

Спектральная кластеризация

Алим Бухараев



Кластеризация исполнителей

Цель

Уметь быстро разбивать множество исполнителей на кластеры

Зачем?

- Чтобы успевать подсовывать пользователю песни, которые ему больше всего нравятся
- Таким образом, он задержится у нас на сайте подольше

Как это нам выгодно?

- Чем больше времени пользователь проведёт на сайте, тем больше он прослушает рекламы

Алгоритм спектральной кластеризации

- 1) Взять матрицу схожести
 - 2) Построить по ней Лапласиан
 - 3) Посчитать k первых собственных векторов
 - 4) Составить из них матрицу X (собств. вектора - столбцы)
 - 5) Пусть каждая строка - точка в k мерном пространстве.
Тогда мы можем кластеризовать N точек просто используя, например, K-means
-

$A =$

	1	0.6	0	0	0	0	0
	0.6	1	0	0	0	0	0
	0	0	1	0.5	0.3	0	0
	0	0	0.5	1	0.7	0	0
	0	0	0.3	0.7	1	0	0
	0	0	0	0	0	1	0.9
	0	0	0	0	0	0.9	1

D =

	1.6	0	0	0	0	0	0
	0	1.6	0	0	0	0	0
	0	0	1.8	0	0	0	0
	0	0	0	2.2	0	0	0
	0	0	0	0	2.0	0	0
	0	0	0	0	0	1.9	0
	0	0	0	0	0	0	1.9
	0	0	0	0	0	0	1.9

$$L = D - A$$

- Симметрична, а значит имеет n вещественных собственных значений
- Положительно полуопределенна
- Для её собственных значений справедливо неравенство:

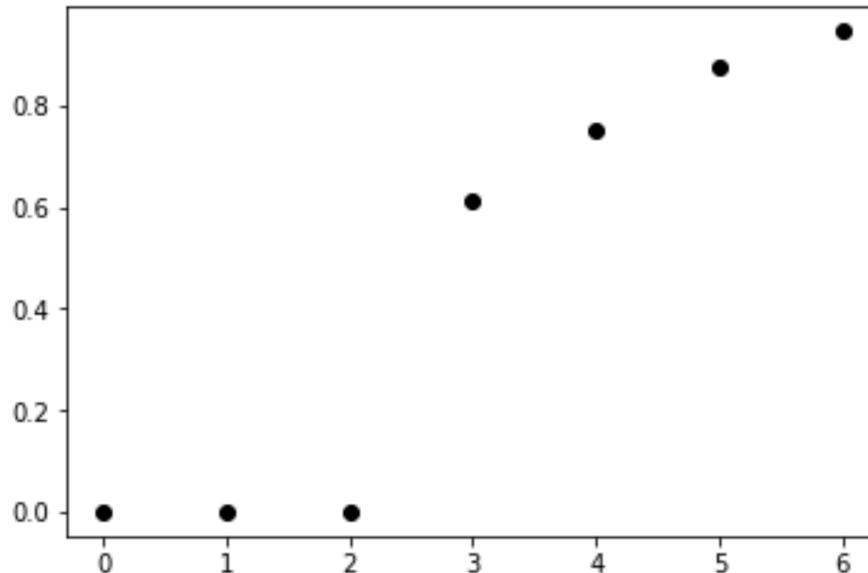
$$\lambda_1 = 0 \leq \lambda_2 \leq \dots \leq \lambda_n$$

- Количество компонент связности L равно размерности собственного подпространства, соответствующего нулевому собственному значению

$L =$

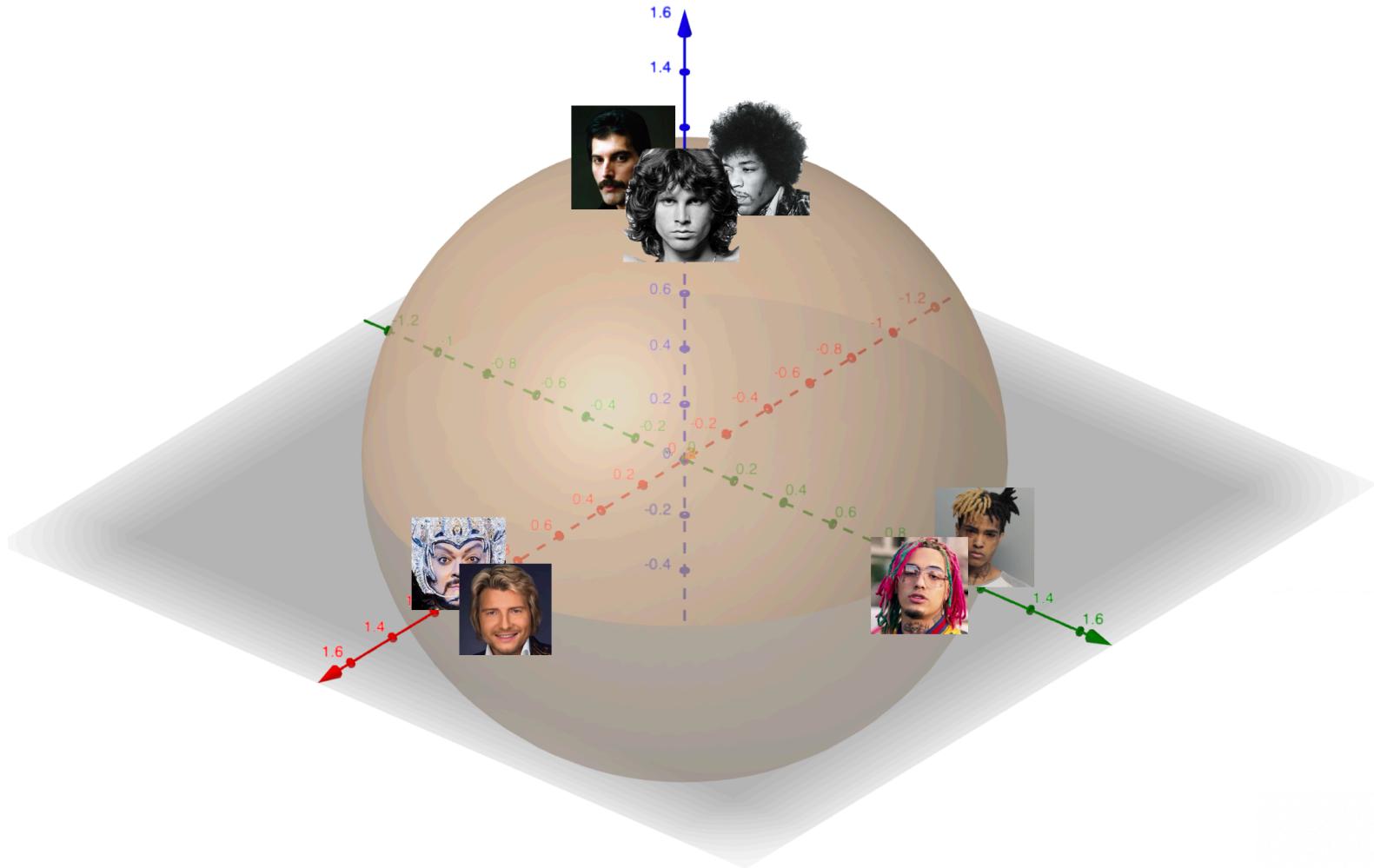
	0.6	-0.6	0	0	0	0	0	0
	-0.6	0.6	0	0	0	0	0	0
	0	0	0.8	-0.5	-0.3	0	0	0
	0	0	-0.5	1.2	-0.7	0	0	0
	0	0	-0.3	-0.7	1.0	0	0	0
	0	0	0	0	0	0.9	-0.9	
	0	0	0	0	0	-0.9	0.9	

Собственные значения матрицы L



$X =$

1	0	0
1	0	0
0	1	0
0	1	0
0	1	0
0	0	1
0	0	1



$$f^T L f = f^T D f - f^T A f = \sum_{i=1}^n d_i f_i^2 - \sum_{i,k=1}^n f_i f_k a_{ik} =$$

$$= \frac{1}{2} \left(\sum_{i=1}^n d_i f_i^2 - 2 \sum_{i,k=1}^n f_i f_k a_{ik} + \sum_{k=1}^n d_k f_k^2 \right) =$$

$$= \frac{1}{2} \sum_{i,k=1}^n a_{ik} (f_i - f_k)^2$$

$$L=D^{-\frac{1}{2}}(D-A)D^{-\frac{1}{2}}$$

$$f^T L f = f^T D^{-\frac{1}{2}} D D^{-\frac{1}{2}} f - f^T D^{-\frac{1}{2}} A D^{-\frac{1}{2}} f =$$

$$= \sum_{i=1}^n d_i \left(\frac{f_i}{\sqrt{d_i}} \right)^2 - \sum_{i,k=1}^n \frac{f_i}{\sqrt{d_i}} \frac{f_k}{\sqrt{d_k}} a_{ik} =$$

$$= \frac{1}{2} \left(\sum_{i=1}^n d_i \left(\frac{f_i}{\sqrt{d_i}} \right)^2 - 2 \sum_{i,k=1}^n \frac{f_i}{\sqrt{d_i}} \frac{f_k}{\sqrt{d_k}} a_{ik} + \sum_{k=1}^n d_k \left(\frac{f_k}{\sqrt{d_k}} \right)^2 \right) =$$

$$= \frac{1}{2} \sum_{i,k=1}^n a_{ik} \left(\frac{f_i}{\sqrt{d_i}} - \frac{f_k}{\sqrt{d_k}} \right)^2$$

- У $D^{-1}(D - A)$ есть собственный вектор из единиц с собственным значением 0
- Если λ и x — собственное значение и соответствующий собственный вектор матрицы $D^{-1}(D - A)$, то у матрицы $L = D^{-\frac{1}{2}}(D - A)D^{-\frac{1}{2}}$ есть тоже собственное значение с собственным вектором $D^{\frac{1}{2}}x$

$$h_{\textcolor{brown}{i}}'Lh_i = \frac{\mathrm{cut}(A_{\textcolor{brown}{i}},\overline{A}_{\textcolor{blue}{i}})}{|A_{\textcolor{brown}{i}}|}.$$

$$\mathrm{RatioCut}(A_1,\ldots,A_{\textcolor{blue}{k}}) = \sum_{i=1}^k h_i'Lh_i:$$

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

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