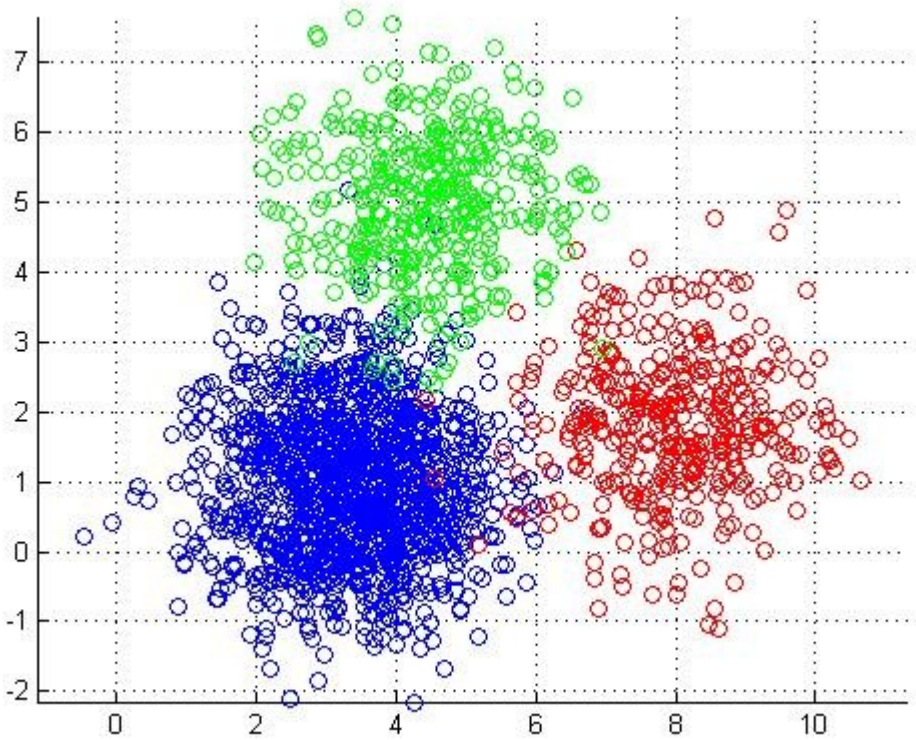
Машинное обучение
Кластеризация



https://yandexdataschool.ru/edu-process/courses/machine-learning

Содержание лекции

- Постановка задачи
- EM-алгоритм
- Метод k-средних
- DBSCAN

Постановка задачи

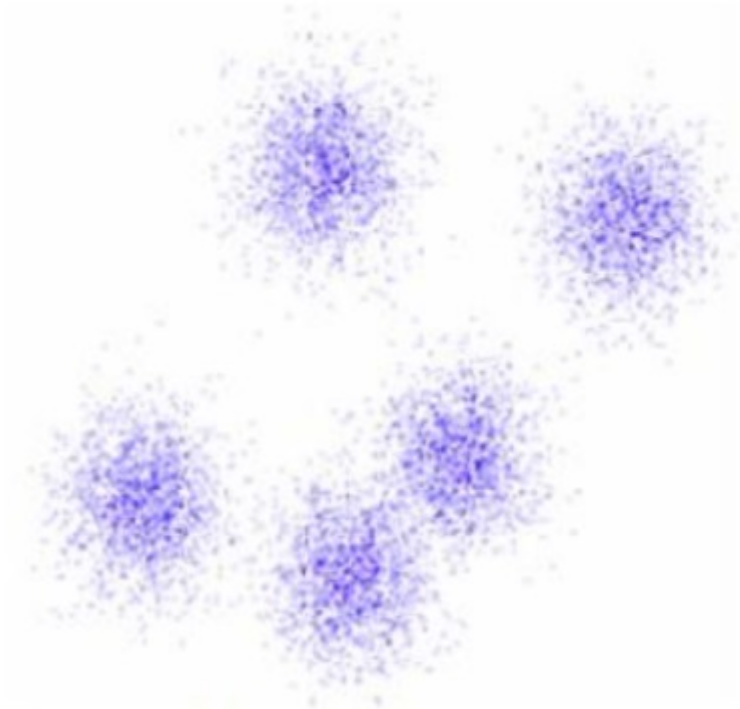
- Дано:
  - пространство объектов X
  - обучающая выборка X^{ℓ}
  - метрика между объектами
- Найти:
  - множество кластеров Y
  - алгоритм кластеризации a : X→Y
- Каждый кластер должен состоять из близких объектов
- Объекты разных кластеров должны быть существенно различны

Классификация и кластеризация

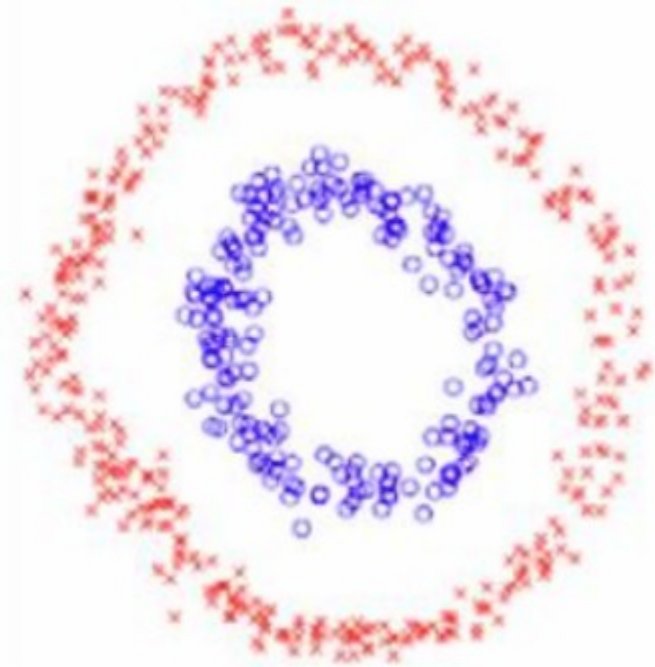
Классификация	Кластеризация
• Известное количество классов	• Неизвестно количество классов
• Классы известны для объектов обучающей выборки	• Нет данных о классах в обучающей выборке
• Используется для классификации объектов “в будущем”	• Используется для исследования множества объектов
• Классификация – это обучение с учителем	• Кластеризация – это обучение без учителя

Близость или связанность?

- Compactness, e.g., k-means, mixture models
- Connectivity, e.g., spectral clustering

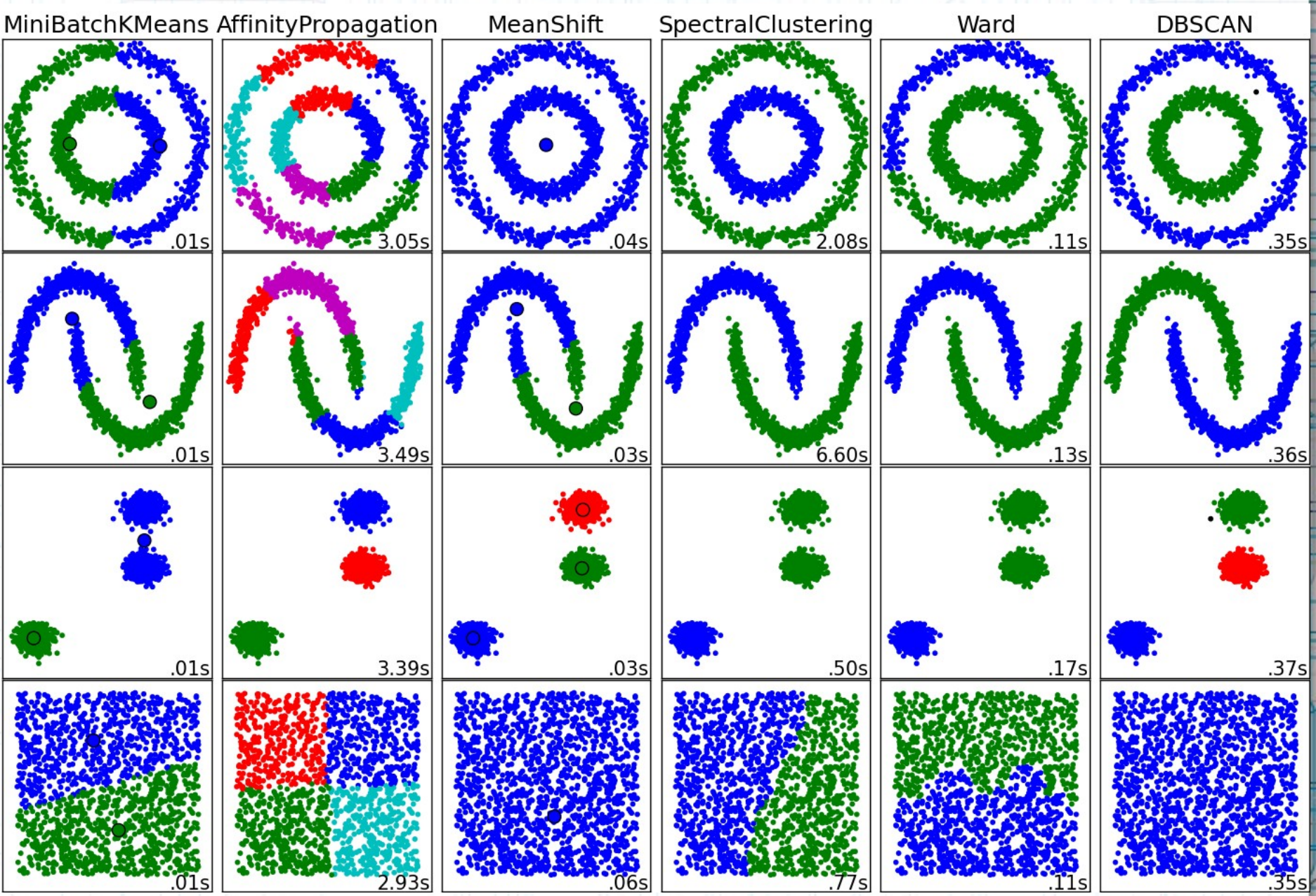


Compactness

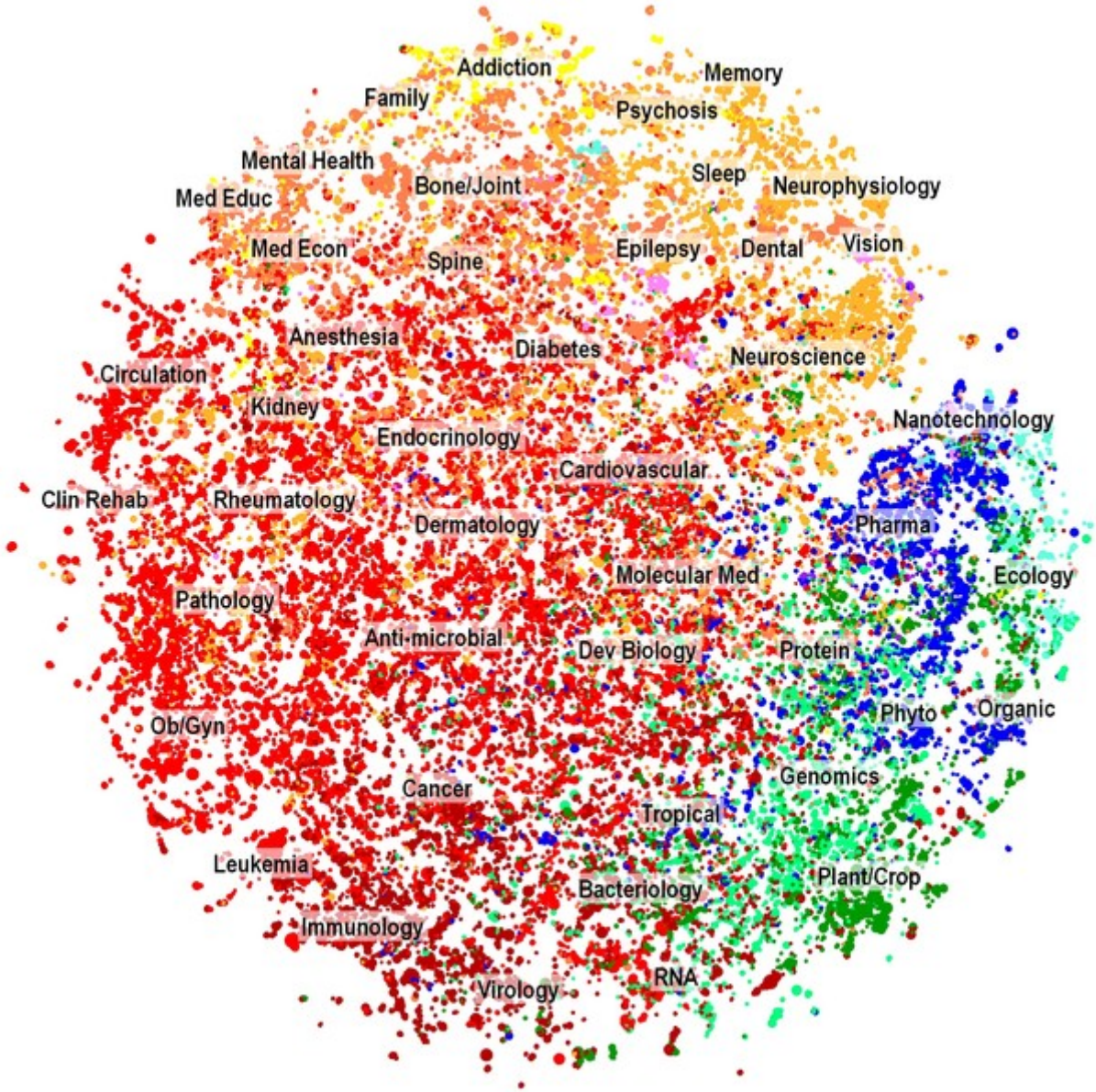


Connectivity

Пример: результаты работы алгоритмов кластеризации



Пример: кластеризация статей по медицине



EM-кластеризация

Гипотеза: выборка X^ℓ порождена смесью гауссовских случайных распределений

$$p(x) = \sum_{y \in Y} w_y p_y(x), \qquad \sum_{y \in Y} w_y = 1,$$

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

$\mu_y = (\mu_{y1}, \ldots, \mu_{yn})$ — центр кластера y;
$\Sigma_y = \operatorname{diag}(\sigma_{y1}^2, \ldots, \sigma_{yn}^2)$ — диагональная матрица ковариаций;

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

EM-кластеризация

1: начальное приближение w_y, μ_y, Σ_y для всех $y \in Y$;

2: **повторять**

3: E-шаг (expectation):

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

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

6: **пока** y_i не перестанут изменяться;

Метод k-средних

1: начальное приближение центров μ_y, $y \in Y$;

2: **повторять**

3: аналог E-шага:

отнести каждый x_i к ближайшему центру:

$y_i := \arg\min_{y \in Y} \rho(x_i, \mu_y)$, $i = 1, \ldots, \ell$;

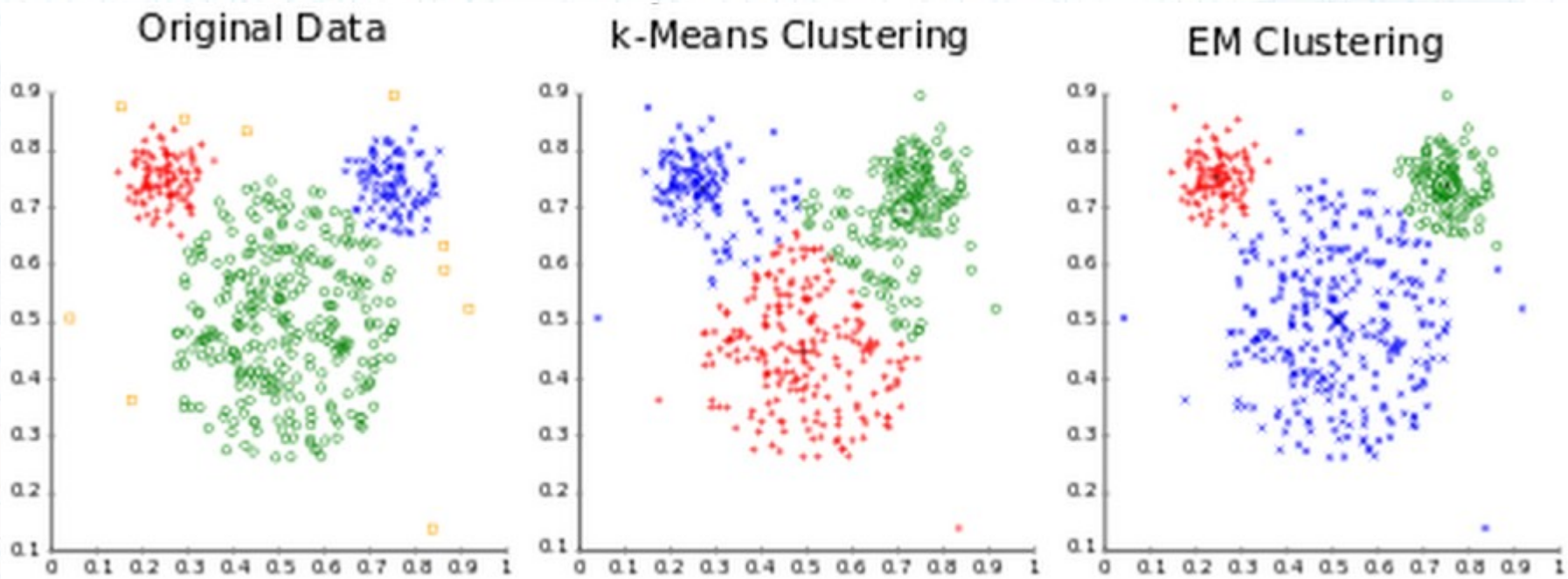
4: аналог M-шага:

вычислить новые положения центров:

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

5: **пока** y_i не перестанут изменяться;

Сравнение k-средних и EM-кластеризации



DBSCAN

- Density-Based Spatial Clustering of Applications with Noise – самый популярный алгоритм кластеризации
- Ключевые понятия:
  - Внутренняя точка – имеет более MinPts соседей (r<Eps)
  - Граничная точка – имеет меньше соседей, но является соседней к какой-либо внутренней точке
  - Остальные точки - шумовые
  - Достижимость по плотности: точка q достижима из внутренней точки p, если существует последовательность Eps-соседних внутренних точек от p к q

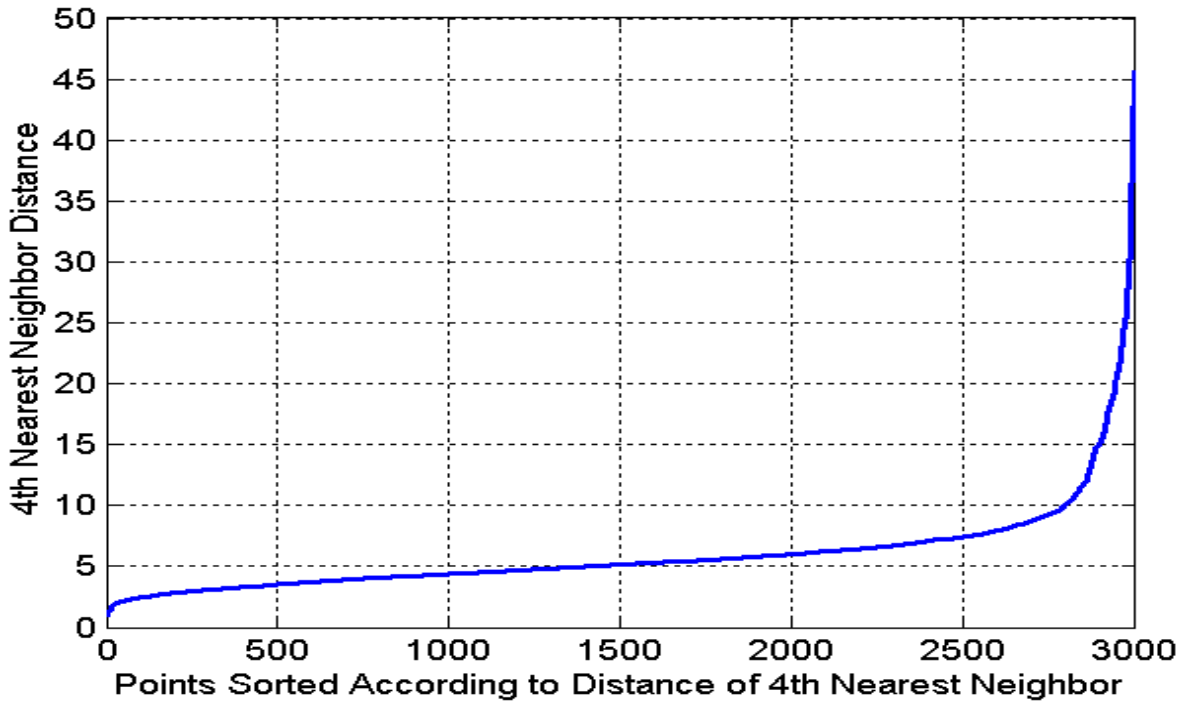
Алгоритм DBSCAN

- Выбрать точку p
- Если p-внутренняя, то
  - Найти все достижимые по плотности точки из p
  - Сформировать кластер
- Иначе – перейти к следующей точке

- Результат не зависит от порядка просмотра точек

DBSCAN: выбор Eps и MinPts

- Ключевая идея: для всех точек одного кластера их k-тый сосед (k<размера кластера) находится на приблизительно одном и том же расстоянии
- Соседи шумовых точек – далеко
- График отсортированных расстояний:



DBSCAN: выбор Eps и MinPts

- Искомое Eps - начало крутого подъема на графике расстояний до соседа с фиксированным номером
- MinPts – номер соседа