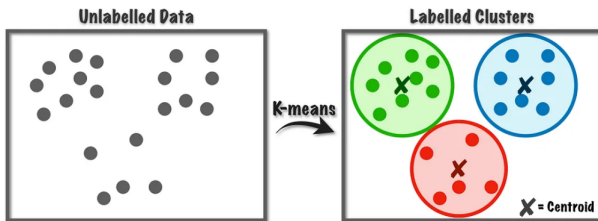


Кластеризация K представителями

Виктор Китов

v.v.kitov@yandex.ru

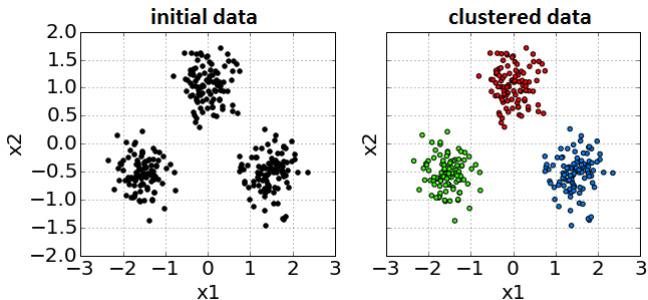


Содержание

- 1 Введение
- 2 Кластеризация, основанная на представителях
- 3 K-средних
- 4 Расширения K представителей
- 5 Определение качества кластеризации и #кластеров

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"

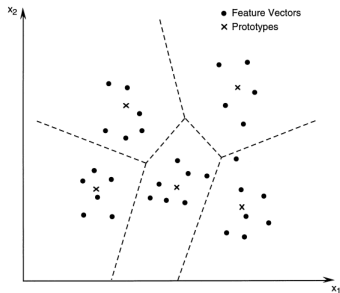
- разные метрики приводят к разным результатам

Применения кластеризации

- **классификация без обучающей выборки**
- **сегментация клиентов**
 - например, для более таргетированных 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**
 - используется в KNN, частичном обучении, активном обучении, рекомендательных системах
- **извлечение новых признаков**
 - номер кластера, расстояние до своего и ближайшего чужого кластера

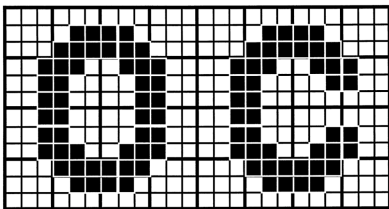
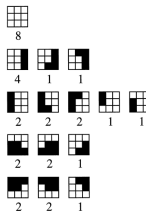
Сжатие данных¹

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



¹Источник иллюстраций.

Сжатие данных²

Feature Vectors with
No. OccurrencesPrototype
No. Occurrences

- В каждом фрагменте 3×3 : $2^9 = 512$ 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

²Источник 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

- 1 Введение
- 2 Кластеризация, основанная на представителях
- 3 K-средних
- 4 Расширения K представителей
- 5 Определение качества кластеризации и #кластеров

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

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

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

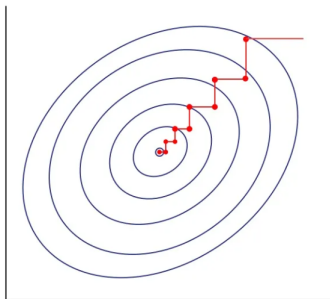
$$\mathcal{L}(z_1, \dots, z_N; \mu_1, \dots, \mu_K) = \sum_{n=1}^N \rho(x_n, \mu_{z_n}) \rightarrow \min_{z_1, \dots, z_N; \mu_1, \dots, \mu_K}$$

Метод оптимизации

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

Находится локальный оптимум методом покоординатного
спуска (μ, z, μ, z, \dots)

Coordinate Descent Convergence



Общий алгоритм

инициализировать μ_1, \dots, μ_K
(случайными объектами выборки)

ПОВТОРЯТЬ по сходимости:

для $n = 1, 2, \dots, N$:

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

для $k = 1, 2, \dots, K$:

$$\mu_k = \arg \min_{\mu} \sum_{n \in C_k} \rho(x_n, \mu)$$

ВЕРНУТЬ z_1, \dots, 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:
 - $\rho(x, x') = \|x - x'\|_2^2 \Rightarrow$ K-средних
 - μ_k - среднее
 - неустойчиво к выбросам
 - $\rho(x, x') = \|x - x'\|_1 \Rightarrow$ K-медиан
 - μ_k - медиана
 - устойчива к выбросам
- μ_k может выбираться только среди существующих объектов
 - например, временные ряды разной длины - не можем усреднять
- Форма кластеров определяется $\rho(\cdot, \cdot)$

Комментарии

Условия сходимости:

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

Оптимальность:

- критерий содержит много локальных оптимумов
- можно запустить 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

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

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

Содержание

- 1 Введение
- 2 Кластеризация, основанная на представителях
- 3 K-средних**
- 4 Расширения K представителей
- 5 Определение качества кластеризации и #кластеров

K-средних - алгоритм

Инициализировать μ_k , $k = 1, 2, \dots, K$.

ПОВТОРЯТЬ до сходимости:

для $n = 1, 2, \dots, N$:

определить кластер для x_i :

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

для $k = 1, 2, \dots, K$:

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 :

$$\mu_k, \quad k = 1, 2, \dots, K, \quad z_n = -1, \quad n = 1, 2, \dots, N.$$

ПОВТОРЯТЬ до сходимости :

для $n = 1, 2, \dots, N$:

определить кластер для x_n :

$$z'_n = \arg \min_{k \in \{1, 2, \dots, K\}} \|x_n - \mu_k\|_2^2$$

если $z'_n \neq z_n$:

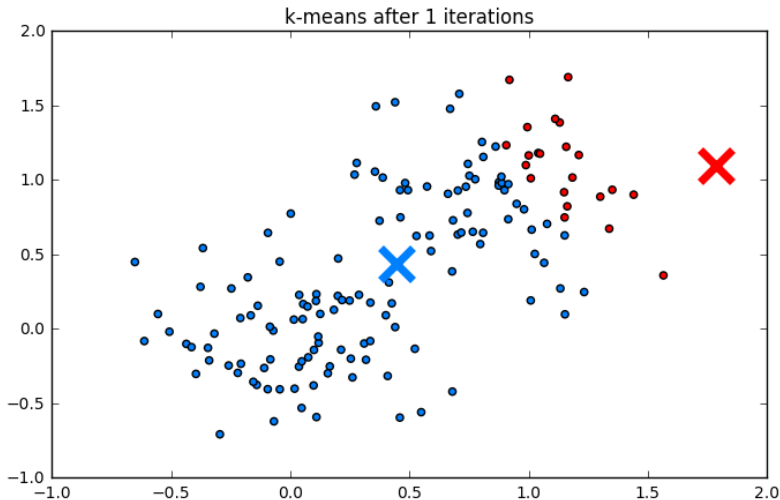
пересчитать центры кластеров

$$z_n \text{ и } z'_n \# \text{ (как средние за } O(1))$$

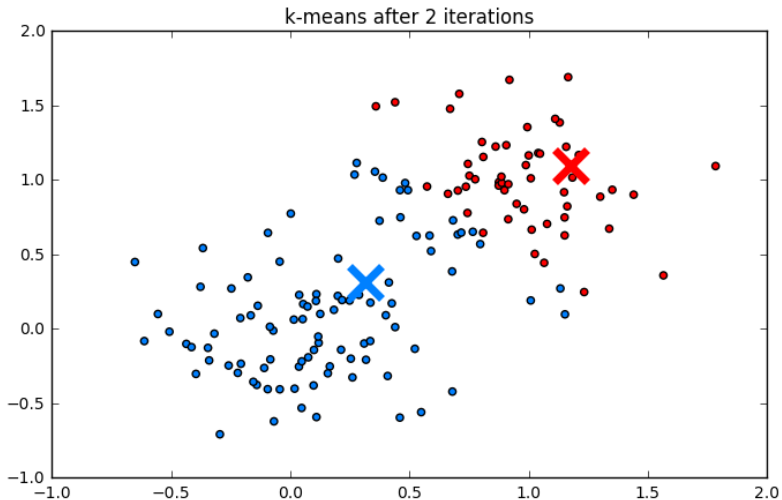
$$z_n = z'_n$$

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

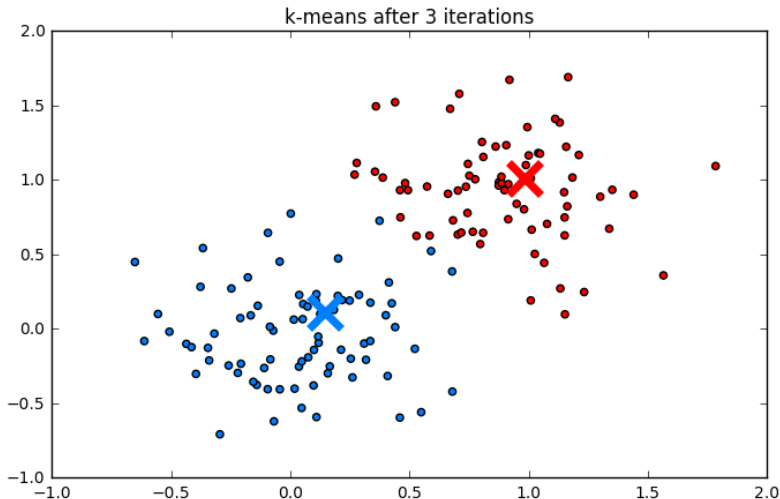
Пример работы К-средних



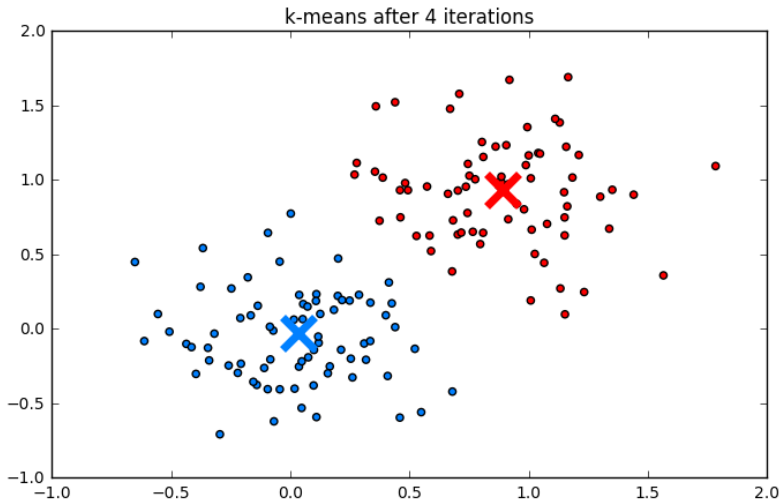
Пример работы K-средних



Пример работы К-средних



Пример работы К-средних

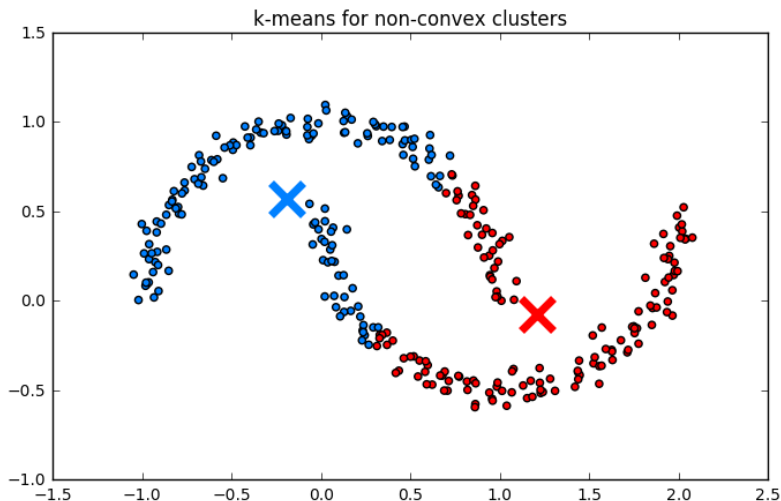


Пример кластеризации рукописных цифр

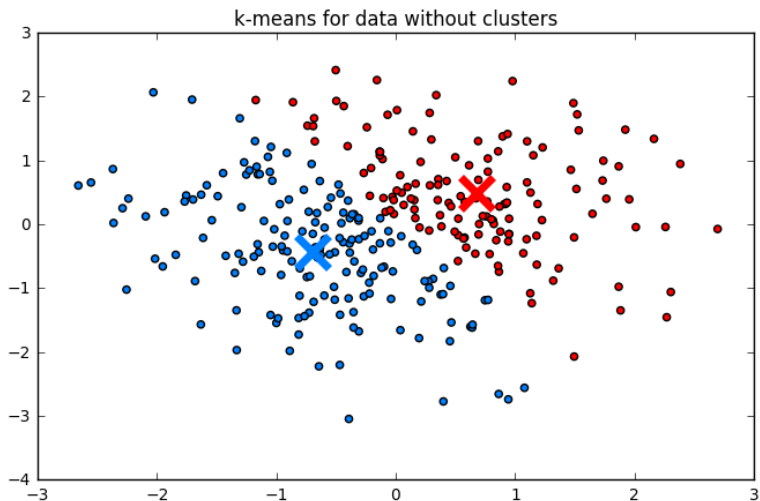
K-means clustering on the digits dataset (PCA-reduced data)
Centroids are marked with white cross



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 , $k = 1, 2, \dots, K$.

ПОВТОРЯТЬ до сходимости:

сэмплируем минибатч случайных объектов x'_b , $b = 1, 2, \dots, B$
для $b = 1, 2, \dots, B$:

определить кластер z_b для x'_b

для $b = 1, 2, \dots, B$:

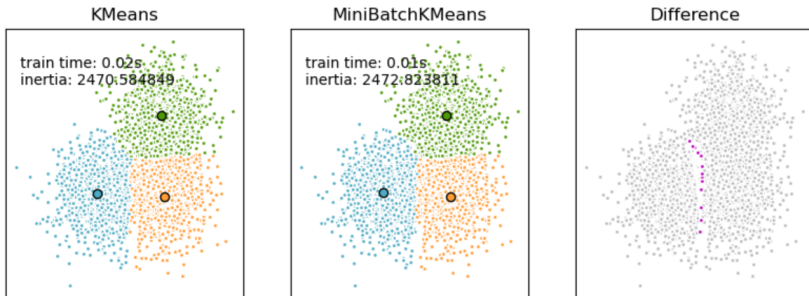
обновить размер кластера: $N(z_b) := N(z_b) + 1$

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

K-means vs. mini-batch K-means

Mini-batch K-means ускоряет сходимость для больших данных

- ценой небольшого ↓ качества



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

Содержание

- 1 Введение
- 2 Кластеризация, основанная на представителях
- 3 K-средних
- 4 Расширения K представителей**
- 5 Определение качества кластеризации и #кластеров

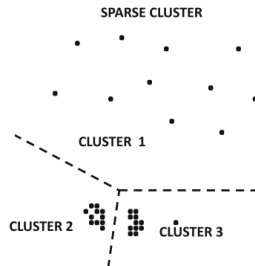
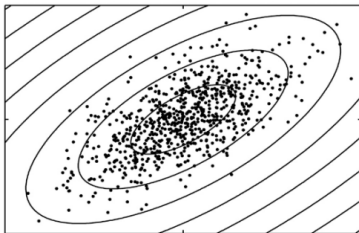
Расстояние Махаланобиса

- 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.

Mahalanobis distance contours



K-медоид - идея

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

K-медоид - алгоритм

инициализировать μ_1, \dots, μ_K из случайных объектов

ПОВТОРЯТЬ до сходимости:

для $n = 1, 2, \dots, N$:

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

для $k = 1, 2, \dots, K$:

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

ВЕРНУТЬ z_1, \dots, 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 .

$$\begin{aligned}
 \rho(x, \mu_k)^2 &= \|x - \mu_k\|^2 = \langle \varphi(x) - \frac{1}{|C_k|} \sum_{i \in C_k} \varphi(x_i), \varphi(x) - \frac{1}{|C_k|} \sum_{i \in C_k} \varphi(x_i) \rangle \\
 &= \langle \varphi(x), \varphi(x) \rangle - 2 \langle \varphi(x), \frac{1}{|C_k|} \sum_{i \in C_k} \varphi(x_i) \rangle + \frac{1}{|C_k|^2} \sum_{i, j \in C_k} \langle \varphi(x_i), \varphi(x_j) \rangle \\
 &= K(x, x) - \underbrace{2 \frac{1}{|C_k|} \sum_{i \in C_k} K(x, x_i)}_{\text{average similarity to cluster}} + \underbrace{\frac{1}{|C_k|^2} \sum_{i, j \in C_k} K(x_i, x_j)}_{\text{cluster compactness}}
 \end{aligned}$$

инициализировать C_1, \dots, C_K

ПОВТОРЯТЬ до сходимости:

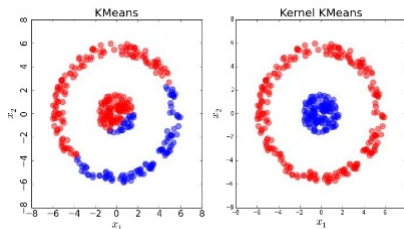
для $n = 1, 2, \dots, N$:

$$z_n = \arg \min_k \rho(x_n, \mu_k)^2$$

ВЕРНУТЬ z_1, \dots, z_N

Ядерное обобщение K-средних

Kernel K-means vs. K-means



- 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 $\langle \cdot, \cdot \rangle$)

Содержание

- 1 Введение
- 2 Кластеризация, основанная на представителях
- 3 K-средних
- 4 Расширения K представителей
- 5 Определение качества кластеризации и #кластеров

Алгоритм Monti consensus clustering³

- Генерируем H псевдовыборок D_1, D_2, \dots, D_H из X , кластеризуем каждую.
- На D_h (x_i, x_j) кластеризуются как (z_i^h, z_j^h) , $h \in M(i, j)$.
 - $M(i, j) = \{h : x_i \in D_h \ \& \ x_j \in D_h\}$
- Определим матрицу консенсуса (consensus matrix)

$$M(i, j) = \frac{\sum_{h \in M(i, j)} \mathbb{I}[z_i^h = z_j^h]}{|M(i, j)|}, \quad M \in \mathbb{R}^{N \times N}$$

- $M(i, j) \in \{0, 1\} \Rightarrow$ у точек (i, j) устойчивая класт-ция.
- $M(i, j) \in (0, 1) \Rightarrow$ у точек (i, j) неустойчивая класт-ция.
 - тем неустойчивее, чем ближе $M(i, j)$ к 0.5.

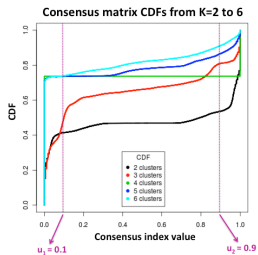
³https://en.wikipedia.org/wiki/Consensus_clustering

Алгоритм Monti consensus clustering

- Для каждого #кластеров K посчитаем пропорцию пар точек с неопределённой кластеризацией (proportion of ambiguous clustering, PAC)

$$PAC(K) = \frac{|\{(i,j) : i < j \text{ \& } 0.1 \leq M(i,j) \leq 0.9\}|}{C_2^N}$$

- $PAC(K) \in [0, 1]$ - мера неустойчивости кластеризации на K кластеров; $K^* = \arg \min_K PAC(K)$.



PAC for each K is $CDF_K(u_2) - CDF_K(u_1)$.
According to this criterion, optimal K here is 4.

Заключение

- Кластеризация - метод обучения без учителя со многими приложениями.
- К представителей - самый популярный метод кластеризации.
- Число кластеров К - 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: К-медиан, К-медоид, ядерное обобщение.
- Monti consensus clustering - определение качества кластеризации и оптимального К.