Основы машинного обучения

Лекция 15

Градиентный бустинг. Кластеризация.

Евгений Соколов

esokolov@hse.ru

НИУ ВШЭ, 2025

Градиентный бустинг в общем виде

Задача обучения базовой модели

- Ошибка на объекте x_i при прогнозе новой модели, равном z :

$$L(y_i, a_{N-1}(x_i) + z)$$

- Посчитаем производную:

$$s_i^{(N)} = -\frac{\partial}{\partial z} L(y_i, z)\Big|_{z=a_{N-1}(x_i)}$$

Задача обучения базовой модели

- Посчитаем производную:

$$s_i^{(N)} = -\frac{\partial}{\partial z} L(y_i, z)\Big|_{z=a_{N-1}(x_i)}$$

- Знак показывает, в какую сторону сдвигать прогноз на x_i, чтобы уменьшить ошибку композиции на нём
- Величина показывает, как сильно можно уменьшить ошибку, если сдвинуть прогноз
- Если ошибка почти не сдвинется, то нет смысла что-то менять

Градиентный бустинг

- Обучение N-й модели:

$$\frac{1}{\ell}\sum_{i=1}^{\ell}\left(b_N(x_i) - s_i^{(N)}\right)^2 \to \min_{b_N(x)}$$

$s_i^{(N)} = -\frac{\partial}{\partial z} L(y_i, z)\Big|_{z=a_{N-1}(x_i)}$ — сдвиги

Резюме

- Чтобы учесть особенности функции потерь, можно посчитать её производные в точке текущего прогноза композиции
- Базовую модель будем обучать на эти производные (со знаком минус)

Гиперпараметры и регуляризация в бустинге

Градиентный бустинг

$$a_N(x) = a_{N-1}(x_i) + b_N(x_i)$$

- Обучение N-й модели:

$$\frac{1}{\ell}\sum_{i=1}^{\ell}\left(b_N(x_i) - s_i^{(N)}\right)^2 \to \min_{b_N(x)}$$

- $s_i^{(N)} = -\frac{\partial}{\partial z}L(y_i, z)\Big|_{z=a_{N-1}(x_i)}$ — сдвиги

Глубина деревьев

- Градиентный бустинг уменьшает смещение базовых моделей
- Разброс может увеличиться
- Поэтому в качестве базовых моделей стоит брать неглубокие деревья

Гиперпараметры

- Глубина базовых деревьев
- Число деревьев N

Проблемы бустинга

- Сдвиги показывают направление, в котором надо сдвинуть композицию на всех объектах обучающей выборки
- Базовые модели, как правило, очень простые
- Могут не справиться с приближением этого направления

Проблемы бустинга

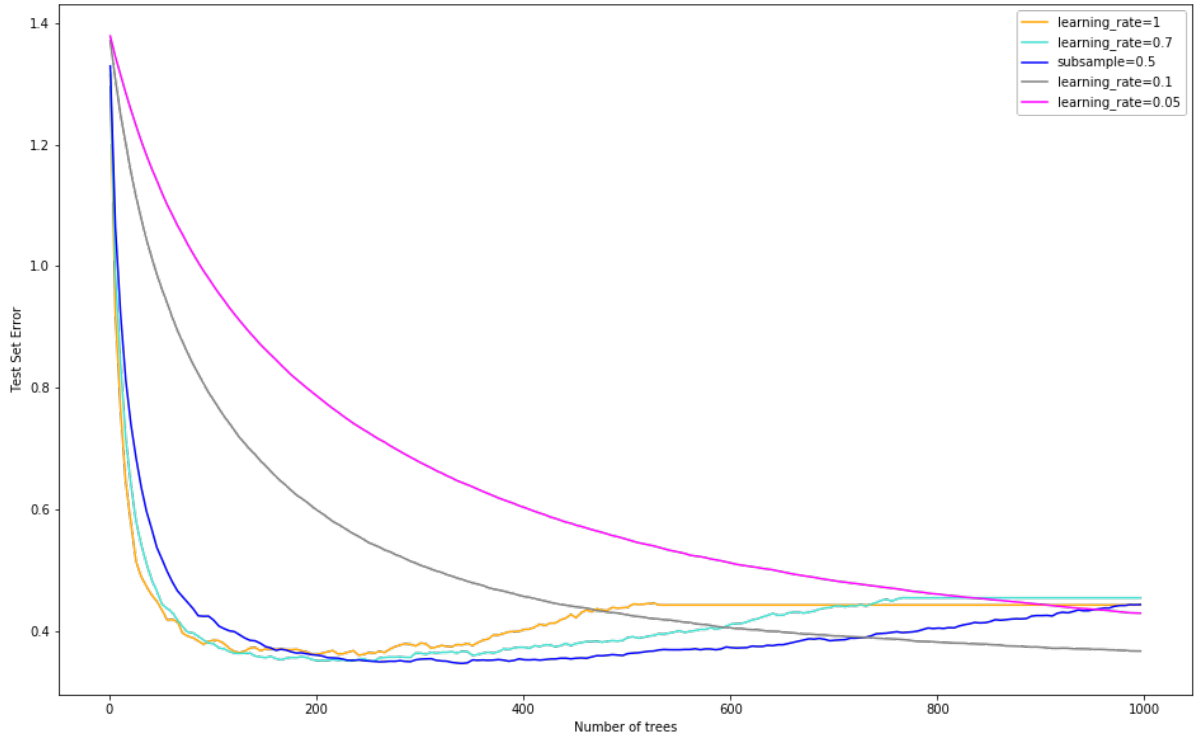
- Сдвиги показывают направление, в котором надо сдвинуть композицию на всех объектах обучающей выборки
- Базовые модели, как правило, очень простые
- Могут не справиться с приближением этого направления
- Выход: добавлять деревья в композицию с небольшим весом

Длина шага

$$a_N(x) = a_{N-1}(x_i) + \eta b_N(x_i)$$

- $\eta \in (0, 1]$ — длина шага
- Можно сказать, что это регуляризация композиции
- Снижает вклад каждой модели в композицию
- Чем меньше η, тем больше надо деревьев

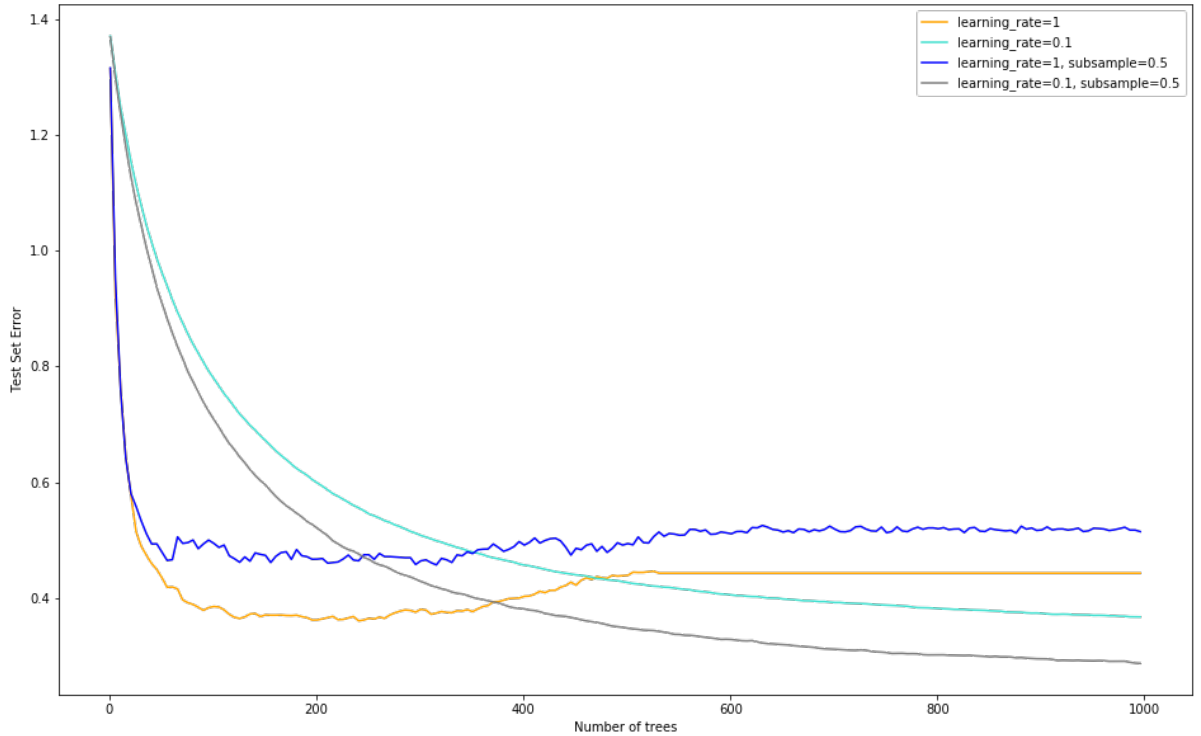
Длина шага


learning_rate=1
learning_rate=0.7
subsample=0.5
learning_rate=0.1
learning_rate=0.05
Test Set Error
Number of trees
1.4
1.2
1.0
0.8
0.6
0.4
0
200
400
600
800
1000

Рандомизация

- Можно обучать деревья на случайных подмножествах признаков
- Бустинг уменьшает смещение, поэтому итоговая композиция всё равно получится качественной
- Может снизить переобучение
- Можно обучать деревья на подмножествах объектов — способ борьбы с шумом в данных

Рандомизация



Гиперпараметры

- Глубина базовых деревьев
- Число деревьев N
- Длина шага
- Размер подвыборки для обучения
- и т.д.

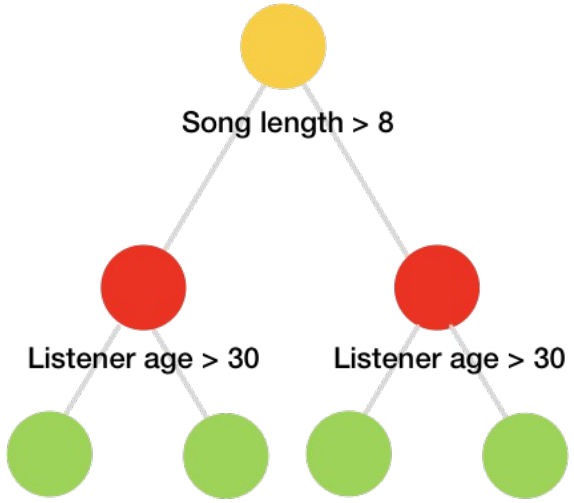
Резюме

- Чтобы снизить переобучение, можно добавлять модели в композицию с небольшими весами
- Также может помочь обучение моделей на подвыборках

Вариации бустинга

ODT

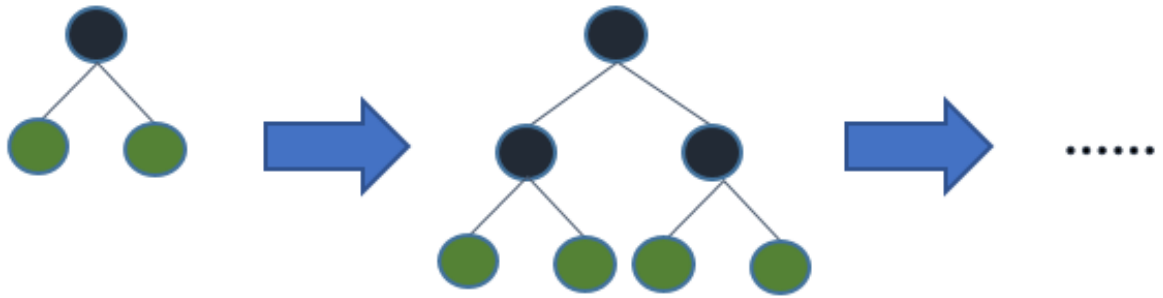
- Oblivious decision trees
- Ограничение: на одном уровне дерева используется один и тот же предикат



https://catboost.ai/

Способ построения дерева

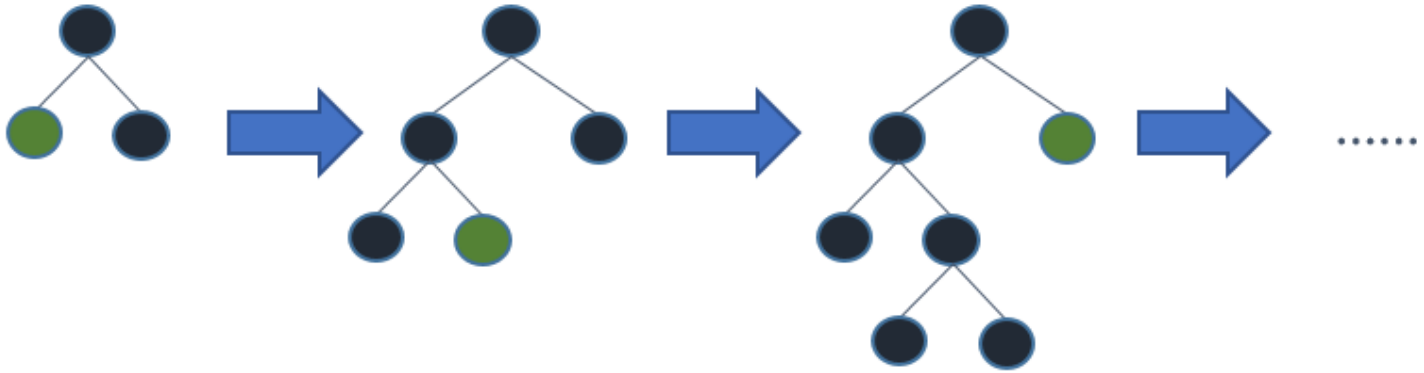
- Level-wise: дерево строится рекурсивно до тех пор, пока не достигнута максимальная глубина



https://lightgbm.readthedocs.io/

Способ построения дерева

- Level-wise: дерево строится рекурсивно до тех пор, пока не достигнута максимальная глубина
- Leaf-wise: среди текущих листьев выбирается тот, чьё разбиение сильнее всего уменьшает ошибку



Leaf-wise tree growth

https://lightgbm.readthedocs.io/

Выбор лучшего порога для предиката

- $[x_j < t]$ — как выбрать t?
- Вариант 1: перебрать все известные значения признака
- Вариант 2: построить гистограмму для признака и искать пороги среди границ на гистограмме
- Вариант 3: просемплировать объекты с близкими к нулю значениями производной

Регуляризация деревьев

- Базовая регуляризация: введение длины шага и семплирования признаков
- Штрафы за число листьев в дереве
- Штрафы за величину прогнозов в листьях дерева

Улучшенное обучение

- Мы обучаем деревья на сдвиги, ошибка измеряется с помощью MSE
- Когда дерево построено, можно подобрать оптимальные значения в листьях с точки зрения исходной функции потерь

Имплементации

- XGBoost
- LightGBM: leaf-wise growth, поиск порогов на основе производных
- CatBoost: ODT

Кластеризация

На прошлых лекциях

- Методы обучения с учителем: линейные модели, решающие деревья, случайные леса, …
- Дано: матрица «объекты-признаки» X и ответы y
- Найти: модель $a(x)$

Обучение с учителем (supervised learning)

- Для каждого объекта известен ответ (класс или число)
- Даны примеры объектов с ответами
- Нужно построить модель, которая будет предсказывать ответы для новых объектов

Обучение без учителя (unsupervised learning)

- Даны объекты
- Нужно найти в них внутреннюю структуру
- Примеры:
  - Кластеризация
  - Обнаружение аномалий
  - Тематическое моделирование
  - Визуализация
  - Предсказание следующего кадра видео
  - …
- Ближе к обучению в реальной жизни

Обучение без учителя: кластеризация

Case 2. Оптимизация воронки продаж

СБЕРБАНК

ШАГ I
Анализ данных,
в т.ч. транзакционных
Way4, ЦОД, кред. фабрика

ШАГ II
Выявление паттернов и сегментация клиентов по характеристикам

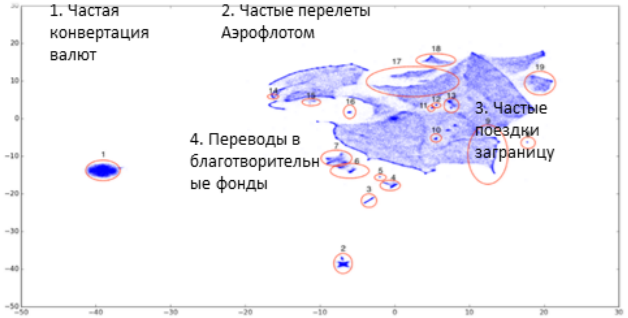
ШАГ III
Формирование продуктовых предложений на базе характеристик клиента

МЕТОДЫ алгоритмы кластеризации, визуализация данных большой размерности с использованием LargeVis

ЭКОНОМИЧЕСКИЙ ЭФФЕКТ

- ✓ Рост эффективности воронки продаж
- ✓ Рост лояльности клиентов

КЛАСТЕРИЗАЦИЯ КЛИЕНТОВ ПО ХАРАКТЕРУ ТРАНЗАКЦИЙ



В ЗАВИСИМОСТИ ОТ КЛАСТЕРА КЛИЕНТА ПРЕДЛОЖИТЬ РЕЛЕВАНТНЫЙ ПРОДУКТ

Паттерн	Продукт
1. Частая конвертация валют	Мультивалютный счет
2. Частые перелеты Аэрофлотом	Карта «Аэрофлот Бонус»
3. Частые поездки заграницу	Страховка для выезжающих за рубеж
4. Переводы в благотворительные фонды	Карта «Подари жизнь»

https://habrahabr.ru/article/318152/

Кластеризация

- Дано: матрица «объекты-признаки» X
- Найти:
  1. Множество кластеров Y
  2. Алгоритм кластеризации $a(x)$, который приписывает каждый объект к одному из кластеров

- Каждый кластер состоит из похожих объектов
- Объекты из разных кластеров существенно отличаются

Отличия

Обучение с учителем

- Цель: минимизация функционала ошибки
- Множество ответов известно заранее
- Конкретные способы измерения качества

Кластеризация

- Нет строгой постановки
- Множество кластеров неизвестно
- Правильные ответы отсутствуют (в большинстве случаев) — нельзя измерить качество

Зачем кластеризовать?

- Маркетинг: искать похожих клиентов
- Модерация: проверять только одно сообщение из кластера
- Соц. опросы: выделять группы схожих анкет
- Соц. сети: искать сообщества

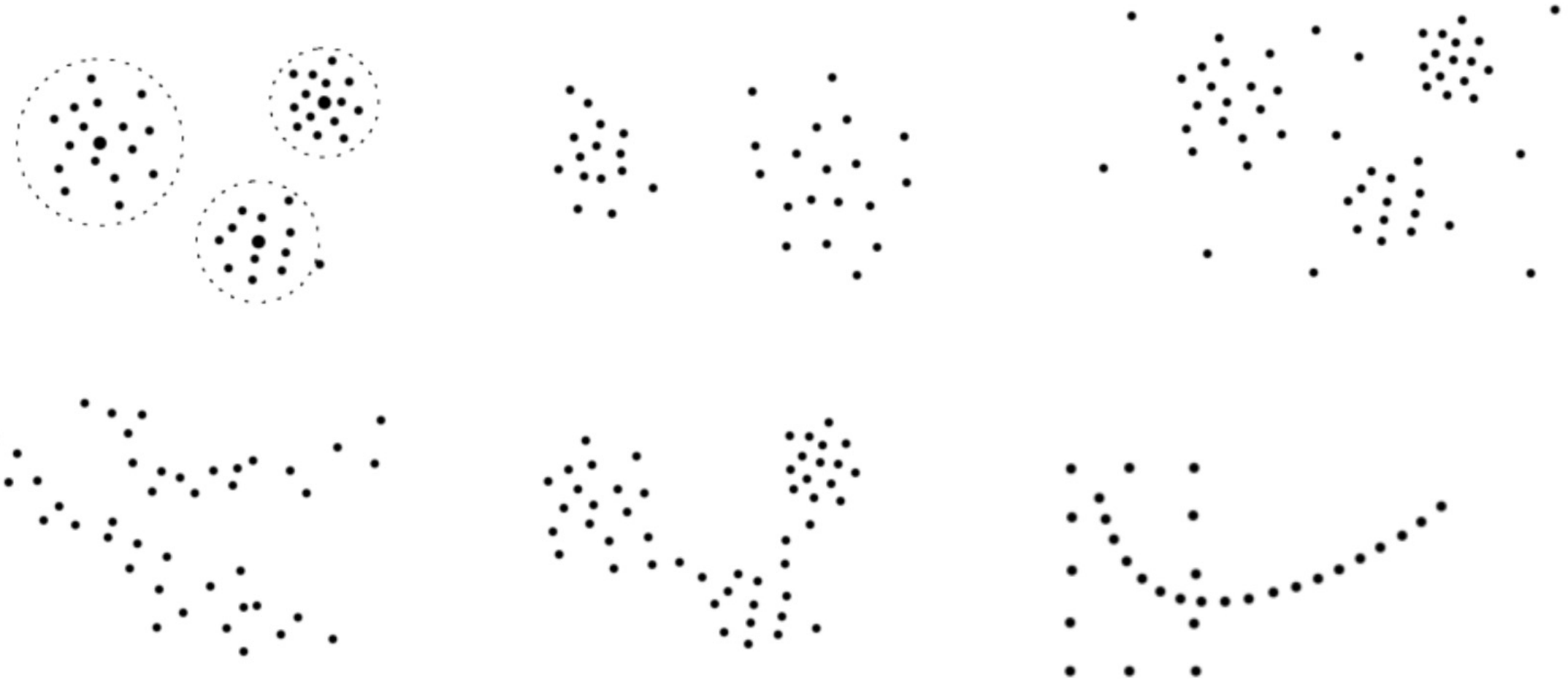
- Выявлять типы людей и формировать поведенческие паттерны для каждого типа

Важно

- Алгоритм кластеризации не знает, чего вы хотите
- Не стоит ожидать, что при кластеризации текстов вы получите разбиение именно по темам
- Нередко кластеры оказываются неинтерпретируемыми

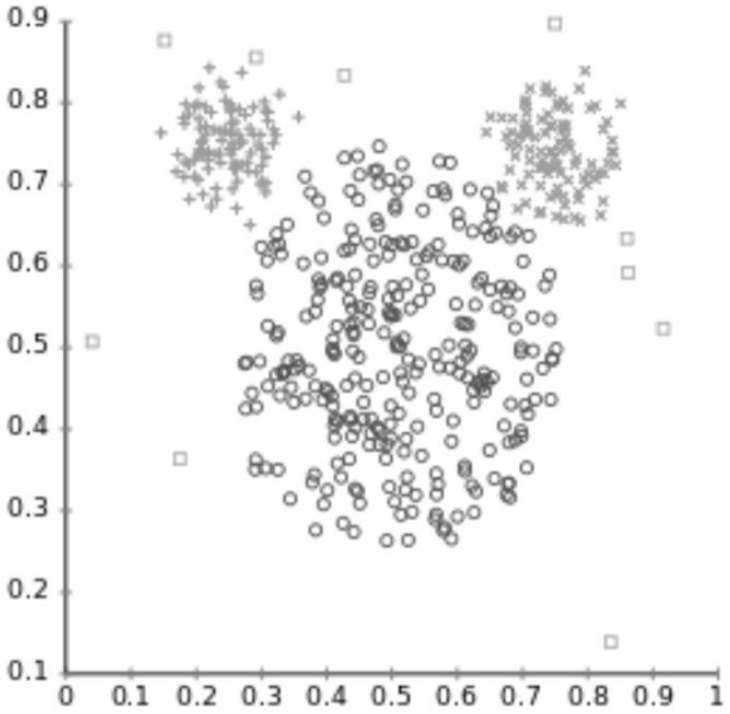
Виды кластеризации

Форма кластеров

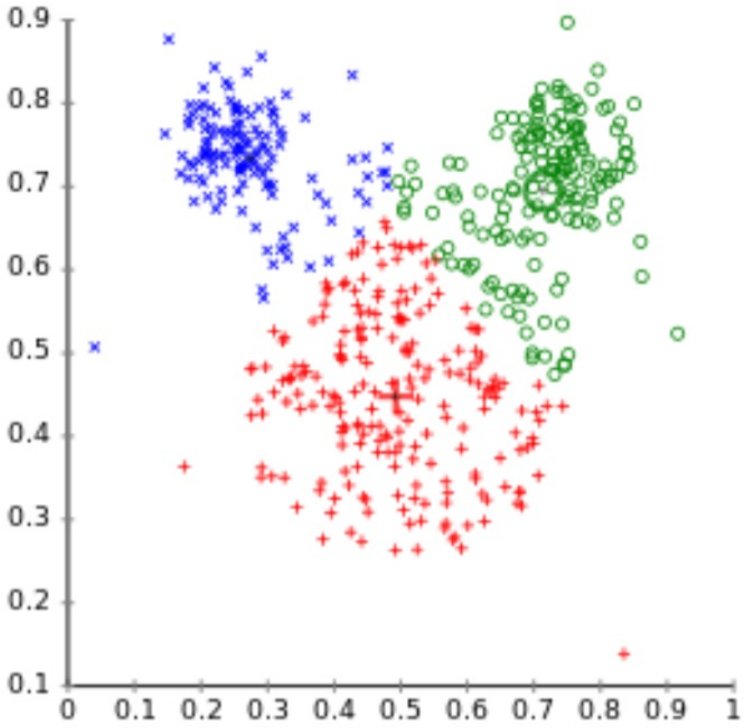


Форма кластеров

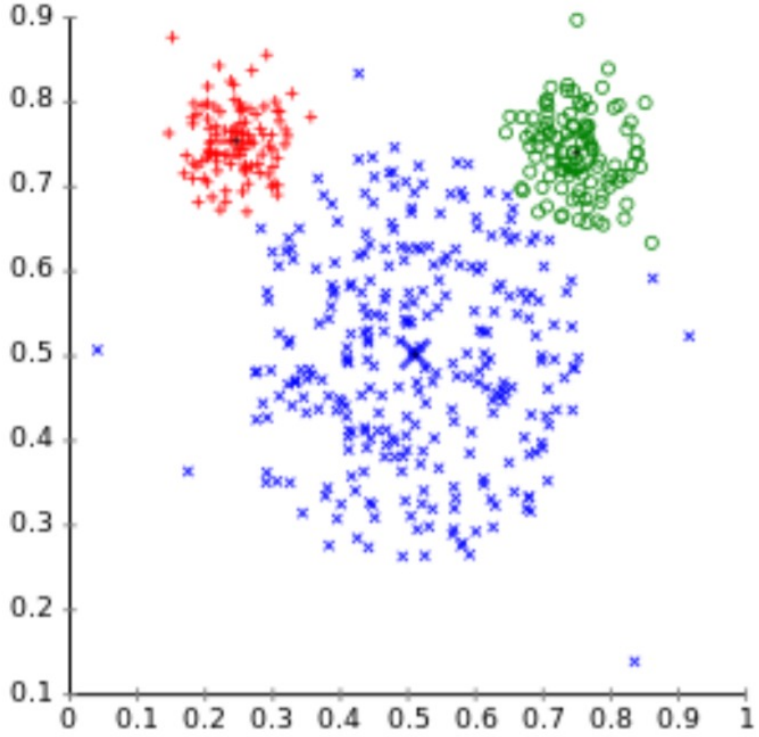
Различия в результатах работы



Исходная выборка
("Mouse" dataset)

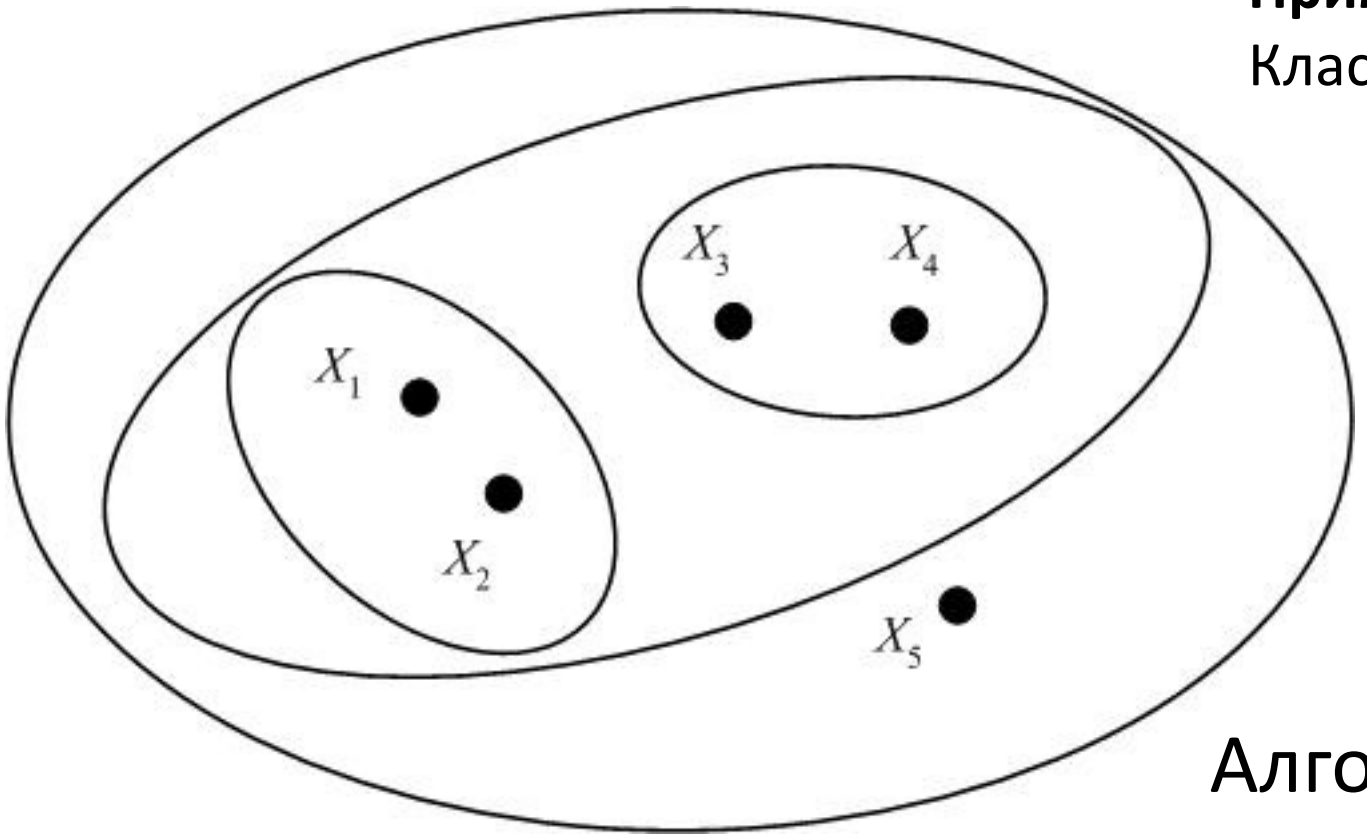


Метод 1

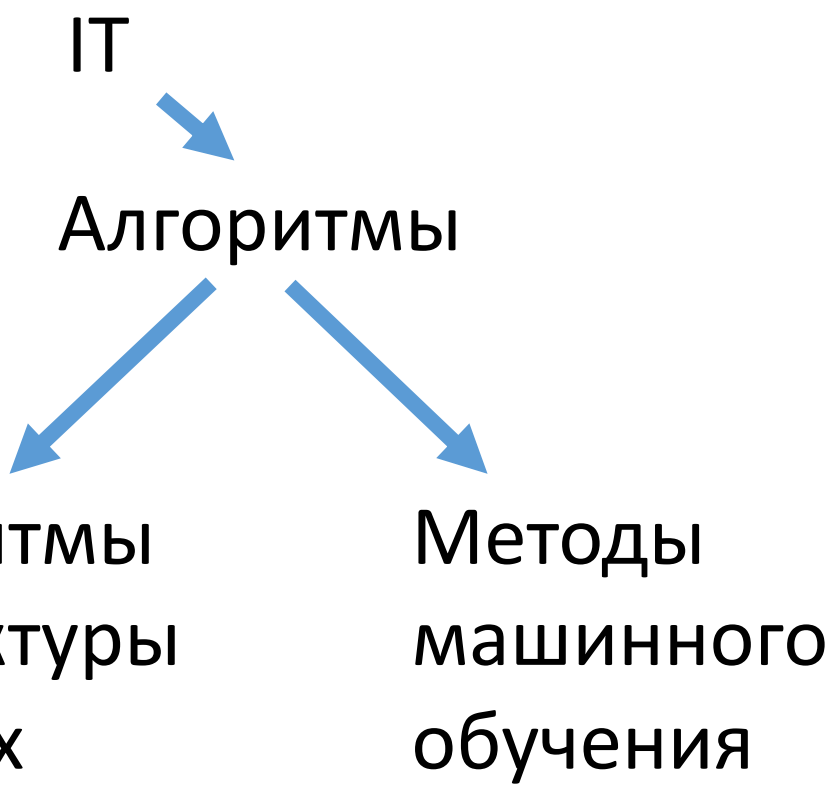


Метод 2

Иерархическая кластеризация



Пример:
Кластеризация статей с Хабра



Требования к кластерам

- Задача кластеризации новостей по содержанию.
- Постановка 1: в один кластер должны попадать новости на одну тему

Батыршин сыграет вместо Хабарова у «Магнитки» в матче с «Салаватом»

Место в третьей паре защиты «Магнитки» на третью встречу плей-офф Кубка Гагарина с «Салаватом Юлаевым» занял защитник Рафаэль Батыршин, сообщает из Уфы корреспондент «Чемпионата» Павел Панышев. Травмированный Ярослав Хабаров выбыл на неопределённый срок. Для форварда Оскара Осалы сезон закончен.

Футболисты ЦСКА проиграли «Долгопрудному» в товарищеском матче

Футболисты московского ЦСКА со счетом 2:3 проиграли клубу второго дивизиона "Долгопрудный" в товарищеском матче, который состоялся в Москве на стадионе "Октябрь". У армейцев забитыми мячами отличились Александр Цауня (15-я минута) и Сергей Ткачев (54).

Скриншот с сайта Яндекс.Новости (news.yandex.ru)

Требования к кластерам

- Задача кластеризации новостей по содержанию.
- Постановка 2: в один кластер должны попадать новости об одном «большом» событии

Скриншот с сайта РИА Новости (ria.ru)

Требования к кластерам

- Задача кластеризации новостей по содержанию.
- Постановка 3: в один кластер должны попадать тексты об одной и той же новости

11:41, 08 ФЕВРАЛЯ 2014

Открытие Олимпиады в Сочи посмотрели несколько миллиардов человек

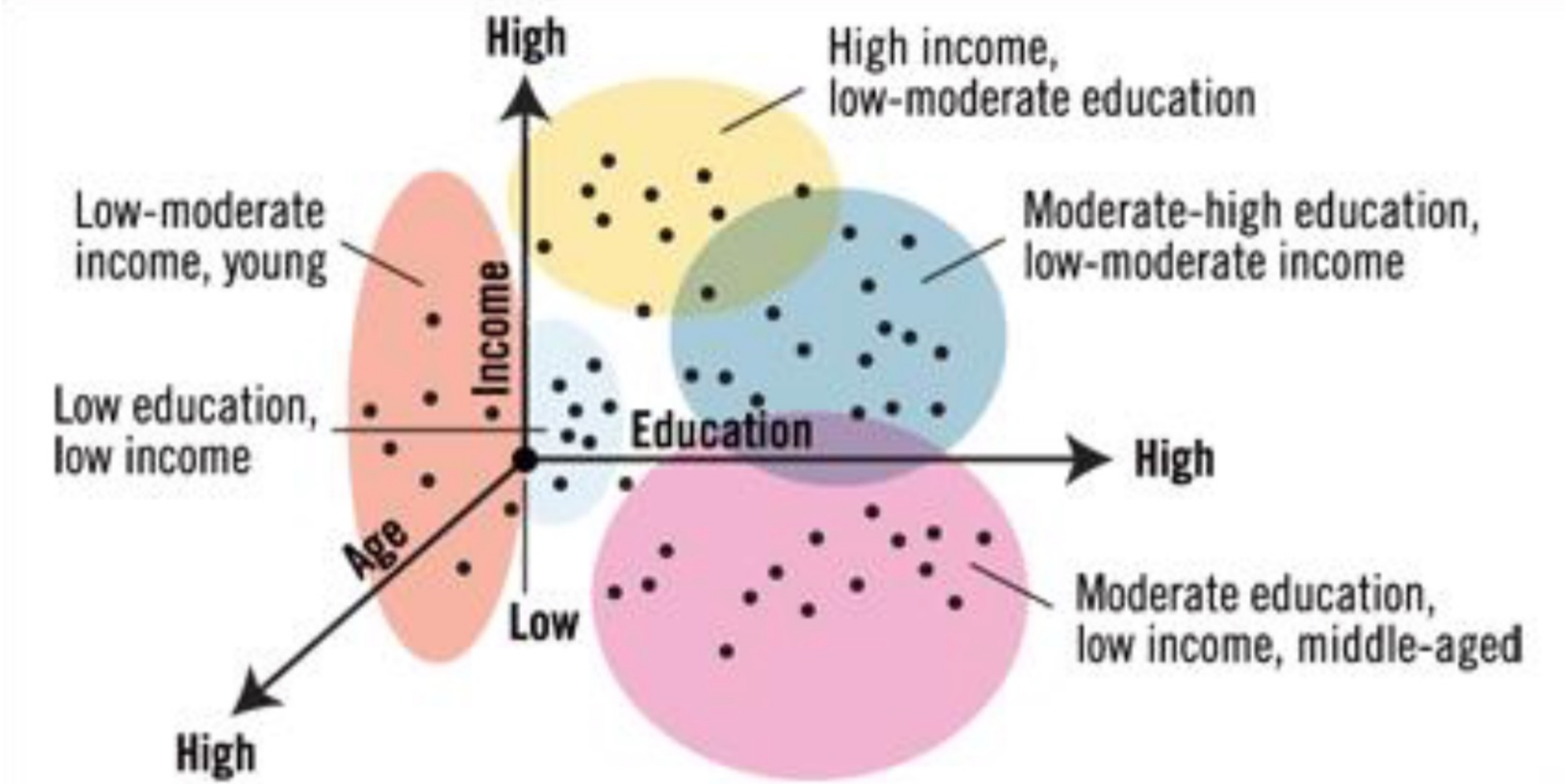
Олимпиада в Сочи открыта

Церемония открытия Олимпиады в Сочи. Онлайн-репортаж

Требования к кластерам

- Чтобы проверить, выполняются ли требования, нужно делать разметку данных
- Для новостей: показывать асессору пары документов и спрашивать, относятся ли они к одному кластеру

Кластеризация как основная задача

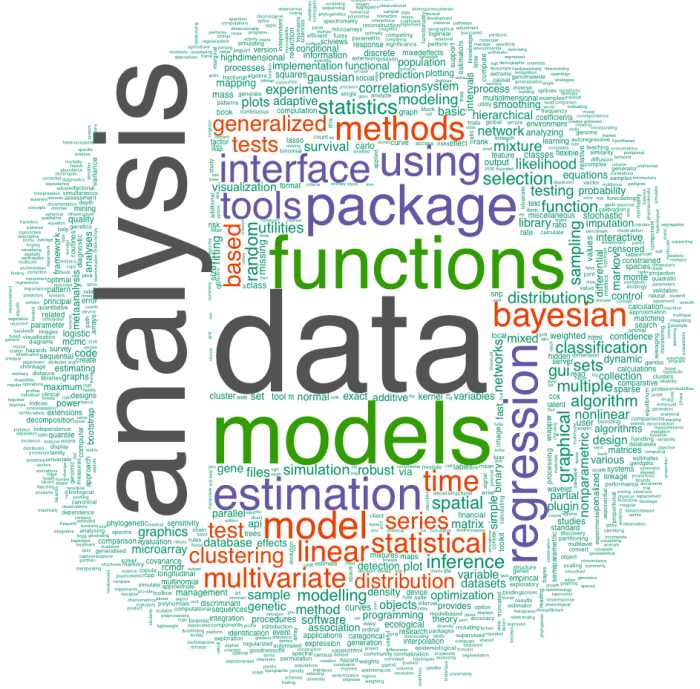


Кластеризация как вспомогательная задача

Цель: улучшение распознавания

«Жёсткая» и «мягкая» кластеризации

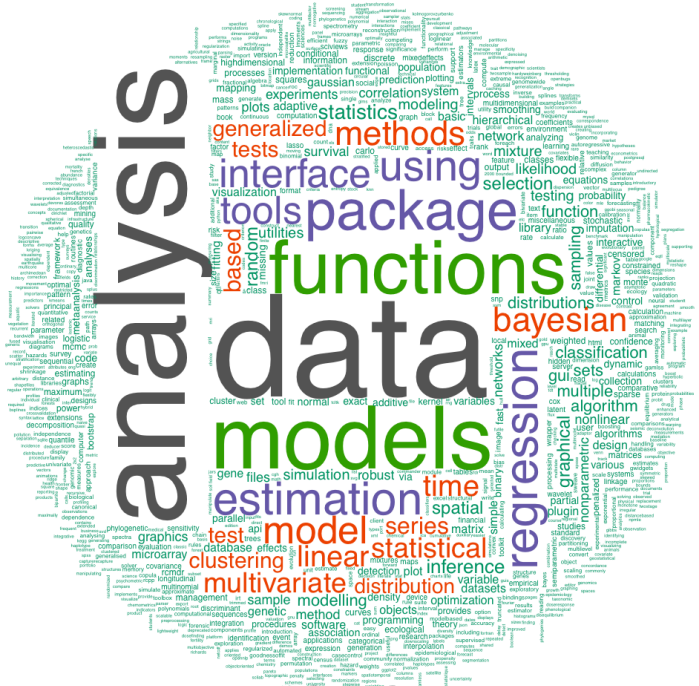
Кластеризация для выделения «тем»



«Жесткая» и «мягкая» кластеризации

Кластеризация для выделения «тем»

0.2



0.3

0.5

Типы задач кластеризации

- Форма кластеров, которые нужно выделять
- Плоская или древовидная структура
- Размер кластеров
- Конечная задача или вспомогательная
- Жесткая или мягкая кластеризация

K-Means

K-Means

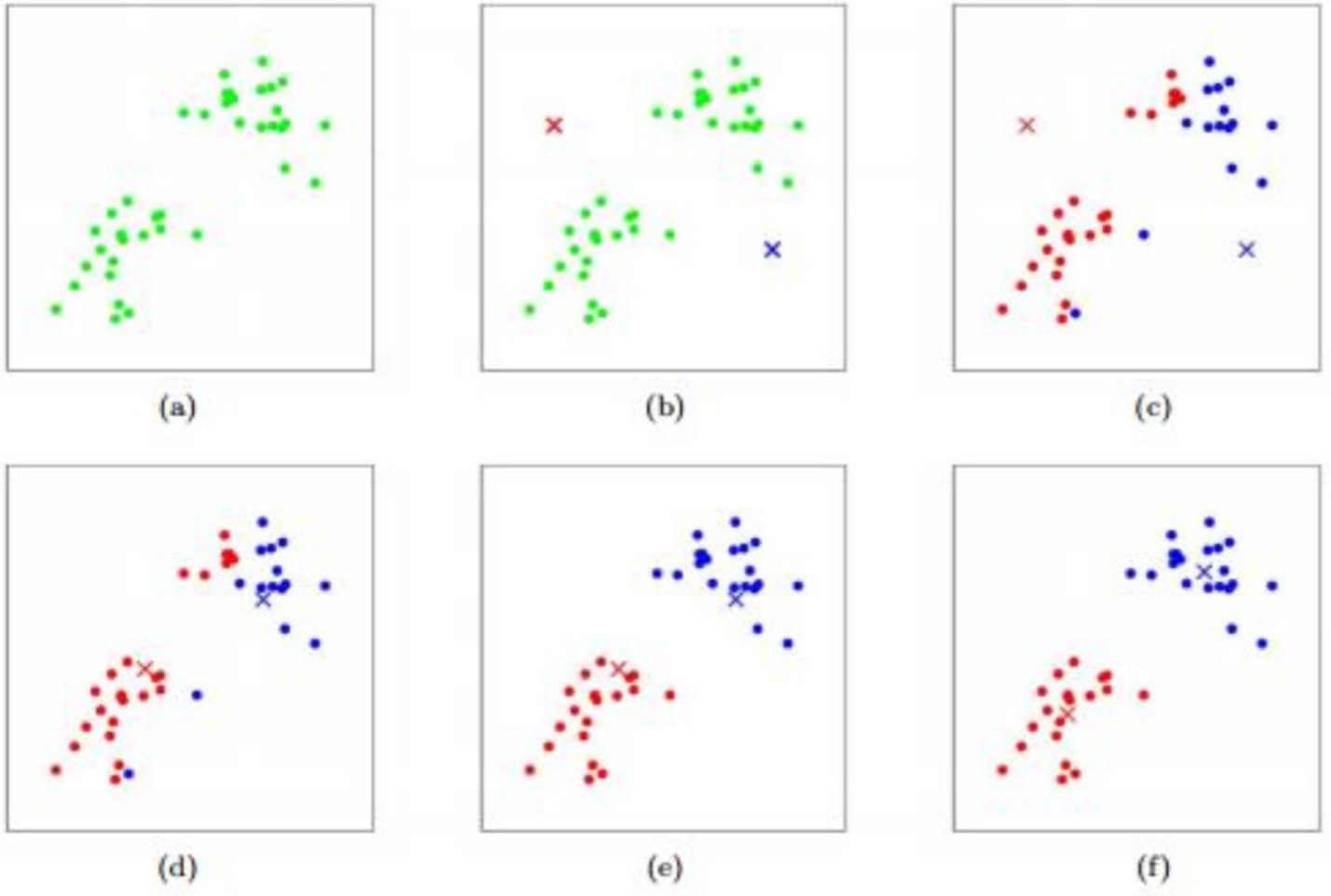
- Дано: выборка $x_1, \dots, x_\ell$
- Параметр: число кластеров K
- Начало: случайно выбрать K центров кластеров $c_1, \dots, c_K$
- Повторять по очереди до сходимости:
  - Шаг А: отнести каждый объект к ближайшему центру

$$y_i = \arg\min_{j=1,\dots,K} \rho(x_i, c_j)$$

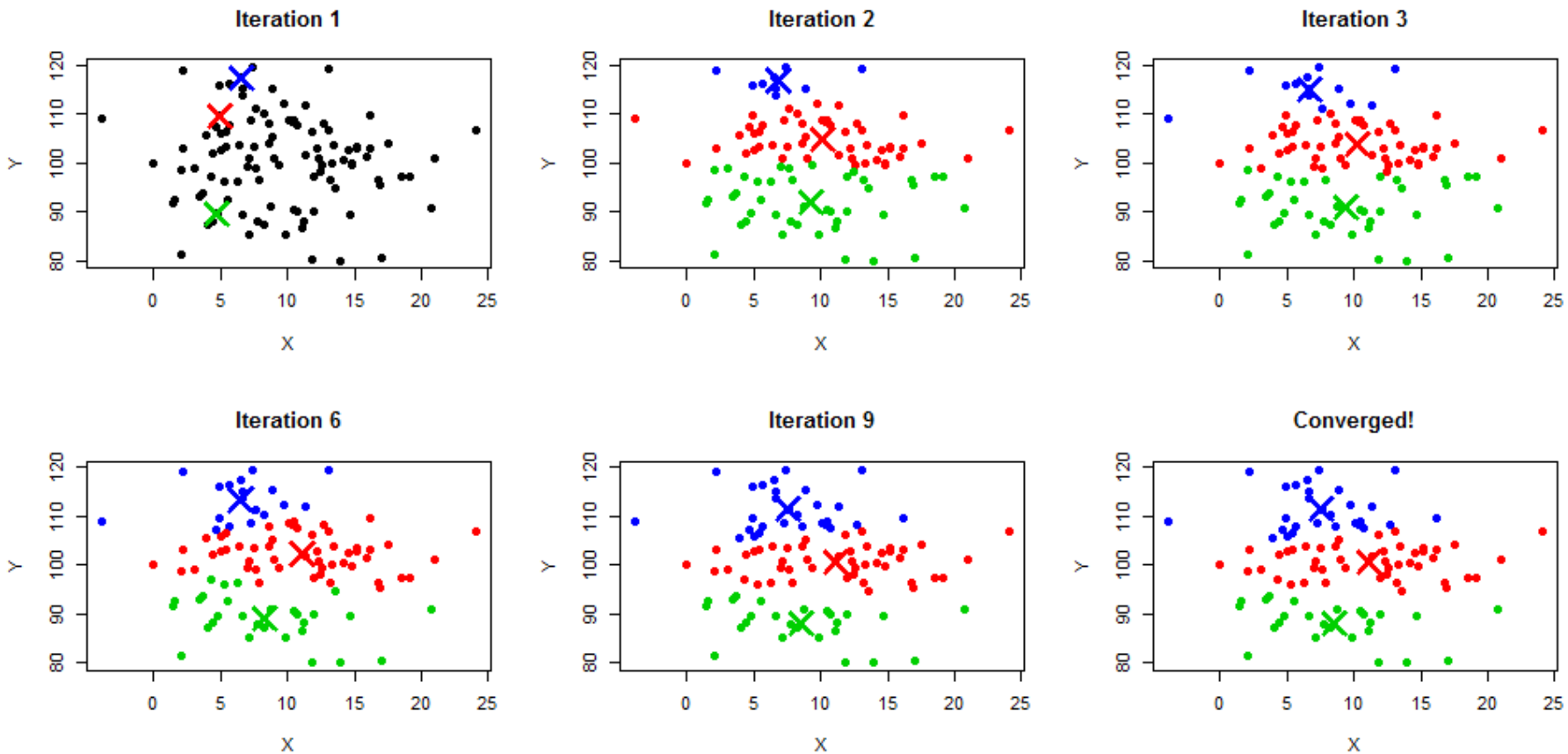
  - Шаг Б: переместить центр каждого кластера в центр тяжести

$$c_j = \frac{\sum_{i=1}^{\ell} x_i [y_i = j]}{\sum_{i=1}^{\ell} [y_i = j]}$$

K-Means



K-Means


Iteration 1
Iteration 2
Iteration 3
Iteration 6
Iteration 9
Converged!
X
Y
0
5
10
15
20
25
80
90
100
110
120

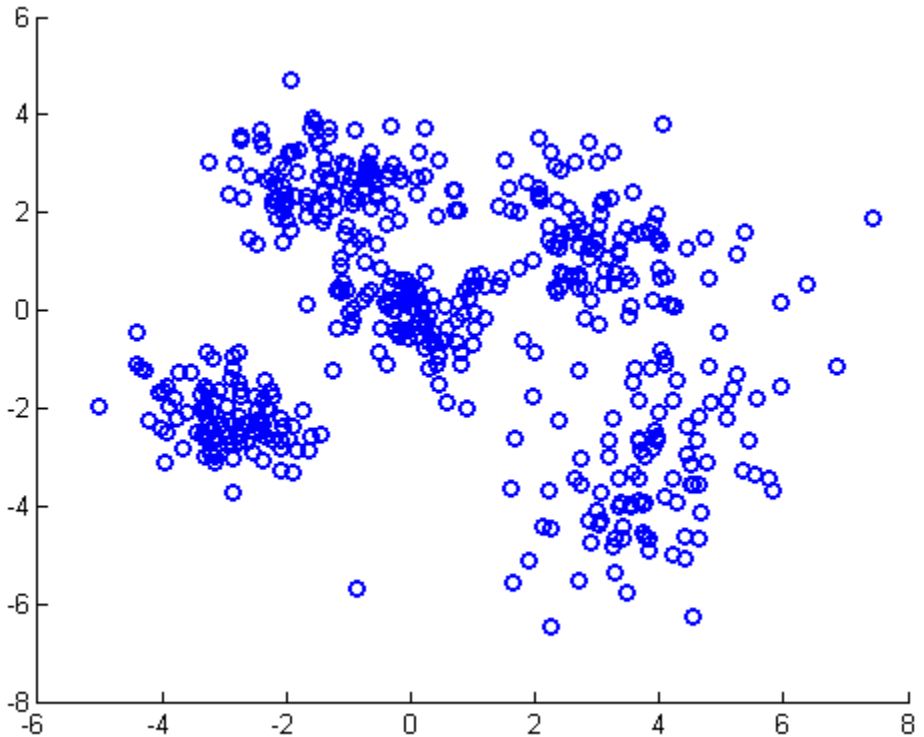
Выбор числа кластеров

- Качество кластеризации: внутрикластерное расстояние

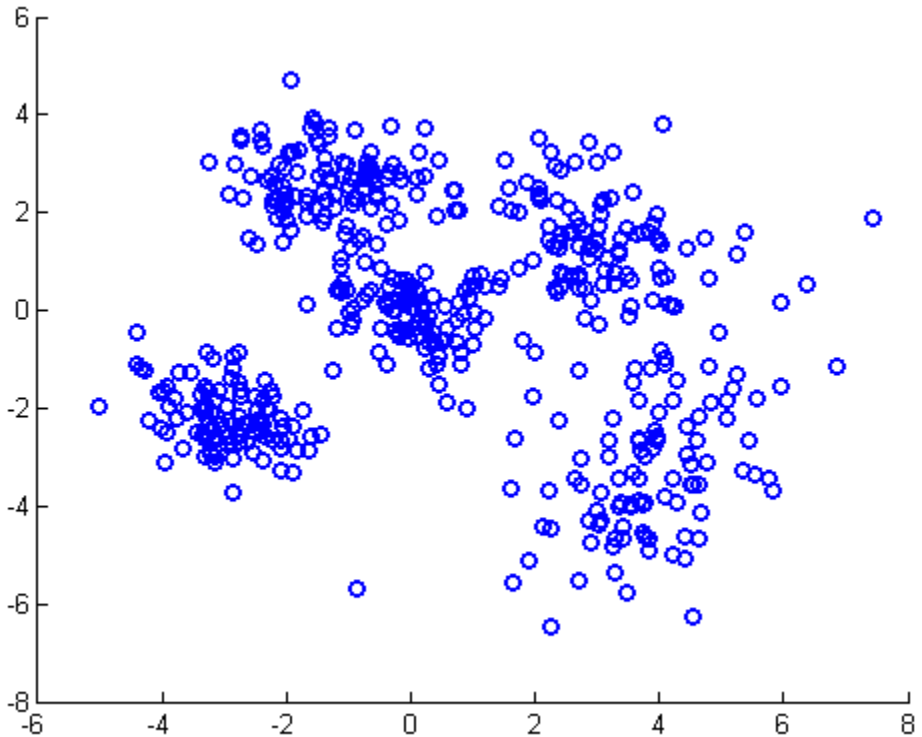
$$J(C) = \sum_{i=1}^{\ell} \rho\left(x_i, c_{y_i}\right)$$

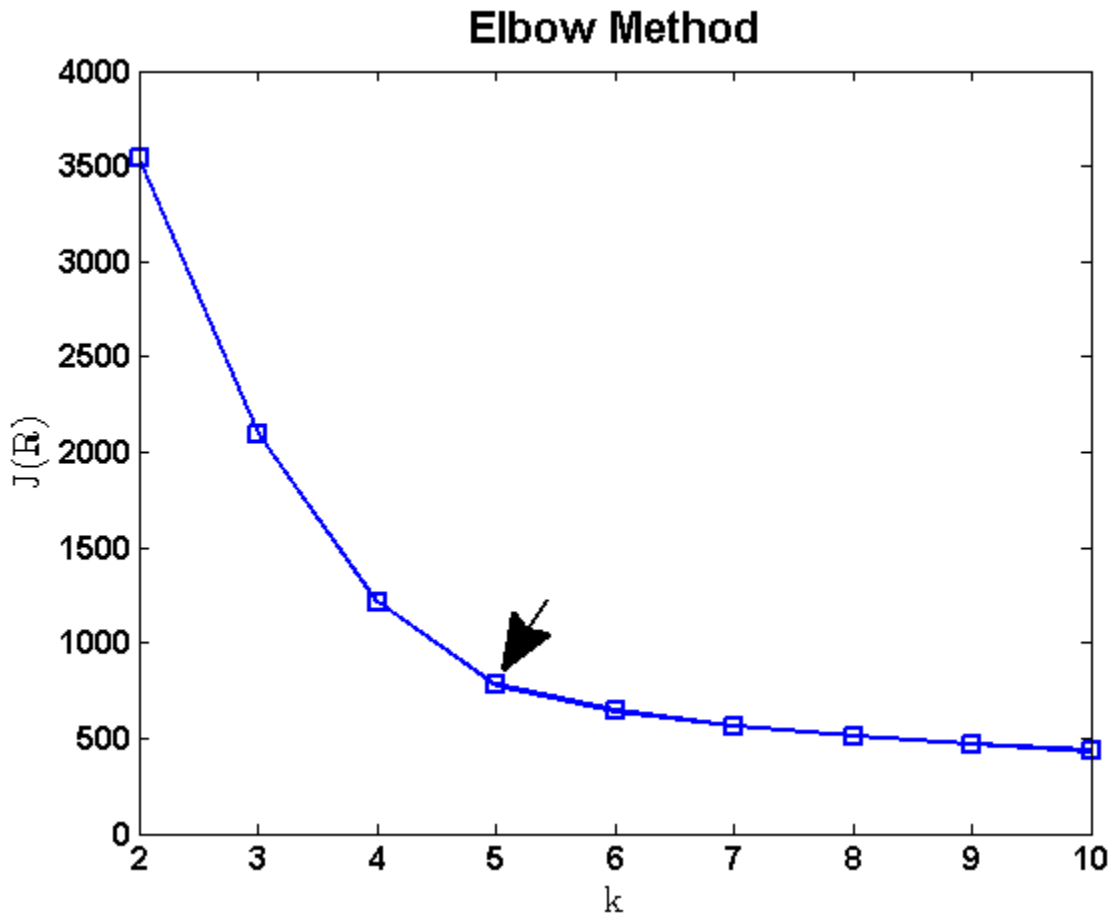
- Зависит от K
- Нужно подобрать такое K, после которого качество меняется не слишком сильно

Выбор числа кластеров



Выбор числа кластеров





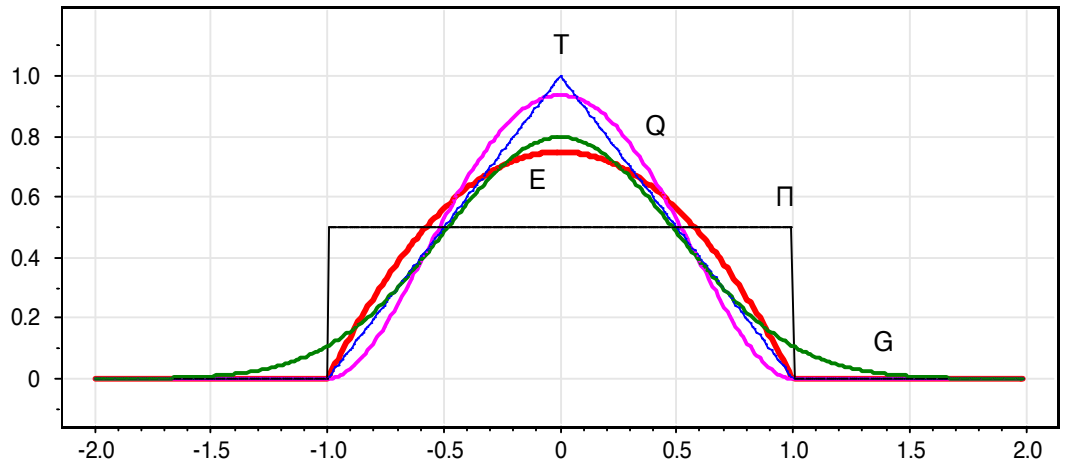
Особенности K-Means

- Может работать с большими объёмами данных
- Подходит для кластеров с простой геометрией
- Требует выбора числа кластеров

Mean Shift

Парзеновское ядро

- Обозначение: $K(z)$
- Чем ближе к нулю z, тем больше $K(z)$
- Примеры:
  - $K(z) = \exp\left(\frac{z}{h}\right)$
  - $K(z) = [z \leq h]$



http://www.machinelearning.ru/wiki/index.php?title=MO

Mean Shift

Пусть дан объект x

1. Вычисляем окрестность $N(x)$ — объекты, где $K\big(\rho(x, x_i)\big) > \varepsilon$
2. Находим градиент плотности:

$$m(x) = \frac{\sum_{x_i \in N(x)} x_i K\big(\rho(x, x_i)\big)}{\sum_{x_i \in N(x)} K\big(\rho(x, x_i)\big)} - x$$

3. Делаем шаг:

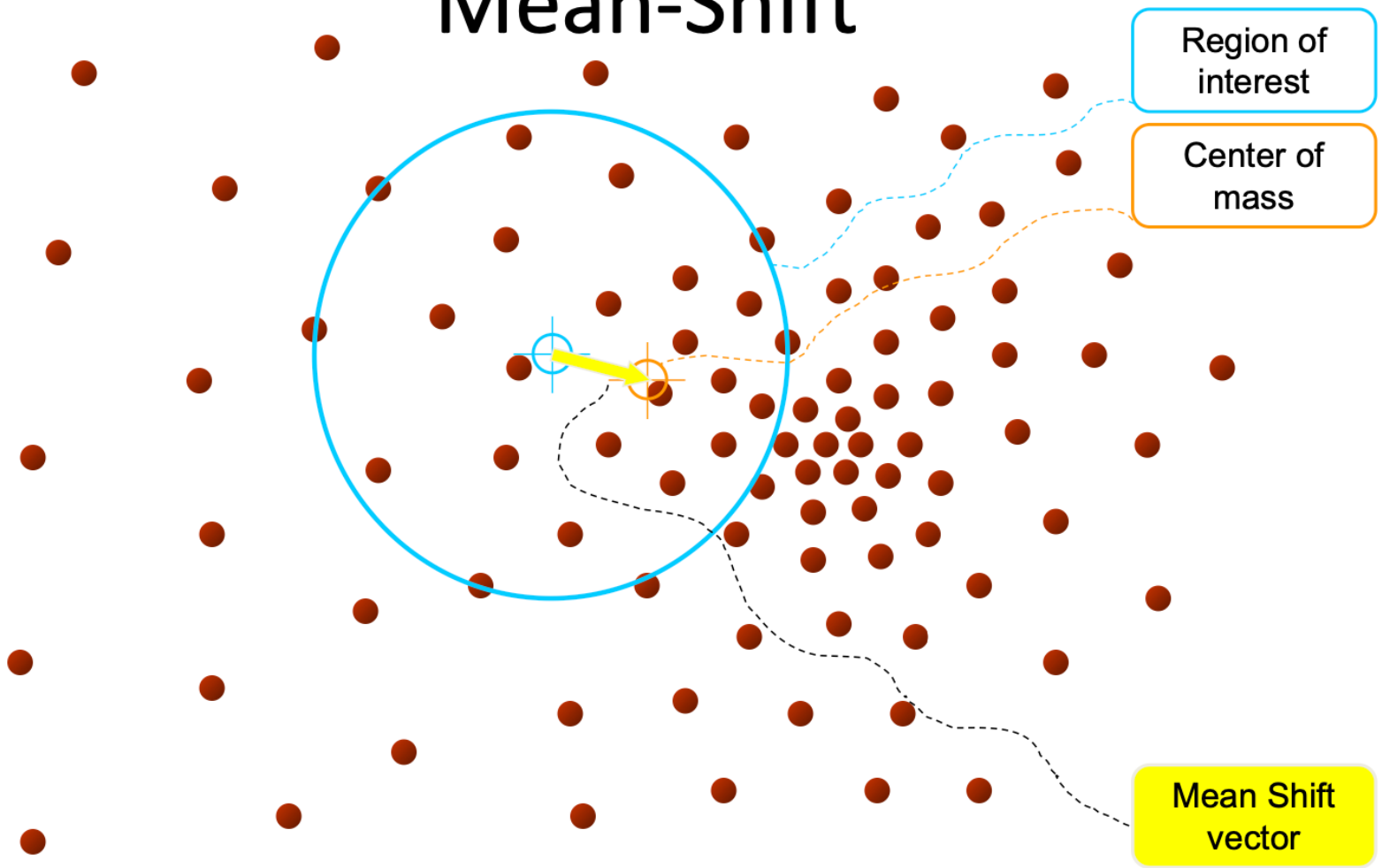
$$x \leftarrow x + m(x)$$

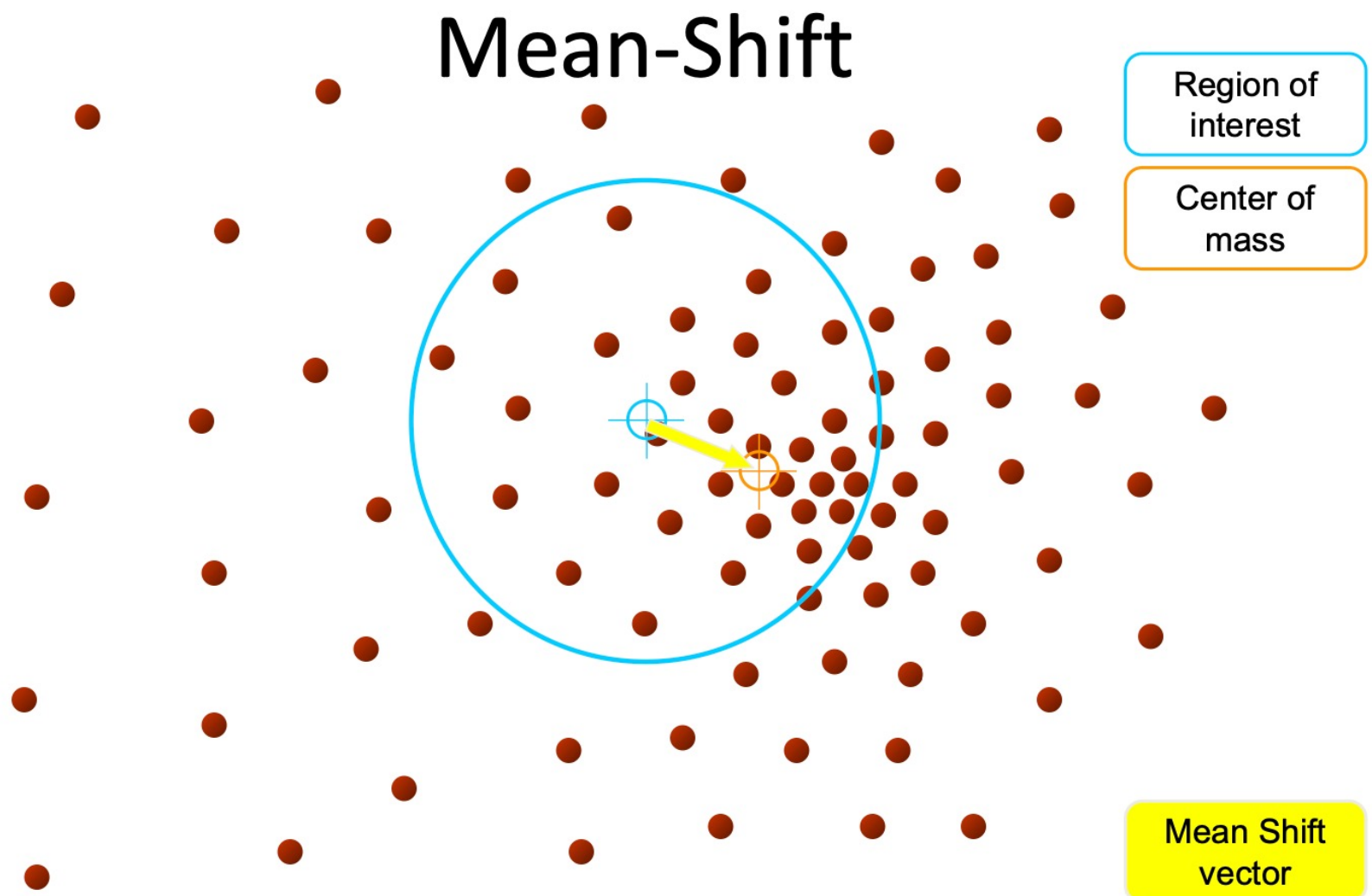
4. Повторяем до сходимости

Mean Shift

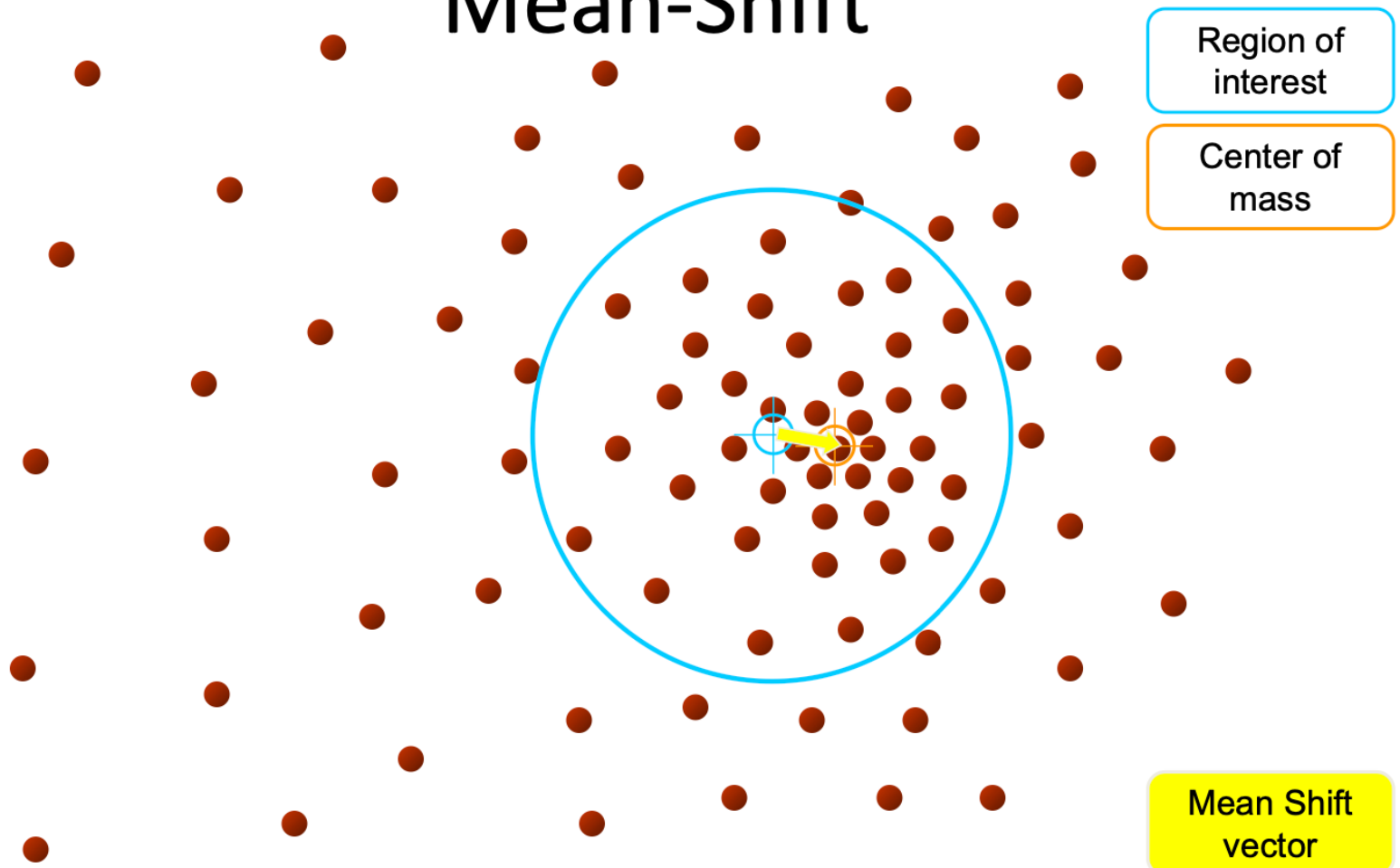
- Алгоритм запускается для каждого объекта
- Те объекты, для которых алгоритм сходится в общую точку, относятся к одному кластеру

Mean-Shift

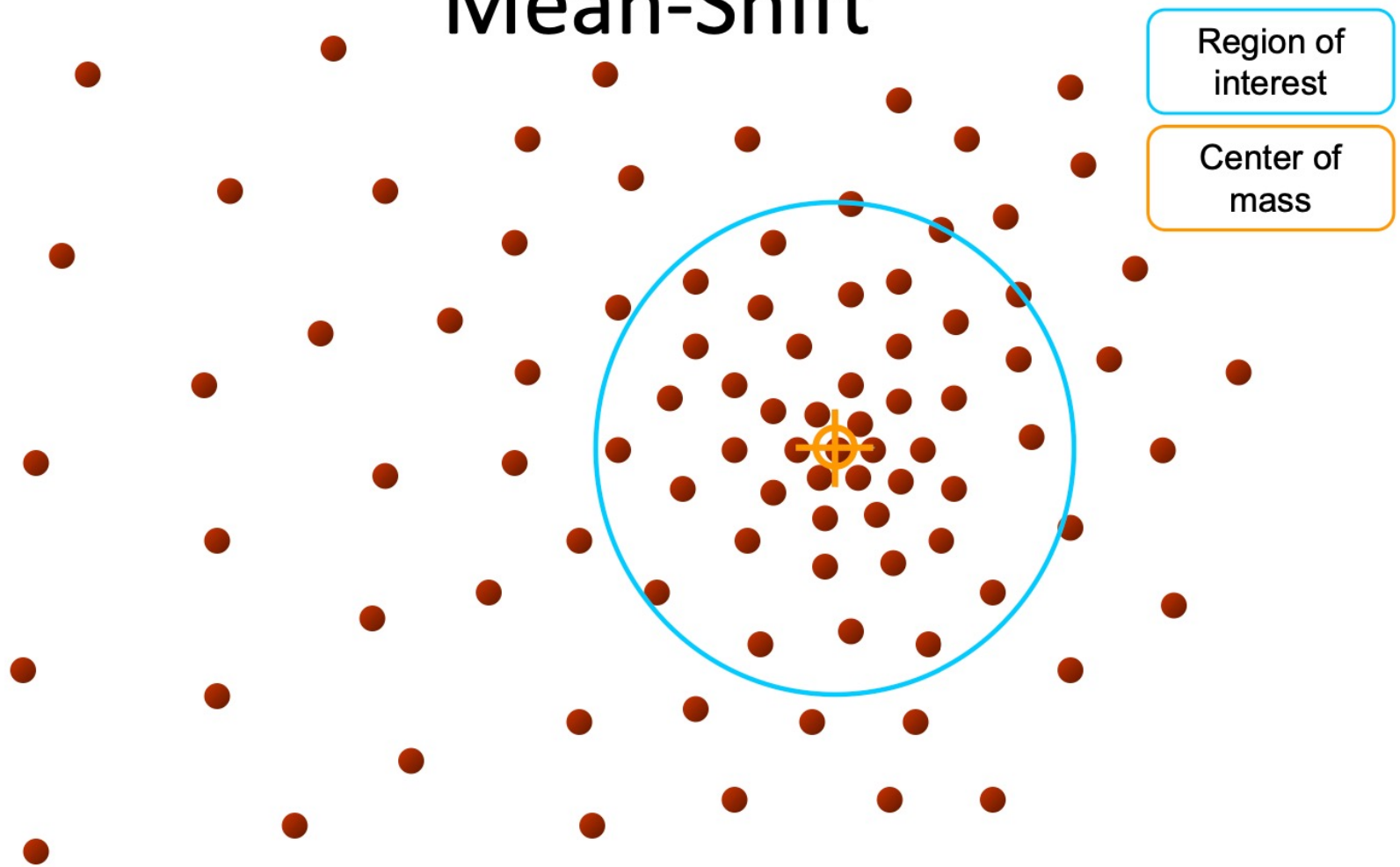



Mean-Shift
Region of interest
Center of mass
Mean Shift vector

Mean-Shift



Mean-Shift



Mean Shift