Fast Approximate Spectral Clustering

Rhian Davies

December 11, 2014

Engineering and Physical Sciences Research Council

excellence with impact

Algorithm 1. NJW spectral clustering algorithm

Input: Dataset $S = \{x_1, \ldots, x_n\}$ in $\Re^1$ and the number of clusters k

Output: k-way partition of the input data

(1) Construct the affinity matrix A by the following Gaussian kernel function:

$$A_{ij} = \begin{cases} \exp\left(\frac{-\|x_i - x_j\|^2}{\delta^2}\right) & \text{for } i \neq j, \\ 0 & \text{for } i = j, \end{cases} \tag{1}$$

where δ is a scale parameter to control how fast the similarity attenuates with the distance between the data points x_i and x_j.

(2) Compute the normalized affinity matrix $\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{A}\ \mathbf{D}^{-1/2}$, where $\boldsymbol{D}$ is the diagonal matrix with $D_{ii} = \sum_{j=1}^{n} A_{ij}$.

(3) Compute the k eigenvectors of $\mathbf{L}$, $v_1, v_2, \ldots, v_k$, which are associated with the k largest eigenvalues, and form the matrix $X = [v_1 v_2, \ldots, v_k]$.

(4) Renormalize each row to form a new matrix $Y \in \Re^{n \times k}$ with $Y_{ij} = X_{ij} / \left(\sum_j X_{ij}^2\right)^{1/2}$, so that each row of $\boldsymbol{Y}$ has a unit magnitude.

(5) Treat each row of $\boldsymbol{Y}$ as a point in $\Re^k$ and partition the n points (n rows) into k clusters via a general cluster algorithm, such as the K-means algorithm.

(6) Assign the original point x_i to the cluster c if and only if the corresponding row i of the matrix $\boldsymbol{Y}$ is assigned to the cluster c.

KASP

- Create a representative set
- Spectral cluster representative set
- Assign labels to all original data

Algorithm KASP $(\mathbf{x}_1, \ldots, \mathbf{x}_n, k)$

Input: n data points $\{\mathbf{x}_i\}_{i=1}^n$, number of representative points k
Output: m-way partition of the input data

1. Perform k-means with k clusters on $\mathbf{x}_1, \ldots, \mathbf{x}_n$ to:
   a) Compute the cluster centroids $\mathbf{y}_1, \ldots, \mathbf{y}_k$ as the k representative points.
   b) Build a correspondence table to associate each $\mathbf{x}_i$ with the nearest cluster centroid $\mathbf{y}_j$.
2. Run a spectral clustering algorithm on $\mathbf{y}_1, \ldots, \mathbf{y}_k$ to obtain an m-way cluster membership for each of $\mathbf{y}_i$.
3. Recover the cluster membership for each $\mathbf{x}_i$ by looking up the cluster membership of the corresponding centroid $\mathbf{y}_j$ in the correspondence table.

Perturbation Analysis

Assume $x_1, \ldots, x_n$ i.i.d according to a probability distribution G

$$\tilde{x}_i = x_i + \epsilon_i \tag{1}$$

$\tilde{x}$ distributed by $\tilde{G}$.

- ϵ_i independent of x_i
- ϵ_i are i.i.d. according to a symmetric dist (mean zero, bounded support)
- $\text{Var}(\epsilon)$ small relative to $\text{Var}(X)$.

Mis-clustering Rate

$$\rho = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(l_i \neq \tilde{l}_i) \tag{2}$$

Theorem

Under the assumptions(*), the mis-clustering rate ρ of a spectral bi-partitioning algorithm on the perturbed data satisfies

$$\rho \leq \|\tilde{v}_2 - v_2\|^2 \tag{3}$$

Lemma

Let g denote the eigengap between the second and the third eigenvalues of L. Then the following holds:

$$\|\tilde{v}_2 - v_2\| \leq \frac{1}{g}\|\tilde{L} - L\| + O(\|\tilde{L} - L\|^2). \tag{4}$$

Theorem

Assume assumptions hold throughout recursive invocation of the Ncut algorithm, g_0 is bounded away from zero and Frobenius norm of perturbation of Laplacian matricies along the recursion is bounded by $c\|\tilde{L} - L\|_F^2$ for some constant $c \geq 1$. Then the mis-clustering rate for an m-way spectral clustering solution can be bounded by:

$$\rho \leq \| \frac{m}{g_0^2} \cdot c\|\tilde{L} - L\|_F^2. \tag{5}$$

Theorem

Assume assumptions hold throughout recursive invocation of the Ncut algorithm, g_0 is bounded away from zero and Frobenius norm of perturbation of Laplacian matricies along the recursion is bounded by $c\|\tilde{L} - L\|_F^2$ for some constant $c \geq 1$. Then the mis-clustering rate for an m-way spectral clustering solution can be bounded by:

$$\rho \leq \| \frac{m}{g_0^2} \cdot c\|\tilde{L} - L\|_F^2. \qquad (5)$$

$$r_1 \leq n \cdot \frac{L_1^2}{g_1^2}$$

$$r_i \leq (n - n_{i-1}) \cdot \frac{L_i^2}{g_i^2}, i = 2, \ldots, m-1.$$

Linking Laplacian to ϵ

Theorem

$$\|\tilde{L} - L\|_F^2 \leq_p c_1 \sigma_\epsilon^{(2)} + c_2 \sigma_\epsilon^{(4)} \tag{6}$$

If we know the probability distribution of the original data, it is possible to characterise the exact amount of distortion.
f is the density function of G

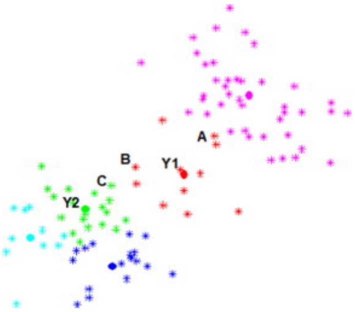
Theorem

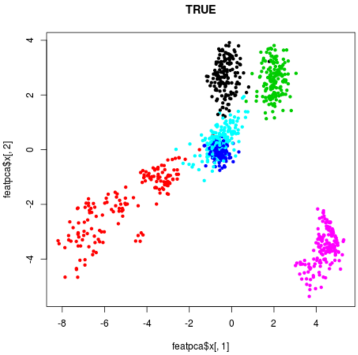
Let data be distributed with density f.

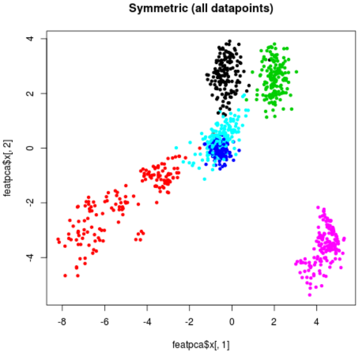
$$\rho = c \cdot b_{2,d} \cdot \|f\|_{\frac{d}{d+2}} \cdot k^{-\frac{2}{d}} + O(k^{-\frac{4}{d}})$$

where c is constant determined by number of clusters, variance of original data, bandwidth of Gaussian kernel and the minimum eigengap of all affinity matrices used in Ncut.

FAIL




TRUE
featpca$x[, 2]
featpca$x[, 1]
4
2
0
-2
-4
-8
-6
-4
-2
0
2
4


Symmetric (all datapoints)
featpca$x[, 2]
featpca$x[, 1]
4
2
0
-2
-4
-8
-6
-4
-2
0
2
4