Кластеризация

Виктор Китов

v.v.kitov@yandex.ru

Содержание

Ядерное обобщение K средних

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

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

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

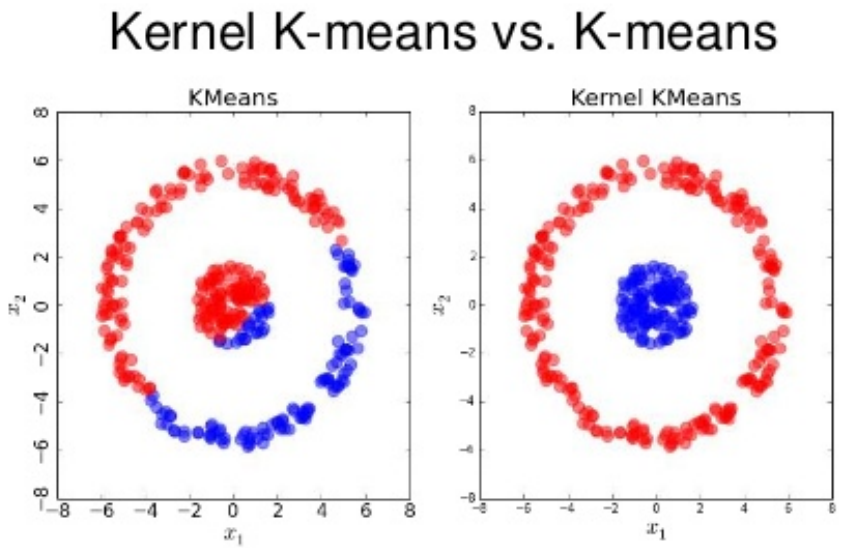
$$= K(x,x) - \underbrace{2\frac{1}{|C_k|}\sum_{i\in C_k}K(x,x_i)}_{\text{average similarity to cluster}} + \underbrace{\frac{1}{|C_k|^2}\sum_{i,j\in C_k}K(x_i,x_j)}_{\text{cluster compactness}}$$

инициализировать $C_1, \dots C_K$

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



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

K-медоид

- 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

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

K-медоид: рандомизированный

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

ПОВТОРЯТЬ до сходимости:
    сгенерировать кандидаты для замены $R = (\mu_{k(i)}, x_{n(i)})_{i=1}^{S}$
    выбрать замену из $\sum_{n=1}^{N} \min_k \rho(x_n, \mu_k) \to \min$

    если нет улучшения:
        восстановить предыдущую конфигурацию
        выйти

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

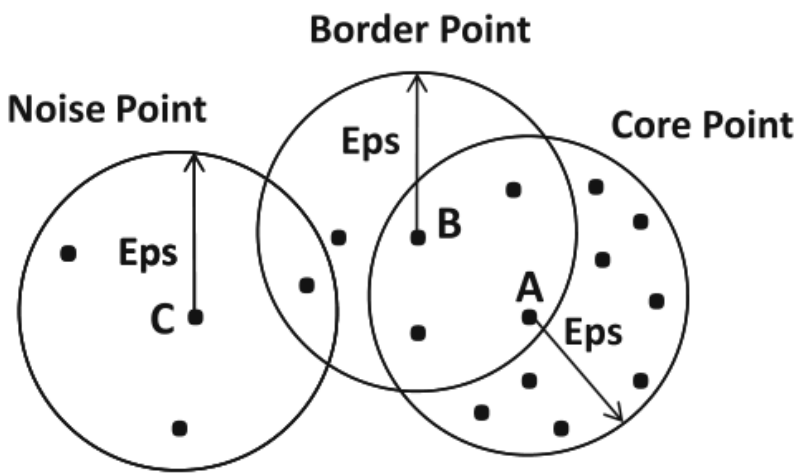
ВЕРНУТЬ $z_1, ...z_N$

Содержание

DBScan

k, ε - параметры метода.
Разделим множество объектов на 3 категории:

- основные точки: имеющие$\geq k$ точек внутри ε-окрестности
- пограничные точки: не основные, но содержащие хотя бы одну основную внутри ε-окрестности
- шумовые точки: не основные и не пограничные



Алгоритм

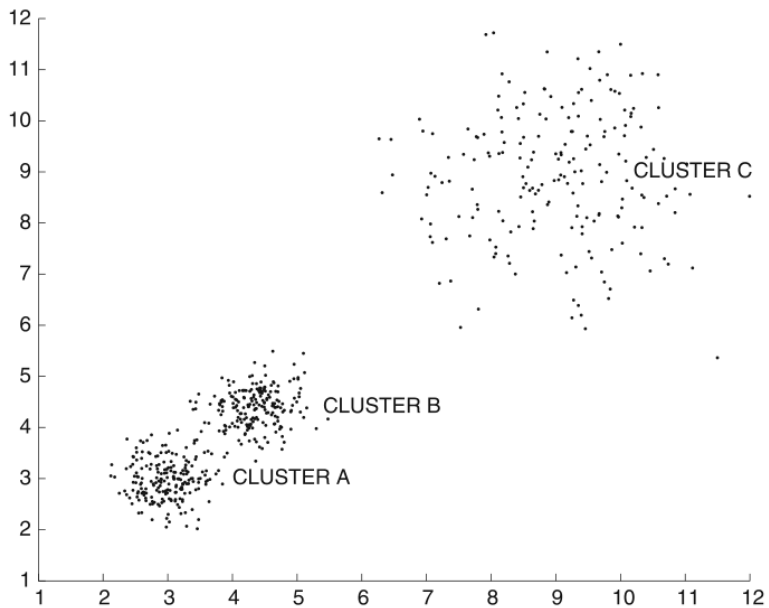
ВХОД: выборка, параметры ε, k.

1) Определить основные/пограничные/шумовые точки, используя ε, k.
2) Создать граф: узлы-основные точки, связи - если точки на расстоянии $\leq \varepsilon$ друг от друга.
3) Определить компоненты связности в графе =кластеры (методом распространения).
4) Соотнести основные точки кластерам=компонентам связности, а пограничные-по основным в их ε окрестности.

ВЫХОД: разбиение на кластеры (основных и пограничных точек)

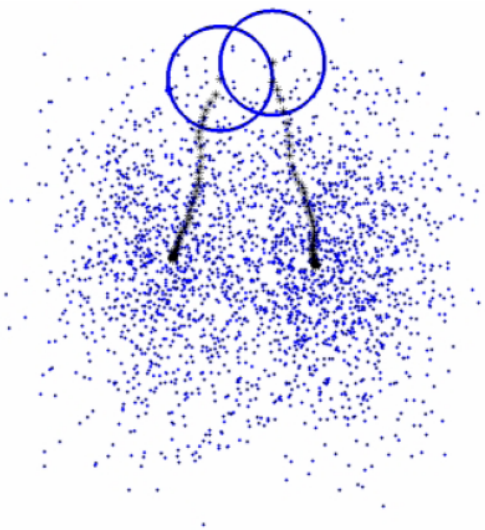
Комментарии

- Соединение основных точек - метод одиночной связи в агломеративной кластеризации с остановкой $\rho > \varepsilon$.
- Преимущества: автоматически определяется # кластеров, устойчиво к выбросам.
- Недостаток: не работает с кластерами разной плотности
  - высокое k-пропустим C; низкое k-A и B объединяться:



Кластеризация сдвигом среднего значения

Кластеризация сдвигом среднего значения (mean shift): точки итеративно сдвигаются в направлении локального увеличения плотности по правилу



Пример сходимости для top-hat ядра $K = \mathbb{I}\left[\frac{\rho(z,x)}{h} \leq 1\right]$

Кластер - итоговый локальный максимум плотности.

Комментарии

- Правило сдвига:

$$z_0 = x_n, \quad z = \frac{\sum_{k=1}^{N} K(\rho(z_i, x_k)/h) x_k}{\sum_{k=1}^{N} K(\rho(z, x_k)/h)}$$

- Ядро $K(\cdot)$ - убывающая ф-ция расстояния.
- Пример: Гауссово ядро

$$K(\rho(x, x')/h) = e^{-\rho(x,x')^2/h^2}$$

- Преимущества:
  - автоматически определяется #кластеров, кластеры могут быть произвольной формы
- Недостаток: вычислительная сложность, нет фильтрации выбросов

Кластеризация mean shift

ВХОД: выборка $x_1, ...x_N$, ядро $K(\cdot)$, ширина окна h.

ДЛЯ $n = 1, ...N$:

$z_0 = x_n, i = 0$

ПОВТОРЯТЬ до сходимости:

$$z_{i+1} = \frac{\sum_{k=1}^{N} K(\rho(z_i, x_k)/h) x_k}{\sum_{k=1}^{N} K(\rho(z, x_k)/h)}$$

$i = i + 1$

ассоциировать x_n пику z_i

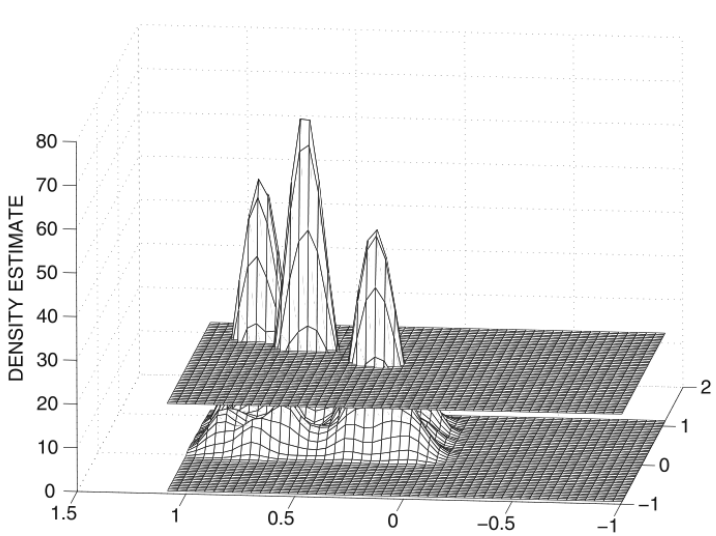
Объединить почти одинаковые расположения пиков $z_1, ...z_N$.

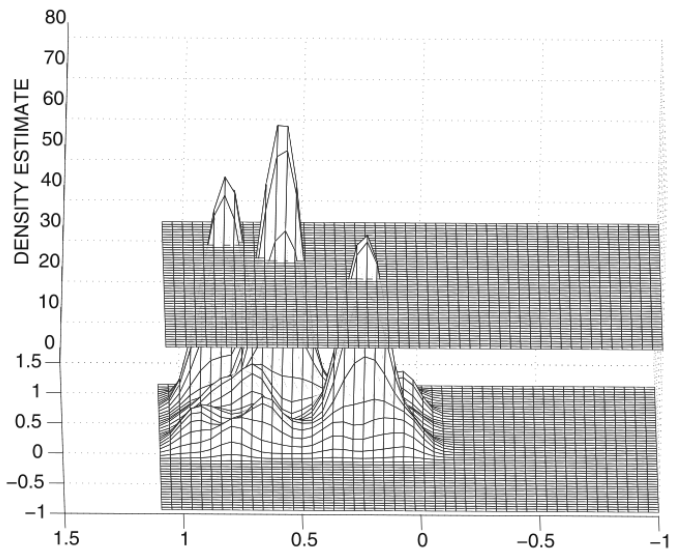
ВЕРНУТЬ кластеры точек, отнесенных одинаковым пикам плотности.

Расширение метода сдвига среднего значения: DENCLUE

1. Производим кластеризацию сдвигом среднего значения.
2. Отбрасываем кластеры с $p(x) < \tau$
3. Объединяем кластеры с пиками, соединяемые цепочкой высоко вероятных значений плотности $p(x_{i(k)}) \geq \tau$.

- варьируя τ можем получить иерархическую кластеризацию (без шага 2)





Содержание

Мотивация иерархической кластеризации

- #кластеров K заранее неизвестно.
- Кластеризация обычно не плоская, а иерархическая с разными уровнями детализации:
  - сайты в интернете
  - книги в библиотеке
  - животные в природе
- Подходы к иерархической кластеризации:
  - сверху вниз
    - более естественное для людей
  - снизу вверху (аггломеративная кластеризация)

Алгоритм

ВХОД:
выборка объектов, алгоритм плоской кластеризации A,
правила выбора листа и остановки

инициализировать дерево корнем, содержащим все объекты

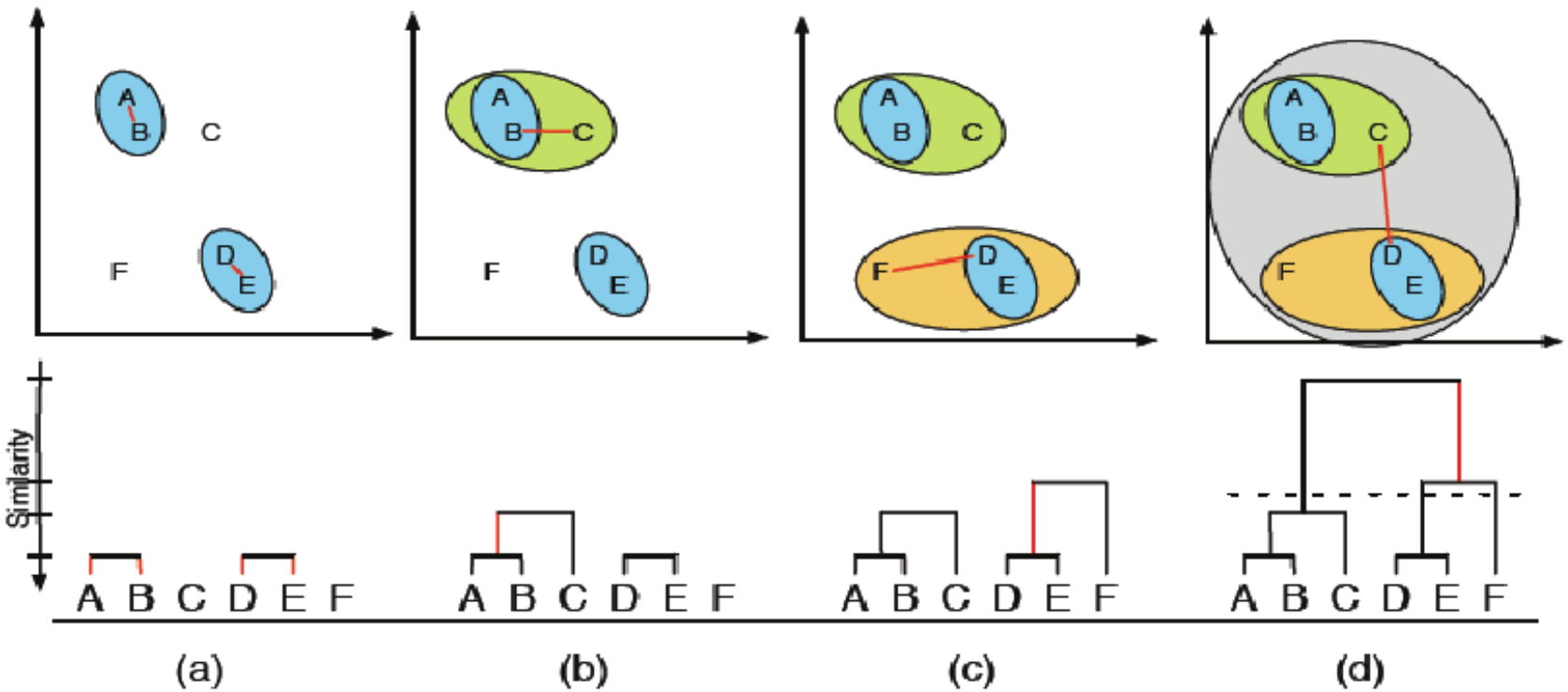
ПОВТОРЯТЬ
    выбрать лист L по правилу выбора листа
    используя A разбить L на кластеры $L_1, ...L_K$
    добавить листы к T, соответствующие $L_1, ...L_K$
ПОКА выполнено условие остановки

Комментарии

- Алгоритм выбора листа:
  - ближайший к корню
    => сбалансированное дерево по высоте
  - с максимальным числом элементов
    => сбалансированное дерево по #объектов в листах

3 Иерархическая кластеризация
- Иерархическая кластеризация сверху вниз
- Иерархическая кластеризация снизу вверх

Иерархическая кластеризация снизу вверх



Алгоритм

инициализировать матрицу попарных расстояний $M \in \mathbb{R}^{N \times N}$ между кластерами из отдельных объектов $\{x_1\}, ...\{x_N\}$

ПОВТОРЯТЬ:

1) выбрать ближайшие кластеры i и j
2) объединить $i, j \to \{i + j\}$
3) удалить строки/столбцы i, j из M
4) добавить строку/столбец для нового $\{i + j\}$

ПОКА не выполнено условие остановки

ВЕРНУТЬ иерархическую кластеризацию

- Условие остановки:
  - Остался 1 кластер либо осталось $\leq K$ кластеров
  - расстояние между ближайшими кластерами$\geq$порога.

Расстояние между кластерами

- Расстояние между объектами => расстояние между кластерами:
  - Метод одиночной связи (single linkage)
$$\rho(A,B) = \min_{a\in A, b\in B} \rho(a,b)$$
  - Метод полной связи (complete linkage)
$$\rho(A,B) = \max_{a\in A, b\in B} \rho(a,b)$$
  - Метод средней связи (group average link)
$$\rho(A,B) = \text{mean}_{\,a\in A, b\in B}\rho(a,b)$$
  - Центроидный метод (pair-group method using the centroid average)
$$\rho(A,B) = \rho(\mu_A,\mu_B)$$
где $\mu_U = \frac{1}{|U|}\sum_{x\in U} x$ или $m_U = median_{x\in U}\{x\}$

Свойства межкластерных расстояний[2]

- Метод одиночной связи
  - извлекает кластеры произвольной формы
  - может случайно объединить разные кластеры цепочкой выбросов
  - $M_{(i\cup j)k} = \min\{M_{ik}, M_{jk}\}$
- Метод полной связи
  - создает компактные кластеры
  - $M_{(i\cup j)k} = \max\{M_{ik}, M_{jk}\}$
- Метод средней связи[1] и центроидный метод-компромисс между одиночной и полной связью.

[1] Как $M_{(i\cup j)k}$ будет пересчитываться для него?

[2] Пусть мы модифицируем $\rho(x, x')$ монотонным преобразованием F: $\rho'(x, x') = F(\rho(x, x'))$. Which of the cluster distances will not be affected by this change?

Свойства межкластерных расстояний

Метод средней связи предпочтительнее центроидного, поскольку

- центроидный метод может приводить к немонотонной последовательности расстояний дендрограммы.
  - методы одиночной, полной и средней связи дают монотонную последовательность
- представление кластера его центром не учитывает структуру кластера
- центроидный метод предпочитает более крупные кластера, для которых центроиды получаются в среднем ближе

Комбинация K-средних и агломеративной

- Сложность агломеративной кластеризации K объектов: $O(K^2 \ln K)$
  - через алгоритм кучи
- Для снижения вычислений:
  1. применим K средних к N объектам (сложность $O(N)$)
  2. применим агломеративную кластеризацию к найденным K кластерам

  - она позволяет выделять невыпуклые кластера

Содержание

Оценка качества кластеризации

Оценка качества кластеризации:

- если кластеризация-промежуточный этап, то см. качество итоговой задачи
- если есть разметка
  - нужно учитывать инвариантность к переименованию кластеров
  - имеет смысл для небольшого #размеченных объектов
    - иначе решить средствами классификации
- Критерии без разметки
  - используют идею, что кластеризация хороша, если:
    - объекты одного кластера похожи
    - объекты разных кластеров непохожи

Коэффициент силуэта[3]

Качество кластеризации каждого объекта x_i определим по формуле:

$$Silhouette_i = \frac{d_i - s_i}{\max\{d_i, s_i\}}$$

где среднее расстояние от x_i до объектов

- s_i- того же кластера
- d_i-ближайшего чужого кластера

Общее качество классификации (коэффициент силуэта):

$$Silhouette = \frac{1}{N}\sum_{i=1}^{N}\frac{d_i - s_i}{\max\{d_i, s_i\}}$$

[3] Rousseeuw (1987). "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". Computational and Applied Mathematics 20: 53–65.

Обсуждение

- Преимущества
  - Интерпретируемость: *Silhouette* $\in [-1, 1]$,
    - 1: идеальная кластеризация
    - 0: случайная кластеризация
    - -1:послонстью некорректная (инвертированная) кластеризация
- Недостатки
  - сложность $O(N^2 D)$
    - можно рассчитывать по случайной подвыборке
  - поощряет выпуклые кластеры

Разброс ковариационной матрицы

Для случайной величины $x \in \mathbb{R}^D$, $x \sim F(\mu,\Sigma)$, и $\forall \alpha \in \mathbb{R}^D$:

$$var(\alpha^T x) = \mathbb{E}\left\{\left(\alpha^T x - \alpha^T \mu\right)^2\right\}$$
$$= \mathbb{E}\left\{\left(\alpha^T x - \alpha^T \mu\right)\left(x^T \alpha - \mu^T \alpha\right)\right\}$$
$$= \alpha \mathbb{E}\left\{(x-\mu)(x-\mu)^T\right\}\alpha = \alpha^T \Sigma \alpha$$
$$= /\text{спектральное разложение}/ = \alpha^T P \Lambda P^T \alpha =$$
$$= \left(\Lambda^{1/2} P^T \alpha\right)^T \left(\Lambda^{1/2} P^T \alpha\right) = \left\|\Lambda^{1/2} P^T \alpha\right\|^2$$

Дисперсия вдоль СВ Σ равна $\lambda_1, \lambda_2, ...\lambda_D$.
Следовательно, разброс Σ определяется $\sum_i \lambda_i = \text{tr}\,\Sigma$

Индекс Калинского[4]

- Рассмотрим K кластеров. Для кластера $k = 1, 2, ...K$ определим
  - I_k - индексы объектов кластера k, $N_k = |I_k|$, $N = \sum_k N_k$
  - μ_k - центроид кластера k, $\mu = \frac{1}{N}\sum_{n=1}^{N} x_n = \frac{\sum_{k=1}^{K} N_k \mu_k}{\sum_{k=1}^{K} N_k}$ - общий центр

[4] Caliński, T., & Harabasz, J. (1974). "A dendrite method for cluster analysis". Communications in Statistics-theory and Methods 3: 1-27.

Индекс Калинского

- Внутрикластерная (within cluster) ковариационная матрица

$$W = \frac{1}{N-K} \sum_{k=1}^{K} \sum_{x \in I_k} (x - \mu_k)(x - \mu_k)^T$$

- Межкластерная (between cluster) ковариационная матрица

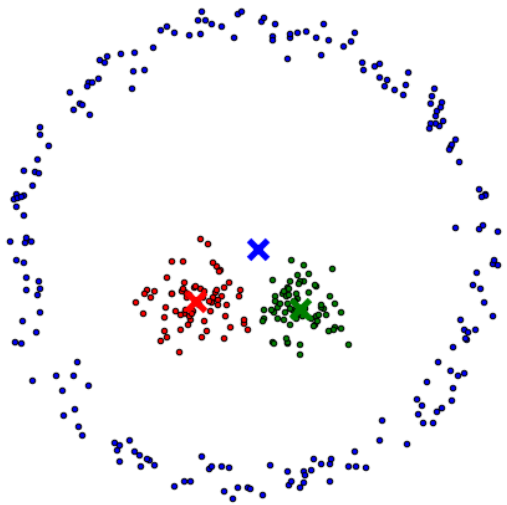
$$B = \frac{1}{K-1} \sum_{k=1}^{K} N_k (\mu_k - \mu)(\mu_k - \mu)^T$$

- Индекс Калинского:

$$I = \frac{\operatorname{tr} B}{\operatorname{tr} W}$$

- Сложность $O(ND)$, но поощряет выпуклые кластеры.

Ограничение для невыпуклого кластера



Из-за невыпуклости синего кластера коэффициент силуэта и индекс Калинского будут занижать хорошее качество кластеризацииб т.к.

- s_i велико, а d_i - мало
- $\text{tr}\, B$ мало, а $\text{tr}\, W$ велико

Заключение

- Плоская кластеризация:
  - K представителей
    - μ_k - вычисляемый (среднее: K-means [доступно ядерное обобщение], медиана: K medians)
    - μ_k - существующий объект
  - Основанная на плотности
    - DB-scan, mean-shift, DENCLUE
- Иерархическая кластеризация
  - сверху-вниз: рекурсивная плоская кластеризация
  - снизу-вверх (агломеративная)
- Оценка качества кластеризации:
  - коэффициент силуэта, индекс Калинского