

谱聚类：spectral clustering

一种基于图论的聚类方法

- 基本思想
 - 将原始数据集转换成为图
 - ε -邻域图
 - kNN 图(互 kNN 图)
 - 全连接图：对于给定数据集中的每个数据对象，计算出该对象与其它对象之间的相似性，得到相似性矩阵 W (也可设置阈值)
 - W 为图的邻接矩阵，则 W_{ij} 为边(v_i, v_j)上的权值
 - 由邻接矩阵计算得到拉普拉斯矩阵

$$L = D - W$$

其中 D 为对角矩阵， $D_{ii} = \sum_j W_{ij}$

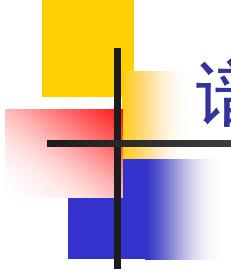
谱聚类：spectral clustering

L 的性质

- 对任意的 n 维向量 $f \in R^n$, 都有

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

- L 是对称的半正定矩阵
- L 的最小特征值为 0, 对应的特征向量为全 1 的向量
- L 拥有 n 个非负的实特征值 $0 = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$
- 谱聚类方法的理论依据是基于上述特征的



谱聚类：spectral clustering

- 对 L 进行特征值分解，求出其全部特征值和特征向量
- 将 L 的特征值从小到大排列，特征向量对应重排
- 取 L 的前 k 个特征值对应的特征向量，将其按列向量形式排列得到一个 $n \times k$ 的矩阵 M
- 将 M 的每一行看做一个新的数据点，对这 n 个数据点使用 k -Means 方法进行聚类
- k 的取值可以与 k -Means 中的 k 一致，也可不同

谱聚类：spectral clustering

算法

Input: Similarity matrix $S \in R^{n \times n}$, number k of clusters to construct

Output: Clusters A_1, \dots, A_k with $A_i = \{j \mid y_j \in C_i\}$

Method:

- Construct a similarity graph, let W be its weighted adjacency matrix
- Compute the Laplacian L
- Compute the first k eigenvectors u_1, \dots, u_k of L
- Let $U \in R^{n \times k}$ be the matrix containing the vectors u_1, \dots, u_k as columns
- For $i=1, \dots, n$, let $y_i \in R^k$ be the vector corresponding to the i -th row of U
- Cluster the points $(y_i)_{i=1,\dots,n}$ in R^k with the k -Means algorithm into clusters C_1, \dots, C_k



谱聚类：spectral clustering

变形

- Normalized spectral clustering

- $L_{sym} = D^{-1/2} L D^{-1/2} = I - D^{-1/2} W D^{-1/2}$
- $L_{rw} = D^{-1} L = I - D^{-1} W$

谱聚类：spectral clustering

算法

Input: Similarity matrix $S \in R^{n \times n}$, number k of clusters to construct

Output: Clusters A_1, \dots, A_k with $A_i = \{j \mid y_j \in C_i\}$

Method:

- Construct a similarity graph, let W be its weighted adjacency matrix
- Compute the normalized Laplacian L_{sym}
- Compute the first k eigenvectors u_1, \dots, u_k of L_{sym}
- Let $U \in R^{n \times k}$ be the matrix containing the vectors u_1, \dots, u_k as columns
- Form the matrix $T \in R^{n \times k}$ from U by normalizing the rows to norm 1, that is set $t_{ij} = u_{ij}/(\sum_k u_{ik})^{1/2}$
- For $i = 1, \dots, n$, let $y_i \in R^k$ be the vector corresponding to the i -th row of T
- Cluster the points $(y_i)_{i=1, \dots, n}$ with the k-Means algorithm into clusters C_1, \dots, C_k

谱聚类：spectral clustering

算法

Input: Similarity matrix $S \in R^{n \times n}$, number k of clusters to construct

Output: Clusters A_1, \dots, A_k with $A_i = \{j \mid y_j \in C_i\}$

Method:

- Construct a similarity graph, let W be its weighted adjacency matrix
- Compute the Laplacian L
- Compute the first k generalized eigenvectors u_1, \dots, u_k of the generalized eigenproblem $Lu = \lambda Du$
- Let $U \in R^{n \times k}$ be the matrix containing the vectors u_1, \dots, u_k as columns
- For $i = 1, \dots, n$, let $y_i \in R^k$ be the vector corresponding to the i -th row of U
- Cluster the points $(y_i)_{i=1, \dots, n}$ in R^k with the k -Means algorithm into clusters C_1, \dots, C_k