

Машинное обучение

Лекция 3. Методы кластеризации

Катя Тузова

Быстрый поиск ближайшего соседа

k-d дерево

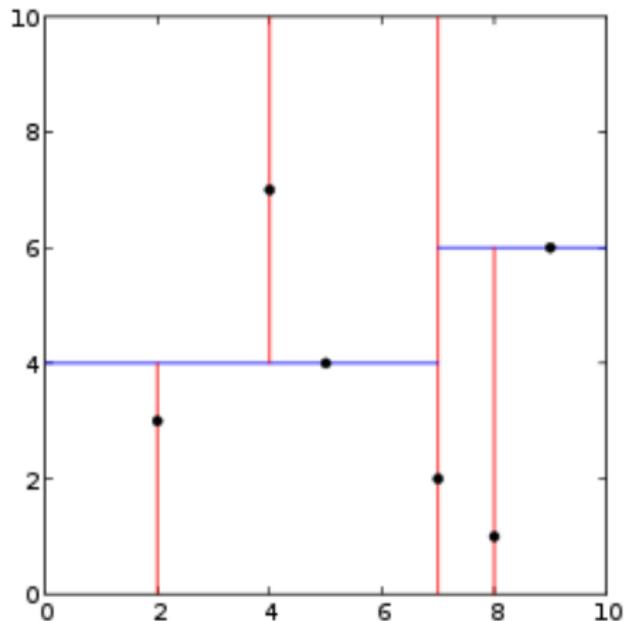
Идея: разложим множество по поторому будем искать в бинарное дерево с простыми условиями и конкретными точками в узлах.

1. По циклу, или рандомно выбираем ось.
2. Ищем точку, разбивающую множество на как можно более равные части.
3. Повторяем 1-2 для каждого из получившихся подмножеств

Сложность построения: $O(n \log n)$

Сложность поиска: в лучшем случае $O(\log n)$, в худшем — $O(n)$

2-d дерево



k-d дерево. Особенности

- + Один из наиболее простых методов
- Работает только при малом количестве параметров
- Затратный алгоритм перестроения

Locality Sensitive Hash

R -соседи – соседи в радиусе R от объекта.

Хэш-функция $h(R, cR, p_1, p_2)$:

$$\|u - v\| \leq R \Rightarrow p(h(u) = h(v)) \geq p_1$$

$$\|u - v\| \geq cR \Rightarrow p(h(u) = h(v)) \leq p_2$$

Постановка задачи кластеризации

Кластеризация – задача разделения объектов одной природы на несколько групп так, чтобы объекты в одной группе обладали одним и тем же свойством.

Кластеризация – это обучение без учителя.

Постановка задачи кластеризации

X – пространство объектов

$X^l = \{x\}_{i=1}^l$ – обучающая выборка

$\rho : X \times X \rightarrow [0, \infty)$ – функция расстояния между объектами

Найти:

Y – множество кластеров

$a : X \rightarrow Y$ – алгоритм кластеризации

Степени свободы в постановке задачи

Степени свободы в постановке задачи

- Критерий качества кластеризации
- Число кластеров неизвестно заранее
- Результат кластеризации существенно зависит от метрики

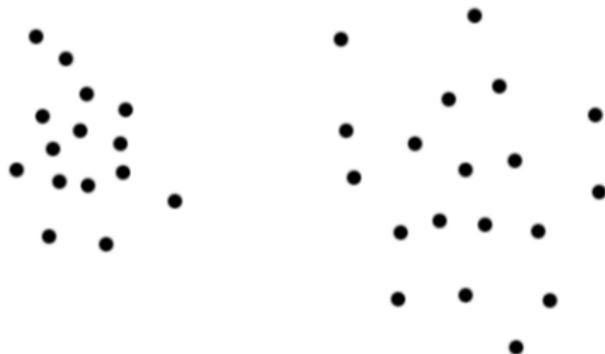
Цели кластеризации

Цели кластеризации

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

Какие бывают кластеры?

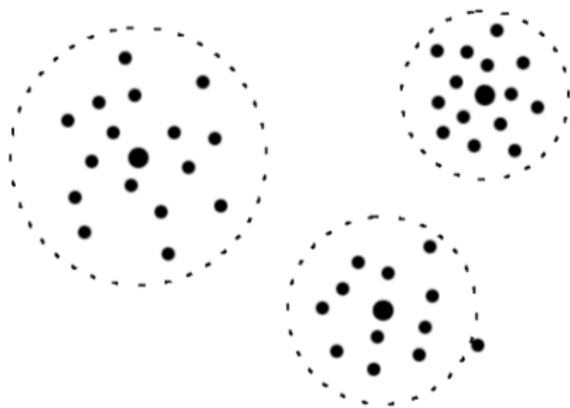
Типы кластерных структур. Сгущения



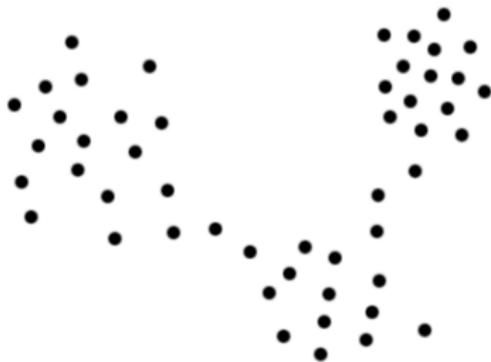
Типы кластерных структур. Ленты



Типы кластерных структур. С центром



Типы кластерных структур. С перемычками



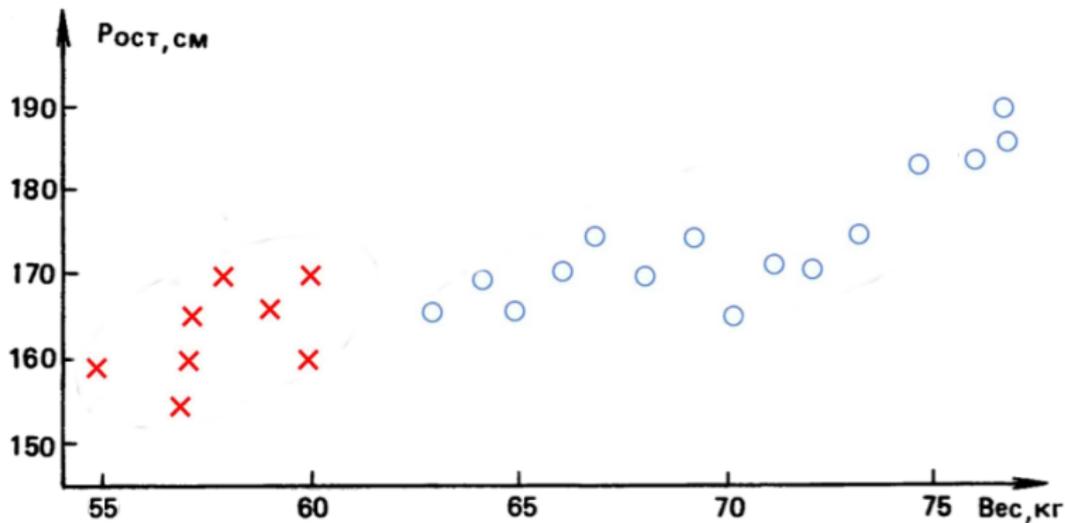
Типы кластерных структур. На фоне



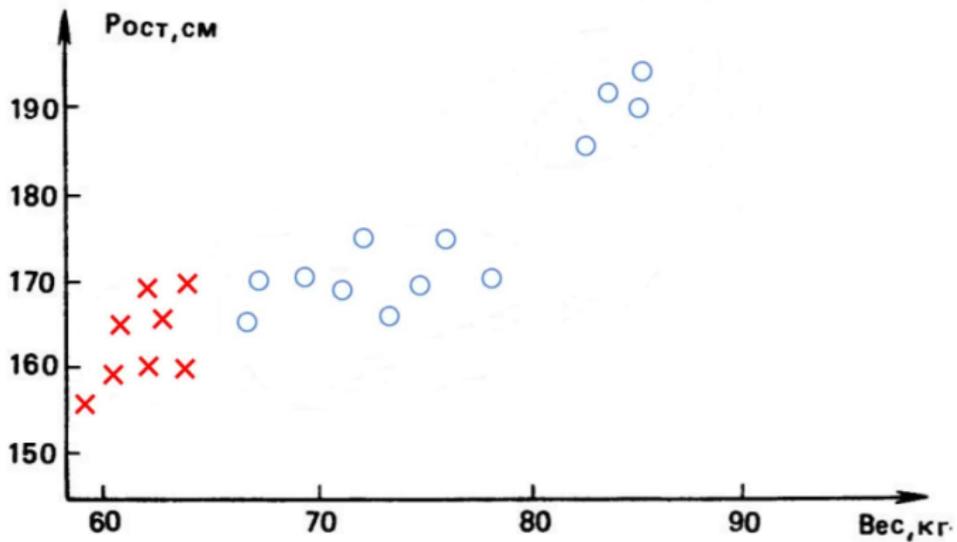
Типы кластерных структур. Перекрывающиеся



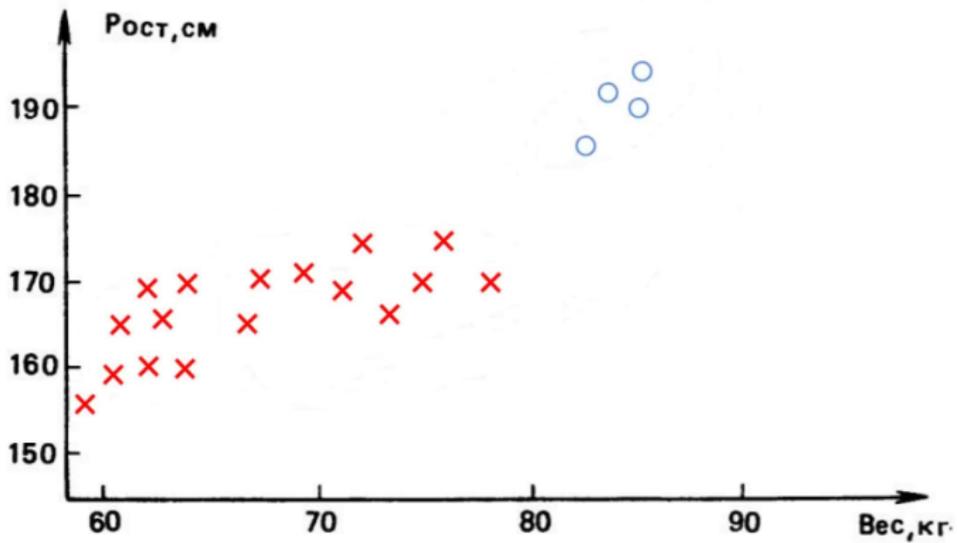
Чувствительность к выбору метрики



Чувствительность к выбору метрики



Чувствительность к выбору метрики



Оценка качества кластеризации

Есть несколько разбиений на кластеры. Как их сравнить?

Оценка качества кластеризации

- Минимизировать среднее внутрикластерное расстояние

$$\frac{\sum_{a(x_i)=a(x_j)} \rho(x_i, x_j)}{\sum_{a(x_i)=a(x_j)} 1} \rightarrow \min$$

- Максимизировать среднее межкластерное расстояние

$$\frac{\sum_{a(x_i) \neq a(x_j)} \rho(x_i, x_j)}{\sum_{a(x_i) \neq a(x_j)} 1} \rightarrow \max$$

Методы кластеризации

- Иерархические
- Графовые
- Статистические

Агломеративный алгоритм Ланса-Уильямса

Идея:

- Считаем каждую точку кластером.
- Затем объединяем ближайшие точки в новый кластер.
- Повторяем.

Алгоритм Ланса-Уильямса

$$C_1 = \{\{x_1\}, \{x_2\}, \dots, \{x_l\}\}$$

for $t = 2, \dots, l$:

$$(U, V) = \arg \min_{U \neq V} R(U, V)$$

$$W = U \cup V$$

$$C_t = C_{t-1} \cup \{W\} \setminus \{U, V\}$$

foreach $S \in C_t$

 вычислить $R(W, S)$

Алгоритм Ланса-Уильямса

$$C_1 = \{\{x_1\}, \{x_2\}, \dots, \{x_l\}\}$$

for $t = 2, \dots, l$:

$$(U, V) = \arg \min_{U \neq V} R(U, V)$$

$$W = U \cup V$$

$$C_t = C_{t-1} \cup \{W\} \setminus \{U, V\}$$

foreach $S \in C_t$

 вычислить $R(W, S)$

Алгоритм Ланса-Уильямса

Чего не хватает?

Формула Ланса-Уильямса

Расстояние $R(W, S)$?

$$W = \{U \cup V\}$$

Знаем:

$$R(U, S), R(V, S), R(U, V)$$

Формула Ланса-Уильямса

Расстояние $R(W, S)$?

$$W = \{U \cup V\}$$

Знаем:

$$R(U, S), R(V, S), R(U, V)$$

$$R(U \cup V, S) = \alpha_U R(U, S) + \alpha_V R(V, S) + \\ + \beta R(U, V) + \gamma |R(U, S) - R(V, S)|$$

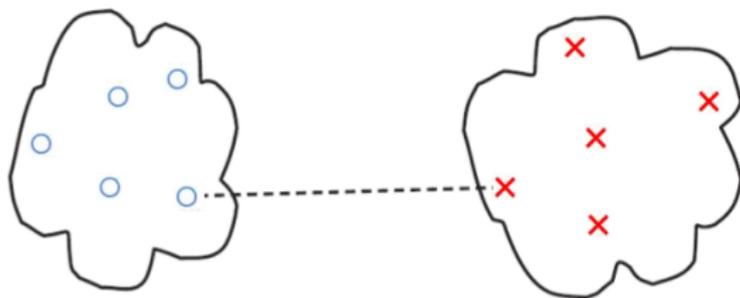
$\alpha_U, \alpha_V, \beta, \gamma$ – числовые параметры

Формула Ланса-Уильямса

Значения параметров $\alpha_U, \alpha_V, \beta, \gamma$?

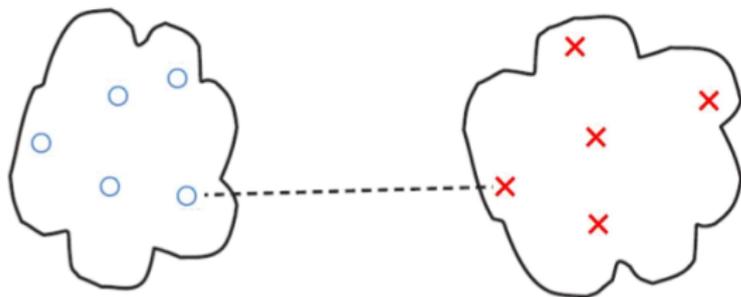
Формула Ланса-Уильямса

Расстояние ближнего соседа:



Формула Ланса-Уильямса

Расстояние ближнего соседа:



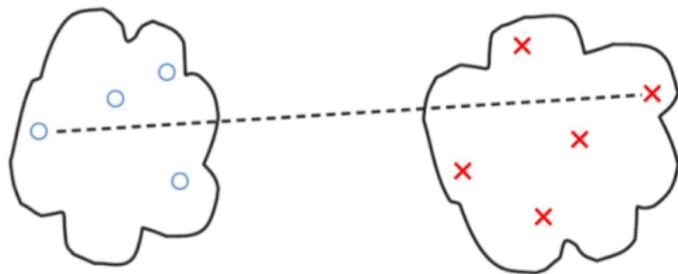
$$\alpha_U = \alpha_V = \frac{1}{2}$$

$$\beta = 0$$

$$\gamma = -\frac{1}{2}$$

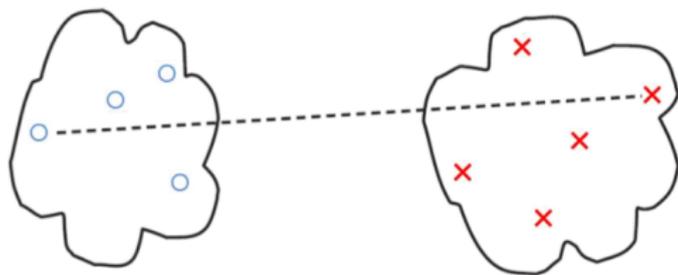
Формула Ланса-Уильямса

Расстояние дальнего соседа:



Формула Ланса-Уильямса

Расстояние дальнего соседа:



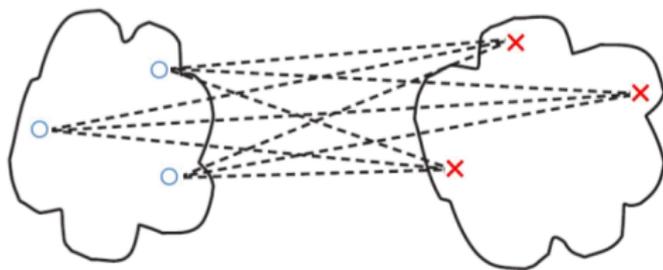
$$\alpha_U = \alpha_V = \frac{1}{2}$$

$$\beta = 0$$

$$\gamma = \frac{1}{2}$$

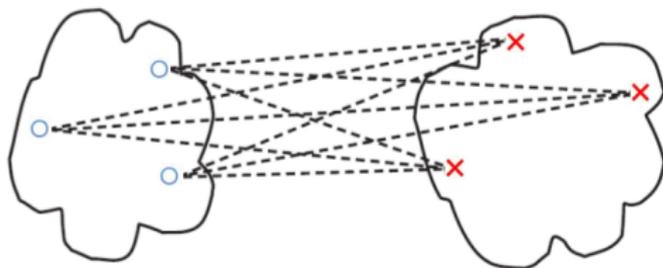
Формула Ланса-Уильямса

Групповое среднее:



Формула Ланса-Уильямса

Групповое среднее:



$$\alpha_U = \frac{|U|}{|W|}$$
$$\alpha_V = \frac{|V|}{|W|}$$
$$\beta = \gamma = 0$$

Графовые алгоритмы

Какие есть две очевидные идеи?

Графовые алгоритмы

Очевидные:

- Выделение связных компонент
- Минимальное покрывающее дерево

Выделение связанных компонент

- Рисуем полный граф с весами, равными расстоянию между объектами
- Выбираем лимит расстояния r и выкидываем все ребра длиннее r
- Компоненты связности полученного графа – наши кластеры

Выделение связанных компонент

Как искать компоненты связности?

Минимальное покрывающее дерево

Минимальное остовное дерево – дерево, содержащее все вершины графа и имеющее минимальный суммарный вес ребер.

Как найти?

Минимальное покрывающее дерево

Как использовать минимальное остовное дерево для разбиения на кластеры?

Минимальное покрывающее дерево

Строим минимальное остовное дерево, а потом выкидываем из него ребра максимального веса.

Сколько ребер выбросим – столько кластеров получим.

Алгоритм FOREL

$$U = X, C = \emptyset$$

while $U \neq \emptyset$:

 выбрать случайную точку x_0

 Повторять пока x_0 не стабилизируется:

$$c = \{x \in X \mid \rho(x, x_0) < R\}$$

$$x_0 = \frac{1}{|c|} \sum_{x \in c} x$$

$$U = U \setminus c, C = C \cup \{c\}$$

Метод k -средних

Идея:

минимизировать меру ошибки

$$E(X, C) = \sum_{i=1}^n \|x_i - \mu_i\|^2$$

μ_i – ближайший к x_i центр кластера

Метод k -средних

Инициализировать центры k кластеров

Пока c_i не перестанет меняться:

$$c_i = \arg \min_{c \in C} \rho(x_i, \mu_c) \quad i = 1, \dots, l$$

$$\mu_c = \frac{\sum_{c_i=c} f_j(x_i)}{\sum_{c_i=c} 1} \quad j = 1, \dots, n, c \in C$$

μ_c – новое положение центров кластеров

c_i – принадлежность x_i к кластеру

$\rho(x_i, \mu_c)$ – расстояние от x_i до центра кластера μ_c

Особенности метода k -средних

- Чувствительность к начальному выбору μ_c
- Необходимость задавать k

Как устранить эти недостатки?

Устранение недостатков

- Несколько случайных кластеризаций
- Постепенное наращивание числа k

На следующей лекции

- Линейные методы классификации
- Метод опорных векторов
- Выбор ядра для метода опорных векторов
- Мультиклассовый классификатор