Notes on papers read

Sourabh Antani

On Spectral Clustering: Analysis and an algorithm - Ng, Jordan, Weiss

Summary

Algorithm:

1. Form Affinity Matrix $A \in \mathbb{R}^{n \times n}$ defined by $A_{ij} = exp(-||s_i - s_j||^2/2\sigma^2)$ if $i \neq j$, and $A_{ii} = 0$. σ^2 is the scaling parameter that determines how rapidly the affinity A_{ij} falls off with distance between points s_i and s_j.

2. Define D to be diagonal matrix such that D_{ii} = row sum of i'th row of A. Construct $L = D^{-1/2}AD^{-1/2}$

3. Form matrix X whose columns are the k largest eigenvectors of L. If eigenvalues are repeated, choose eigenvectors to be mutually orthogonal.

4. Form matrix Y by normalizing rows of X to have unit length ($Y_{ij} = X_{ij}/(\sum_j X_{ij}^2)^{1/2}$)

5. Cluster rows of Y into k clusters, treating each row as a point in $\mathbb{R}$. Assign the point s_i to cluster j if row i of Y was assigned to cluster j.

Matrix perturbation theory [G. W. Stewart and J. G. Sun. Matrix Perturbation Theory. Academic Press, 1990.] indicates that the stability fo the eigenvectors of a matrix is determined by the *eigengap*. i.e. the *subspace* spanned by first k eigenvectors of $\hat{L}$ will be stable to small changes to $\hat{L}$ if $\delta = |\lambda_k - \lambda_{k+1}|$ is large. This algorithm assumes that the clusters are dense and that no cluster can be subdivided into two distinct dense clusters. Also that no point is 'too much less connected' then other points in that cluster.

Points to be noted

1. What matrix is used: $L = D^{-1/2}AD^{-1/2}$

2. Which eigenvectors are kept : corresponding to highest eigenvalues

Spectral Grouping Using the Nyström Method - Charless Fowlkes, Serge Belongie, Fan Chung, and Jitendra Malik

Pairwise methods (methods that rely on pairwise comparison of all elements) have certain advantages like flexibility of definition of affinity between elements and that they do not assume the features to have a certain distribution (e.g. gaussian). However, they are computationally intensive. This paper presents an alternate way of determining approximate eigenvectors of the affinity matrix using sampled points only.

Nyström method is a technique for finding numerical approximations for eigen function problems of the form

$$\int_a^b W(x,y)\phi(y)dy = \lambda\phi(x)$$

by evaluating the equation at a set of evenly spaced points $\xi_1, \ldots, \xi_n$ on interval $[a, b]$ and employing quadrature rule

$$\frac{(b-a)}{n}\sum_{j=1}^{n} W(x,\xi_j)\hat{\phi}(\xi_j) = \lambda\hat{\phi}(x)$$

where $\hat{\phi}(x)$ is approximation to $\phi(x)$. To solve this, we set $x = \xi_i$ resulting in a system of equations, with one equation for each ξ_i. Let $[a, b] = [0, 1]$, we get the matrix eigenvalue problem $A\hat{\Phi} = n\hat{\Phi}\Lambda$ where $A_{ij} = W(\xi_i, \xi_j)$ and $\Phi = [\phi_1, \ldots, \phi_n]$ are eigenvectors of A with eigenvalues $\lambda_1, \ldots, \lambda_n$. Thus Nyström extension

$$\hat{\phi}_i(x) = \frac{1}{n\lambda_i}\sum_{j=1}^{n} W(x,\xi_j)\hat{\phi}_i(\xi_j)$$

This extends the eigenvector for sample points to any arbitrary point using $W(\cdot, \xi_i)$ as interpolation weights.

Let $A = U\Lambda U^T$ be the affinity matrix between sampled points and B be affinity matrix between sampled points and all other points. The Nyström extnesion is $B^T U\Lambda^{-1}$ where B^T corresponds to $W(\xi_i, \cdot)$, U corresponds to $\hat{\phi}(\xi_i)$ and Λ^{-1} corresponds to $\frac{1}{\lambda_i}$. Thus

$$W = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}$$

Let $\bar{U}$ denote the approximate eigenvectors of W, the Nyström extension gives

$$\bar{U} = \begin{bmatrix} U \\ B^T U\Lambda^{-1} \end{bmatrix}$$

Thus the approximation of W is

$$\begin{aligned}
\hat{W} &= \bar{U}\Lambda\bar{U}^T \\
&= \begin{bmatrix} U \\ B^T U \Lambda^{-1} \end{bmatrix} \Lambda \begin{bmatrix} U^T & \Lambda^{-1} U^T B \end{bmatrix} \\
&= \begin{bmatrix} U\Lambda U^T & B \\ B^T & B^T A^{-1} B \end{bmatrix} \\
&= \begin{bmatrix} A & B \\ B^T & B^T A^{-1} B \end{bmatrix} \\
&= \begin{bmatrix} A \\ B^T \end{bmatrix} A^{-1} \begin{bmatrix} A & B \end{bmatrix} \\
&= \begin{bmatrix} A & B \\ B^T & B^T A^{-1} B \end{bmatrix} \\
\therefore C &\approx B^T A^{-1} B
\end{aligned}$$

If A is positive definite: Let $A^{1/2}$ be symmetric positive definite square root of A. Define $S = A + A^{-1/2} B B^T A^{-1/2}$ and diagonalize it as $S = U_S \Lambda_S U_S^T$. In the appendix of the paper authors show that $\hat{W}$ can be diagonalized by V and Λ_S, i.e. $\hat{W} = V \Lambda_S V^T$ and $V^T V = I$, where

$$V = \begin{bmatrix} A \\ B^T \end{bmatrix} A^{-1/2} U_S \Lambda_S^{-1/2}$$

If A is indefinite: let $\bar{U}_S^T = [U_S^T \quad \Lambda_S^{-1} U_S^T B]$, define $Z = \bar{U}_S \Lambda^{1/2}$ such that $\hat{W} = Z Z^T$. Let $f \sigma F^t$ be diagnonalization of $Z^T Z$. Then $V = Z F \Sigma^{-1/2}$ contains the leading orthonormalized eigenvectors of $\hat{W}$, i.e. $\hat{W} = V \Sigma V^T$ and $V^T V = I$. However, this two step procedure is expensive ($O(n^3)$) and leads to loss of significant figures. Hence the one shot method (above) must be applied when Kernel is positive definite.

In both above cases, if A has linearly dependent columns, pseudoinverse can be used instead of regular inverse.

The authors also show that for the purpose of image segmentation, less than 1 percent of pixels need to be sampled to achieve sufficient performance.

Points to be noted

1. What matrix is used: $L = I - D^{-1/2} A D^{-1/2}$. eigenvalues of L lie in [0,2] and those $D^{-1/2} A D^{-1/2}$ lie in [-1,1]

2. Which eigenvectors are kept : greatest eigenvectors of $D^{-1/2} A D^{-1/2}$

A unifying theorem for spectral embedding and clustering - Matthew Brand, Kun Huang

Summary

In this paper the authors present a theorem that explains why spectral methods succeed in clustering by using selected eigenvectors. Authors begin by stating that the underlying notion behind various methods is that truncated eigenbasis amplifies structure in the data so that any heuristic post-processing is more likely to succeed. Following is a summary of the main theoretical result:

1. An eigenvalue-scaled eigenvector representation of the data encodes angles (equivalently, correlations) between points embedded in the surface of a hypersphere.
2. When the representation is truncated by suppressing the smallest magnitude eigenvalues, the angles (equiv., correlations) between high-affinity points are least distorted, highlighting the manifold structure of the data.
3. As the representation is further truncated, the angles (equiv., correlations) decrease between points having high affinity and increase between points having low affinity, highlighting the cluster structure of the data.

In short, nonlinear dimensionality reduction and clustering can be obtained from the same process. The theorem is limited to symmetric non-negative definite affinity matrices, but a corollary establishes relevance to non-positive matrices as well, and to asymmetric matrices (e.g., $\mathbf{B}$) via their Grams ($\mathbf{B}^\top\mathbf{B}$ or $\mathbf{B}\mathbf{B}^\top$).

Theorem (polarization): As positive (resp., nonnegative) $\mathbf{A}$ is projected to successively lower ranks $\mathbf{A}_{(D-1)}$,$\mathbf{A}_{(D-2)}\ldots$,$\mathbf{A}_{(d)}$,$\ldots$,$\mathbf{A}_{(2)}$, $\mathbf{A}_{(1)}$, the sum of squared angle-cosines $\sum_{i\neq j}\cos\theta_{ij}^2$ (equivalently squared correlations $||\mathbf{Y}_{(d)}^\top\mathbf{Y}_{(d)}||_F^2$) is strictly increasing (resp., non-decreasing).

Corollary (clustering): Truncation of the eigenbasis amplifies any unevenness in the distribution of points on the d-dimensional hypersphere by causing points of high affinity to move toward each other and other to move apart.

Using a subset of all the eigenvectors emphasizes the data's cluster structure, improving the output of any heuristic clustering procedure. This does not mean that the lowest-dimensional embedding is the best one for clustering; there is a tradeoff between amplifying cluster structure and losing information.

Prior to this paper, Fiedler first showed that the eigenvector of the Laplacian matrix corresponding to the second eigenvalue gives an embedding of the graph in a real line; cutting this embedding at the origin gives a bipartitioning of the graph. This was extended to k-way partitioning using normalized row vectors of matrix formed by the first k eigenvectors of affinity matrix. Similarly Ng. et al. used k-means clustering on normalized row vectors of the first k weighted eigenvectors. Results are stable if the data is nearly clustered. Chan et al. used directrional angle between row vectors of the first k eigenvectors as distance measure for partitioning.

The basic strategy is to use two alternating projections: projection to low-rank and projection to set of zero-diagonal doubly stochastic matrices. The projection to low-rank matrix $\mathbf{A}$ ($\mathbf{P}$)$\rightarrow \mathbf{A}_d$ is application

of polarization theorem with minimal loss of energy $||\mathbf{A} - \mathbf{A}_{(d)}||_F^2$. The projection to zero-diagonal doubly stochastic matrix $\mathbf{A}_{(d)} \rightarrow \mathbf{P} = \text{diag}(\mathbf{d})(\mathbf{A}_{(d)} - \text{diag}(\text{diag}(\mathbf{A}_{(d)})))\text{diag}(\mathbf{d})$ suppresses any difference sin the stationary probability of points induced by projection to low rank. Suppressing diagonal induces negative eigenvalues in the spectrum of $\mathbf{P}$, these eigenvalues account for less than half of the energy in $\mathbf{P}$. Subsequent projection to lower rank matrix suppresses these negative and unit eigenvalues. This gives an automatic determination of d and bound on loss of variance. This alternating projections stop when $\mathbf{P}$ has two or more stochastic (unit) eigenvalues.

Scalable Spectral Clustering Using Random Binning Features - Lingfei Wu, Pin-Yu Chen, Ian En-Hsu Yen, Fanli Su, Yinlong Xia, Charu Aggarwal

Summary

In spectral clustering (SC), the challenges faced are memory and computational complexity in formation of the pairwise graph and Laplacian construction and the eigendecomposition there of. This paper uses *Random Binning Features* (RB) to produce $\mathbb{R}^{N\times D}$ matrix instead of $\mathbb{R}^{N\times N}$ reduce the computational cost from $O(N^2d)$ to $O(NRd)$ and memory from $O(N^2)$ to $O(NR)$. PRIMME eigensolver is used to compute the eigenvectors of graph Laplacian L without explicit formulation. The computational complexity is reduced from $O(KN^2m)$ to $O(NKRm)$, where m is number of iterations of the eigensolver. Finally, the authors show that $R = \Omega(1/\kappa\epsilon)$ RB features are sufficient for uniform convergence to ϵ precision of exact SC.

Mathematically, RB considers feature map of the the form

$$k(x_1, x_2) = \int_\omega p(\omega)\phi_{B_\omega}(x_1)^T\phi_{B_\omega}(x_2)d\omega$$

where B_ω is the random grid determined by $\omega = (\omega_1, u_1, \cdots, \omega_d, u_d)$ where (ω_i, u_i) are width and bias (offset of first bin from origin in the direction of i-th axis) in the i-th dimension of the grid. The feature vector $\phi_{B_\omega}(x)$ for a bin $b \in B_\omega$ is

$$\phi_b(x_i) = 1, if b = (\lfloor\frac{x_i^{(1)} - u_1}{\omega_1}\rfloor, \ldots, \lfloor\frac{x_i^{(d)} - u_d}{\omega_d}\rfloor)$$

In simple terms this means that the feature vector ϕ_{B_ω} have a component for each bin containing at lease one point and the feature vector for a given point has 1 in the component corresponding to the bin that contains that point. The size of the feature vectors is the number of occupied bins D, which depends upon the width of bins in each dimension (large ω_i means less bin count means small R). The RB feature vectors $\phi_{B_\omega}(x_i)$ stacked side by side will form a sparse $\mathbb{R}^{N\times D}$ matrix Z.

Next, to compute the approximate Lapacian $\hat{L}$, use approximate affinity matrix $\hat{W} = ZZ^T$. Thus, $\hat{L} = I - \hat{D}^{-1/2}\hat{W}\hat{D}^{-1/2} = \hat{D}^{-1/2}ZZ^T\hat{D}^{-1/2}$. The approximate degree matrix $\hat{D} = \text{diag}(\hat{W}\mathbf{1}) = \text{diag}(Z(Z^T\mathbf{1}))$, which can be computed as two matrix-vector multiplications instead of a matrix-matrix multiplications. Finally, define $\hat{Z} = \hat{D}^{-1/2}Z$ so $\hat{L} = I - \hat{Z}\hat{Z}^T$.

The largest left singular vectors of $\hat{Z}$ are the smallest eigenvectors of $\hat{L}$. The authors suggest using PRIMME which is in classes of Generalized Davidson type methods which enjoy benefits of advanced subspace restarting and preconditioning techniques to accelerate the convergence. Finally the K left singular vectors are normalized and clustered using K-means, where K is the desired number of clusters.

Total computational and memory consumption are $O(NRd + NKRm + NK^2t)$ and $O(NR)$, thus, linear in number of datapoints N. Thus scalable for large datasets compared to quadratic ($O(N^2)$ for SC.

Points to be noted

1. What matrix is used: $L = I - D^{-1/2}WD^{-1/2}$
2. Which eigenvectors are kept : smallest eigenvectors of L
3. Code: https://github.com/IBM/SpectralClustering_RandomBinning
4. Code: https://github.com/teddylfwu/RandomBinning
5. Code: https://github.com/primme/primme