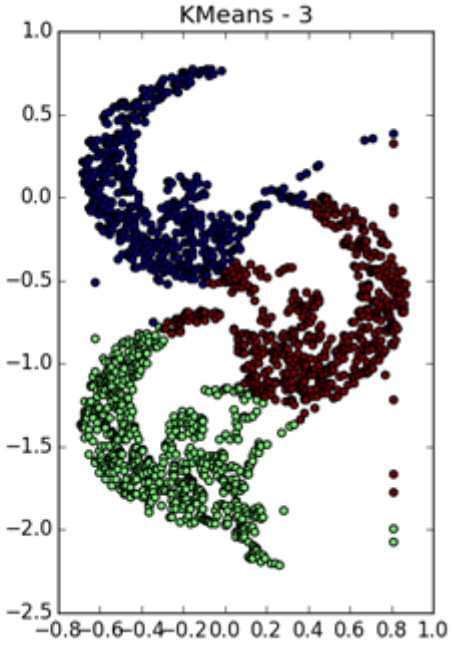
Задача кластеризации данных

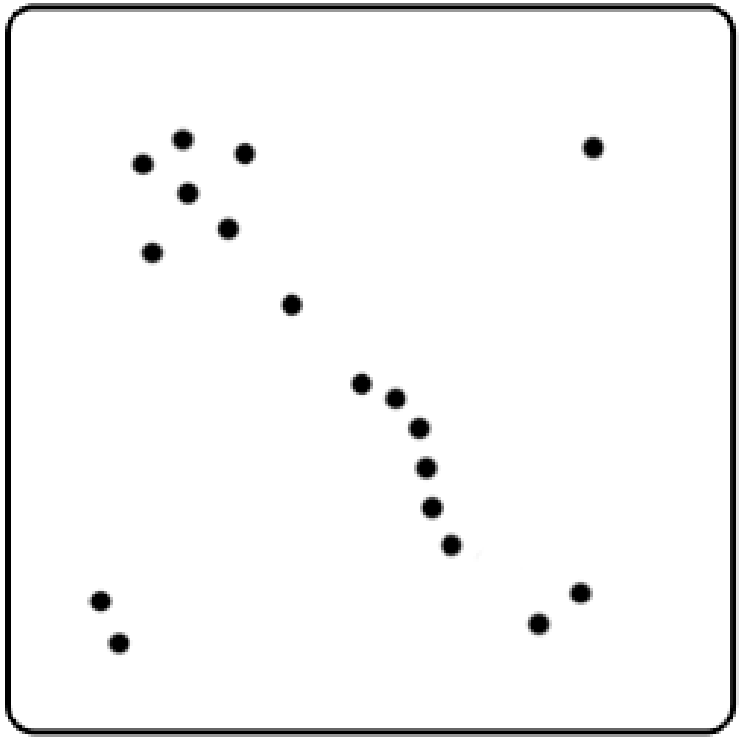


Группа людей, действуя совместно, может свершить такое, о чем поодиночке они не могли бы и мечтать.

Франклин Рузвельт

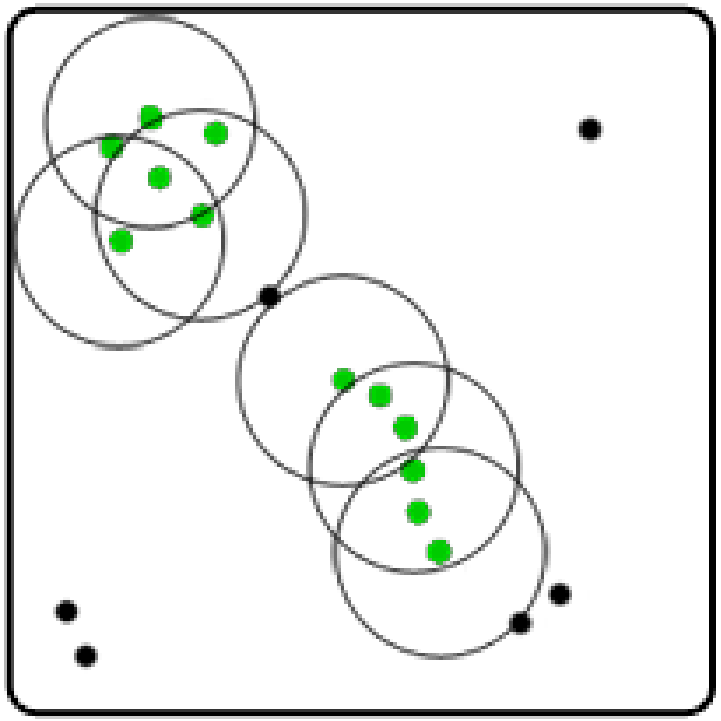
Содержание

Алгоритм DBSCAN: базовые идеи



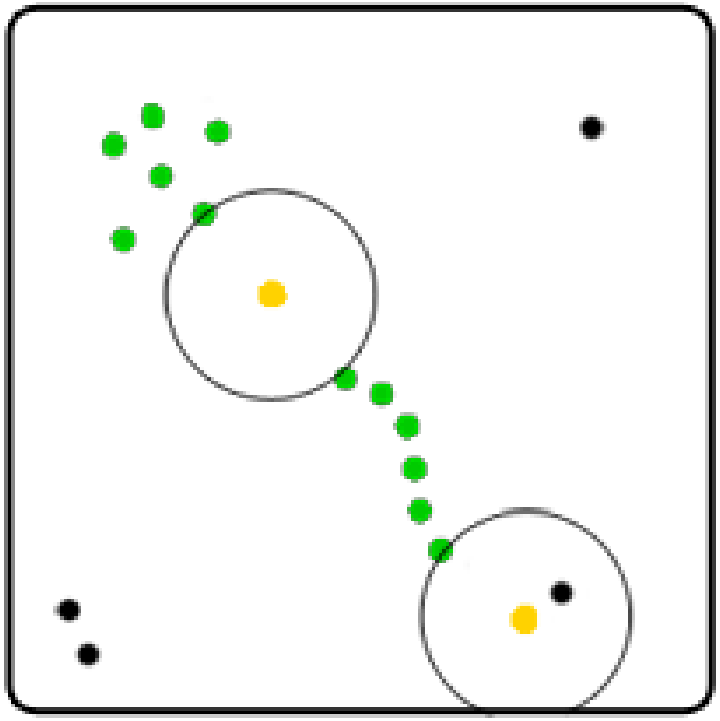
- Для каждой точки оценить плотность (количество) других точек (соседей) в ее окрестности и итеративно формировать кластеры
- Выбранная точка входит в кластер, если *близко* от нее находится *несколько* соседей
  - «Близко» – параметр Eps
  - «Несколько» – параметр $MinPts$

Алгоритм DBSCAN: корневые точки



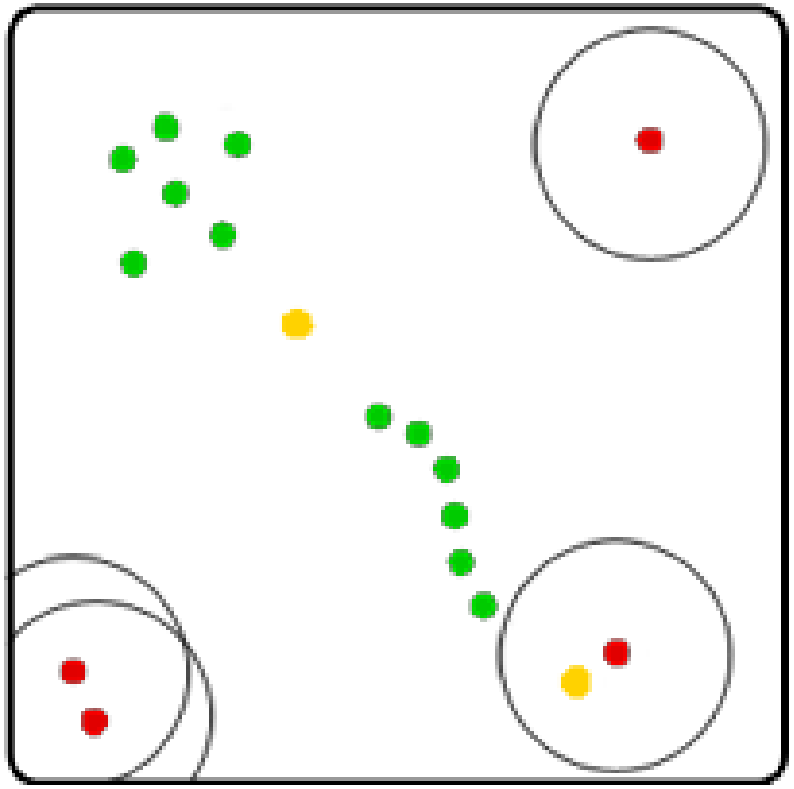
- Для каждой точки подсчитать, сколько соседей в ее Eps-окрестности
- Если соседей не менее $MinPts$, точка является *корневой (core point)*: она будет формировать кластер

Алгоритм DBSCAN: граничные точки



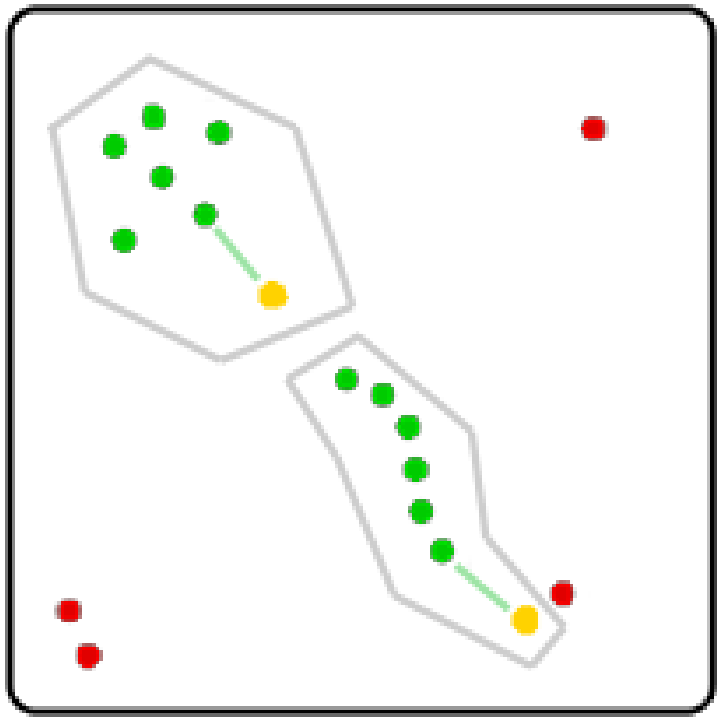
- Точка, у которой менее $MinPts$ соседей, хотя бы один из которых корневой, является *граничной (border point)*: она будет на границе кластеров

Алгоритм DBSCAN: точки-выбросы (шум)



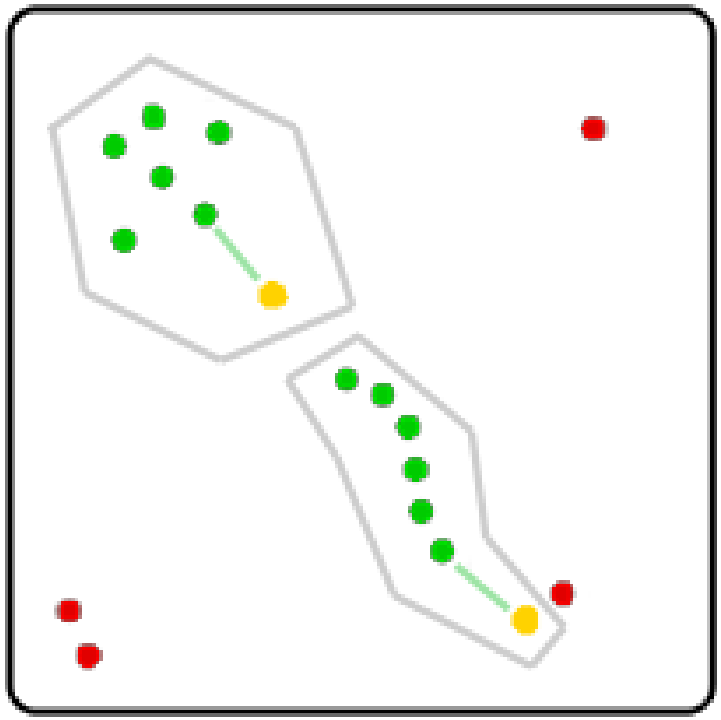
- Выбросы (*noise point*) – оставшиеся точки, не принадлежащие ни одной группе

Алгоритм DBSCAN



- Если между двумя точками есть цепочка корневых соседей, то эти точки из одного кластера
  - Подобные цепочки разделены пустым пространством или граничными точками
  - Цепочки корневых соседей можно занумеровать (это кластеры)

Алгоритм DBSCAN



- Если между двумя точками есть цепочка корневых соседей, то эти точки из одного кластера
  - Подобные цепочки разделены пустым пространством или граничными точками
  - Цепочки корневых соседей можно занумеровать (это кластеры)
- Если у граничной точки только один корневой сосед, то точка будет в том же кластере, что и этот сосед. Если несколько – можно выбрать кластер ближайшего корневого соседа

Алгоритм DBSCAN

- Параметры: метрика $d(\cdot,\cdot)$, Eps, $MinPts$
- Eps-окрестность
  - $E(p) = \{q \mid d(p,q) \leq Eps\}$
- Корневая точка
  - $|E(p)| \geq MinPts$
- Непосредственная достижимость
  - точка p *непосредственно достижима (directly density-reachable)* из точки q, если q – корневая точка и $p \in E(q)$
- Достижимость
  - точка p *достижима (density-reachable)* из точки q, если $\exists\, p_1, p_2, \ldots, p_n$, что $p_1 \equiv q$, $p_n \equiv p$ и $\forall\, 1 \leq i \leq n-1$ p_{i+1} непосредственно достижима из p_i

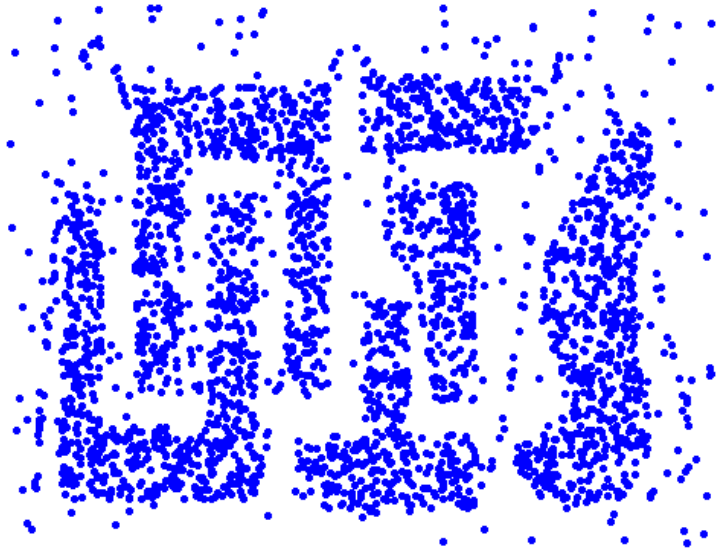
Алгоритм DBSCAN

- Выбрать произвольную корневую точку, пометить ее как обработанную.
- Поместить всех непосредственно достижимых соседей корневой точки в список обхода.
- Для каждой точки из списка обхода:
  - пометить эту точку как обработанную
  - если она тоже корневая, добавить всех ее соседей в список обхода.
- Кластеры помеченных точек, сформированные в ходе этого алгоритма, максимальны и связны в смысле достижимости
- Если обойдены не все точки, можно перезапустить обход из другой корневой точки, и новый кластер не поглотит предыдущий

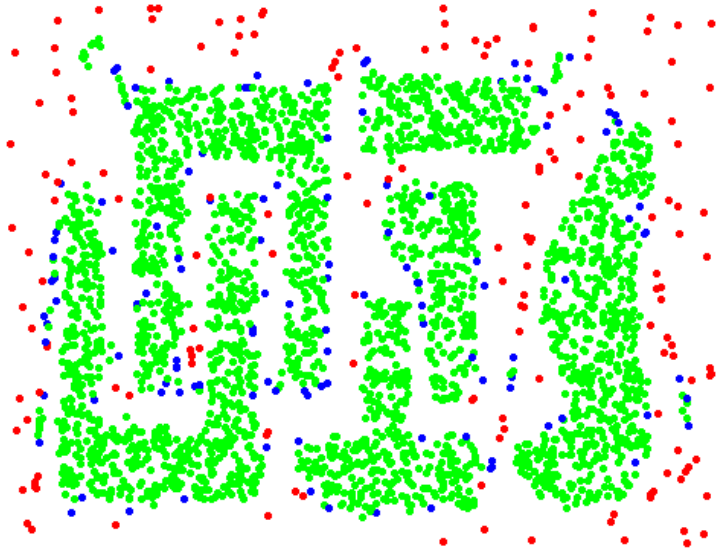
DBSCAN(D, $MinPts$, Eps)

```
(1)  mark all objects as unvisited;
(2)  do
(3)      randomly select an unvisited object p;
(4)      mark p as visited;
(5)      if the ϵ-neighborhood of p has at least MinPts objects
(6)          create a new cluster C, and add p to C;
(7)          let N be the set of objects in the ϵ-neighborhood of p;
(8)          for each point p′ in N
(9)              if p′ is unvisited
(10)                 mark p′ as visited;
(11)                 if the ϵ-neighborhood of p′ has at least MinPts points,
                     add those points to N;
(12)             if p′ is not yet a member of any cluster, add p′ to C;
(13)         end for
(14)         output C;
(15)     else mark p as noise;
(16) until no object is unvisited;
```

DBSCAN: корневые, граничные точки и шум



Исходное множество точек



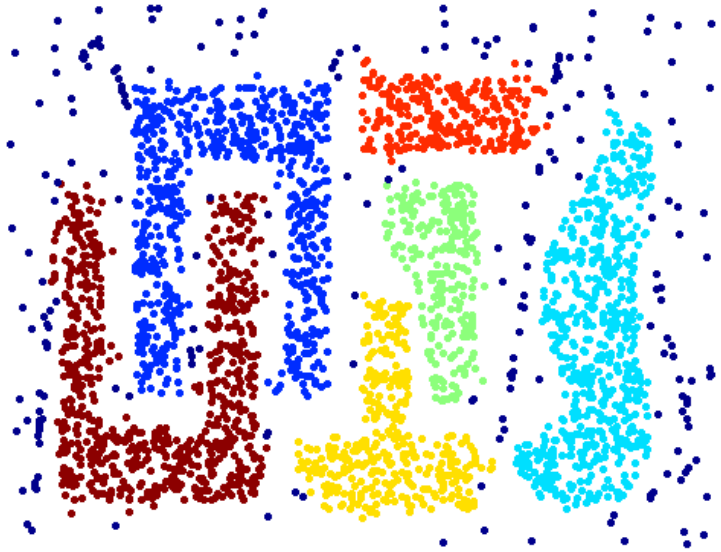
Виды точек:
корневая, **граничная** и **шум**

$$Eps = 10, MinPts = 4$$

DBSCAN: преимущества

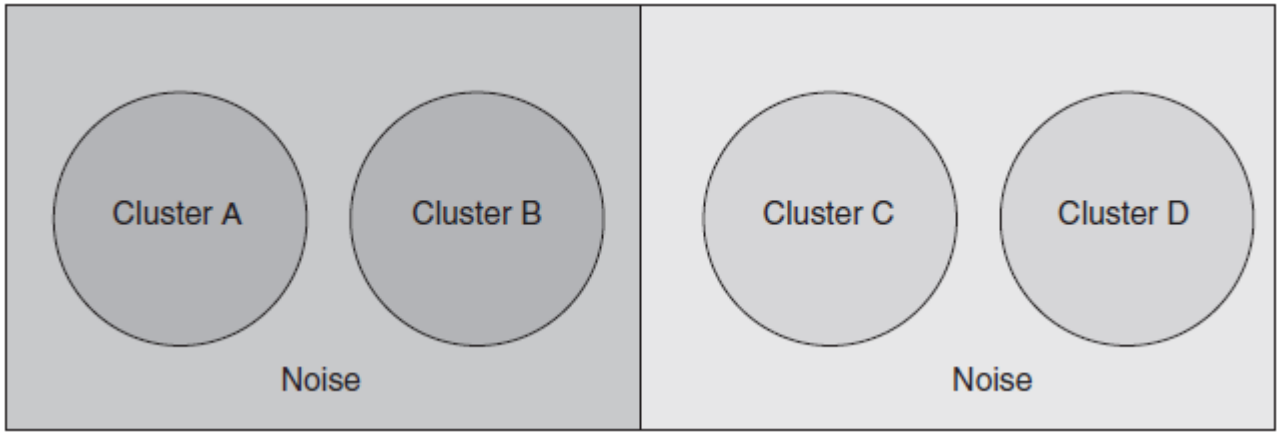
- Устойчивость к шумам
- Выявление кластеров различных форм и мощности

Исходное множество



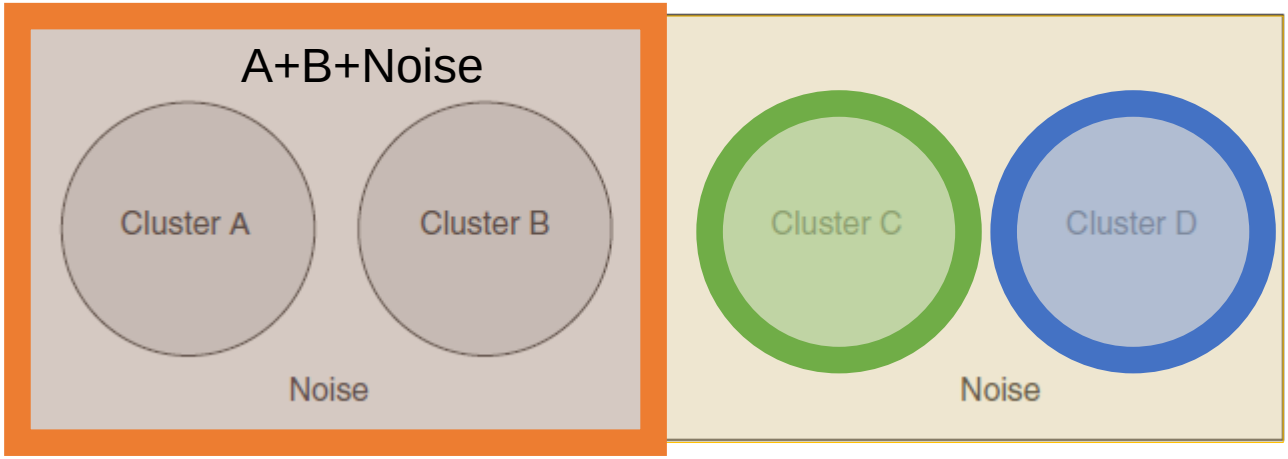
Кластеры

DBSCAN: недостатки

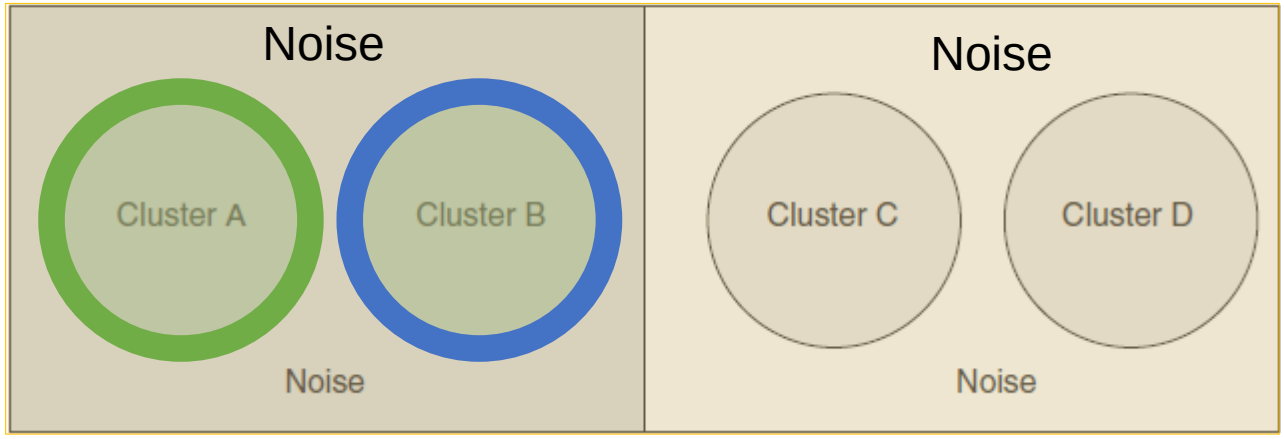


- Кластеры A, B, C, D погружены в шум. Плотность кластеров и шума отражается их темнотой
- Шум вокруг пары более плотных кластеров, A и B, имеет ту же плотность, что и кластеры C и D

DBSCAN: недостатки



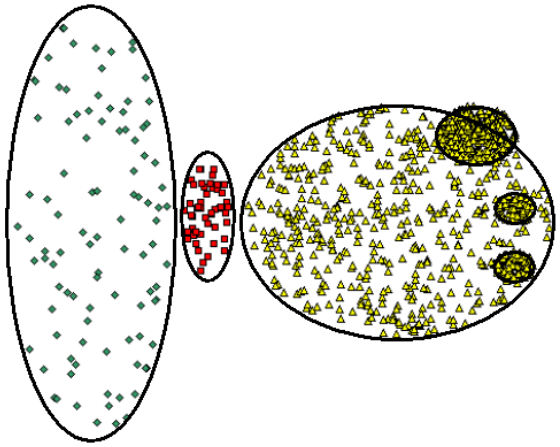
$MinPts = x,$
$Eps = y$



$MinPts = x,$
$Eps = z$

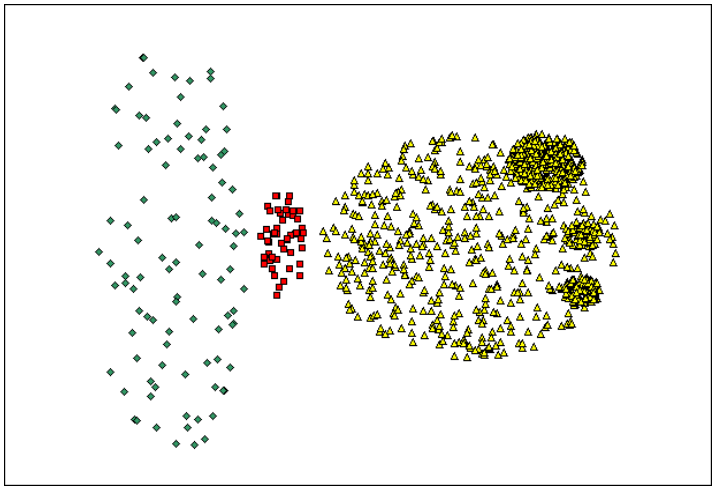
DBSCAN: недостатки

- Варьирующаяся плотность

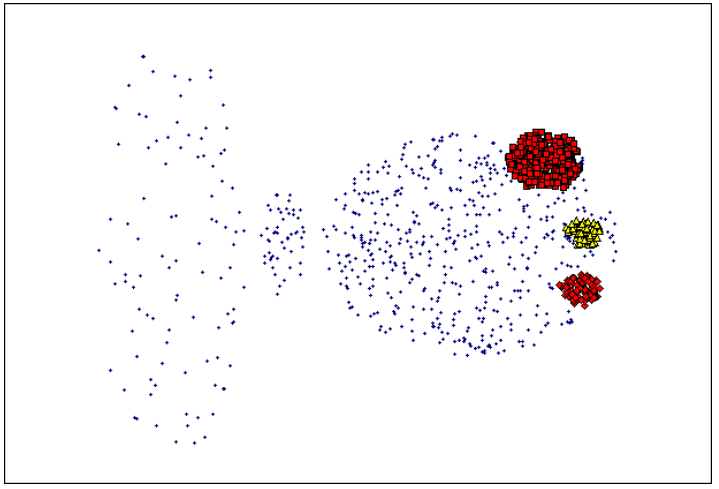


Исходное множество

$MinPts = 4,$
$Eps = 9.75$



$MinPts = 4,$
$Eps = 9.92$

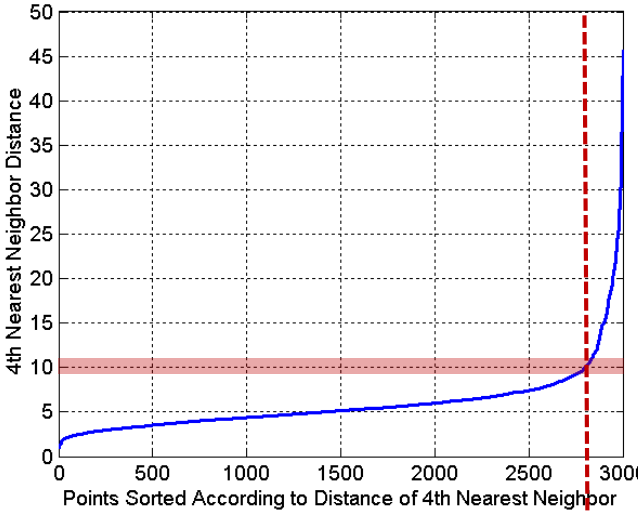


- Затраты на вычисление расстояний для точек большой размерности

DBSCAN: выбор параметров $MinPts$ и Eps

- Интуитивный выбор $MinPts$:
  - $3 \leq MinPts \leq 9$, типичное значение: 4
  - Чем больше неоднородность или/и шумов в данных, тем больше $MinPts$
- Выбор Eps на основе расстояния до $MinPts$-го ближайшего соседа
  - Для всех значений-кандидатов $MinPts$
    - для каждой точки вычислить расстояние до его $MinPts$-го ближайшего соседа
    - упорядочить точки по возрастанию расстояния, нанести на график
  - Среди всех $MinPts$ выбрать такое значение, при котором происходит наиболее сильный перегиб графика
  - Выбрать Eps из интервала, в котором происходит наиболее сильный перегиб

$MinPts = 4, 9 \leq Eps \leq 11$



корневые и граничные точки выбросы

Литература

- Han J., Kamber M., Pei J. Data Mining: Concepts and Techniques. 3rd Edition. Morgan Kaufmann, 2011. 740 p. ISBN 978-0123814791
  - 10.4 Density-Based Methods, pp. 471-479
- Tan P.-N., Steinbach M., Karpatne A., Kumar V. Introduction to Data Mining. 2nd Edition. Pearson, 2019. 839 p. ISBN: 978-0-13-312890-1
  - 7.4 DBSCAN, pp. 565-570