Spectral Clustering

На основе статьи "A Tutorial on Spectral Clustering", Ulrike von Luxburg, 2007

Загонов Дмитрий

Национальный исследовательский университет
"Высшая школа экономики"

15 декабря 2025 г.

Метод k-средних (k-means)

Задача. Даны точки $x_1, \ldots, x_n \in \mathbb{R}^d$. Требуется разбить их на k кластеров.

Оптимизационная постановка.

$$\min_{C_1,\ldots,C_k} \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2, \qquad \mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$$

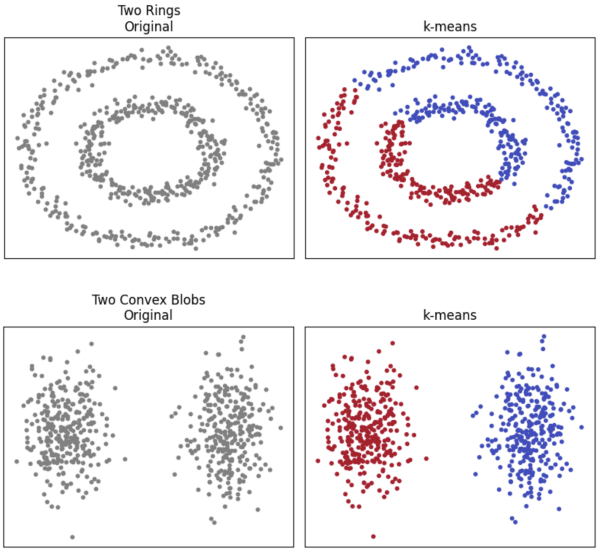
Алгоритм (Ллойд).

- Инициализировать центры $\mu_1, \ldots, \mu_k$
- *Assignment:* каждая точка $\to$ ближайший центр
- *Update:* μ_i — среднее точек кластера
- Повторять до сходимости

Геометрическая интерпретация.

- Кластеры — выпуклые области
- Границы — линейные (ячейки Вороного)
- **Главный минус: не может распознавать кластеры сложной формы.**

Демонстрация работы k-means



Мотивация спектральной кластеризации

Проблема k-means.

- Использует только координаты точек в $\mathbb{R}^d$
- Кластеры должны быть выпуклыми
- Плохо работает для сложной геометрии данных

Ключевая идея.

- Важны не координаты, а *связи* между точками
- Опишем данные как граф сходства

Граф сходства.

- Вершины — точки данных
- Рёбра — мера близости (например, kNN или Gaussian kernel)

Интуиция кластеризации.

- Внутри кластера — сильные связи
- Между кластерами — слабые связи

Построение графа связности

Исходные данные. Даны точки $x_1, \ldots, x_n$ и функция сходства $s(x_i, x_j)$. Пример функции сходства - гауссовское ядро: $s(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$.

Способы построения графа.

- **ε-граф:**

$$w_{ij} = \begin{cases} s(x_i, x_j), & \|x_i - x_j\| \leq \varepsilon \\ 0, & \text{иначе} \end{cases}$$

- **k ближайших соседей (kNN-граф):**
  - соединяем x_i с его k ближайшими соседями
  - граф может быть ориентированным или симметризованным
- **Полносвязный граф:**

$$w_{ij} = s(x_i, x_j) \quad \text{для всех } i \neq j$$

Результат. На основе полученного графа строится *матрица смежности* $W = (w_{ij})$.

Ненормированный лапласиан графа

Матрицы графа.

- Матрица смежности: $W = (w_{ij})$
- Матрица степеней:

$$D = \mathrm{diag}(d_1, \ldots, d_n), \qquad d_i = \sum_{j=1}^{n} w_{ij}$$

Определение.

$$L := D - W$$

называется ненормированным лапласианом графа.

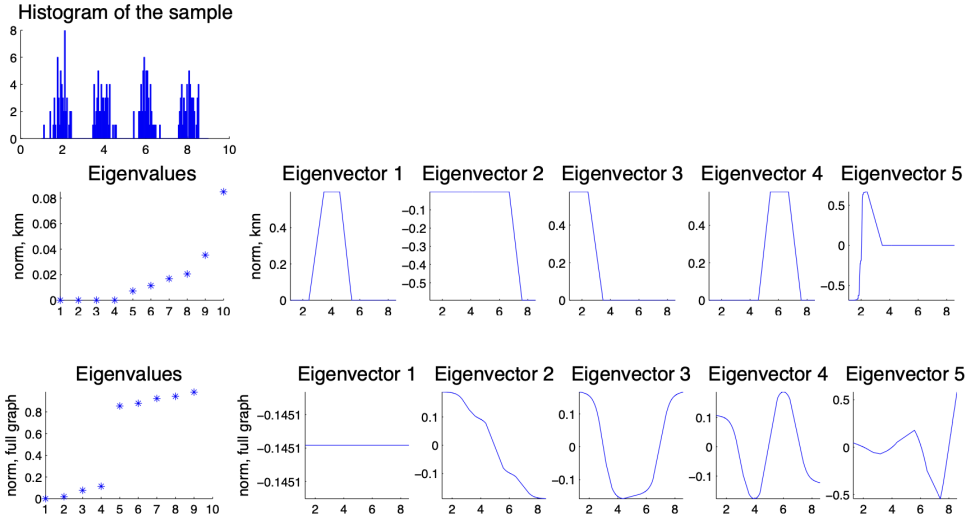
Свойства ненормированного лапласиана

- L — симметричная и неотрицательно определённая матрица
- $$f^\top L f = \frac{1}{2} \sum_{i,j} w_{ij}(f_i - f_j)^2, \quad f \in \mathbb{R}^n$$
- Наименьшее собственное значение - 0, соответствующий собственный вектор — константный вектор $\mathbb{1}$.
- L имеет n неотрицательных собственных значений: $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$.

Теорема.

Пусть G - неориентированный граф с неотрицательными весами рёбер. Тогда число компонент связности в графе равно кратности собственного значения 0 его лапласиана L. Собственное подпространство, соответствующее собственному значению 0, порождается индикаторными векторами этих компонент.

Демонстрация собственных векторов лапласиана



Кластеризация как разбиение графа

Интуиция кластеризации.

- Точки внутри одного кластера должны быть похожи
- Точки из разных кластеров — непохожи

Графовое представление данных.

- Вершины — объекты данных
- Веса рёбер w_{ij} — мера сходства

Задача. Найти разбиение графа на группы так, чтобы

- рёбра между группами имели малый вес,
- рёбра внутри групп — большой вес.

Подход. Рассматривать кластеризацию как задачу разбиения графа.

Минимальный разрез (mincut)

Обозначения. Для $A, B \subset V$ определим

$$W(A, B) := \sum_{i \in A} \sum_{j \in B} w_{ij}.$$

Разбиение на k кластеров. Пусть $V = A_1 \cup \cdots \cup A_k$.

Определение разреза.

$$\text{cut}(A_1, \ldots, A_k) := \frac{1}{2} \sum_{i=1}^{k} W(A_i, \bar{A}_i).$$

Идея mincut. Найти разбиение, минимизирующее суммарный вес рёбер между различными кластерами.

Проблема. На практике mincut часто отделяет одну вершину от остального графа, что не соответствует интуиции кластеризации.

Балансировка кластеров: RatioCut

Причина проблемы mincut.

- Размер кластеров никак не учитывается
- Малые множества оказываются выгодными

Идея. Явно требовать, чтобы кластеры были *достаточно большими*.

Определение RatioCut.

Для разбиения $V = A_1 \cup \cdots \cup A_k$:

$$\mathrm{RatioCut}(A_1, \ldots, A_k) := \sum_{i=1}^{k} \frac{\mathrm{cut}(A_i, \bar{A}_i)}{|A_i|}.$$

Интуиция.

- Малые кластеры сильно штрафуются
- Предпочтение отдаётся сбалансированным разбиениям

Минимизация RatioCut: переход к вектору f (случай $k = 2$)

Задача. Пусть $A \subset V$, $\bar{A} = V \setminus A$.

$$\mathrm{RatioCut}(A, \bar{A}) := \mathrm{cut}(A, \bar{A})\Big(\frac{1}{|A|} + \frac{1}{|\bar{A}|}\Big).$$

Дискретный вектор, кодирующий разбиение. Определим $f \in \mathbb{R}^n$:

$$f_i = \begin{cases} \sqrt{\frac{|\bar{A}|}{|A|}}, & i \in A, \\ -\sqrt{\frac{|A|}{|\bar{A}|}}, & i \in \bar{A}. \end{cases}$$

Свойства f.

$$f \perp 1, \qquad \|f\| = \sqrt{n}.$$

Идея. Минимизацию RatioCut по множествам A можно переписать как минимизацию квадратичной формы $f^\top L f$ по таким дискретным f.

Связь $f^\top Lf$ и RatioCut: идея доказательства

Напоминание. Для любого $f \in \mathbb{R}^n$:

$$f^\top Lf = \frac{1}{2} \sum_{i,j} w_{ij}(f_i - f_j)^2.$$

Шаг 1. Внутрикластерные рёбра. Если $i, j \in A$ или $i, j \in \bar{A}$, то

$$f_i = f_j \quad \Rightarrow \quad (f_i - f_j)^2 = 0.$$

Следствие. Рёбра внутри кластеров не дают вклада в $f^\top Lf$.

Шаг 2. Межкластерные рёбра. Ненулевой вклад дают только рёбра, соединяющие A и $\bar{A}$.

Связь $f^\top Lf$ и RatioCut: вычисление

Разность значений на разрезе. Для $i \in A$, $j \in \bar{A}$:

$$f_i - f_j = \sqrt{\frac{|\bar{A}|}{|A|}} + \sqrt{\frac{|A|}{|\bar{A}|}} = \frac{n}{\sqrt{|A||\bar{A}|}} \implies (f_i - f_j)^2 = \frac{n^2}{|A||\bar{A}|}.$$

Суммирование по рёбрам разреза. Так как

$$\sum_{i \in A,\, j \in \bar{A}} w_{ij} = \mathrm{cut}(A, \bar{A}),$$

получаем

$$f^\top Lf = \mathrm{cut}(A, \bar{A}) \cdot \frac{n^2}{|A||\bar{A}|}.$$

Следовательно,

$$f^\top Lf = n \cdot \mathrm{RatioCut}(A, \bar{A}).$$

От дискретной задачи к спектральной релаксации ($k = 2$)

Эквивалентная дискретная оптимизация.

$$\min_{A \subset V} \operatorname{RatioCut}(A, \bar{A}) \quad \Longleftrightarrow \quad \min_{f} f^\top L f$$

при ограничениях

$$f \perp 1, \qquad \|f\| = \sqrt{n}, \qquad f_i \in \{\alpha, -\beta\}.$$

Это дискретная (комбинаторная) задача, в общем случае NP-трудная.

Спектральная релаксация.

Снимаем дискретность и разрешаем $f \in \mathbb{R}^n$:

$$\min_{f \perp 1,\ \|f\| = \sqrt{n}} f^\top L f.$$

Решение. (Rayleigh-Ritz theorem) Минимум достигается на собственном векторе, соответствующем второму наименьшему собственному значению L.

Возврат к дискретному разбиению

Проблема. Вектор f принимает непрерывные значения, а кластеризация требует дискретного разбиения вершин.

Интуиция. Компоненты f близки внутри кластера и существенно различаются между кластерами.

Решение (приводит к ответу).

Используем k-means для разбиения элементов вектора f на два кластера.

Минимизация RatioCut при $k > 2$

Постановка задачи. Пусть $V = A_1 \cup \cdots \cup A_k$ — разбиение вершин графа на k кластеров.

$$\mathrm{RatioCut}(A_1, \ldots, A_k) = \sum_{i=1}^{k} \frac{\mathrm{cut}(A_i, \bar{A}_i)}{|A_i|}.$$

Индикаторная матрица. Разбиение кодируется матрицей

$$H = (h_1, \ldots, h_k) \in \mathbb{R}^{n \times k},$$

где каждый столбец h_i — нормированный индикатор кластера A_i.

Дискретная оптимизация. Минимизация RatioCut эквивалентна задаче

$$\min_{H} \ \mathrm{Tr}(H^\top L H)$$

при ограничениях:

$H^\top H = I, \qquad H_{ji} \in \left\{ 0, \ \frac{1}{\sqrt{|A_i|}} \right\}$, в каждой строке H ровно один ненулевой элемент.

Спектральная релаксация и алгоритм ($k > 2$)

Проблема. Эта задача является комбинаторной и NP-трудной.

Спектральная релаксация.

Снимаем дискретные ограничения и решаем

$$\min_{H \in \mathbb{R}^{n \times k}} \operatorname{Tr}(H^\top L H), \qquad \text{при } H^\top H = I.$$

Решение. (Rayleigh-Ritz theorem) Оптимальное H задаётся k собственными векторами лапласиана L, соответствующими наименьшим собственным значениям.

Спектральное вложение.

- Каждой вершине i сопоставляется строка $H_{i\cdot} \in \mathbb{R}^k$
- Вершины одного кластера имеют близкие строки

Возврат к дискретности (ответ).

Кластеризация строк матрицы H (например, методом k-средних) даёт итоговое разбиение графа.

Алгоритм спектральной кластеризации (ненормированный лапласиан)

Algorithm 1 Спектральная кластеризация (ненормированный лапласиан)

Require: Данные $x_1, \ldots, x_n$, число кластеров k, граф сходства W

Ensure: Разбиение вершин на k кластеров

1: Построить матрицу степеней D, где $D_{ii} = \sum_{j=1}^{n} w_{ij}$
2: Построить лапласиан $L = D - W$
3: Найти k собственных векторов $u_1, \ldots, u_k$ матрицы L, соответствующих наименьшим собственным значениям
4: Сформировать матрицу $U = (u_1, \ldots, u_k) \in \mathbb{R}^{n \times k}$
5: Применить алгоритм k-средних к строкам матрицы U
6: **return** Кластерная принадлежность вершин