ТЕХНОСФЕРА

Лекция №3

Введение в машинное обучение

Спасёнов Алексей

Введение в машинное обучение. Лекция 3

Содержание лекции

1. Задача кластеризации
2. EM-алгоритм
3. Алгоритм K-Means
4. Иерархическая кластеризация
5. Алгоритм DBSCAN
6. Оценка качества

Задача кластеризации (1/3)

Дано:

$x_i, \dots, x_l$ —объекты обучающей выборки X

ρ — функция расстояния между объектами

Задача:

Поиск меток $y_i, \dots, y_l$, таких, чтобы объекты с одинаковыми метками были близки по ρ, а с разными метками существенно различались

Задача кластеризации (2/3)

Цели:

1) Упрощение обработки данных за счёт разбиения исходного набора данных на схожие подгруппы

2) Уменьшение объёма хранимых данных

3) Поиск объектов, не относящихся ни к одному из исследуемых классов

4) Построение иерархии множества объектов

Задача кластеризации (3/3)

Проблемы:

1)Выбор критерия качества кластеризации

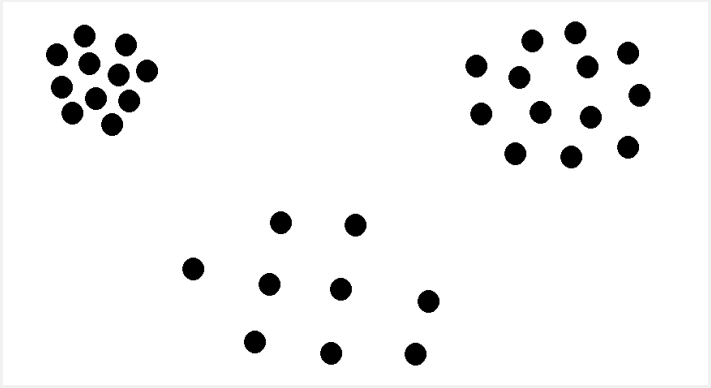
2)Выбор метода кластеризации

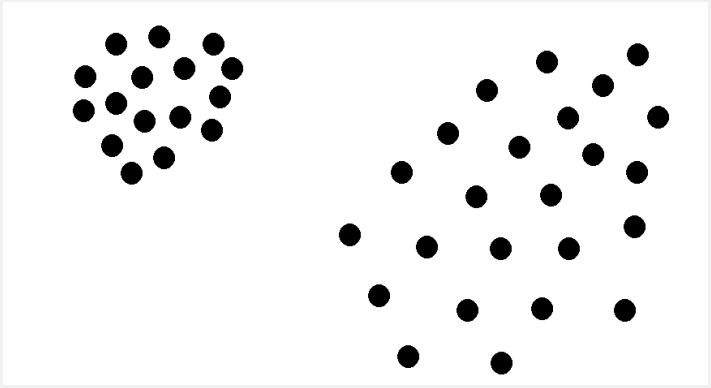
3)Выбор числа кластеров, на которые требуется разбить исходное множество объектов

4)Выбор функции расстояния ρ (метрики)

Конфигурации кластеров (1/4)

Кластеры с центром

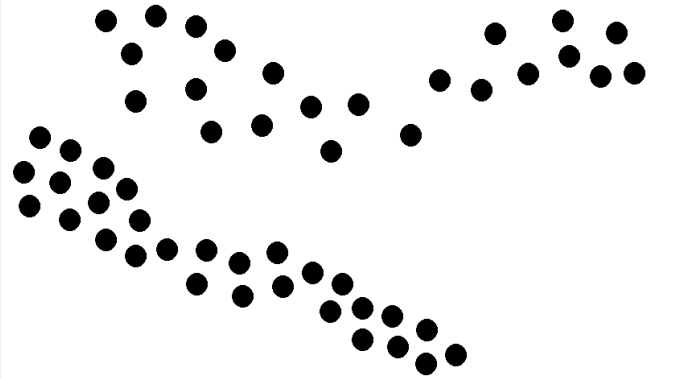




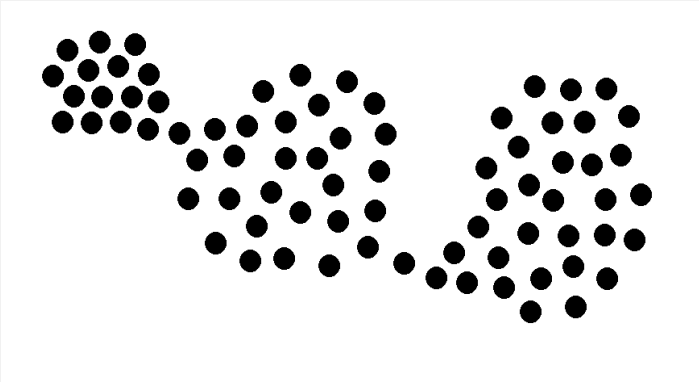
Расстояние между объектами внутри кластера меньше межкластерного

Конфигурации кластеров (2/4)

Ленточные кластеры

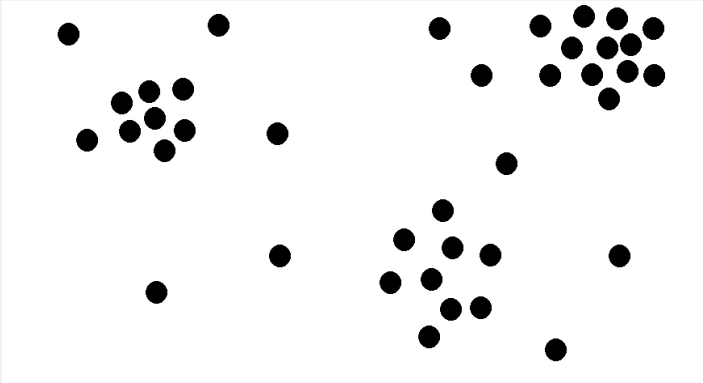


Кластеры с перемычками

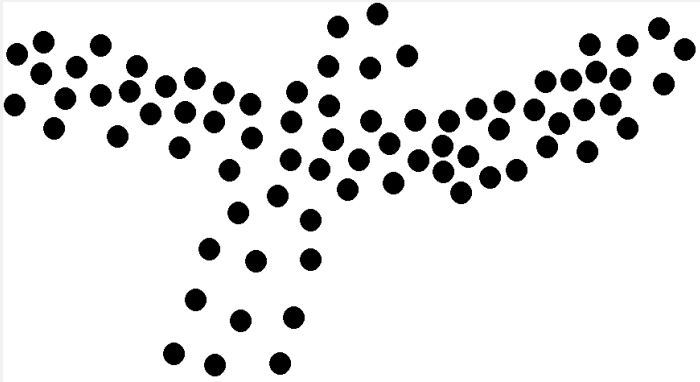


Конфигурации кластеров (3/4)

Присутствие фона

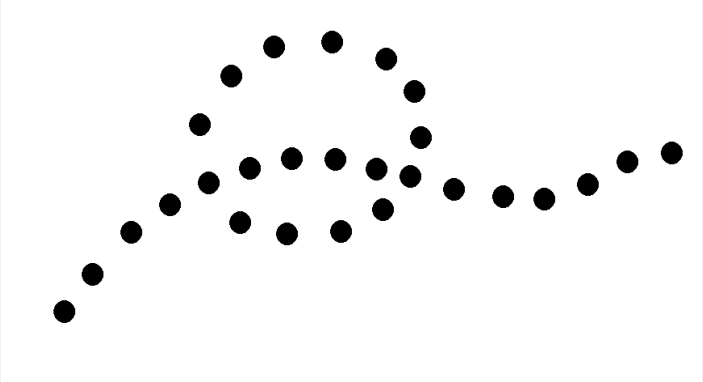


Кластеры с перекрытием

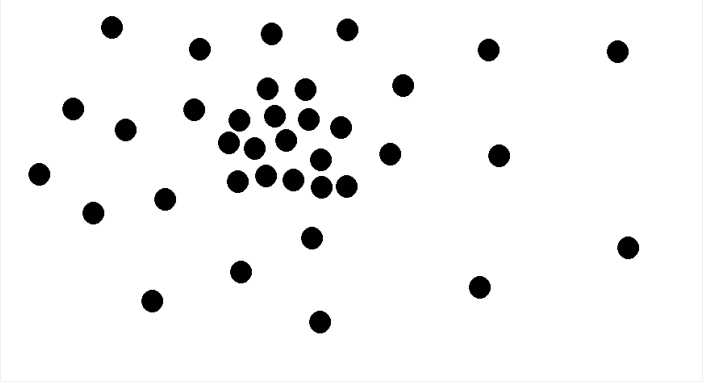


Конфигурации кластеров (4/4)

Внутренние особенности

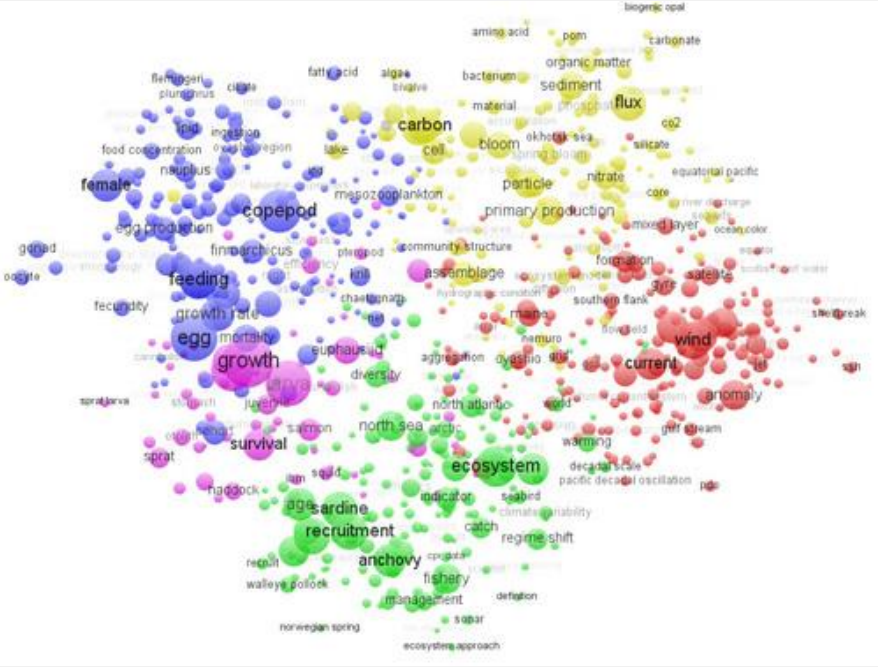


Отсутствие кластеров



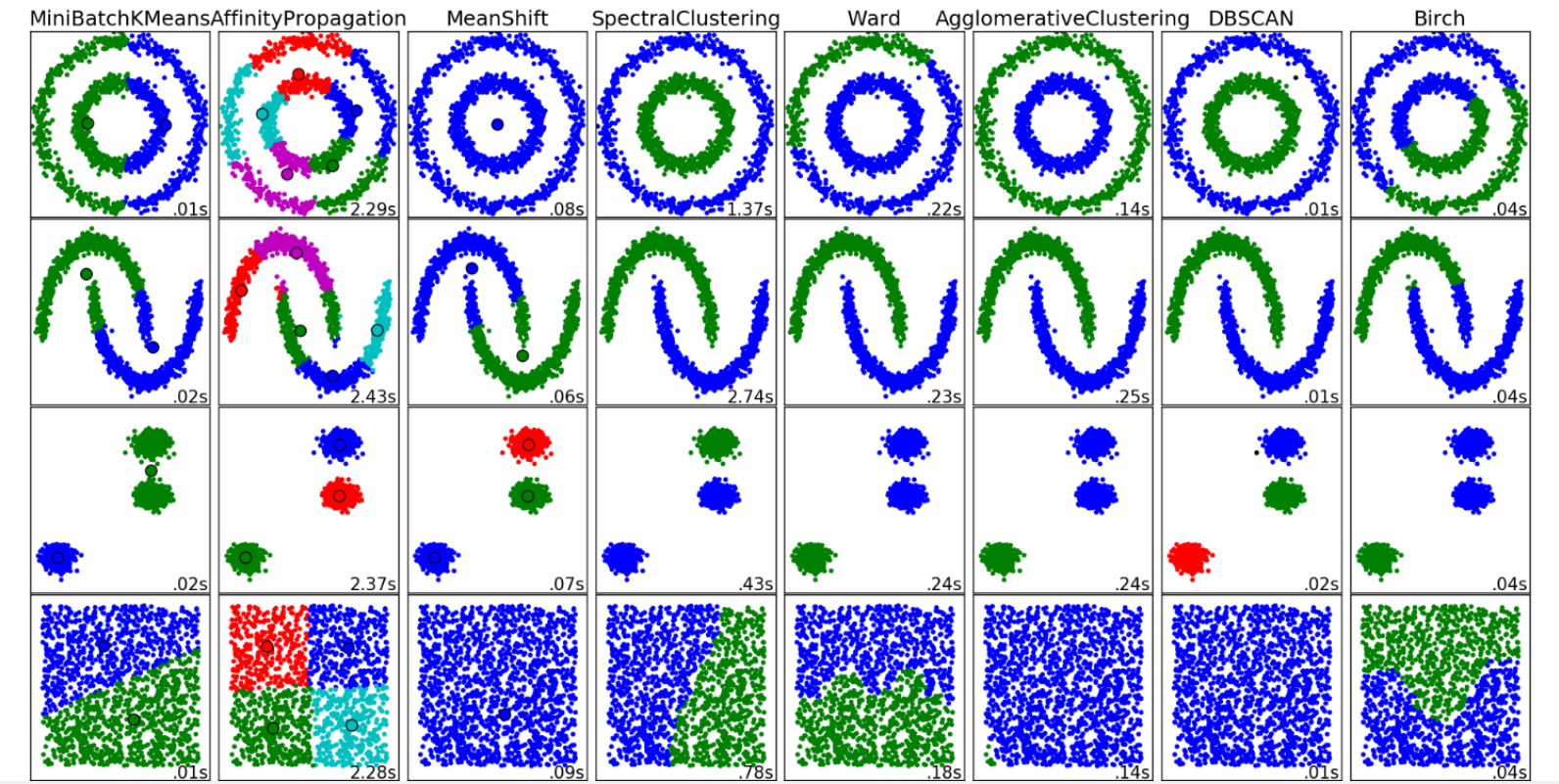
Типы кластеризации

Жёсткая и мягкая кластеризация



В «мягкой» кластеризации объект можно отнести к нескольким кластерам с разным весом

Алгоритмы кластеризации



http://scikit-learn.org/stable/modules/clustering.html#clustering

EM-алгоритм

Имеется выборка X^l, состоящая из смеси распределений

$$p(x) = \sum_{y=1}^{M} w_y p_y(x), \sum_{y=1}^{M} w_y = 1$$

$p_y(x)$ – плотность

w_y - априорная вероятность кластера y

EM-алгоритм

$X = R^n$,

Кластеры n-мерные гауссовские:

$$p_y = (2\pi)^{-\frac{n}{2}}(\sigma_{y1} \dots \sigma_{yn})^{-1}\exp\left(-\frac{1}{2}\rho_y^2(x, \mu_y)\right)$$

$\mu_y = (\mu_{y1} \dots \mu_{yn})$ - центр кластера y,

$\Sigma_y = diag(\sigma_{y1}^2, \dots, \sigma_{yn}^2)$ – диагональная матрица ковариаций,

$$\rho_y^2(x, x') = \sum_{j=1}^{n} \sigma_{yj}^{-1}|f_j(x) - f_j(x')|^2$$

Алгоритмы кластеризации

Шаг 1: Выбираем начальные приближения для w_y, μ_y, Σ_y

Шаг 2: do

Шаг 2.1: E-шаг (expectation)

$$g_{iy} = P(y|x_i) = \frac{w_y p_y(x_i)}{\sum_{j=1}^{M} w_z p_z(x_i)}, y \in Y, i = 1, \dots, l$$

Шаг 2.2: M-шаг (maximization)

$$w_y = \frac{1}{l}\sum_{i=1}^{l} g_{iy}, y \in Y$$

$$\mu_{yj} = \frac{1}{l w_y}\sum_{i=1}^{l} g_{iy} f_j(x_i), y \in Y, j = 1, \dots, n$$

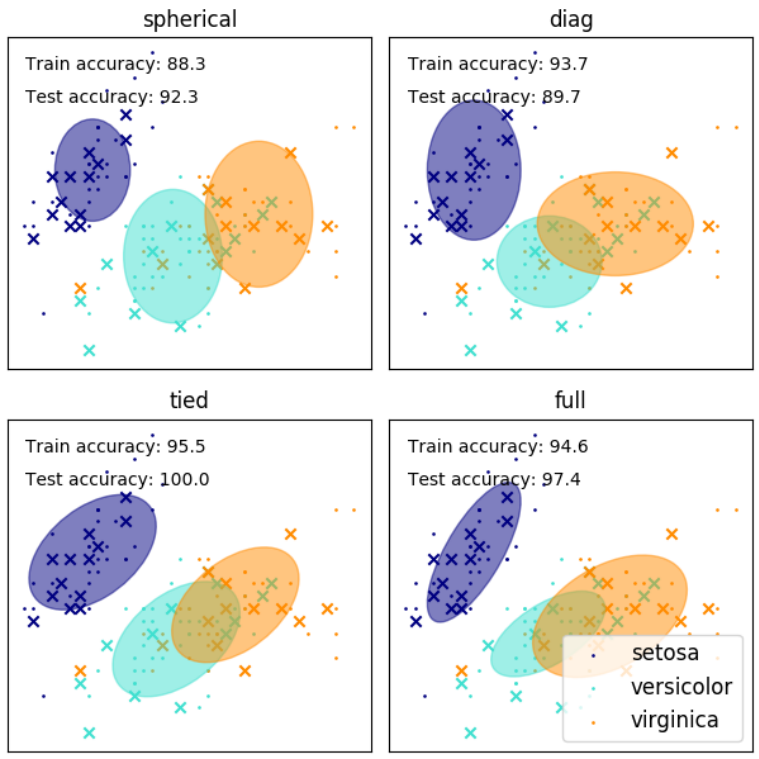
$$\sigma_{yj}^2 = \frac{1}{l w_y}\sum_{i=1}^{l} g_{iy}(f_j(x_i) - \mu_{yj})^2, y \in Y, j = 1, \dots, n$$

$$y_i = \arg\max_{y \in Y} g_{iy}, i = 1, \dots, l$$

while Не будут изменяться y_i

Алгоритмы кластеризации

EM-алгоритм



http://scikit-learn.org/stable/modules/mixture.html

Алгоритм K-Means

Шаг 1: Выбираем начальные приближения для μ_y (положения центров)

Шаг 2: do

Шаг 2.1: Аналог E-шага

Относим каждый объект x_i к ближайшему центру:

$$y_i = \arg\min \rho(x_i, \mu_y), y \in Y, i = 1, \dots, l$$

Шаг 2.2: Аналог M-шага

Вычисляем новые положения центров

$$\mu_y = \frac{\sum_{i=1}^{l} [y_i = y] f_j(x_i)}{\sum_{i=1}^{l} [y_i = y]}, y \in Y, j = 1, \dots, n$$

while Не будут изменяться y_i

Алгоритм K-Means

Оптимизируем среднее внутриклассовое расстояние:

$$F = \frac{\sum_{i<j}[y_i = y_j]\rho(x_i, x_j)}{\sum_{i<j}[y_i = y_j]} \to min$$

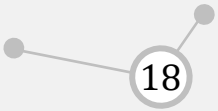
Алгоритм K-Means

Отличия от EM-алгоритма:

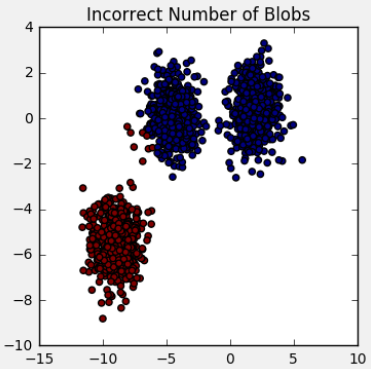
1) В алгоритме EM – мягкая кластеризация, в k-means – жёсткая
2) В алгоритме EM форма кластеров эллиптическая, в k-means зависит от выбора метрики ρ

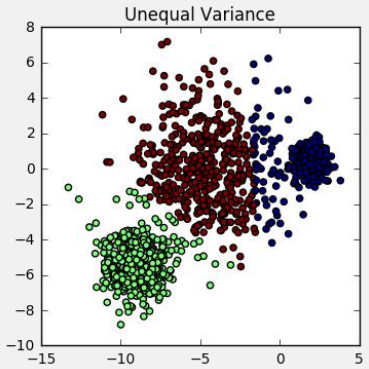
Недостатки алгоритма:

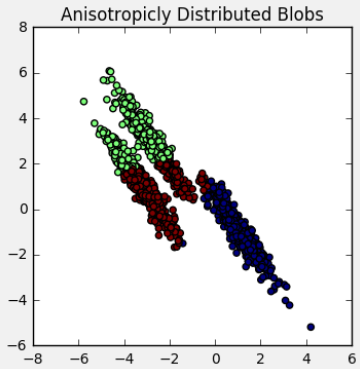
1) Чувствителен к выбору исходному расположению центров кластеров
2) Необходимо выбирать количество кластеров

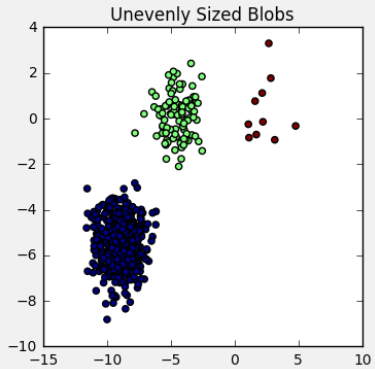


Алгоритм K-Means









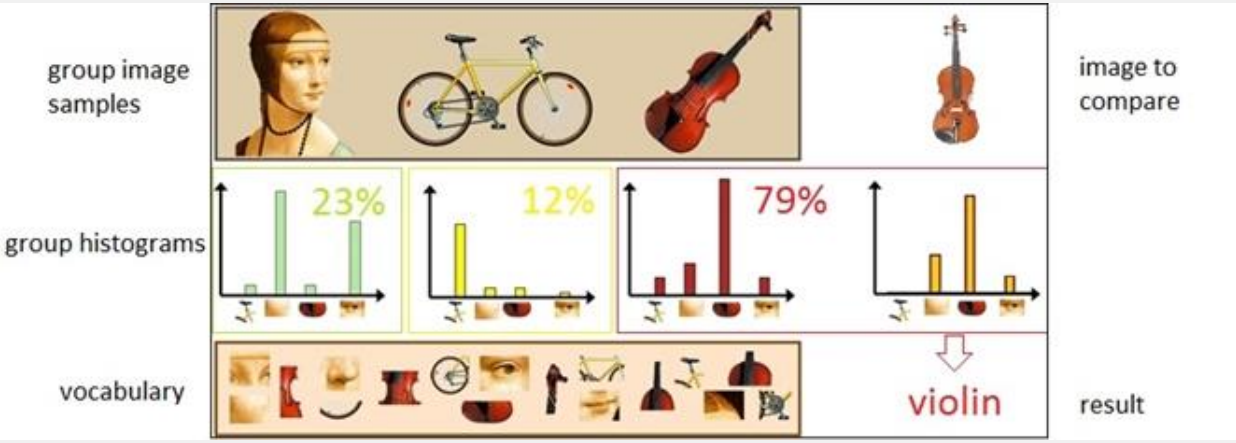
http://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

Алгоритм K-Means

64 цвета
(кластера)

Алгоритм K-Means

Bag of visual words



Модификации K-Means

Mini-Batch K-Means

Если данных достаточно много, то вычисление расстояний от всех объектов до центов кластеров может занять достаточно много времени

Решение: На каждом шаге выбирать из набора данных случайную подвыборку

Модификации K-Means

K-Means++

1)Выбор начального приближения центров кластеров значительно влияет на скорость сходимости алгоритма

2)Выбираем начальные положения центров на максимальном расстоянии друг от друга

3)Решение:

1)Выбираем начальные центры из равномерного распределения на выборке

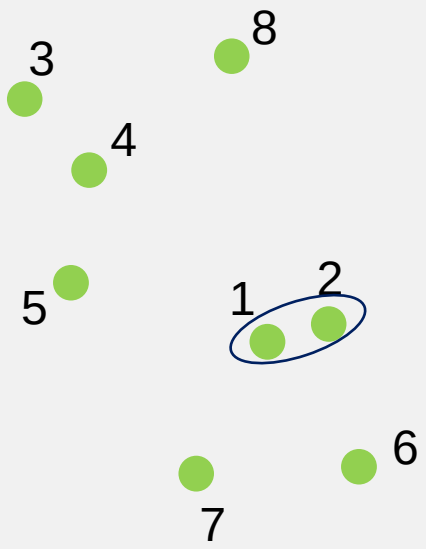
2)Каждый следующий центр выбираем случайно из оставшихся точек так, чтобы вероятность выбора точки была пропорциональна квадрату расстояний от неё до ближайшего центра

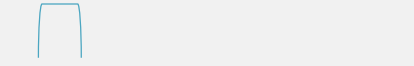
Иерархическая кластеризация

Различают 2 вида алгоритмов иерархической кластеризации

1) **Дивизимные** или нисходящие алгоритмы разбивают выборку на всё более и более мелкие кластеры.
2) **Агломеративные** или восходящие алгоритмы, в которых объекты объединяются во всё более и более крупные кластеры.

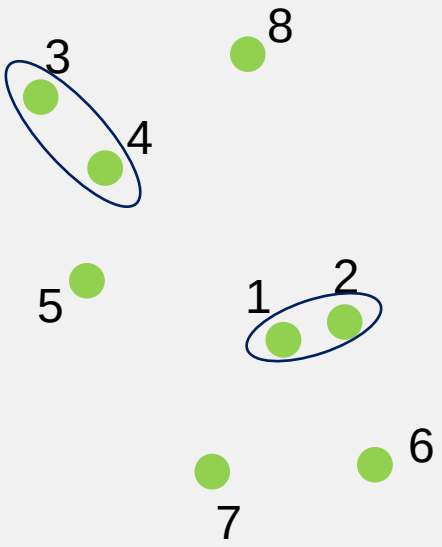
Агломеративная кластеризация

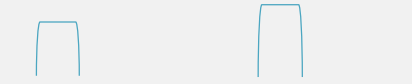




Дендрограмма

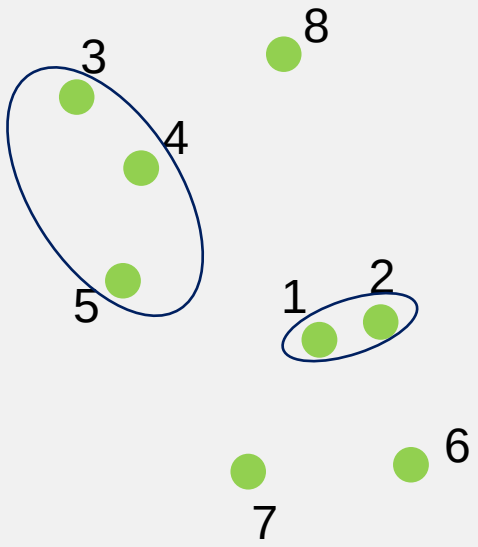
Агломеративная кластеризация

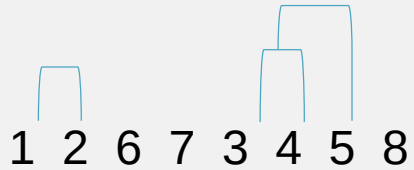




Дендрограмма

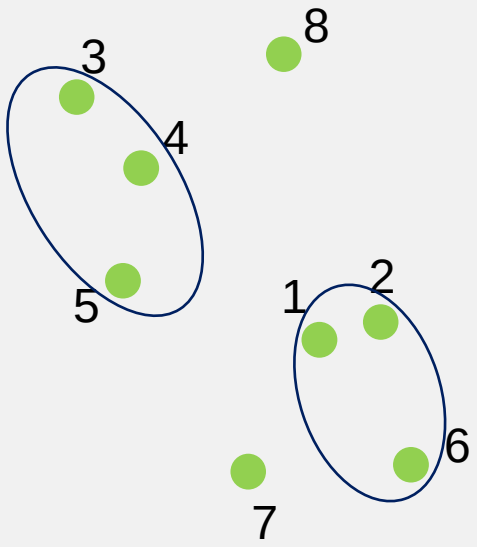
Агломеративная кластеризация

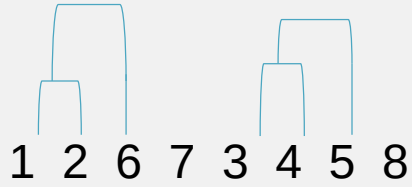




Дендрограмма

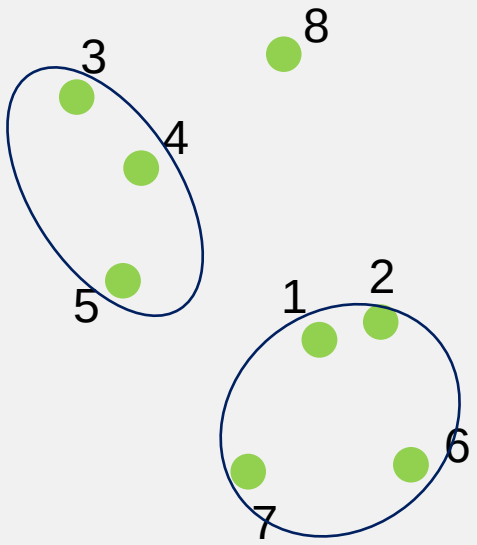
Агломеративная кластеризация

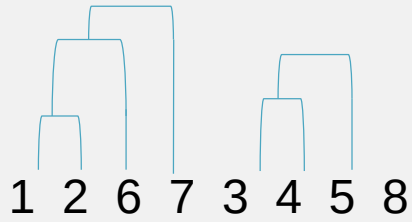




Дендрограмма

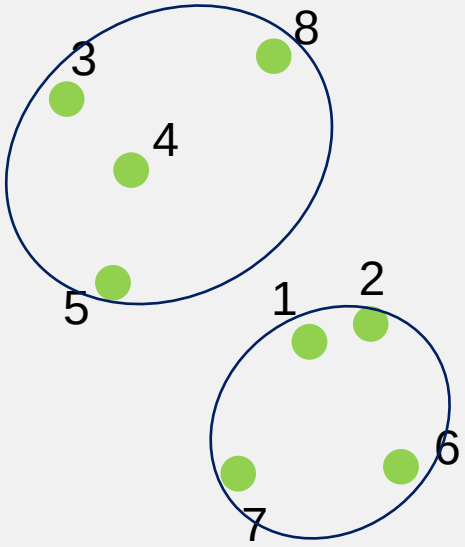
Агломеративная кластеризация

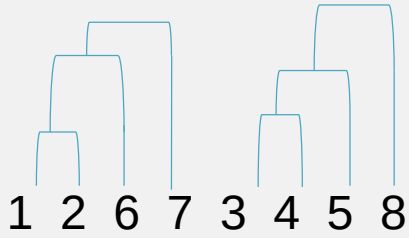




Дендрограмма

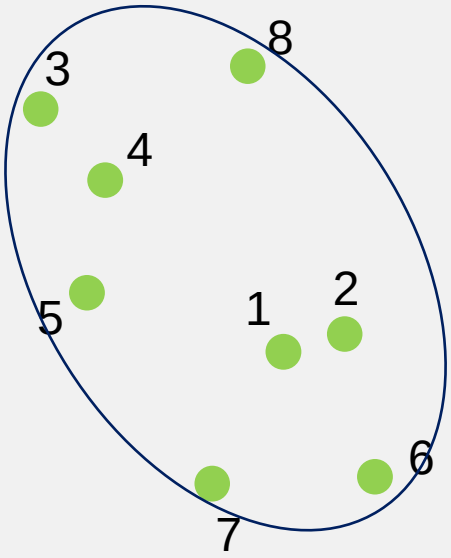
Агломеративная кластеризация

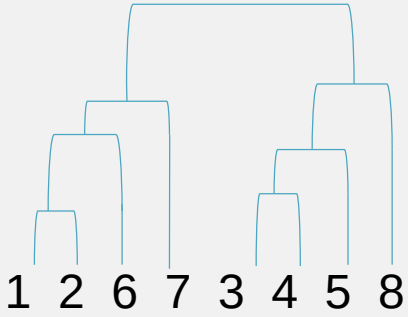




Дендрограмма

Агломеративная кластеризация





Дендрограмма

Формула Ланса-Уильямса

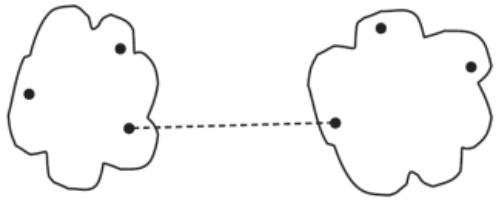
$$
\begin{aligned}
R(U \cup V, S) = {} & \alpha_U \cdot R(U, S) + \\
& + \alpha_V \cdot R(V, S) + \\
& + \beta \cdot R(U, V) + \\
& + \gamma \cdot \left| R(U, S) - R(V, S) \right|,
\end{aligned}
$$

где α_U, α_V, β, γ — числовые параметры.

Алгоритмы кластеризации

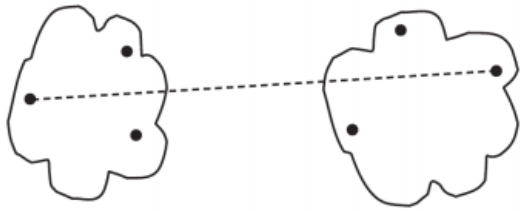
Расстояние ближайшего соседа

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



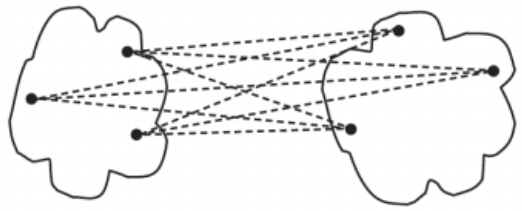
Расстояние дальнего соседа

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

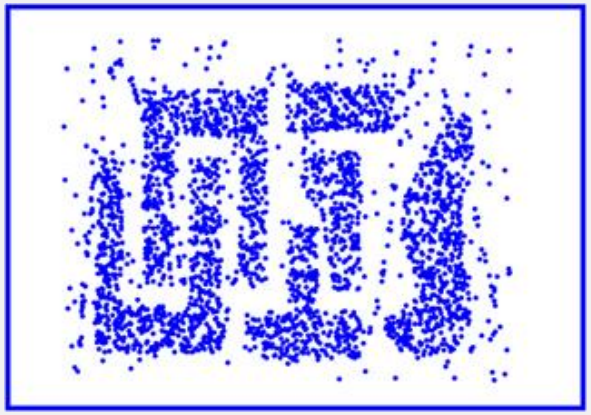


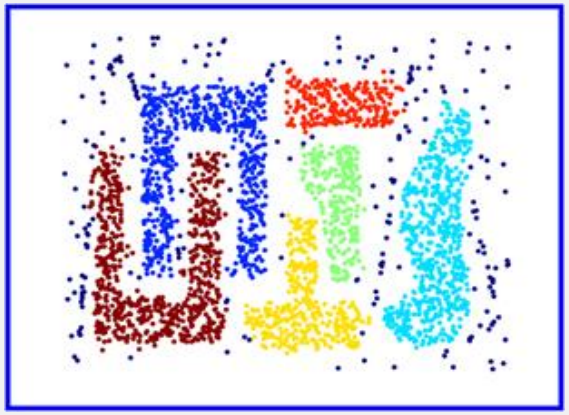
Групповое среднее расстояние

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

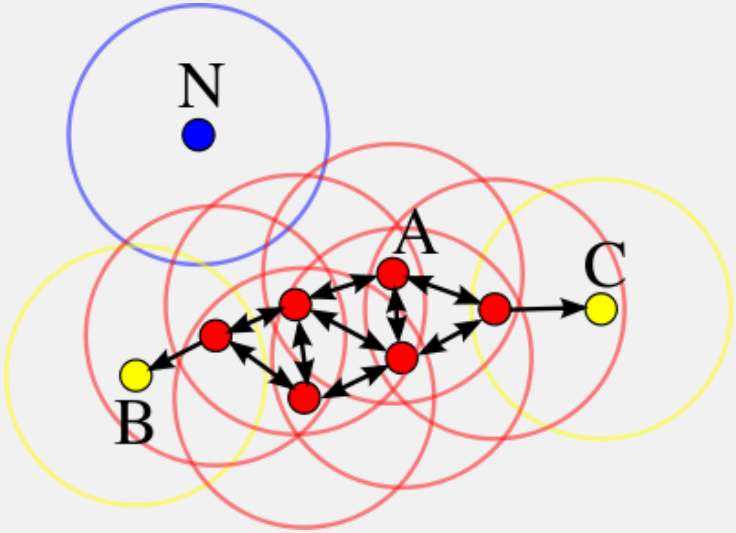


Алгоритм DBSCAN





Алгоритм DBSCAN



Шаг 1: Помечаем все точки как основные, пограничные или шумовые

Шаг 2: Отбрасываем точки шума

Шаг 3: Соединяем все основные точки, находящиеся на расстоянии Eps друг от друга

Шаг 4: Соединяем каждую группу объединённых точек в отдельный кластер

Оценка качества кластеризации

Внутриктерное расстояние:

$$F_s = \frac{\sum_{i,j}[y_i = y_j]\rho(x_i, x_j)}{\sum_{i,j}[y_i = y_j]} \to min$$

Межкластерное расстояние:

$$F_1 = \frac{\sum_{i,j}[y_i \neq y_j]\rho(x_i, x_j)}{\sum_{i,j}[y_i \neq y_j]} \to max$$

Объединение расстояний F_0 и F_1

Алгоритмы кластеризации

Silhouette score

$$F_0 = \frac{F_0 - F_1}{\max(F_0, F_1)}$$

Межкластерное расстояние: F_0

Внутриктерное расстояние: F_1

И другие:
http://scikit-learn.org/stable/modules/clustering.html#clustering-evaluation

Алгоритмы кластеризации

Если известна исходная разметка:

Homogeneity score (однородность)

Кластеры содержат только те элементы, которые являются объектами одного и того же класса.

Completeness score (полнота)

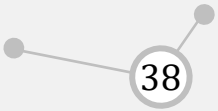
Все объекты данных, являющиеся членами данного класса, являются элементами одного и того же кластера.

V-measure score:

$$v = \frac{2 * (\text{homogeneity} * \text{completeness})}{\text{homogeneity} + \text{completeness}}$$

homogeneity_score(a, b) == completeness_score(b, a)

http://aclweb.org/anthology/D/D07/D07-1043.pdf



ТЕХНОСФЕРА

Спасибо за внимание!

Спасёнов Алексей

a.spasenov@corp.mail.ru