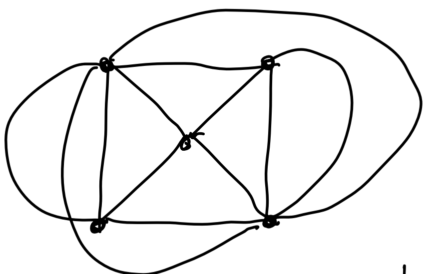
# Графовые методы кластеризации.

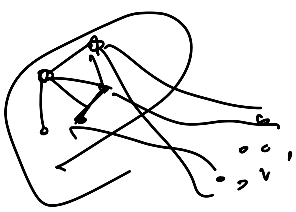
Представим данные в виде графа.

$G = (V, E)$

V — vertices, E — edges

1) Полный взвешенный граф.





1) Adjacency matrix ✓
2) Adjacency list ✗

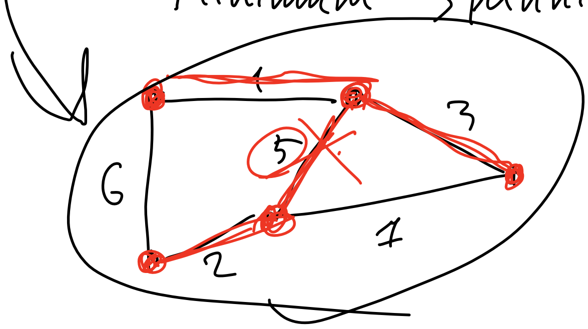
$$
\begin{array}{c c}
 & \xrightarrow{V} \\
V \big\downarrow & \begin{pmatrix} 0 & w_{ji} & \\ w_{ij} & 0 & \\ & & \ddots & 0 \end{pmatrix}
\end{array}
$$

Симметрия.
Неотр. опр.,
$w_{ij} \geq 0$

$w_{ij} = w_{ji}$

Spectral. ...

2) Кластеризация через MST
Minimum Spanning Tree

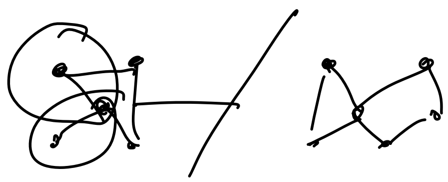



1) Строим граф (полный)
2) Строим MST
3) Убираем k-1 самых больших рёбер

3) $\varepsilon$-графы

Соединим в граф только те точки, которые лежат в $\varepsilon$-окрестности

Вершина $x_i$ может быть связана только с $x_j$ s.t. $\rho(x_i, x_j) < \varepsilon$



4) Соединим точку $x_i$ с её $p$ ближ. соседями $p$-NN граф

## Методы ускорения

1) Считать по батчам (k-means)

Пересчитываем центроиды не по всем данным

2) Считать не по всей выборке ???

3) NSW ?

4) Понизить размерность

5) Доп модель поверх обученной случайными

## Спектральная кластеризация

$G = (V, E)$ $n$ вершин

$W = (w_{ij})_{i,j=1,\dots,n}$

Степень вершины $v_i$:

$$d_i = \sum_{j=1}^{n} w_{ij}$$

$D$ – диагональная матрица $D_{ii} = d_i$

Ненормиров. лапласиан:

$$L = D - W$$

1) $f \in R^n$:

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

$$f^T L f = f^T D f - f^T W f =$$

$$= \sum_{i=1}^{n} d_i f_i^2 - \sum_{i,j} f_i f_j w_{ij} =$$

$$= \frac{1}{2} \left( \sum_{i=1}^{n} d_i f_i^2 - 2 \sum_{i,j} f_i f_j w_{ij} + \sum_{j=1}^{n} d_j f_j^2 \right) =$$

$$= \frac{1}{2} \left( \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \underline{f_i} - 2 \sum_{i,j} \underline{\underline{f_i f_j}} w_{ij} + \sum_{j=1}^{n} \sum_{i=1}^{n} w_{ij} \underline{f_j^2} \right) =$$

$$= \frac{1}{2} \sum_{i,j} \overset{\geqslant 0}{w_{ij}} \underbrace{(f_i - f_j)^2}_{\geqslant 0} \geqslant 0$$

2) $L$ – симметричная, положительно

полуопределённая матрица

3) Наименьшее собств. значение = 0 и соотв вектору $1 = (1, \dots, 1)^T$

Алгоритм:

1) $L = D - W$

2) Находим С.В. $u_1, \dots, u_n$

3) Выбираем $k$ собств. векторов с наим. собств. значениями.

$$U = (u_1, \dots, u_k) \quad \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

4) Кластеризуем выборку $U$ с помощью k-means.

$k = 2$ Graph-cut

$W$ – схожесть объектов в графе

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

$A_k \subset V \qquad A_i \cap A_j = \varnothing$

$$cut(A_1, \dots, A_k) = \frac{1}{2} \sum_{i=1}^{k} W(A_i, \bar{A}_i) \to \min$$

$$RatioCut(A_1, \dots, A_k) = \frac{1}{2} \sum_{i=1}^{k} \frac{W(A_i, \bar{A}_i)}{|A_i|}$$

$$\min_{A \subset V} \mathrm{RatioCut}(A, \bar{A})$$

$$f = (f_1, \dots, f_n)^T$$

$$f_i = \begin{cases} \sqrt{|\bar{A}|/|A|}, & v_i \in A \\ -\sqrt{|A|/|\bar{A}|}, & v_i \in \bar{A} \end{cases}$$

$$f^T L f = n \,\mathrm{RatioCut}(A, \bar{A})$$

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

$$= \frac{1}{2} \sum_{i \in A, j \in \bar{A}} w_{ij} \left( \sqrt{\frac{|\bar{A}|}{|A|}} + \sqrt{\frac{|A|}{|\bar{A}|}} \right)^2 +$$

$$\frac{1}{2} \sum_{i \in \bar{A}\, j \in A} w_{ij} \left( -\sqrt{\frac{|A|}{|\bar{A}|}} - \sqrt{\frac{|\bar{A}|}{|A|}} \right)^2 =$$

$$= \frac{1}{2} \left( \sqrt{\frac{|\bar{A}|}{|A|}} + \sqrt{\frac{|A|}{|\bar{A}|}} \right)^2 \left( W(A, \bar{A}) + W(\bar{A}, A) \right) =$$

$$= \frac{1}{2} \left( \frac{|\bar{A}|}{|A|} + 2 + \frac{|A|}{|\bar{A}|} \right) \left( W(A, \bar{A}) + W(\bar{A}, A) \right)$$

$$= \frac{1}{2} \left( \frac{\overbrace{|\bar{A}| + |A|}^{n}}{|A|} + \frac{\overbrace{|A| + |\bar{A}|}^{n}}{|\bar{A}|} \right) \left( W(A, \bar{A}) + W(\bar{A}, A) \right)$$

$$n \quad \frac{1}{2}\left(\frac{1}{|A|}+\frac{1}{|\bar{A}|}\right)\left(W(A,\bar{A})+W(\bar{A},A)\right)$$

$$\frac{W(A,\bar{A})}{|A|}+\cancel{\frac{W(A,\bar{A})}{|\bar{A}|}}+\cancel{\frac{W(\bar{A},A)}{|A|}}+\frac{W(\bar{A},A)}{|\bar{A}|}$$

$$= n\,\mathrm{RatioCut}(A,\bar{A}) \qquad ??$$

$$\sum_{i=1}^{n} f_i = 0 \qquad \sum_{i=1}^{n} f_i^2 = n$$

$$f^T L f \to \min_f$$

$$\langle f, \vec{1}\rangle = 0$$

$$\|f\| = \sqrt{n}$$

$$\mathcal{L} = f^T L f + \lambda_1 f^T \vec{1} + \lambda_2(\|f\| - \sqrt{n})$$

$$\nabla_f \mathcal{L} = 2Lf + \lambda_1 \vec{1} + \lambda_2 \frac{1}{\|f\|} f = 0 \;\Big|\cdot \vec{1}^T$$

$$2\vec{1}^T L f + \lambda_1 n = 0$$

$$\lambda_1 = -\frac{2}{n}\vec{1}^T L f$$

$$\lambda_1^T = -\frac{2}{n} f^T \underbrace{L\vec{1}}_{=0\,\vec{1}} = 0$$

$$2Lf + \lambda_2 \underbrace{\frac{1}{\|f\|}}_{=\sqrt{n}} f = 0$$

$$Lf = -\frac{\lambda_2}{2\sqrt{n}} f \qquad \lambda =$$

$$-\frac{\lambda_2}{2\sqrt{n}} = \alpha$$

$$f^T L f = f^T \alpha f = \alpha \|f\|^2 = \alpha n \to \min$$

Минимум достигается на 2-ом с.в. $L$ с с.з. $\alpha$

$$\begin{cases} v_i \in A, & \text{если } f_i \geq 0 \\ v_i \in \bar{A}, & \text{если } f_i < 0 \end{cases}$$