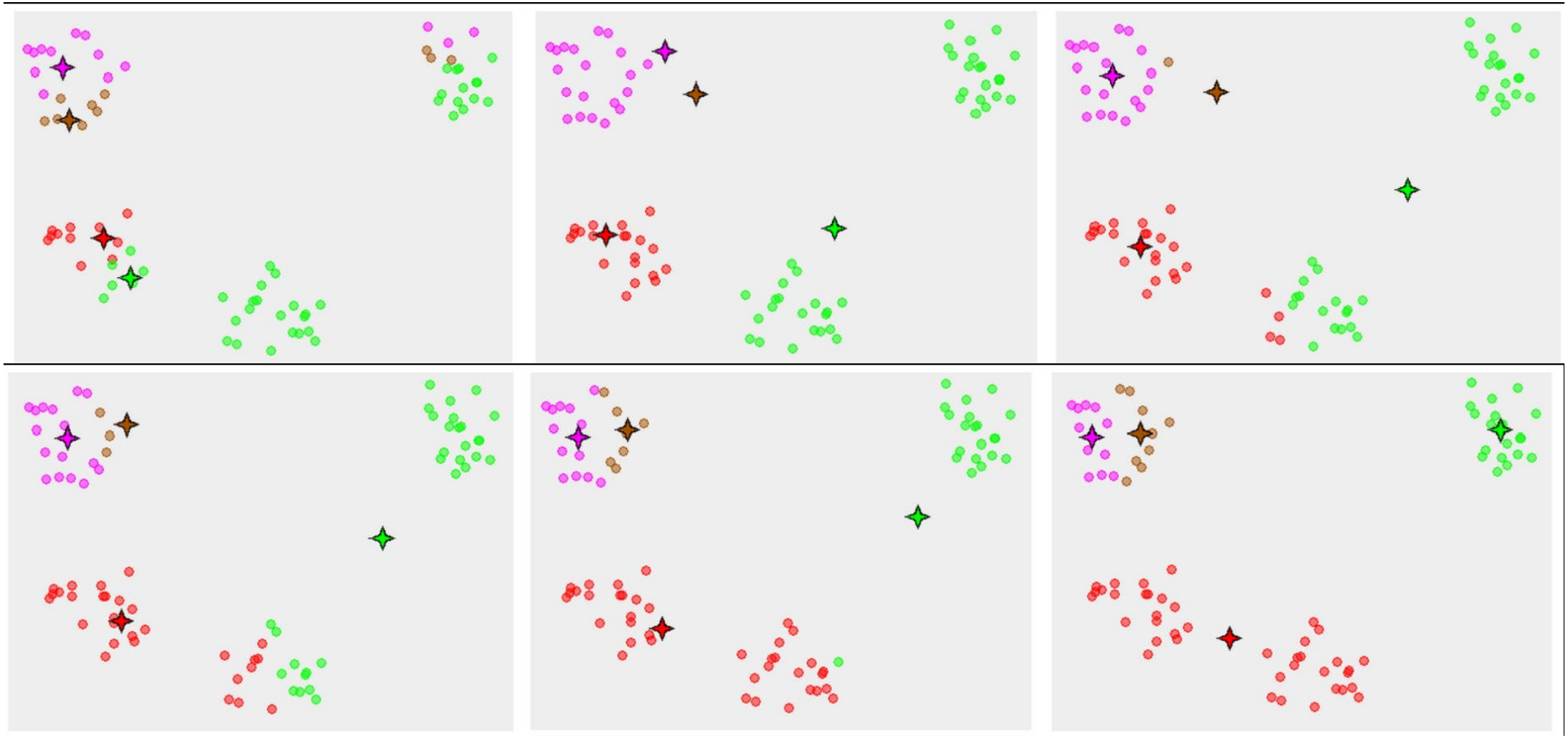








Проблемы K-means



K-means++ (выбор кластеров)

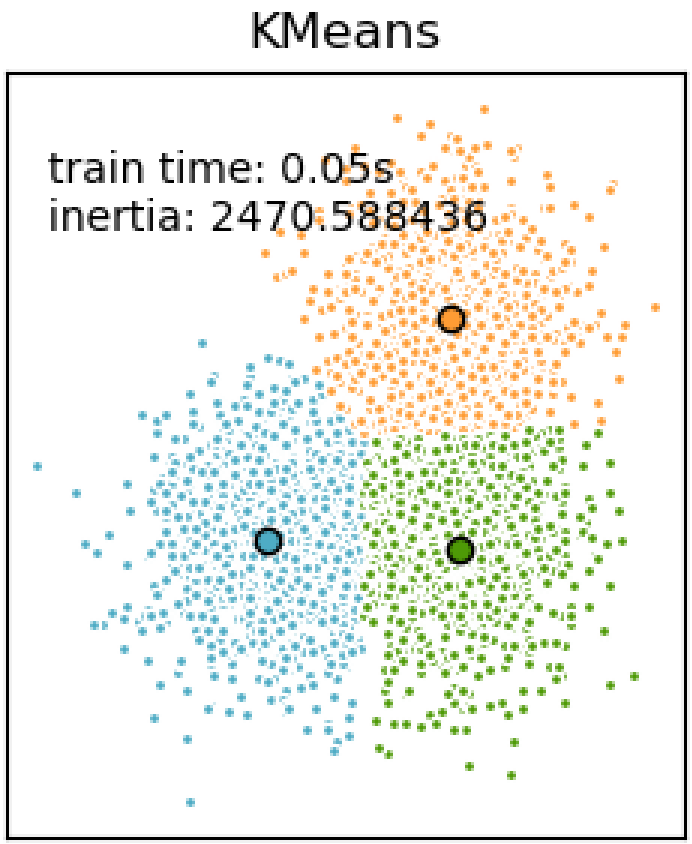
- Результат зависит от выбора исходных центров кластеров, их оптимальный выбор неизвестен
- Алгоритм нестабилен
- Попробуем найти хорошие начальные кластеры
  - Выбираем случаные точки
  - Выбираем случайные элементы
  - Выбираем случайные элементы несколько раз, делая пару итераций смотрим где лучше сходимость
  - kmeans++

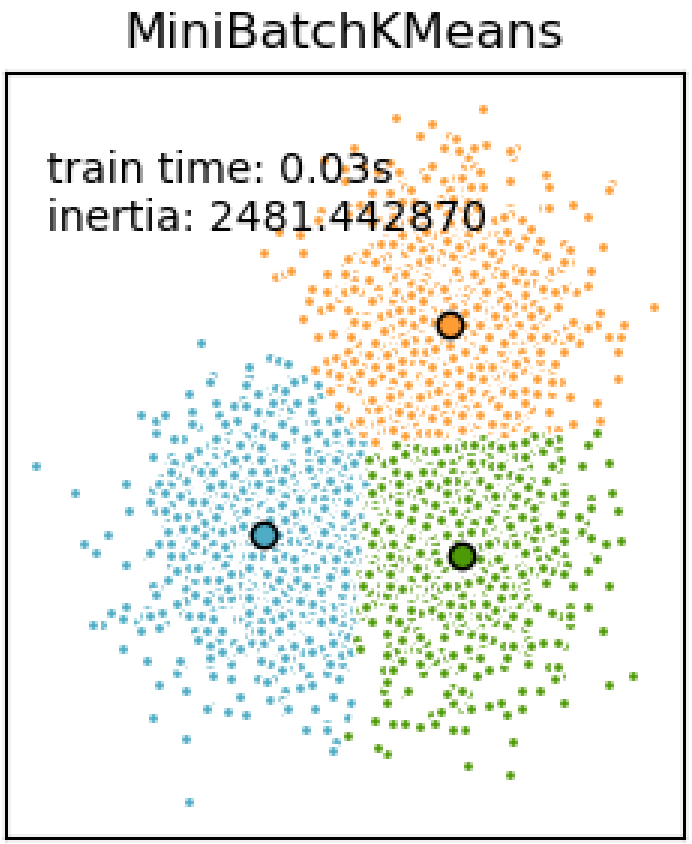
K-means++ (выбор кластеров)

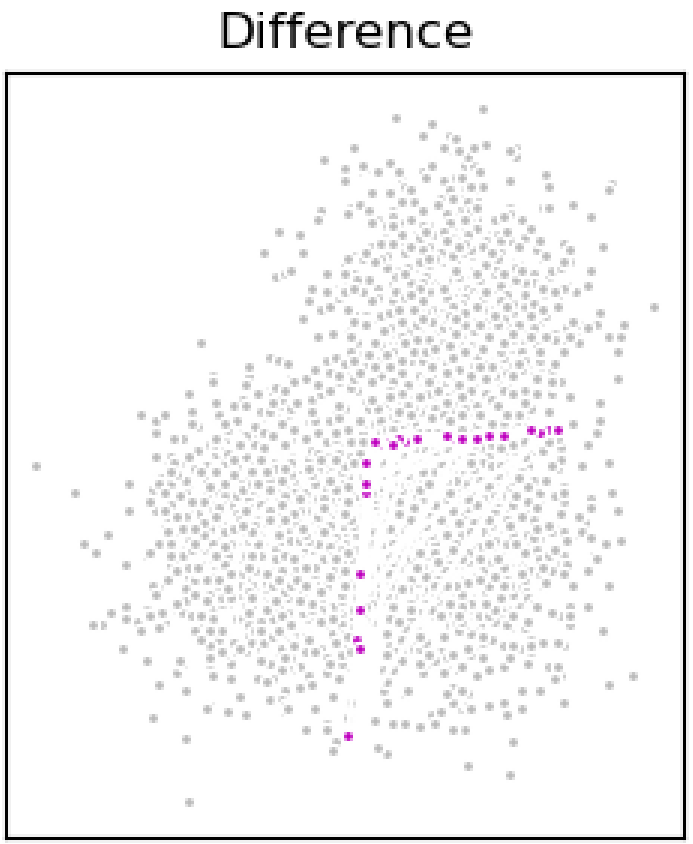
- Попробуем найти хорошие начальные кластеры

1. Выбрать первый центроид случайным образом (среди всех точек)
2. Для каждой точки найти значение квадрата расстояния до ближайшего центроида dx^2
3. Выбираем следующую точку в зависимости от расстояния
4. Повторяем шаги 2-3 пока не выберем все центройды

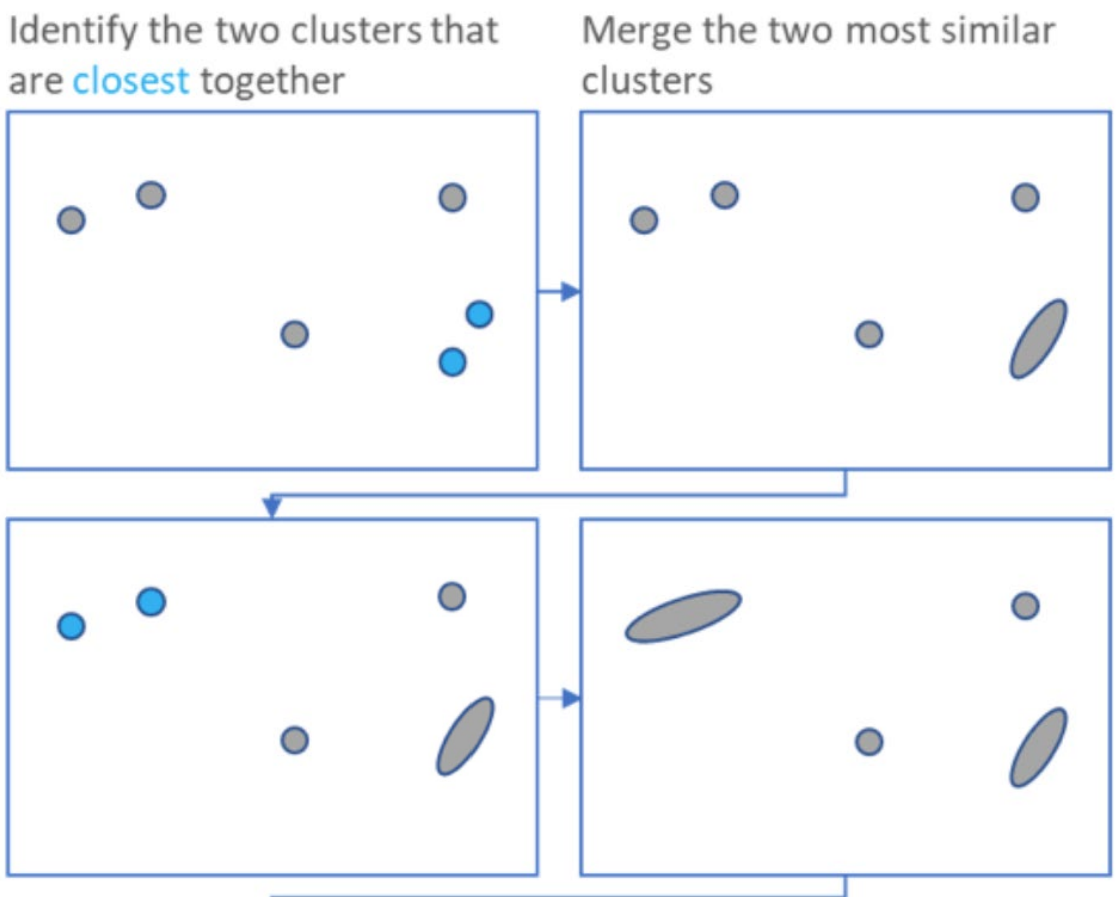
MiniBatch K-means



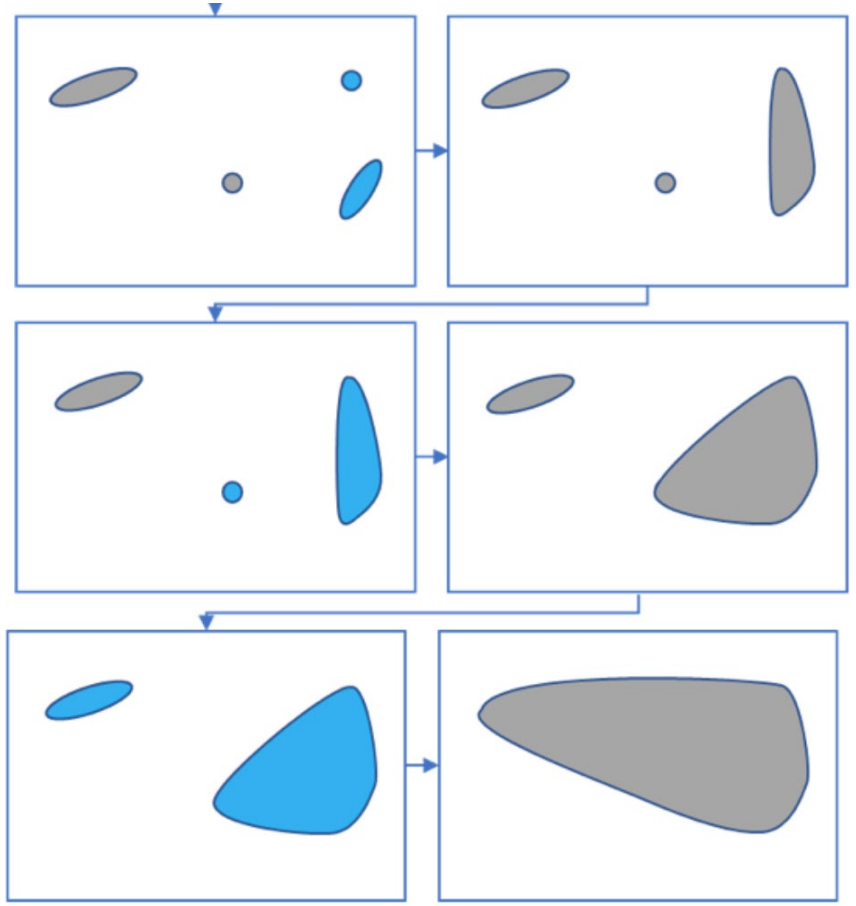




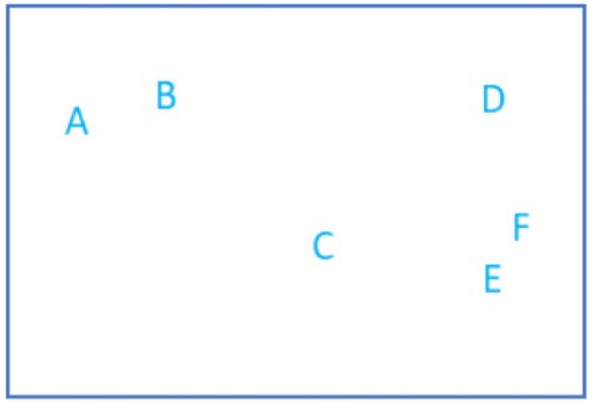
Иерархическая кластеризация

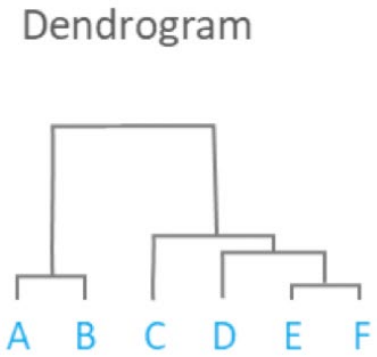


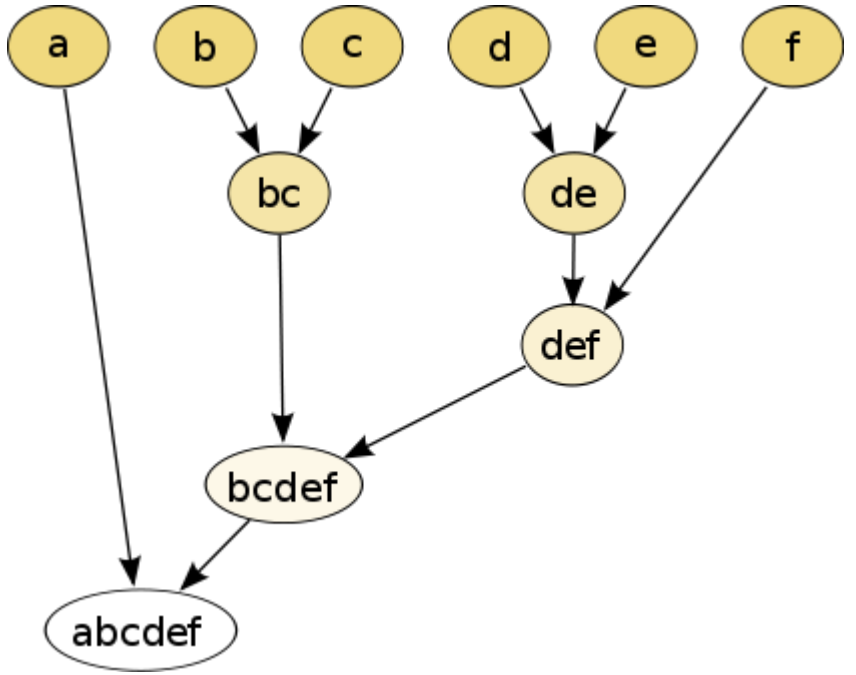
Иерархическая кластеризация

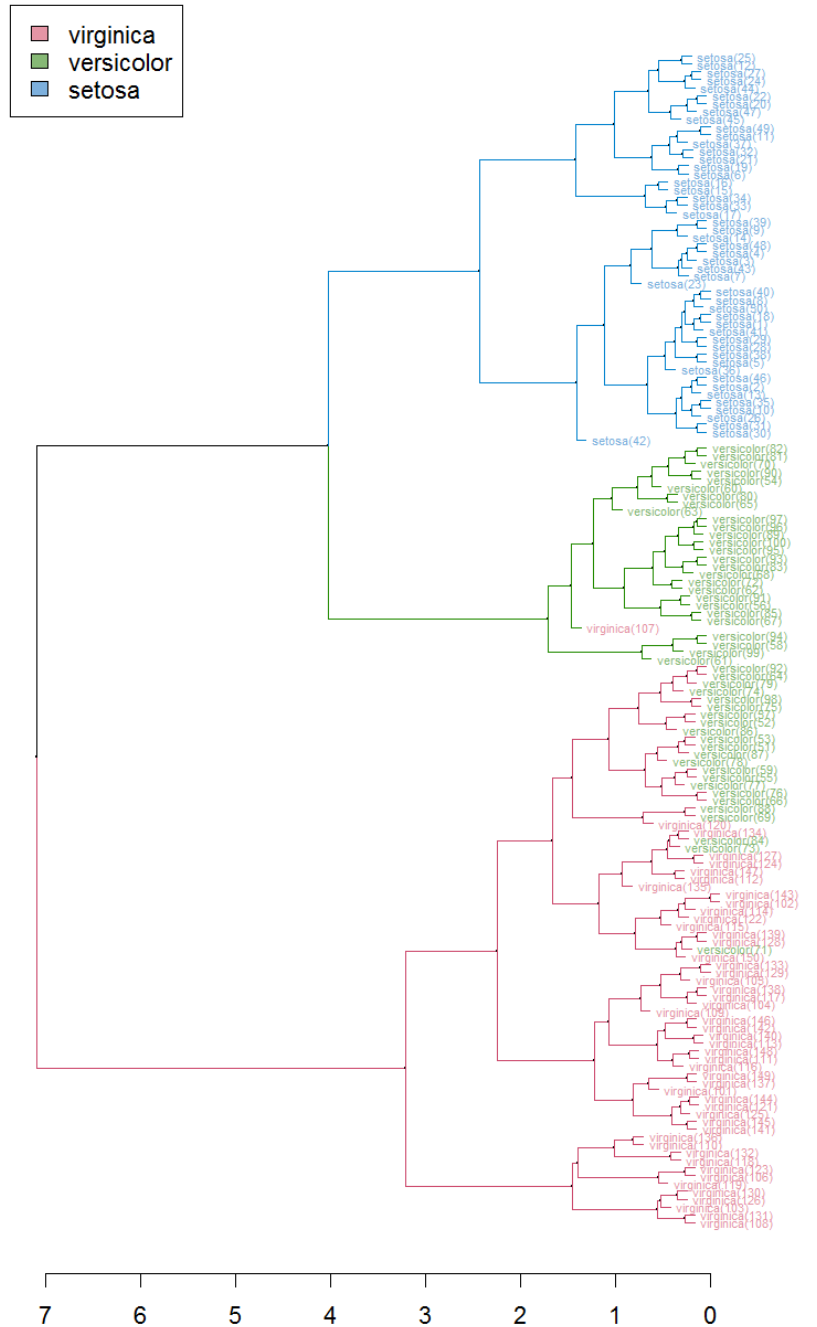


Иерархическая кластеризация







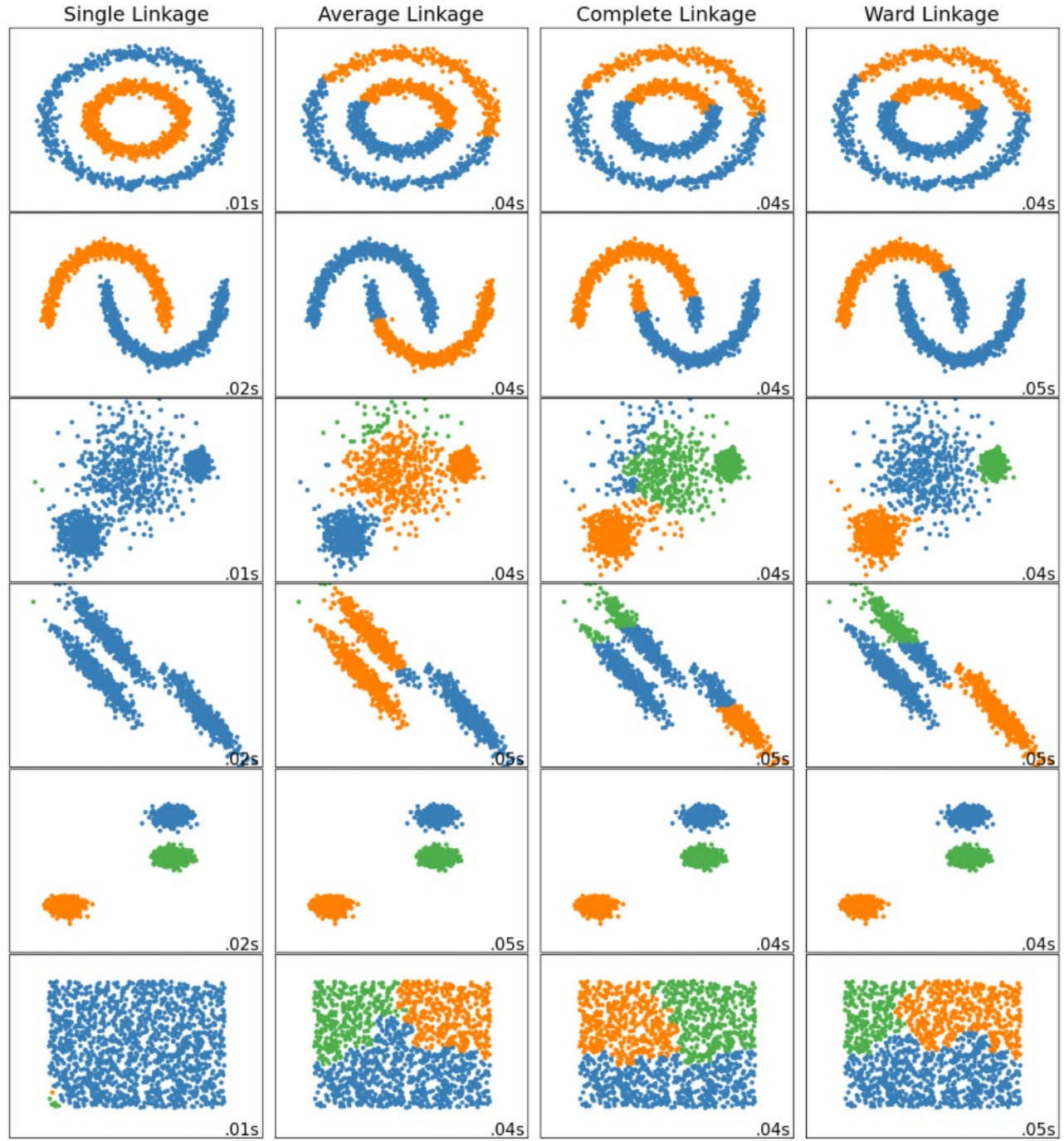

Clustered Iris data set
(the labels give the true flower species)
virginica
versicolor
setosa
7
6
5
4
3
2
1
0

Как понять какие кластеры близки?

- complete-linkage
- single-linkage
- average-linkage
- ward-linkage

- $\max\{d(a,b): a \in A, b \in B\}$
- $min\{d(a,b): a \in A, b \in B\}$
- $\frac{1}{|A|\cdot|B|}\sum_{a\in A}\sum_{b\in B} d(a,b)$

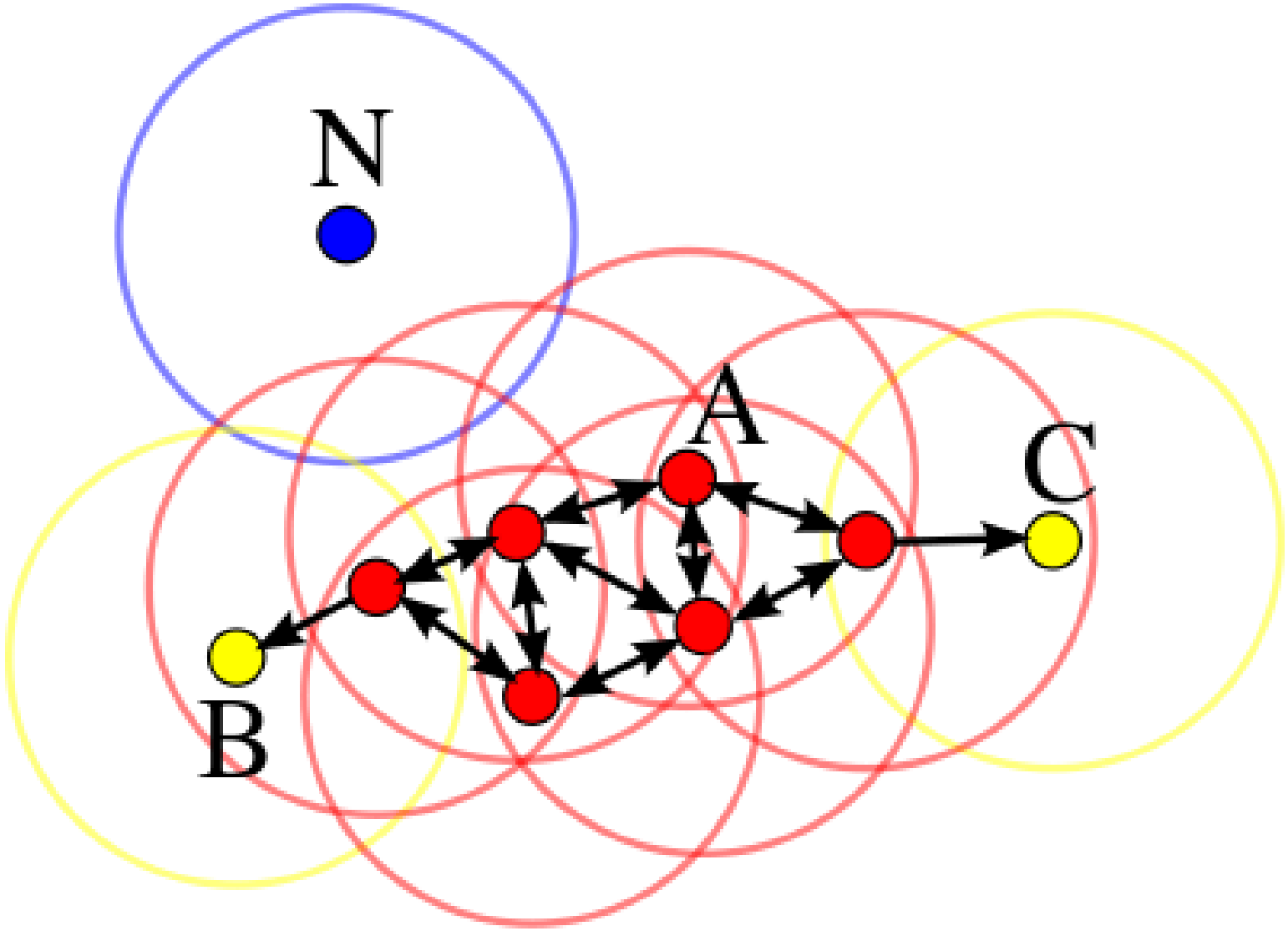
$$\Delta(A,B) = \sum_{i\in A\bigcup B} ||\overrightarrow{x_i} - \overrightarrow{m}_{A\bigcup B}||^2 - \sum_{i\in A} ||\overrightarrow{x_i} - \overrightarrow{m}_A||^2 - \sum_{i\in B} ||\overrightarrow{x_i} - \overrightarrow{m}_B||^2 = \frac{n_A n_B}{n_A + n_B} ||\overrightarrow{m}_A - \overrightarrow{m}_B||^2$$


Single Linkage
Average Linkage
Complete Linkage
Ward Linkage
.01s
.04s
.04s
.04s
.02s
.04s
.04s
.05s
.01s
.04s
.04s
.04s
.02s
.05s
.05s
.05s
.02s
.05s
.04s
.04s
.01s
.04s
.04s
.05s

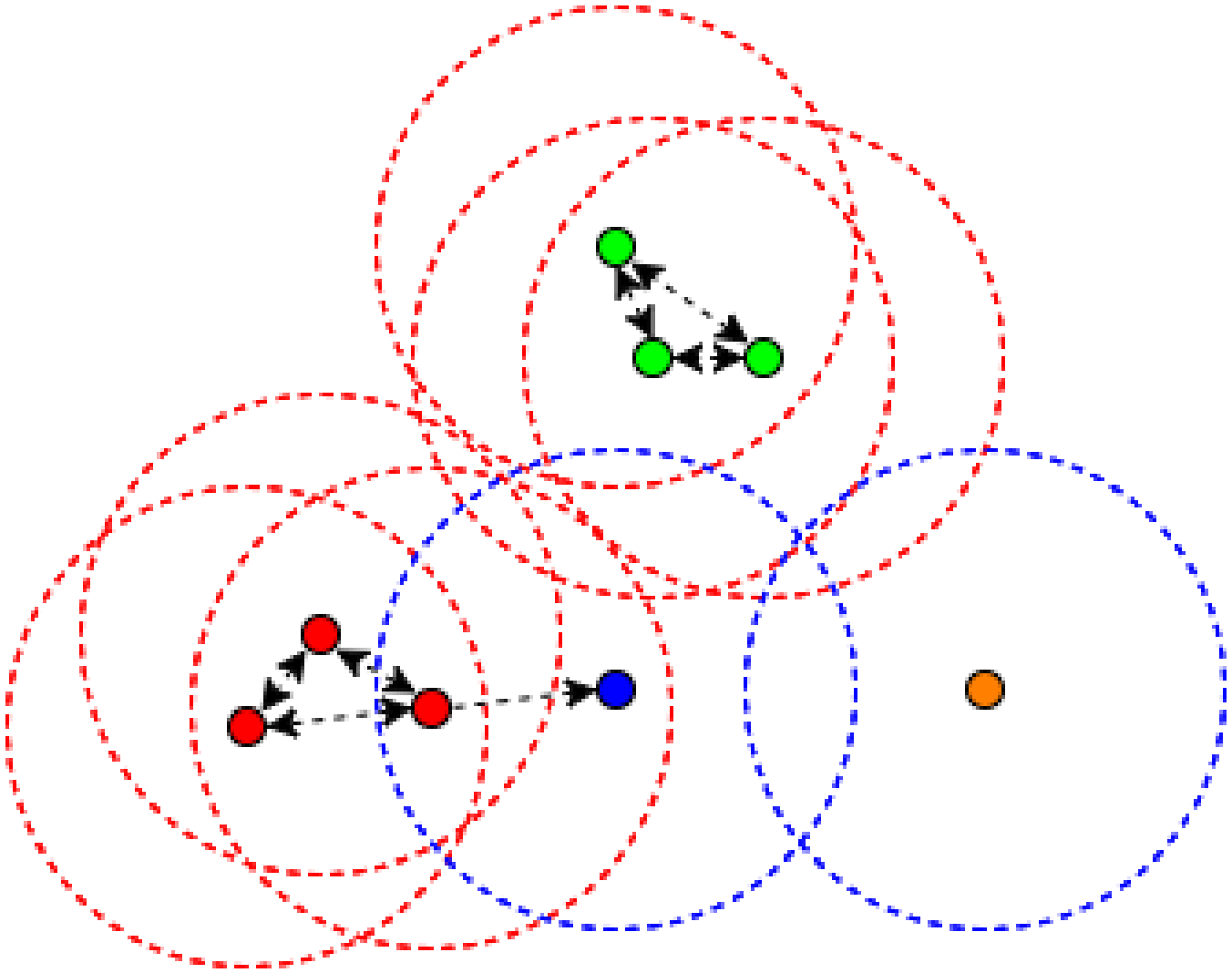
DBSCAN

- Будем объединять точки в областях в высокой плотностью
- Разделим их на основные, достижимые по плотности и выпадающие точки
- Точка p является основной точкой, если как минимум $\min\_samples$ точек находятся на расстоянии, не превосходящем eps
- Точка q достижима из p, если имеется путь $p_1, \ldots, p_n, q$, где каждая точка p_{i+1} достижима прямо из p_i
- Все точки, не достижимые из основных точек, считаются выбросами.

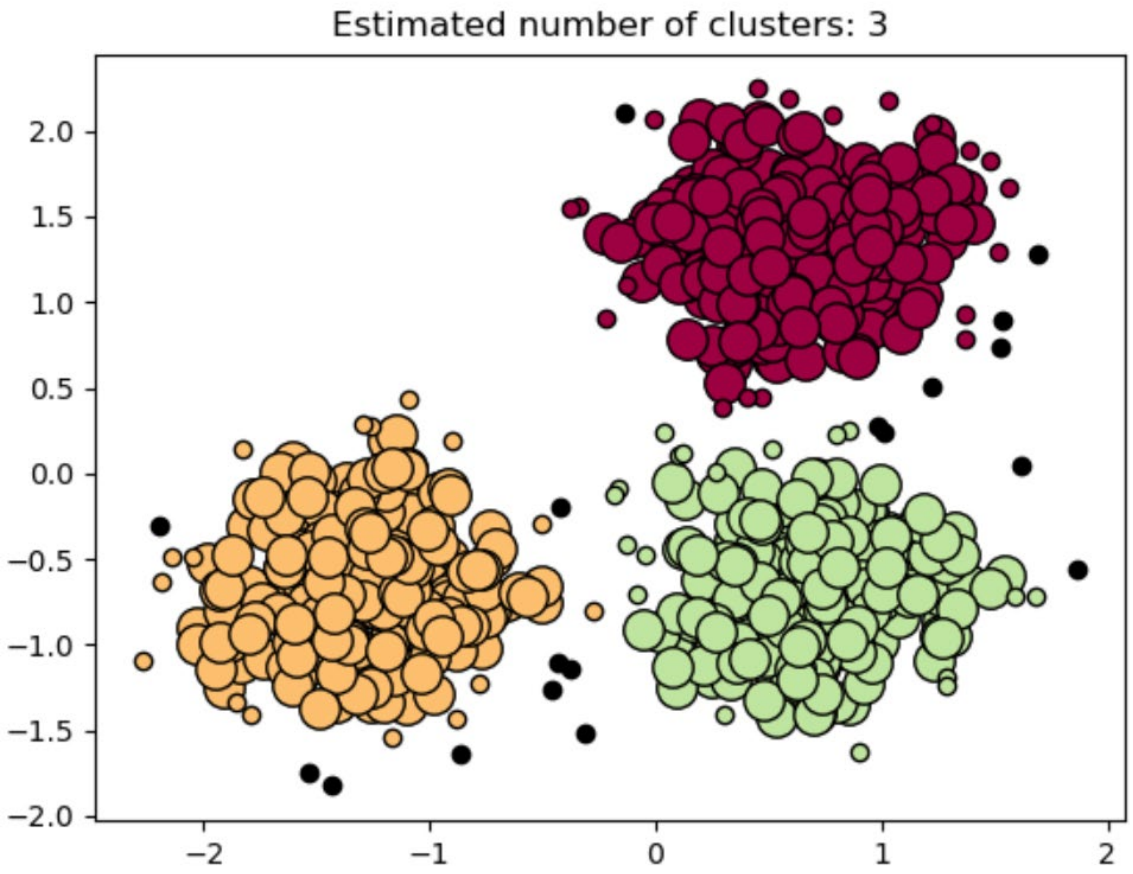
DBSCAN



DBSCAN



DBSCAN



Оценка качества кластеризации

- Нет таргета -> нельзя посчитать стандартные метрики
- Нельзя брать метку кластера как истину
- Будем оценивать как хорошо наши данные разделяются на кластеры
- Помним: объекты в одном классе более похожи, чем в другом

Adjusted Rand index

- Если мы знаем истинные метки
- n – число объектов
- a – число пар с одинаковыми метками в одном кластере
- b – число пар с разными метками в разных кластерах
- $Rand\ Index\ (RI) = \frac{2(a+b)}{n(n-1)}$
- $Adjusted\ Rand\ Index\ (ARI) = \frac{RI - E[RI]}{\max(RI) - E[RI]}$

Adjusted Rand index

- Плюсы

- У случайных (равномерно) меток ARI=0
- От -1 до 1
- Не зависит от структуры кластеров

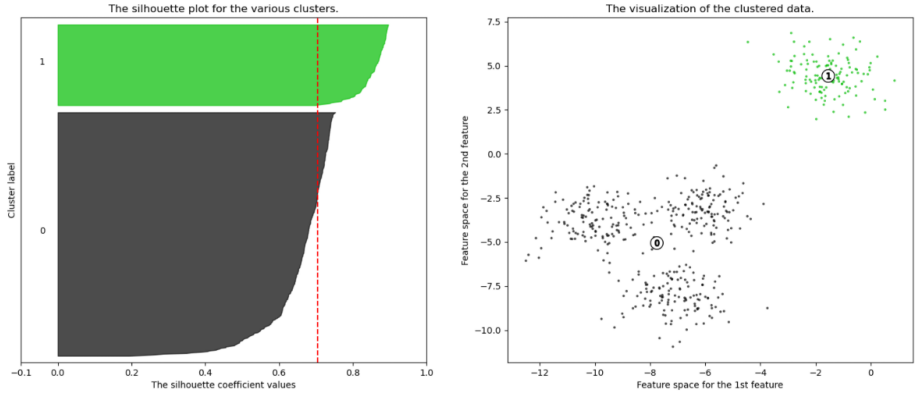
- Минусы

- Нужно знать истинные метки

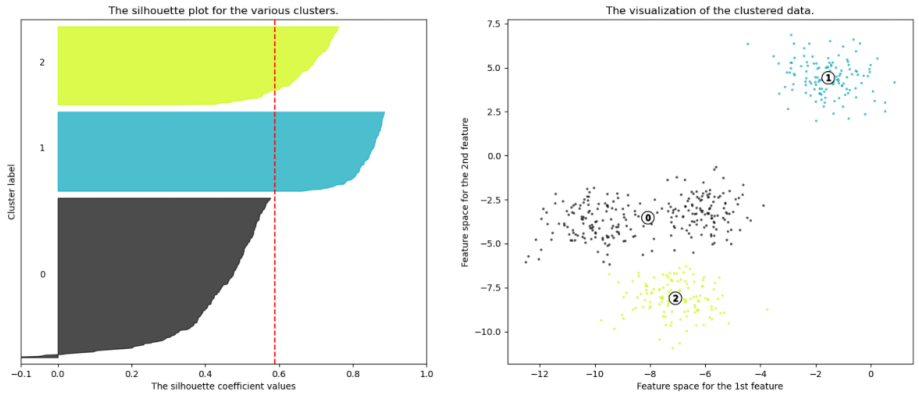
Silhouette

- a – среднее расстояние между объектом и другими в одном кластере
- b – среднее расстояние между объектом и ближайшим другим кластером
- $S = \frac{b-a}{\max(a,b)}$
- Насколько далеко другой кластер от текущего?
- Чем больше, тем лучше

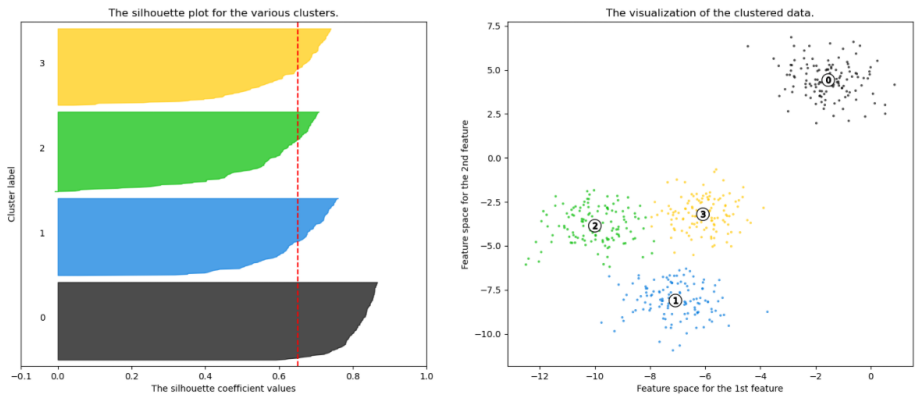
Silhouette analysis for KMeans clustering on sample data with n_clusters = 2



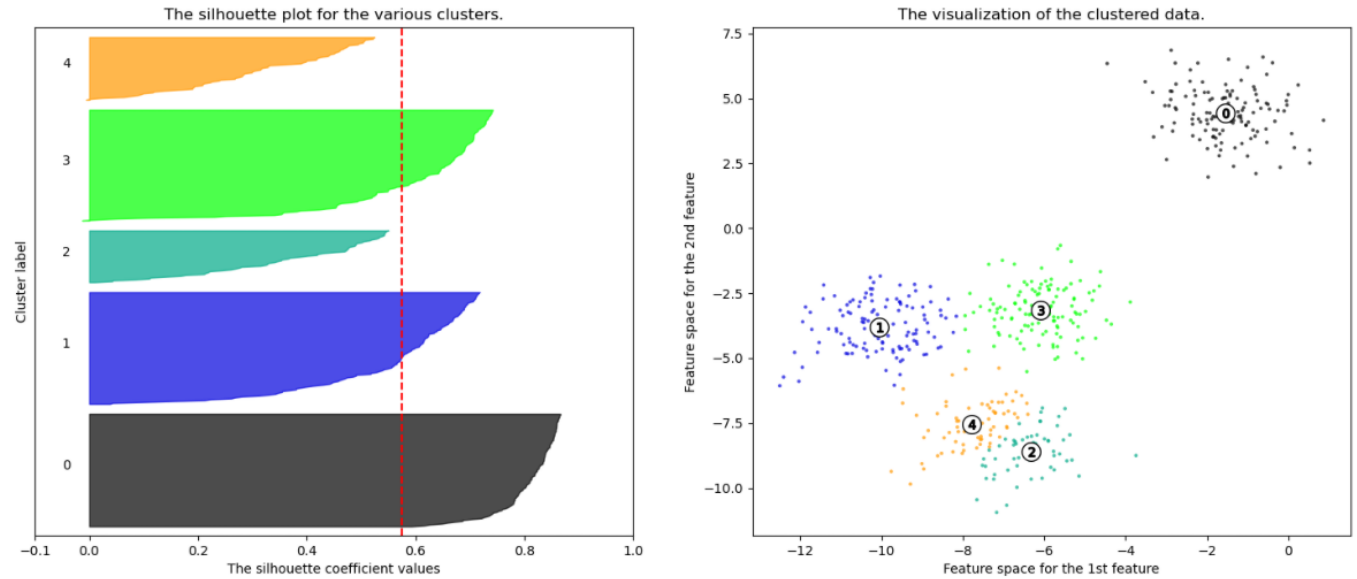
Silhouette analysis for KMeans clustering on sample data with n_clusters = 3



Silhouette analysis for KMeans clustering on sample data with n_clusters = 4



Silhouette analysis for KMeans clustering on sample data with n_clusters = 5



Silhouette analysis for KMeans clustering on sample data with n_clusters = 6