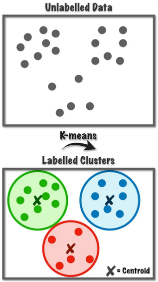
Кластеризация K представителями

Виктор Китов

victorkitov.github.io



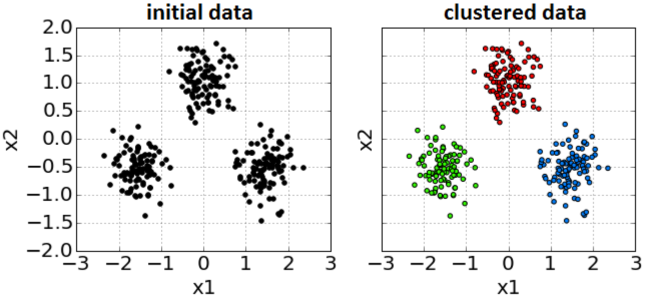
Курс поддержан фондом 'Интеллект'

Победитель конкурса VK среди курсов по IT

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"

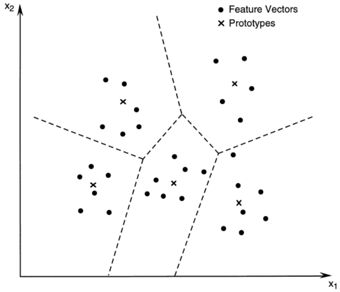
- разные метрики приводят к разным результатам

Применения кластеризации

- **классификация** без обучающей выборки
- **сегментация клиентов**
    - например, для более тагретированных спец. предложений
- **рекомендательная система**
    - рекомендуем клиентам то, что нравится др. клиентам их кластера
- **детекция выбросов**
    - выбросы не принадлежат ни одному кластеру
- **ускорение поиска похожих объектов по кластеру**
    - в KNN, частичном обучении, активном обучении
- **отладка моделей с учителем**
    - на характерных представителях каждого кластера
- **извлечение новых признаков**
    - номер кластера, расстояние до своего и ближайшего чужого кластера

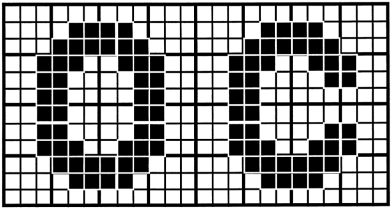
Сжатие данных[1]

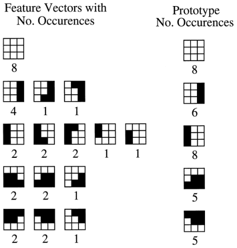
- Для экономии памяти и $\downarrow$переобучения заменим похожие объекты номером их кластера
- Англ. vector quantization, feature space quantization.



[1] Источник иллюстраций.

Сжатие данных[2]





- В каждом фрагменте 3x3: $2^9 = 512$ вариантов.
- Уникальных фрагментов 32, заменим их 5 прототипами.
- Заменяя изображение частотой встречи каждого прототипа получим bag-of-visual-words кодирование.
  - Машина-колесо, окно, фара... Человек: глаз, нос, руки...
  - нейросети извлекают признаки лучше, но им нужно много обучающих данных

[2] Источник иллюстраций.

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 $z_n \in \{1, 2, ...K\}$.
- Индексы кластеров:

$$C_k = \{n : z_n = k\}$$

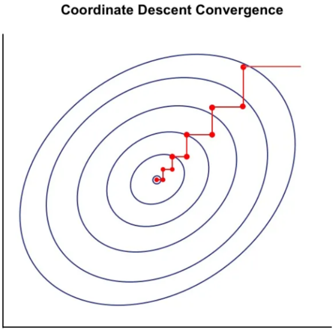
- Каждый кластер k определяется центром μ_k, $k = 1, 2, ...K$.
- Решается задача:

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

Метод оптимизации

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

Находится локальный 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 $k = 1, 2, ...K$:
        $\mu_k = \arg\min_\mu \sum_{n \in C_k} \rho(x_n, \mu)$

ВЕРНУТЬ $z_1, ...z_N$

Число кластеров

- K - гиперпараметр.
  - если малый, то различные кластеры сольются в один
  - лучше взять завышенным, а потом объединить похожие

Комментарии

- разные ф-ции расстояния приводят к разным алгоритмам:
  - $\rho(x, x') = \|x - x'\|_2^2$=> К-средних
    - μ_k - среднее
    - неустойчиво к выбросам
  - $\rho(x, x') = \|x - x'\|_1$=> К-медиан
    - μ_k - медиана
    - устойчива к выбросам
- μ_k -только среди существующих объектов (K medoids)
  - например, временные ряды разной длины - не можем усреднять
- 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)

Оптимальность:

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

Инициализация центров

Инициализация центров:

- $\{\mu_i\}_{i=1}^K$ инициализируются случайными объектами
  - распределение центров=распределению объектов
  - центры могут получиться слишком похожими
  - k-means++: μ_1-случайно, а далее

$$p(\mu_k = x_n) \propto \max_{i=1,2,...k-1} \|x_n - \mu_i\|^2, \quad k = 2, 3, ...K.$$

  - но если выброс, то кластер будет содержать только его
    - инициализировать медианами из нескольких случайных объектов

Содержание

K-средних - алгоритм

Инициализировать $\mu_k,\ k=1,2,...K$.

ПОВТОРЯТЬ до сходимости:

для $n=1,2,...N$:

определить кластер для x_i:

$z_n = \arg\min_{k\in\{1,2,...K\}} ||x_n-\mu_k||_2^2$

для $k=1,2,...K$:

пересчитать центры:

$\mu_k = \frac{1}{|C_k|}\sum_{n\in C_k} x_n$

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

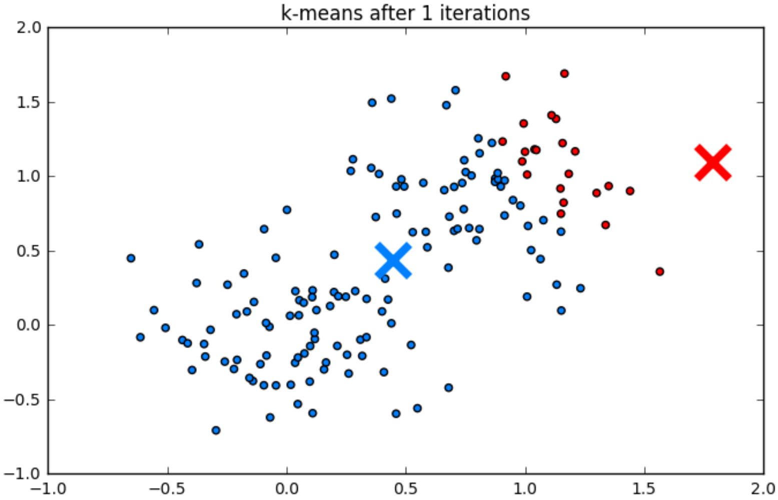
- Частичное обучение: если часть классов известна - фиксируем их и центры инициализируем по ним.

K-средних - динамический алгоритм

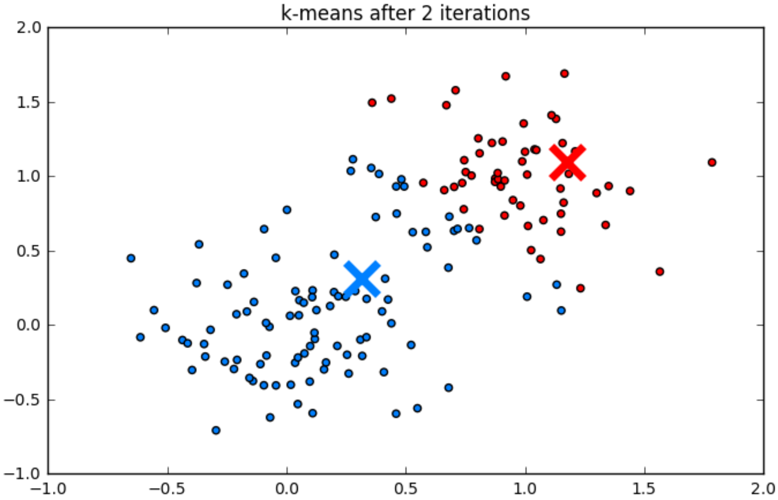
```
Инициализировать:
μ_k, k = 1,2,...K ,      z_n = -1, n = 1,2,...N .

ПОВТОРЯТЬ до сходимости:
    для n = 1,2,...N:
        определить кластер для x_n:
        z'_n = arg min_{k∈{1,2,...K}} ||x_n - μ_k||_2^2
        если z'_n != z_n:
            пересчитать центроиды кластеров
                z_n и z'_n # (как средние за O(1))
            z_n = z'_n
```

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

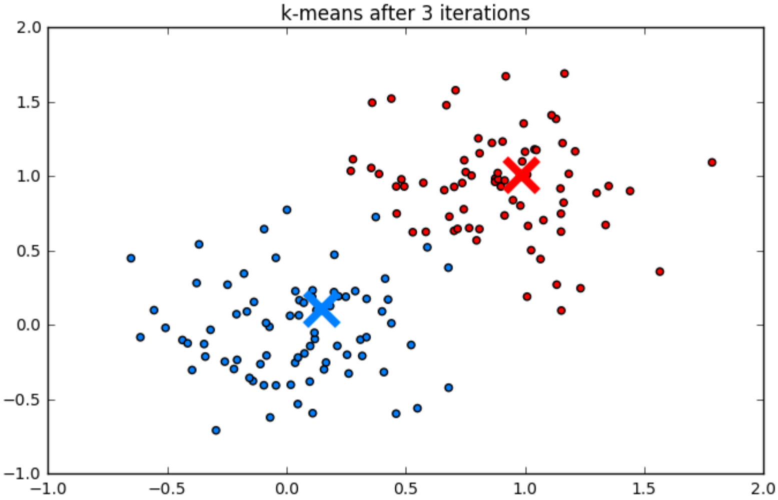
Пример работы K-средних



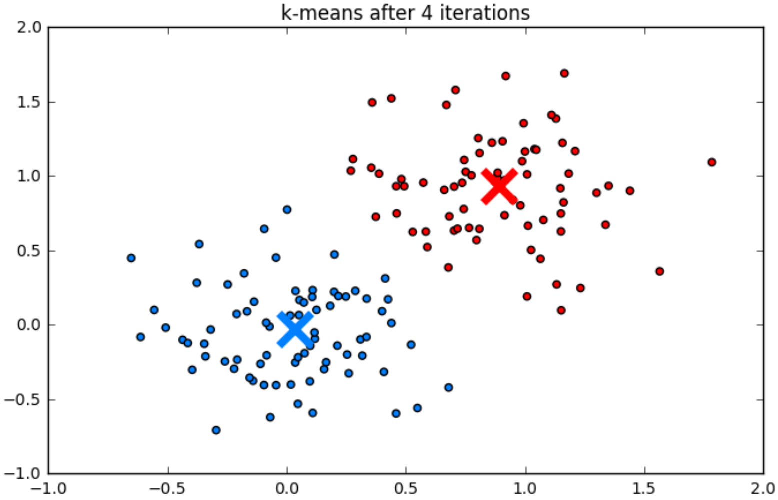
Пример работы K-средних



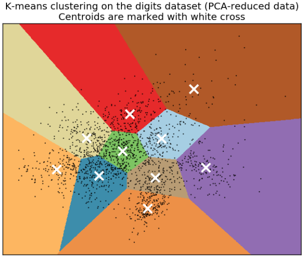
Пример работы K-средних



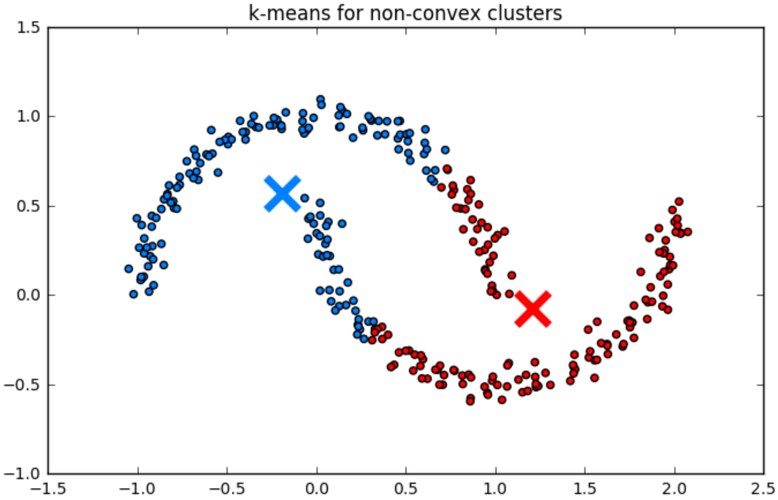
Пример работы K-средних



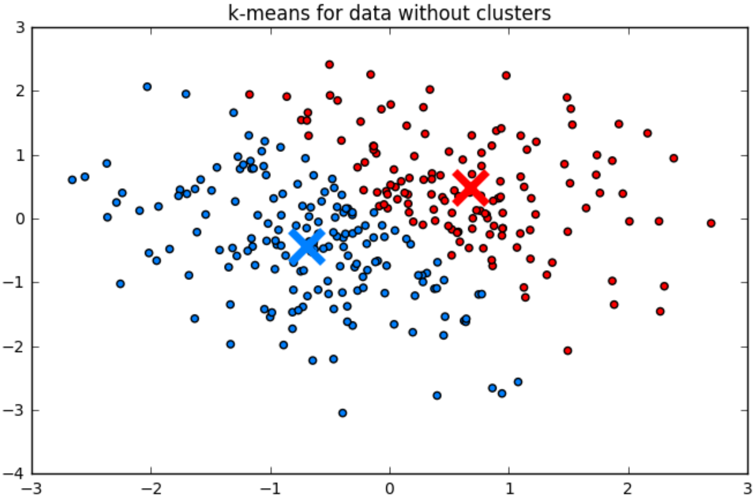
Пример кластеризации рукописных цифр



K-средних для невыпуклых кластеров



К-средних для равномерно распределенных данных



Mini-batch K-means

- Mini-batch K-means - для больших данных (как SGD).
- Обозначим $N(k)$=текущее #элементов кластера k.

```
Инициализировать μ_k, k = 1,2,...K.

ПОВТОРЯТЬ до сходимости:
    сэмлируем минибатч случайных объектов x'_b, b = 1,2,...B
    для b = 1,2,...B:
        определить кластер z_b для x'_b

    для b = 1,2,...B:
        обновить размер кластера: N(z_b) := N(z_b) + 1
        обновить центр кластера:
     μ_z(b) := (1 - 1/N(z_b)) μ_z(b) + 1/N(z_b) x'_b
```

Инициализировать $\mu_k,\ k=1,2,...K$.

ПОВТОРЯТЬ до сходимости:

сэмлируем минибатч случайных объектов $x'_b,\ b=1,2,...B$

для $b=1,2,...B$:

определить кластер z_b для x'_b

для $b=1,2,...B$:

обновить размер кластера: $N(z_b):=N(z_b)+1$

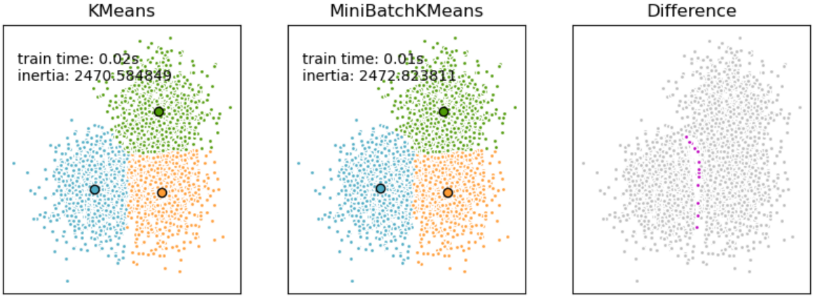
обновить центр кластера:

$$\mu_{z(b)} := \left(1-\frac{1}{N(z_b)}\right)\mu_{z(b)} + \frac{1}{N(z_b)}x'_b$$

K-means vs. mini-batch K-means

Mini-batch K-means ускоряет сходимость для больших данных

- ценой небольшого $\downarrow$ качества



K-means, K-means++, mini-batch K-means есть в sklearn.

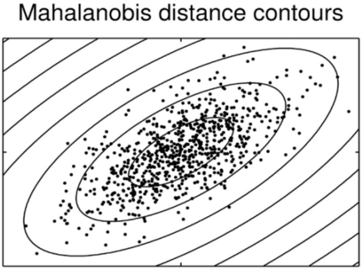
Содержание

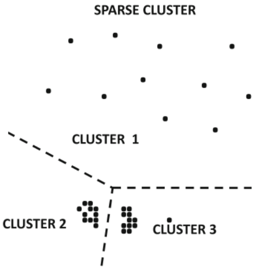
Расстояние Махаланобиса

- Расстояние Махаланобиса учитывает μ_k, Σ_k кластера:

$$\rho(x, \mu_k)^2 = (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)$$

- Это позволяет выделять кластеры эллиптической формы разного размера и плотности.





K-медоид - идея

- K медоид - K представителей, с ограничением, что центроидом м. быть только реальный объект
    - более интерпретируемо
    - если не можем усреднять объекты
        - например, временные ряды разной длины

K-медоид - алгоритм

инициализировать $\mu_1, ...\mu_K$ из случайных объектов

ПОВТОРЯТЬ до сходимости:
    для $n = 1, 2, ...N$:
        $z_n = \arg\min_k \rho(x_n, \mu_k)$

    для $k = 1, 2, ...K$:
        $\mu_k = \arg\min_{\mu \in \{x_n : z_n = k\}} \sum_{n: z_n = k} \rho(x_n, \mu)$

ВЕРНУТЬ $z_1, ...z_N$

сложность одной итерации $O(N^2)$

- из-за поиска центрального объекта каждого кластера

Ядерное обобщение K средних

- Мотивация: строить кластера более общей невыпуклой формы.
- Пусть $C_k := \{n : z_n = k\}$ - индексы объекта в кластере k.

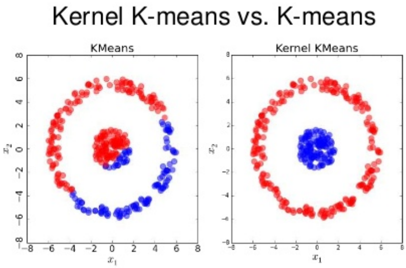
$$\rho(x,\mu_k)^2 = \|x-\mu_k\|^2 = \langle \varphi(x) - \frac{1}{|C_k|}\sum_{i\in C_k}\varphi(x_i),\ \varphi(x) - \frac{1}{|C_k|}\sum_{i\in C_k}\varphi(x_i)\rangle$$

$$= \langle \varphi(x),\varphi(x)\rangle - 2\langle \varphi(x),\ \frac{1}{|C_k|}\sum_{i\in C_k}\varphi(x_i)\rangle + \frac{1}{|C_k|^2}\sum_{i,j\in C_k}\langle \varphi(x_i),\ \varphi(x_j)\rangle$$

$$= K(x,x) - \underbrace{2\frac{1}{|C_k|}\sum_{i\in C_k}K(x,x_i)}_{\text{average similarity to cluster}} + \underbrace{\frac{1}{|C_k|^2}\sum_{i,j\in C_k}K(x_i,x_j)}_{\text{cluster compactness}}$$

```
инициализировать C_1, ...C_K

ПОВТОРЯТЬ до сходимости:
    для n = 1, 2, ...N :
        z_n = arg min_k ρ(x_n, μ_k)^2

ВЕРНУТЬ z_1, ...z_N
```

Ядерное обобщение K-средних



- Гауссово ядро (как пример): $K(x,\mu) = e^{-\gamma\|x-\mu\|^2}$
- Сложность: сложность каждой итерации $O(N^2)$, общая $O(N^2 I)$.
- Центроиды не вычисляются напрямую (не можем, используя $\langle\cdot,\cdot\rangle$)

Заключение

- Кластеризация - метод обучения без учителя со многими приложениями.
- K представителей - самый популярный метод кластеризации.
- Число кластеров K - гиперпараметр (задаётся пользователем)
- Важный параметр - $\rho(x, \mu)$, например

$$\|x - \mu_k\|_1, \quad (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)$$

- Обобщения: K-медиан, K-медоид, ядерное обобщение.