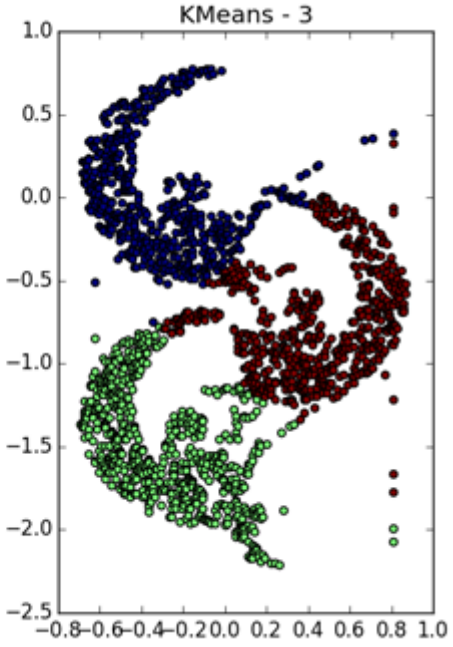
Задача кластеризации данных



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

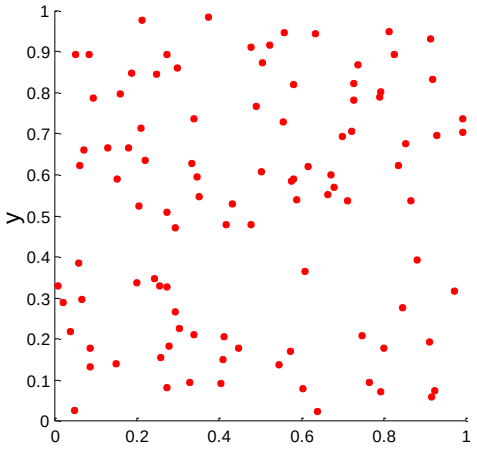
Франклин Рузвельт

Содержание

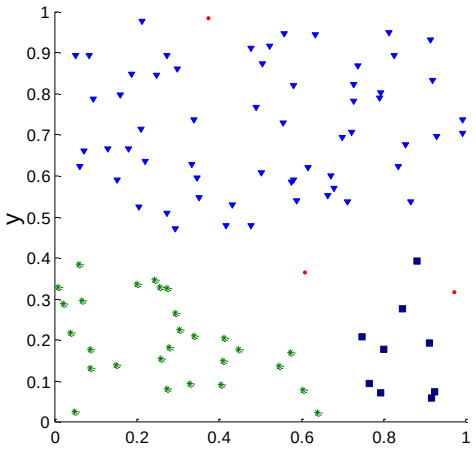
- Разделительная кластеризация
- Иерархическая кластеризация
- Плотностная кластеризация
- Нечеткая кластеризация
- **Оценка качества кластеризации**

Оценка неслучайности кластеризуемых данных

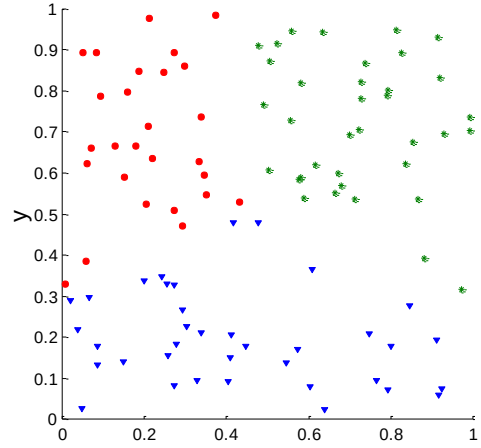
Случайные точки



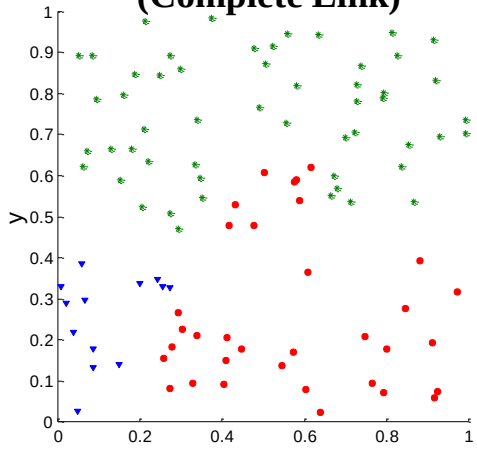
DBSCAN



K-means



Иерарх. кластеризация (Complete Link)



- Кластеризация даст осмысленные результаты, если исходные данные имеют неслучайную структуру
- Насколько близки исходные данные к нормальному распределению?

Критерий Хопкинса

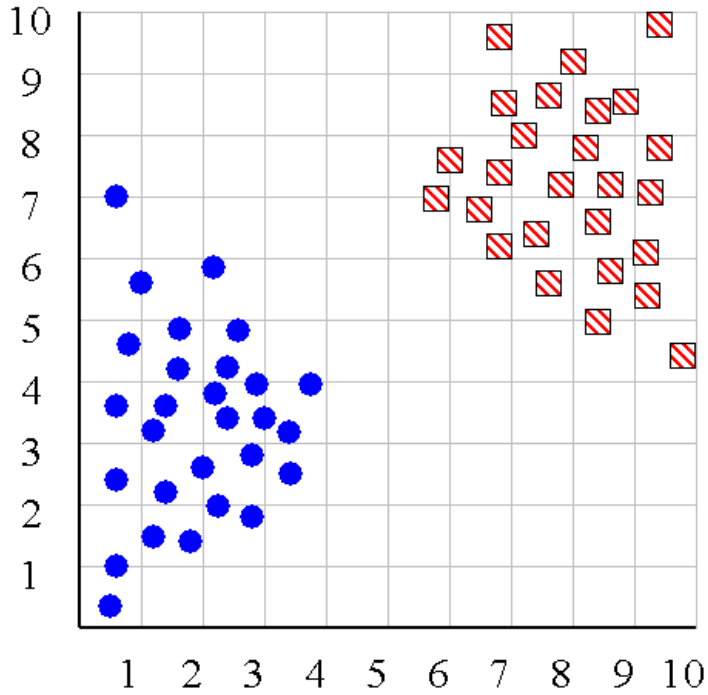
- Равномерные случайные выборки из D: $p_1, \dots, p_n$ и $q_1, \dots, q_n$
  - $x_i = \min\limits_{v \in D} dist(p_i, v)$
  - $y_i = \min\limits_{v \in D} dist(q_i, v)$
- Число Хопкинса

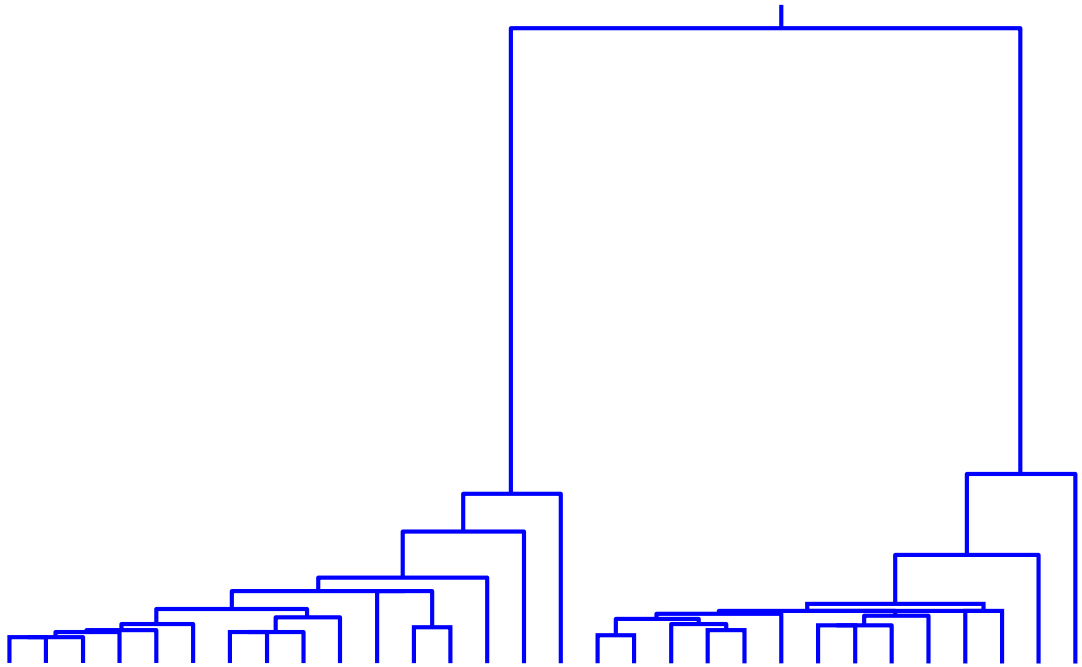
$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} x_i}$$

- Критерий
  - Если D равномерно распределено, то $\sum_{i=1}^{n} y_i \approx \sum_{i=1}^{n} x_i$, т.е. $H \approx 0.5$
  - Если в D существенные перекосы, то $\sum_{i=1}^{n} y_i \ll \sum_{i=1}^{n} x_i$, т.е. $H \approx 0$
- Повторить выборку и вычисления, на основе медианы/среднего сделать вывод о тенденции к группированию в данных:
  - если $H \geq 0.5$, то кластеры не будут осмысленными
  - если $H \leq 0.25$, то кластеры будут осмысленными

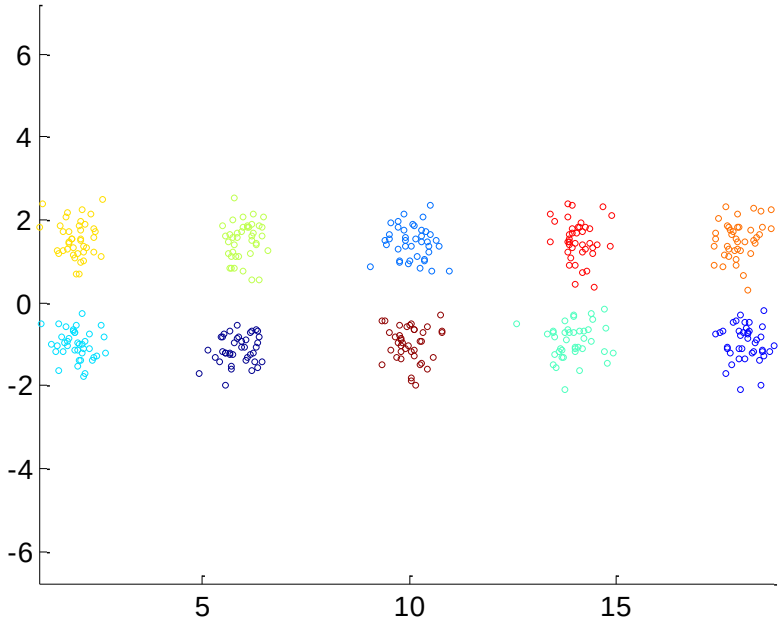
Оптимальное число кластеров: эмпирические методы

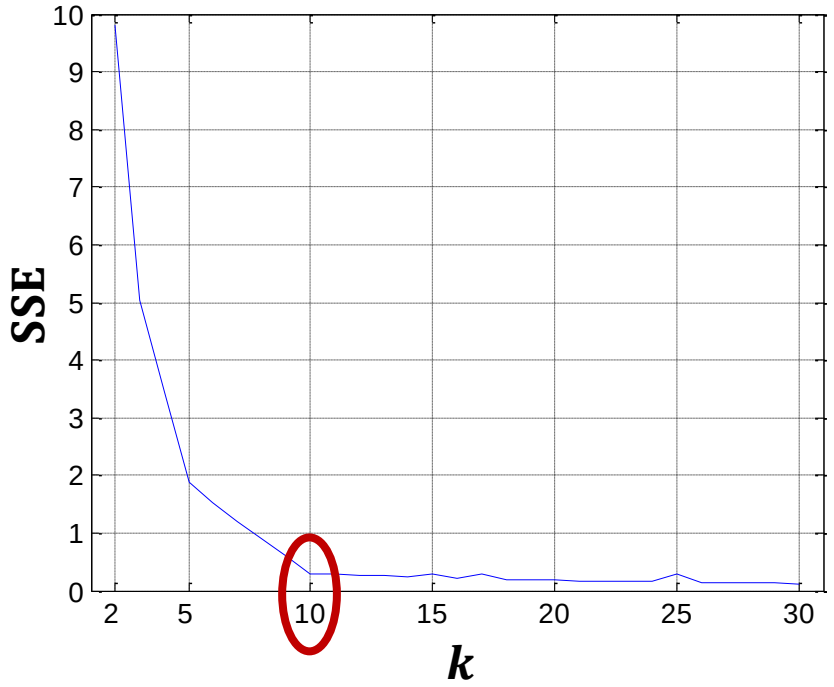
- $k = \sqrt{n/2}$ (надеемся, что в каждом кластере $\sqrt{2n}$ объектов)
- Визуализация иерархической кластеризации





Оптимальное число кластеров: метод локтя





$$\mathrm{SSE}(D) = \sum_{i=1}^{k} \sum_{j=1}^{|C_i|} dist^2(c_i, x_j)$$

$D = \{x_j\}$ – исходное множество, c_i – центроид кластера C_i

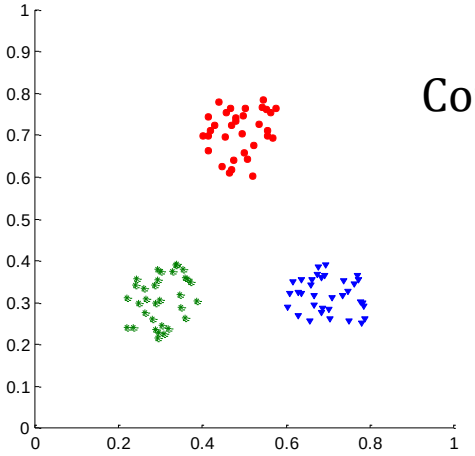
Кросс-валидация для оптимального числа кластеров

- Разбить исходное множество на m частей
- Для различных k
  - Выполнить кластеризацию $m - 1$ частей
  - Для m-й части вычислить $\text{SSE} = \Sigma dist^2(c, x)$, где x – объект этой части, c – ближайший к нему центроид
  - Повторить кластеризацию и вычисление SSE для всех частей
  - Вычислить среднее SSE
- Взять значение k, при котором среднее значение SSE минимально

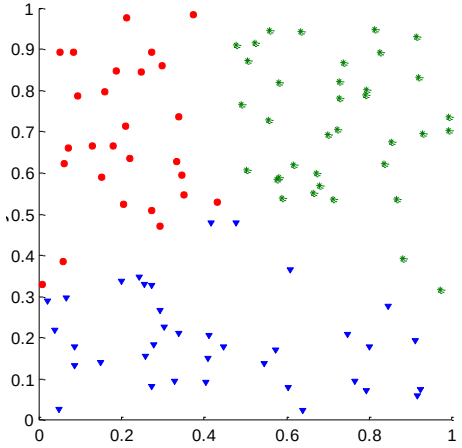
Оценка качества кластеризации через корреляцию

- Матрица расстояний:
  - $(d_{ij}) = dist(x_i, x_j)$
- Матрица идеальной схожести:
  - $(s_{ij}) = \begin{cases} 0, & x_i \in C_p \wedge x_j \in C_q \\ 1, & x_i, x_j \in C_p \end{cases}$
- Вычислить $corr_{d,s}$
  - Высокая корреляция покажет, что объекты одного кластера близки друг к другу

Corr = -0.9235



Corr = -0.5810



$$\text{Corr}_{xy} = \frac{\Sigma(x_i - \bar{x}) \times (y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \times \Sigma(y_i - \bar{y})^2}}$$

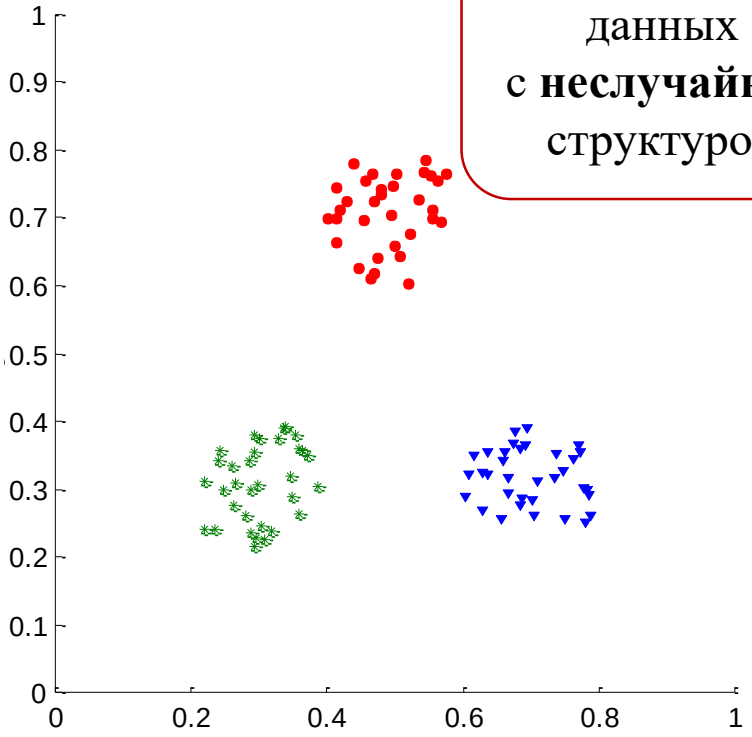
Шкала Чеддока силы корреляции

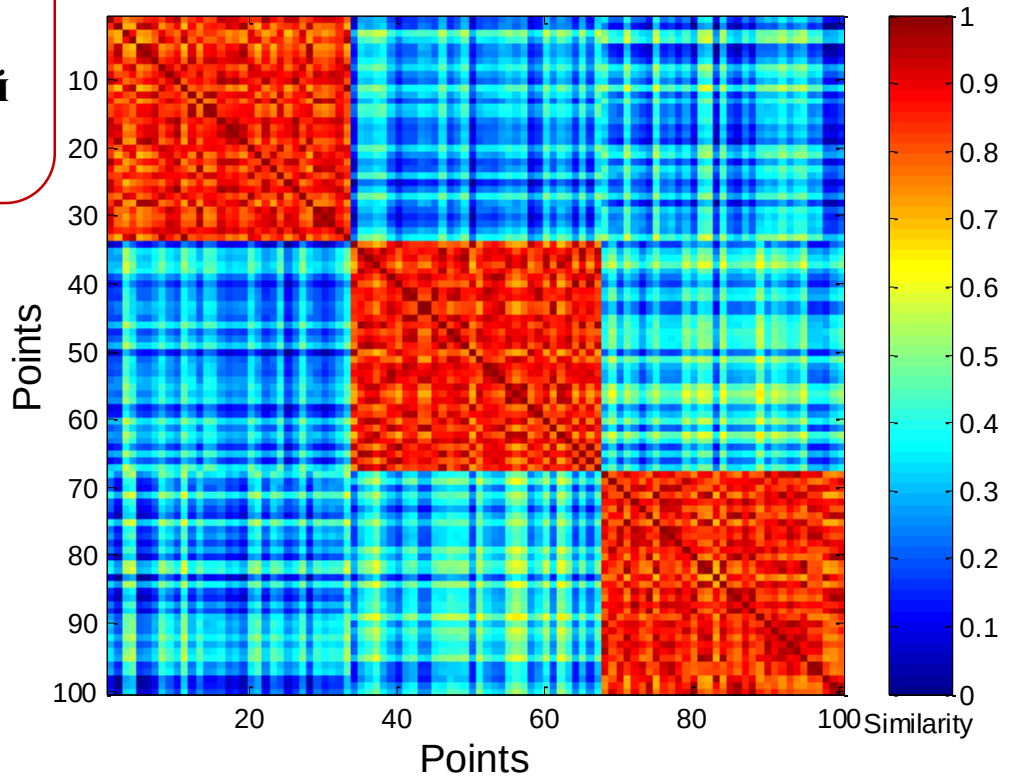
+1.0	идеальная положительная
+0.9	очень сильная положительная
+0.7	сильная положительная
+0.4	умеренная положительная
+0.1	слабая положительная
0.0	отсутствие корреляции
−0.1	слабая отрицательная
−0.4	умеренная отрицательная
−0.7	сильная отрицательная
−0.9	очень сильная отрицательная
−1.0	идеальная отрицательная

Визуальная оценка качества кластеризации по матрице расстояний (схожести)

- Упорядочить объекты кластеризованного множества по меткам кластеров и визуализировать матрицу расстояний (схожести)

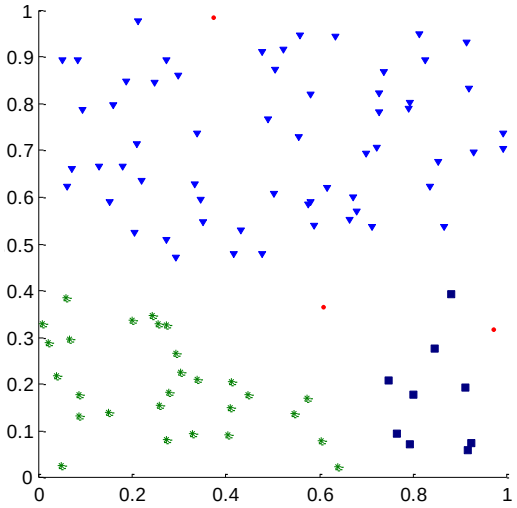
Кластеризация данных **с неслучайной** структурой



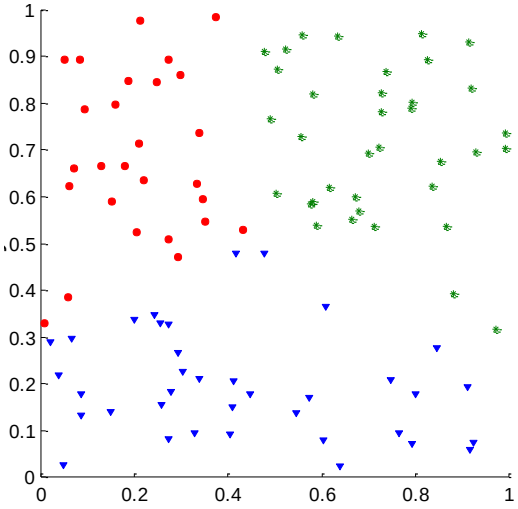


Визуальная оценка качества кластеризации по матрице расстояний (схожести)

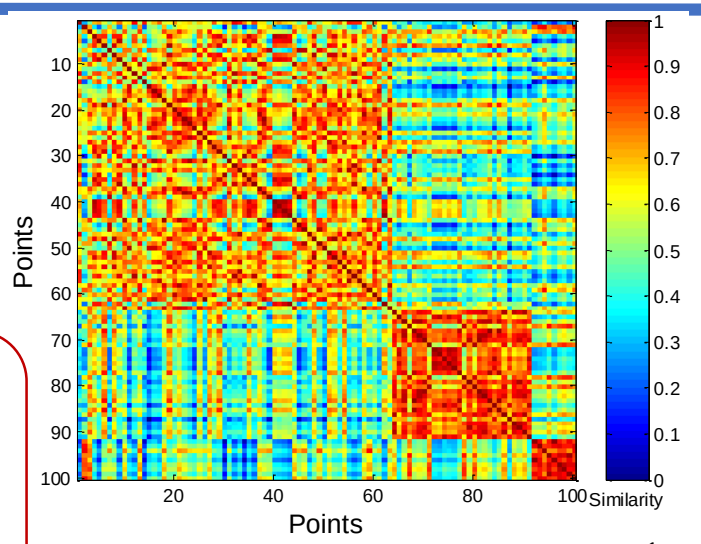
DBSCAN

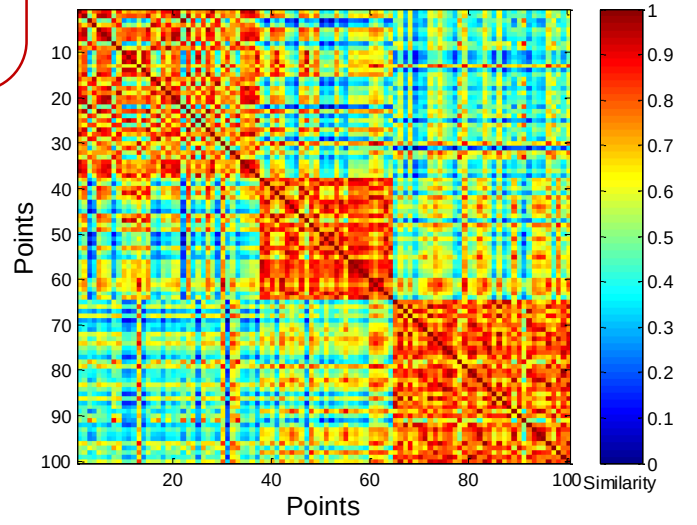


k-Means



Кластеризация данных со случайной структурой

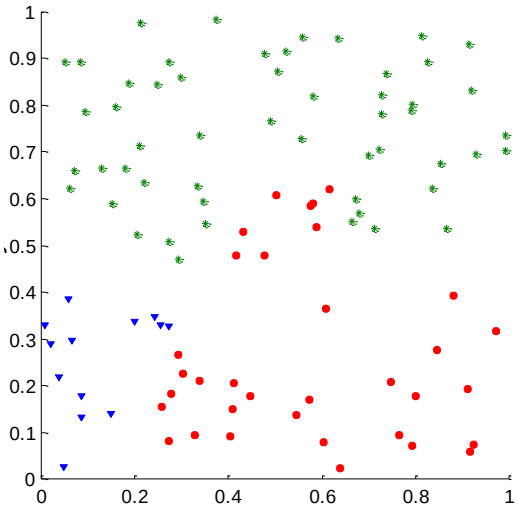


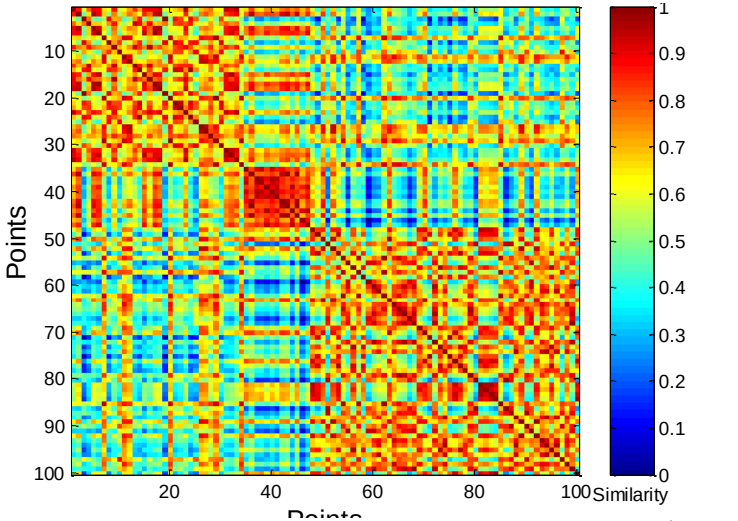


Визуальная оценка качества кластеризации по матрице расстояний (схожести)

Иерарх. класт-я (Complete linkage)

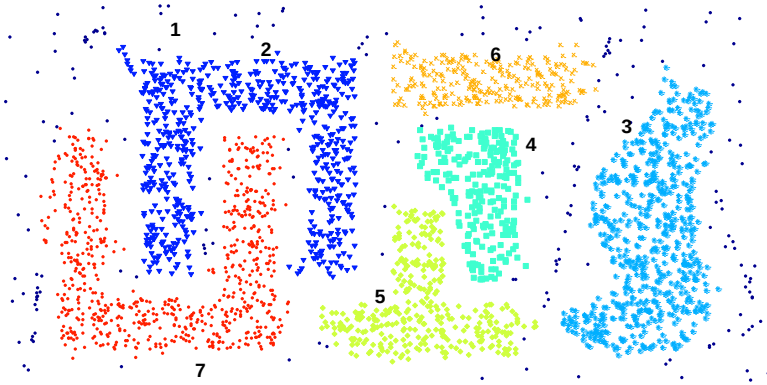
Кластеризация данных со случайной структурой

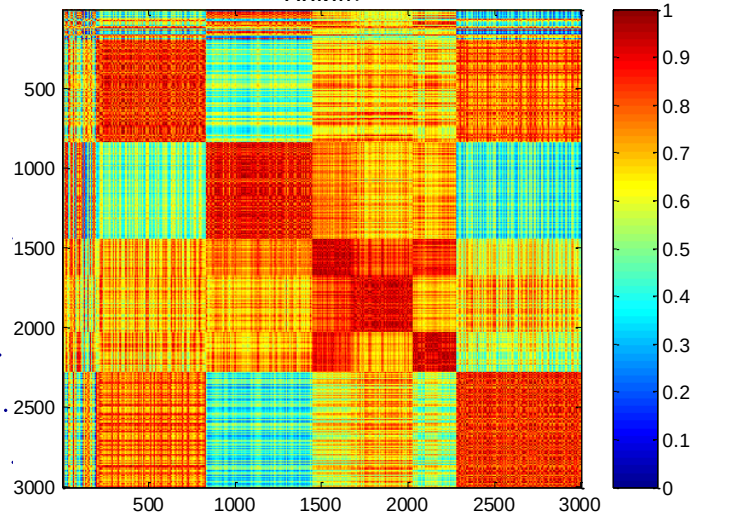




DBSCAN

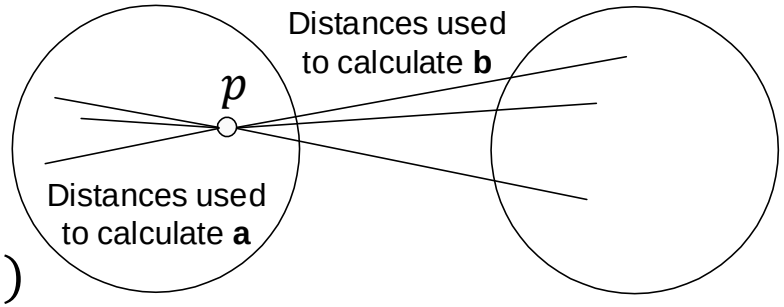
Кластеризация данных с **неслучайной** структурой





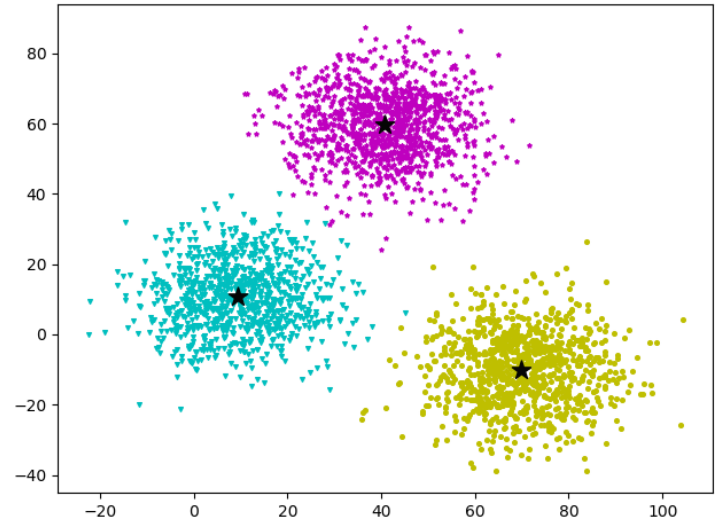
Силуэтный коэффициент

- ***Сцепление*** точки с другими точками того же кластера:
  среднее расстояние от данной точки до других точек кластера
  - $a(p) = \frac{1}{|C_i|-1} \sum_{q \in C_i \wedge q \neq p} dist(p,q)$
- ***Отдаленность*** точки от точек других кластеров:
  минимум среднего расстояния до точек других кластеров
  - $b(p) = \min\limits_{C_j \neq C_i} \frac{1}{|C_j|} \sum_{q \in C_j} dist(p,q)$
- Силуэтный коэффициент
  - для одной точки: $s(p) = \frac{b(p)-a(p)}{\max\{a(p),b(p)\}}$
  - для множества точек: $S(D) = \frac{1}{|D|} \sum_{p \in D} s(p)$

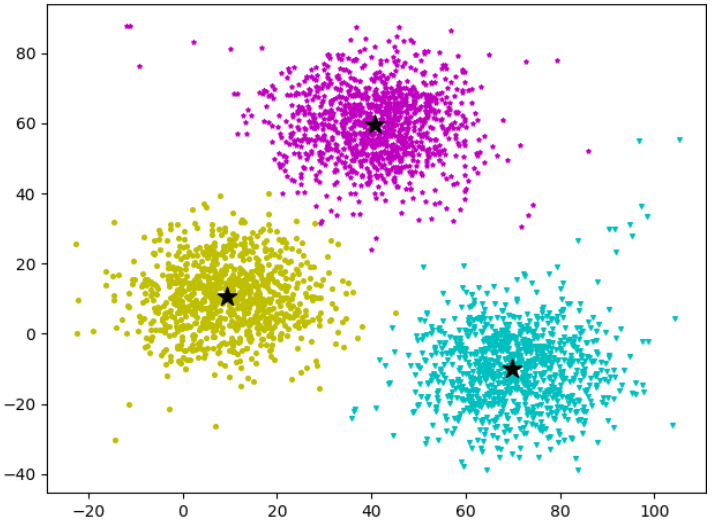


- $-1 \leq s, S \leq 1$
  - Чем ближе к 1, тем выше качество кластеризации
  - При отрицательном значении качество кластеризации низкое

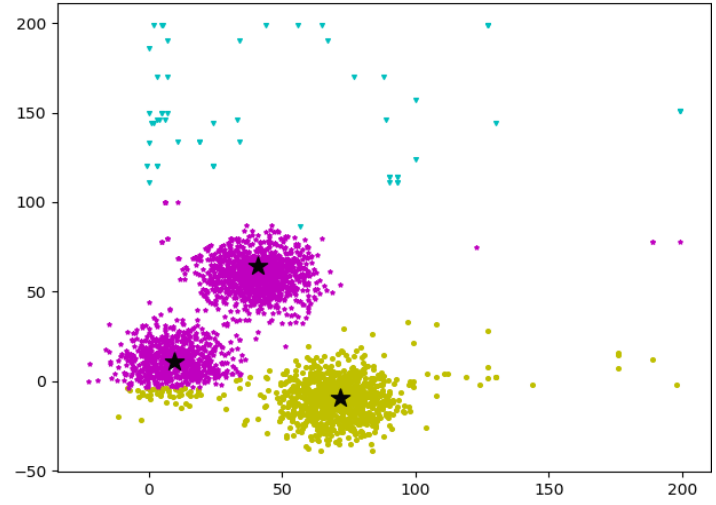
Пример: вычисление силуэтного коэффициента



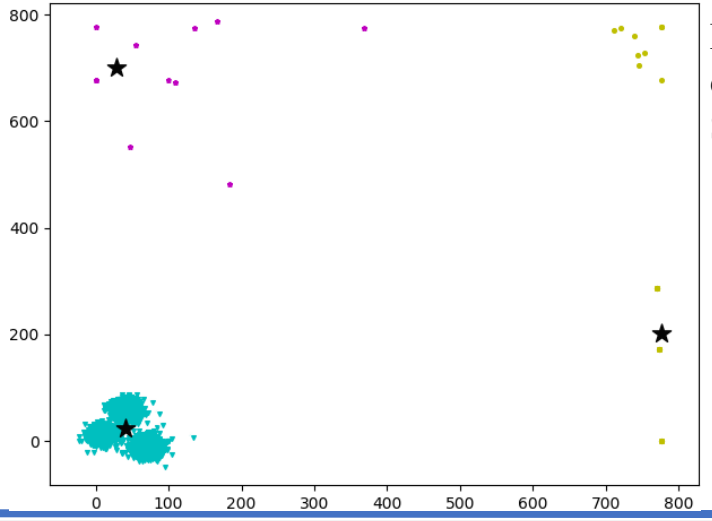
В данных
нет шума
S=0.88



В 3% данных
есть шум
S=0.62



В 5% данных
есть шум
S=0.39



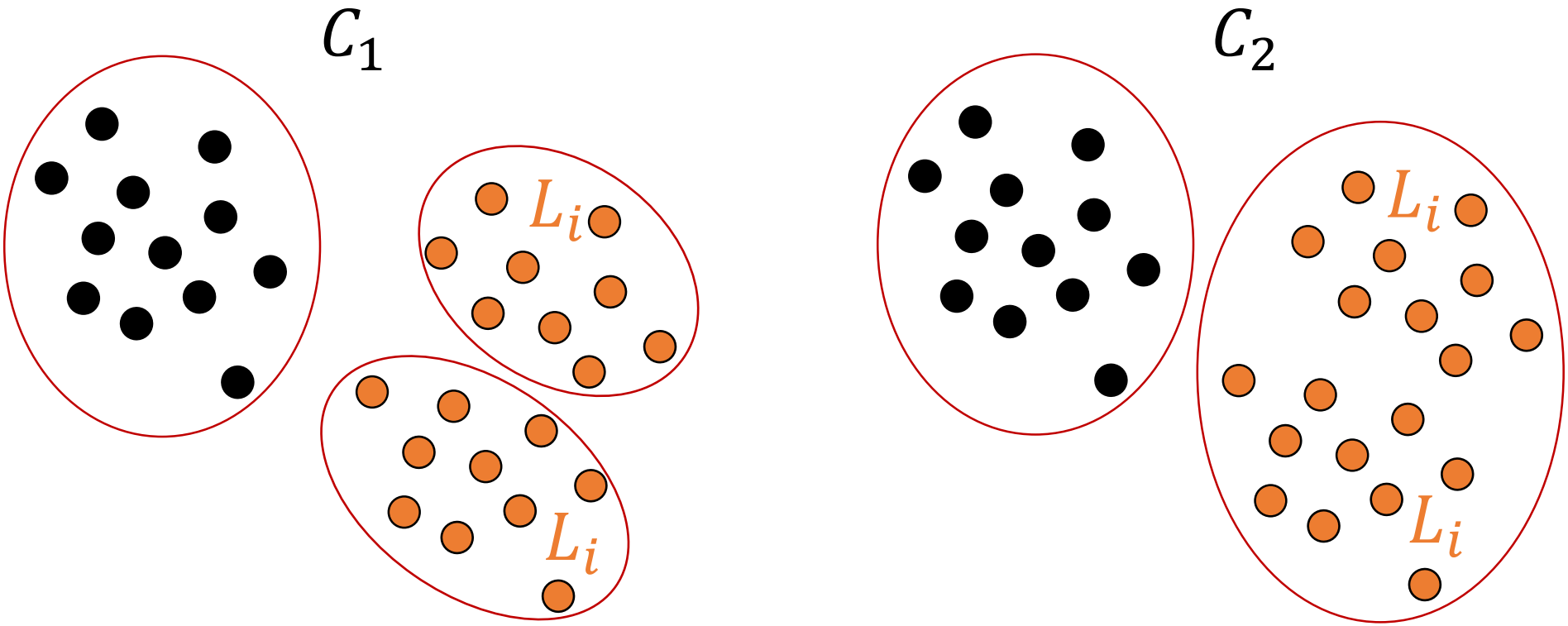
В 10% данных
есть шум
S=-0.05

Оценка качества кластеризации на основе предварительной классификации

- Результат кластеризации множества объектов:
  - C
- Результат классификации множества объектов:
  - C_g (ground truth)
  - Классы: $L_1, \ldots, L_n$
- Мера качества кластеризации:
  - $Q(C, C_g)$

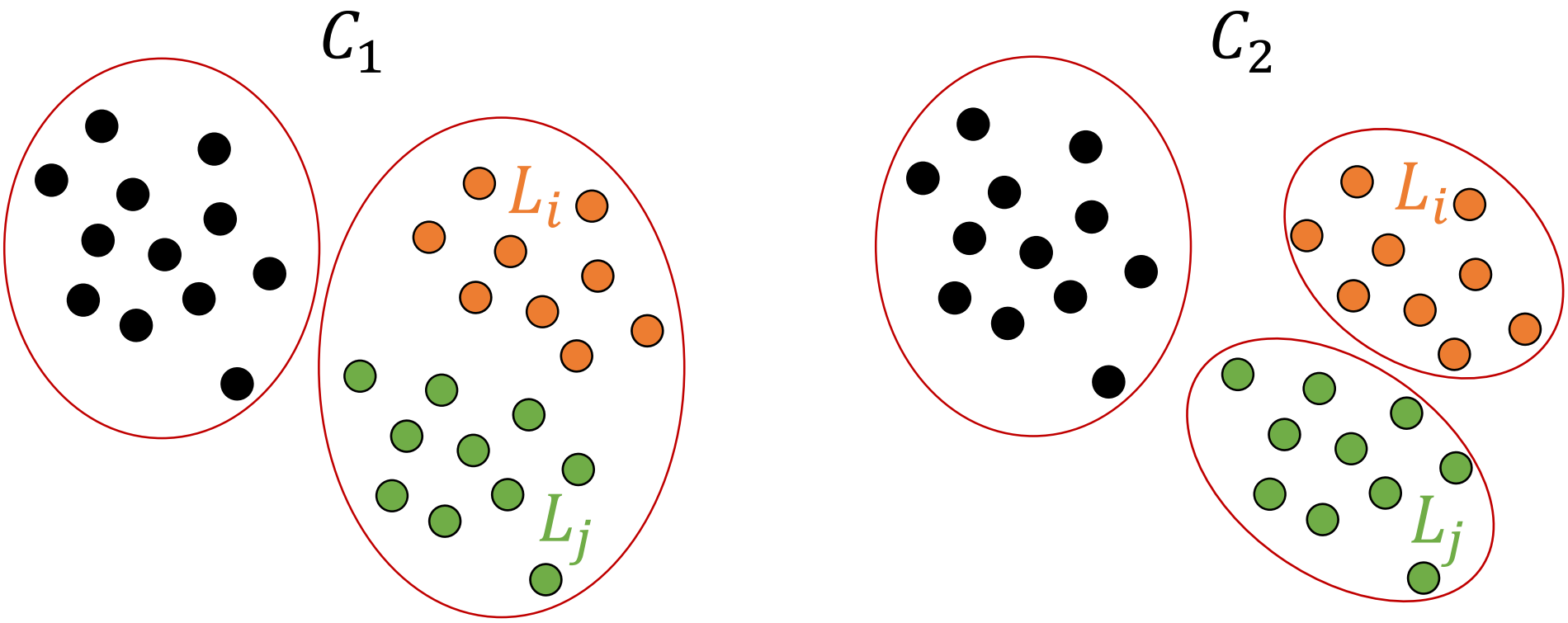
Однородность кластеров

- $Q(C_1, C_g) < Q(C_2, C_g)$



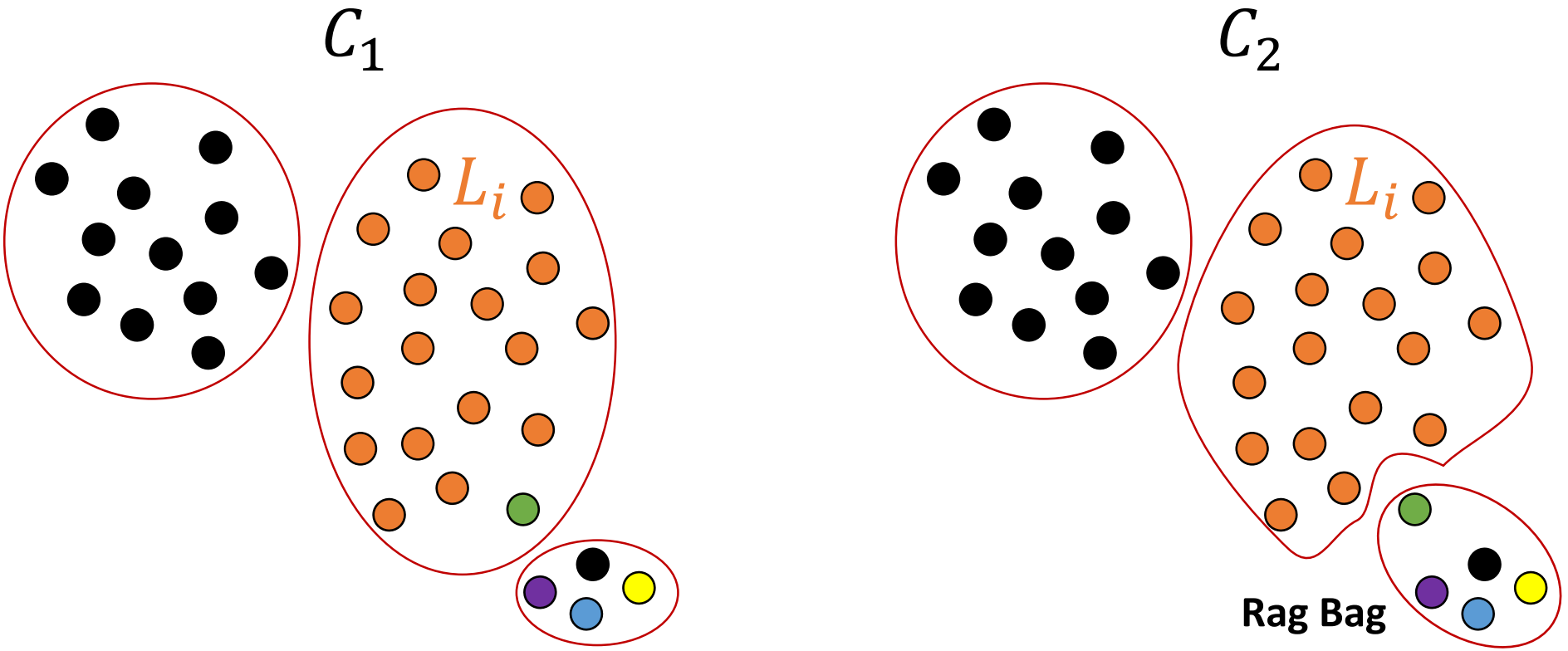
Полнота кластеров

- $Q(C_1, C_g) < Q(C_2, C_g)$



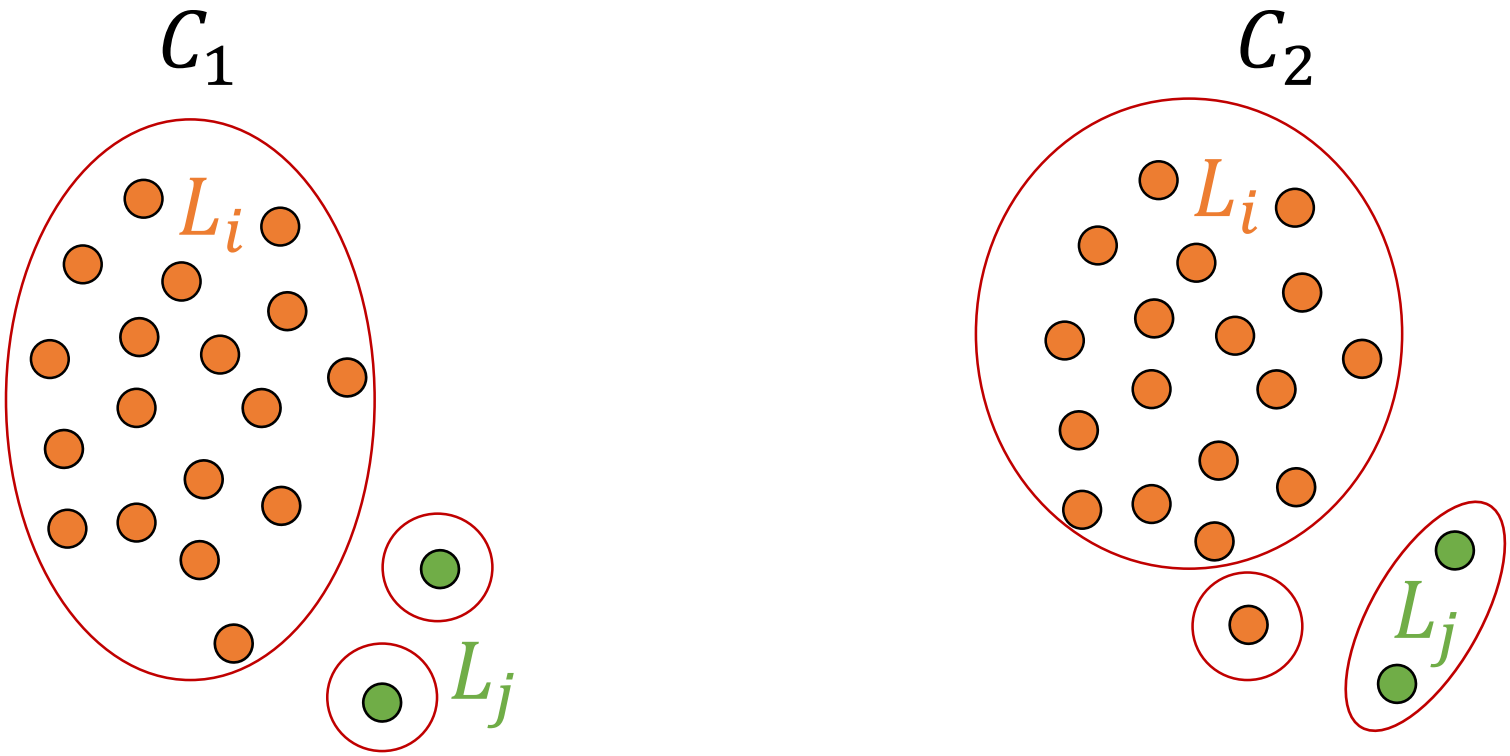
Согласованность объектов в кластерах

- $Q(C_1, C_g) < Q(C_2, C_g)$



Сохранение маломощных кластеров

- $Q(C_1, C_g) < Q(C_2, C_g)$



Меры точности и полноты кластеризации

- Кластеризуемые объекты: $D = \{x_1, \dots, x_n\}$
- Классы объектов: $L(x_i)$
- Кластеры объектов: $C(x_i)$
- Корректность кластеризации пары объектов:

$$Correctness(x_i, x_j) = \begin{cases} 1, L(x_i) = L(x_j) \Leftrightarrow C(x_i) = C(x_j) \\ 0, \text{иначе} \end{cases}$$

Точность и полнота кластеризации

- $BCubed\ precision = \frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{\{x_j \in D | i \neq j \wedge C(x_i)=C(x_j)\}} Correctness(x_i,x_j)}{|\{x_j \in D \mid i \neq j \wedge C(x_i)=C(x_j)\}|}$

- $BCubed\ recall\ = \frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{\{x_j \in D | i \neq j \wedge L(x_i)=L(x_j)\}} Correctness(x_i,x_j)}{|\{x_j \in D \mid i \neq j \wedge L(x_i)=L(x_j)\}|}$

Литература

- Han J., Kamber M., Pei J. Data Mining: Concepts and Techniques. 3rd Edition. Morgan Kaufmann, 2012. 740 p. ISBN 978-0123814791
  - 10.6 Evaluation of Clustering, pp. 483-490
- Tan P.-N., Steinbach M., Karpatne A., Kumar V. Introduction to Data Mining. 2nd Edition. Pearson, 2019. 839 p. ISBN-13: 978-0-13-312890-1
  - 7.5 Cluster Evaluation, pp. 571-612