

Лекция 17

Кластеризация

Е. А. Соколов, А. Гусев, И. Садртдинов
ФКН ВШЭ

21 февраля 2020 г.

Вернемся к одной из задач обучения без учителя — кластеризации. Пусть у нас есть выборка $X = \{x_i\}_{i=1}^l, x_i \in \mathbb{X}$, и мы хотим построить алгоритм $a : \mathbb{X} \rightarrow \{1, \dots, K\}$, который ставил бы в соответствие объекту номер кластера. Ранее предлагалось использовать номера кластеров как новый признак для обучения с учителем. Сегодня мы рассмотрим модель, которая будет генерировать для нас сколько угодно таких кластерных признаков.

1 Графовые методы

§1.1 От выборки к графу

В курсе МО-1 мы рассматривали несколько алгоритмов кластеризации, а именно:

- *K-Means* — метрический алгоритм, оптимизирующий внутрикластерное расстояние;
- *DBSCAN* — алгоритм, основанный на плотности расположения объектов;

Попробуем изобрести немного другой подход. Мы можем представить объекты из выборки в виде вершин некоторого неориентированного графа $G = (\mathcal{V}, \mathcal{E})$, $\mathcal{V} = X = \{x_1, \dots, x_l\}$. Рассмотрим несколько вариантов, как в таком представлении можно задать ребра \mathcal{E} :

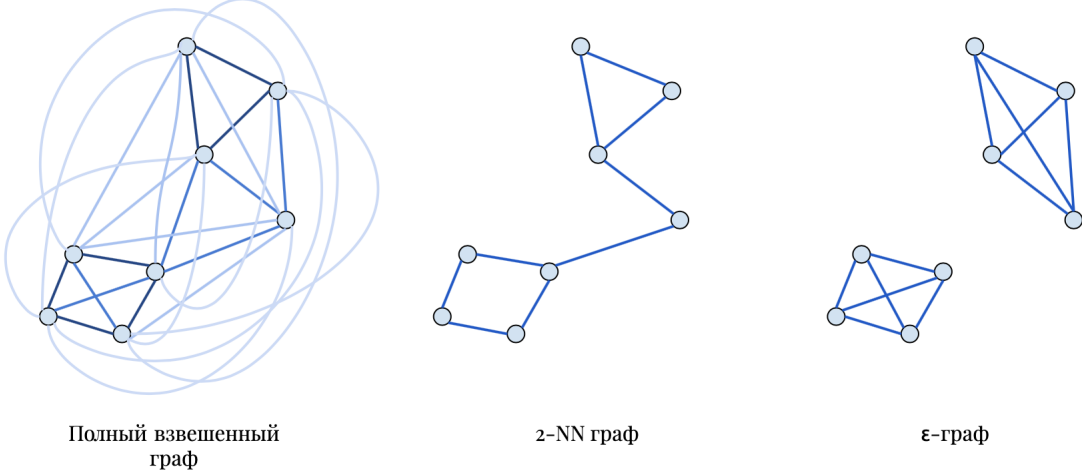
1. Граф G может быть полным с ребрами, вес которых определяется по некоторой формуле, например:

$$w_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$$

Гиперпараметр σ определяет, насколько нам важны далекие объекты.

2. G можно задать как kNN-граф, то есть объект x_i будет связан с k его ближайшими соседями.

3. Вершина x_i может быть связана с теми вершинами, расстояние до которых меньше выбранного ε , то есть $\rho(x_i, x_j) < \varepsilon$. Такой граф будет называться ε -графом.



§1.2 Лапласиан графа

Обозначим через W матрицу смежности графа G . Степени вершин будем считать как $d_i = \sum_{j=1}^l w_{ij}$. Пусть $D = \text{diag}(d_1, \dots, d_l)$. Тогда матрица $L = D - W$ будет называться *лапласианом* графа G . Рассмотрим несколько свойств лапласиана L :

1. Пусть $f \in \mathbb{R}^n$. Тогда имеет место следующая формула:

$$f^T L f = \frac{1}{2} \sum_{i,j=1}^l w_{ij} (f_i - f_j)^2$$

Проверка формулы:

$$\begin{aligned} f^T L f &= f^T D f - f^T W f = \sum_{i=1}^l d_i f_i^2 - \sum_{i,j=1}^l w_{ij} f_i f_j = \\ &= \sum_{i=1}^l \left(\sum_{j=1}^l w_{ij} \right) f_i^2 - \sum_{i,j=1}^l w_{ij} f_i f_j = \sum_{i,j=1}^l w_{ij} f_i^2 - \sum_{i,j=1}^l w_{ij} f_i f_j = \\ &= \frac{1}{2} \sum_{i,j=1}^l w_{ij} f_i^2 + \frac{1}{2} \sum_{i,j=1}^l w_{ij} f_i^2 - \sum_{i,j=1}^l w_{ij} f_i f_j = \\ &= \frac{1}{2} \sum_{i,j=1}^l w_{ij} f_i^2 + \frac{1}{2} \sum_{i,j=1}^l w_{ij} f_j^2 - \sum_{i,j=1}^l w_{ij} f_i f_j = \\ &= \frac{1}{2} \sum_{i,j=1}^l w_{ij} (f_i^2 - 2f_i f_j + f_j^2) = \frac{1}{2} \sum_{i,j=1}^l w_{ij} (f_i - f_j)^2 \end{aligned}$$

2. L — симметричная неотрицательно определенная матрица. Симметричность вытекает из неориентированности графа. Свойство неотрицательной определенности легко следует из первого пункта. Действительно, в обсужденных методах построения графа $w_{ij} \geq 0$, притом $(f_i - f_j)^2 \geq 0$. Следовательно, $f^\top L f \geq 0$, что и означает неотрицательную определенность.

Однако у лапласиана также есть свойство, которое поможет нам в задаче кластеризации. Сформулируем его в виде теоремы.

Теорема 1.1. Пусть L — лапласиан графа G . Тогда выполнены следующие два пункта:

1. Собственное значение $\lambda = 0$ матрицы L имеет кратность, равную числу компонент связности k .
2. Пусть A_1, \dots, A_k — компоненты связности графа G . Тогда векторы f_1, \dots, f_k , определяемые по формуле $f_i = ([x_j \in A_i])_{j=1}^l$, будут являться собственными векторами для $\lambda = 0$.

Доказательство.

Сперва рассмотрим случай $k = 1$. Поймем, почему $\lambda = 0$ вообще является собственным значением матрицы L . Для этого рассмотрим вектор $f = (1, \dots, 1)$:

$$\begin{aligned} Lf &= Df - Wf = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_l \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} - \begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1l} \\ w_{21} & w_{22} & \cdots & w_{2l} \\ \vdots & \vdots & \ddots & \vdots \\ w_{l1} & w_{l2} & \cdots & w_{ll} \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \\ &= \begin{pmatrix} d_1 \\ \vdots \\ d_l \end{pmatrix} - \begin{pmatrix} w_{11} + \cdots + w_{1l} \\ \vdots \\ w_{l1} + \cdots + w_{ll} \end{pmatrix} = 0 \end{aligned}$$

Теперь предположим, что существует собственный вектор $f' \in \mathbb{R}^n$: $\exists p \neq q \rightarrow f'_p \neq f'_q$, то есть неконстантный вектор, соответствующий $\lambda = 0$. Тогда $Lf' = 0 \Rightarrow f'^\top Lf' = 0$. Но рассматриваемый граф является связным. Значит существует путь $p = i_0 \rightarrow i_1 \rightarrow \cdots \rightarrow i_{n-1} \rightarrow i_n = q$. Поскольку вершины i_r и i_{r+1} соединены ребром, то $w_{i_r i_{r+1}} > 0$, а значит, $f'_{i_r} = f'_{i_{r+1}}$ (иначе получим $f'^\top Lf' > 0$). Отсюда $f'_p = f'_{i_1} = \cdots = f'_{i_{n-1}} = f'_q$ — константный вектор. Получили противоречие $\Rightarrow f'$ не является собственным вектором для $\lambda = 0$. Значит, мы доказали оба пункта для $k = 1$.

Теперь пусть $k > 1$. Можно упорядочить вершины так, чтобы лапласиан L стал блочно-диагональной матрицей:

$$L = \begin{pmatrix} L_1 & 0 & \cdots & 0 \\ 0 & L_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & L_k \end{pmatrix}$$

Блоки L_1, \dots, L_k будут являться лапласианами компонент связности графа $G \Rightarrow \lambda = 0$ имеет кратность k , а собственные векторы f_1, \dots, f_k задаются по формуле

$$f_i = ([x_j \in A_i])_{j=1}^l$$

■

Гипотеза 1.1. Пусть есть похожие объекты x_j и x_k , то есть расстояние между ними невелико. Тогда у собственных векторов f_i , соответствующих маленьким собственным значениям, выполнено $f_{ij} \approx f_{ik}$.

Эта гипотеза имеет лишь эмпирическое доказательство. Однако алгоритм, построенный на этой гипотезе, позволяет достигать успеха в задаче кластеризации. Данный алгоритм называется *спектральной кластеризацией*.

§1.3 Спектральная кластеризация

Алгоритм спектральной кластеризации:

Сложность шага

1. Строим по объектам граф G и лапласиан $L = D - W$. $\mathcal{O}(l^2)$
2. Находим нормированные собственные векторы u_1, \dots, u_m матрицы L , соответствующие m наименьшим собственным значениям. $\mathcal{O}(l^3)$
3. Составляем матрицу $U \in \mathbb{R}^{l \times m}$: $U = (u_1 | \dots | u_m)$. $\mathcal{O}(lm)$
4. Обучаем на этой матрице алгоритм К-Means с k кластерами. $\mathcal{O}(l^{mk+1})$

Если мы верим в нашу гипотезу, то для похожих объектов соответствующие координаты векторов u_1, \dots, u_m будут близки. Матрица U имеет размерность $l \times m$, поэтому можно посмотреть на нее, как на матрицу объекты-признаки. При этом новыми признаками будут как раз собственные вектора u_1, \dots, u_m . В описанном выше алгоритме предлагается обучать на них К-Means, но никто не запрещает использовать эти признаки для других задач.

Как видно, алгоритм обладает высокой вычислительной сложностью. Но утверждается, что он позволяет достигать впечатляющих результатов в задаче кластеризации. Однако как оценить, насколько хорошо алгоритм справляется с задачей классификации? Этому вопросу посвящен остаток лекции.

2 Оценка качества кластеризации

§2.1 Метрика на разметке

Пусть существует разметка (y_1, \dots, y_l) , не участвующая при обучении. Мы не использовали эту разметку в качестве дополнительного признака, так как нам не хочется мотивировать модель данным признаком. Тогда предлагается ввести оценку качества алгоритма кластеризации при помощи этой разметки, саму же разметку тогда называют *gold standard*. Введем несколько требований к внешней метрике качества Q :

1. *Гомогенность*. Базовое свойство разделения разных объектов в разные кластеры:

$$Q \left(\begin{array}{cc} & \diamond & \diamond \\ \times & & \diamond \\ \times & \times & \end{array} \right) < Q \left(\begin{array}{cc} & \diamond & \diamond \\ \times & & \diamond \\ \times & \times & \end{array} \right)$$

2. *Полнота*. Один кластер не должен дробиться на несколько маленьких:

$$Q \left(\begin{array}{cc} & \times & \times \\ \times & & \times \\ \times & \times & \end{array} \right) < Q \left(\begin{array}{cc} & \times & \times \\ \times & & \times \\ \times & \times & \end{array} \right)$$

3. *Rag-bag*. Весь мусор должен быть в одном "мусорном" кластере, чтобы остальные кластеры были "чистыми":

$$Q \left(\begin{array}{cc} \times & \times & \bullet & \circ \\ \times & \times & \triangleright & \star \\ \times & * & \odot & \square \end{array} \right) < Q \left(\begin{array}{cc} \times & \times & \bullet & \circ \\ \times & \times & \triangleright & \star \\ \times & * & \odot & \square \end{array} \right)$$

4. *Cluster size vs. quantity*. Лучше испортить один кластер с целью улучшить качество множества других:

$$Q \left(\begin{array}{cc} \times & \circ & \circ \\ \times & \star & \star \\ \times & \triangleright & \triangleright \\ \times & \odot & \odot \end{array} \right) < Q \left(\begin{array}{cc} \times & \circ & \circ \\ \times & \star & \star \\ \times & \triangleright & \triangleright \\ \times & \odot & \odot \end{array} \right)$$

§2.2 BCubed

Единственной известной на данный момент метрикой, обладающей всеми четыремя названными свойствами является BCubed. Она считается следующим способом. Пусть $L(x)$ — gold standard, $C(x)$ — номер кластера, выдаваемый рассматриваемым алгоритмом. Тогда рассмотрим несколько величин:

$$\text{Correctness}(x, x') = \begin{cases} 1 & , C(x) = C(x') \wedge L(x) = L(x') \\ 0 & , \text{otherwise} \end{cases}$$

$$\text{Precision-BCubed} = \text{Avg}_x [\text{Avg}_{x': C(x)=C(x')} \text{Correctness}(x, x')]$$

$$\text{Recall-BCubed} = \text{Avg}_x [\text{Avg}_{x': L(x)=L(x')} \text{Correctness}(x, x')]$$

Тогда F-мера от определенных точности и полноты будет удовлетворять всем нужным нам требованиям.

Пример 2.1.

$$\begin{array}{cc} & \times & \times \\ \times & & \times \\ \times & \times & \end{array} \quad \text{Precision-BCubed} = \frac{1}{2}, \quad \text{Recall-BCubed} = 1, \quad \text{BCubed} = \frac{1 \cdot \frac{1}{2}}{1 + \frac{1}{2}} = \frac{1}{3}$$