Кластеризация K представителями

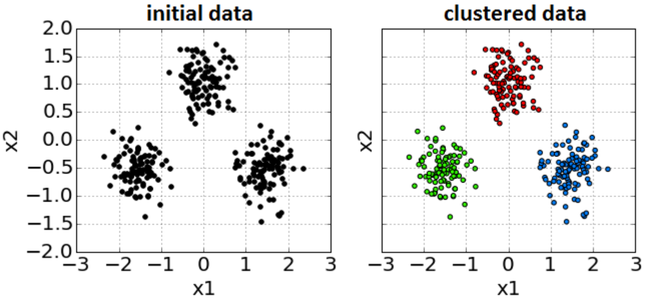
Виктор Китов

v.v.kitov@yandex.ru

Содержание

Идея кластеризации

- Кластеризация - разбиение объектов на группы, такие что
  - внутри групп объекты очень метрически похожи
  - объекты из разных групп метрически непохожи
- Обучение без учителя, нет "золотого стандарта"



Нет единого понятия "похожести"

- разные метрики приводят к разным результатам

Применения кластеризации

- сегментация клиентов
    - например, для более таргетированных спец. предложений
- детекция сообществ в соц. сетях
    - узлы-люди, похожесть-длина мин. пути
        - в графе дружбы, сообщений, лайков
- детекция выбросов
    - выбросы не принадлежат ни одному кластеру
- сжатие данных
    - вектор признаков можно заменить на номер кластера
- извлечение новых признаков
    - номер кластера, расстояние до своего и ближайшего чужого кластера

Характеристики алгоритмов кластеризации

Можем сравнивать различные алгоритмы кластеризации:

- по вычислительной сложности
- #кластеров находится автоматически?
- строится плоская или иерархическая кластеризация?
- гибкость формы кластеров
    - могут ли быть разной плотности, невыпуклые?
- устойчивость алгоритма к наличию выбросов
- используемая метрика похожести

Содержание

Кластеризация, основанная на представителях

Кластеризация, основанная на представителях (representative-based clustering)

- Кластеризация плоская (не иерархическая).
- #кластеров K задается пользователем.
- Каждый объект x_n соотносится кластеру $z_n \in \{1, 2, ...K\}$.
- Каждый кластер k определяется центром μ_k, $k = 1, 2, ...K$.
- Решается задача:

$$\mathcal{L}(z_1, ...z_N; \mu_1, ...\mu_K) = \sum_{n=1}^{N} \rho(x_n, \mu_{z_n}) \to \min_{z_1,...z_N;\mu_1,...\mu_K} \tag{1}$$

- Находится <u>локальный</u> оптимум методом покоординатного спуска $(\mu, z, \mu, z...)$

Общий алгоритм

инициализировать $\mu_1, ...\mu_K$
(случайными объектами выборки)

ПОВТОРЯТЬ по сходиомсти:

  для $n = 1, 2, ...N$:

    $z_n = \arg\min_k \rho(x_n, \mu_k)$

  для $k = 1, 2, ...K$:

    $\mu_k = \arg\min_\mu \sum_{n:z_n=k} \rho(x_n, \mu)$

ВЕРНУТЬ $z_1, ...z_N$

Комментарии

- разные ф-ции расстояния приводят к разным алгоритмам:
    - $\rho(x, x') = \|x - x'\|_2^2$=> K-средних
        - μ_k - среднее
        - неустойчиво к выбросам
    - $\rho(x, x') = \|x - x'\|_1$=> K-медиан
        - μ_k - медиана
        - устойчива к выбросам
- μ_k может выбираться только среди существующих объектов
    - например, временные ряды разной длины - не можем усреднять
- K - гиперпараметр.
    - если малый, то различные кластеры сольются в один
    - лучше взять завышенным, а потом объединить похожие
- Форма кластеров определяется $\rho(\cdot, \cdot)$

Комментарии

Условия сходимости:

- достигнуто максимальное # итераций
- назначения кластеров $z_1, ...z_N$ перестали меняться (полная сходимость)
- изменения $\{\mu_i\}_{i=1}^K$ меньше порога (приближенная сходимость)

Инициализация центров:

- $\{\mu_i\}_{i=1}^K$ инициализируются случайными объектами
    - если взять выброс, кластер будет содержать только его
    - более устойчиво:
        - инициализировать медианами из нескольких случайных объектов

Оптимальность:

- критерий содержит много локальных оптимумов
- можно запусть оптимизацию из разных инициализаций и выбрать лучшее решение

K-средних - алгоритм

Инициализировать μ_j
(случайными объектами выборки).

ПОВТОРЯТЬ до сходимости:

для $i = 1, 2, ...N$:
определить кластер для x_i:
$z_i = \arg\min_{j \in \{1,2,...K\}} ||x_i - \mu_j||_2^2$

для $j = 1, 2, ...K$:
пересчитать центры:
$\mu_j = \frac{1}{\sum_{n=1}^N \mathbb{I}[z_n = j]} \sum_{n=1}^N \mathbb{I}[z_n = j] x_i$

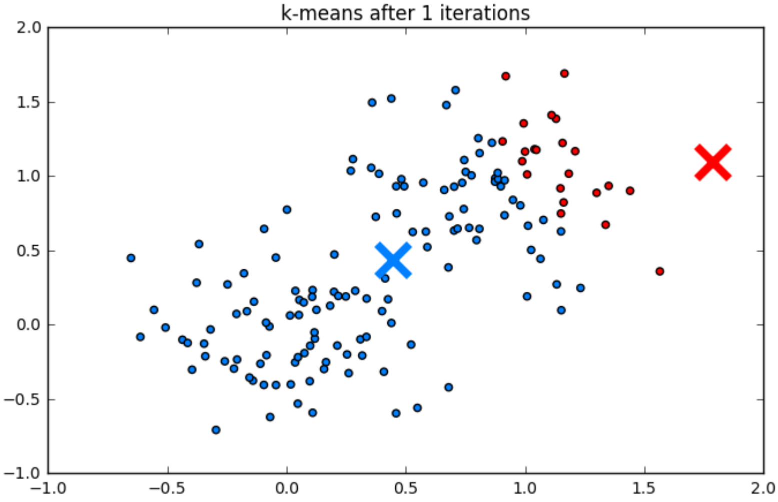
Сложность: $O(NDKI)$, K-#кластеров, I-#итераций.

K-средних - динамический алгоритм

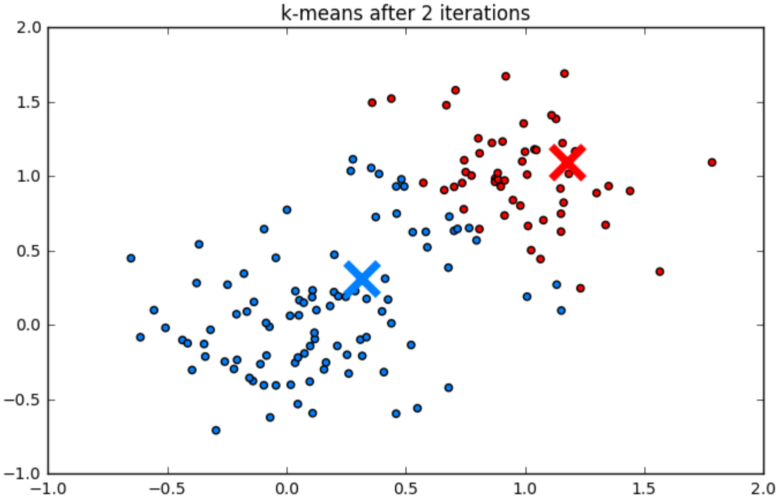
```
Инициализировать μ_j
(случайными объектами выборки).

ПОВТОРЯТЬ до сходимости:
    для i = 1, 2, ...N:
        определить кластер для x_i:
        z_i' = arg min_{j∈{1,2,...K}} ||x_i − μ_j||_2^2
        если z_i' != z_i:
            пересчитать μ_{z_i} и μ_{z_i'}:
            μ_{z_i} = 1/(Σ_{n=1}^N I[z_n'=z_i]) Σ_{n=1}^N I[z_n' = z_i] x_i
            μ_{z_i'} = 1/(Σ_{n=1}^N I[z_n'=z_i']) Σ_{n=1}^N I[z_n' = z_i'] x_i
            z_i = z_i'
```

- Сходится за $\downarrow$#итераций, пустые кластеры невозможны.

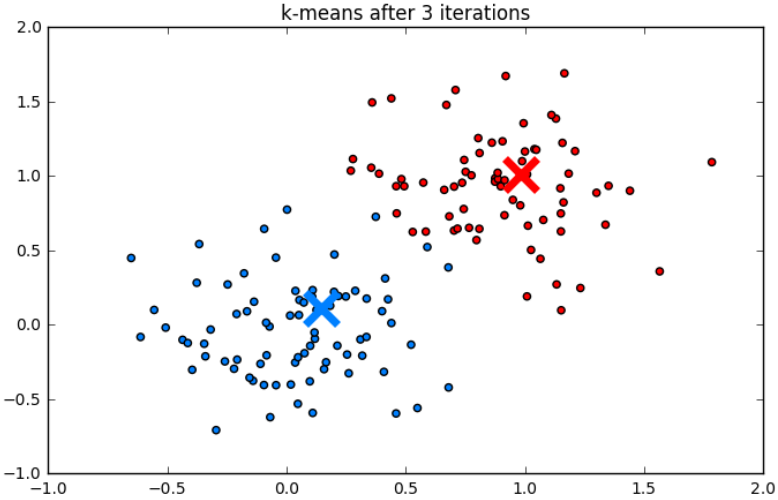
Пример работы K-средних



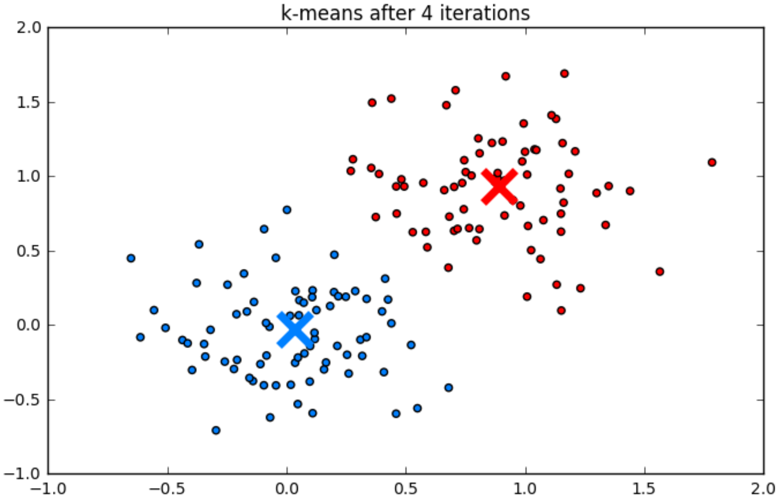
Пример работы K-средних



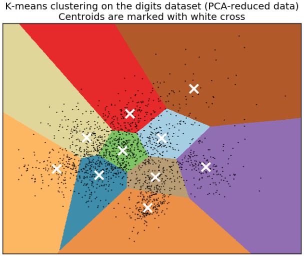
Пример работы K-средних



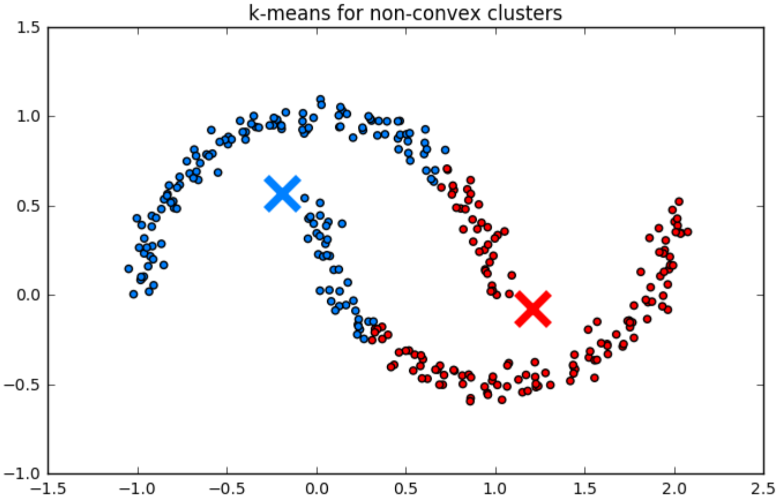
Пример работы K-средних



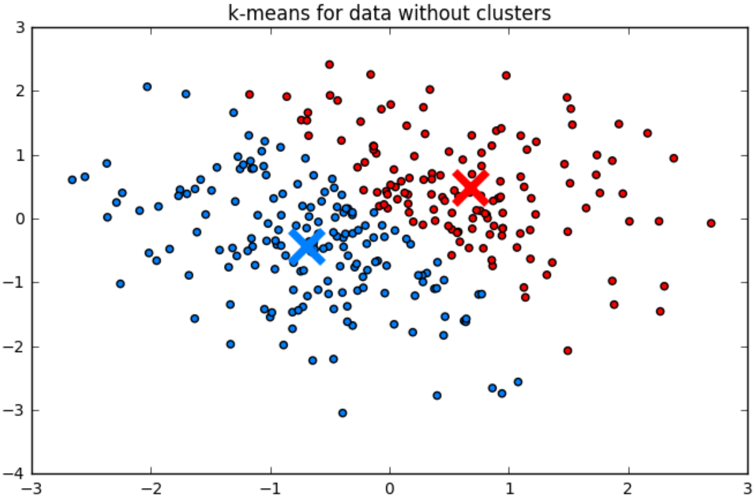
Пример кластеризации рукописных цифр



K-средних для невыпуклых кластеров

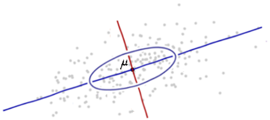


K-средних для равномерно распределенных данных



K-представителей - расстояние Махаланобиса

- Расстояние Махаланобиса:
$\rho(x,\mu_k)^2 = (x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k),\ k=1,2...K.$



- Расстояние Махаланобиса позволяет моделировать кластеры эллиптической формы, разного размера и плотности