4.3 Diffusion Maps

Diffusion Maps is a non-linear dimensionality reduction technique or feature extraction, introduced by Coifman and Lafon [CL06]. With ISOMAP, it is another example of manifold learning algorithms that capture the geometry of the data set. Data are represented by parameters of its underlying geometry in a low dimensional Euclidean space. Main intention is to discover the underlying manifold that the data has been sampled from. The main idea is to construct a kernel based on the connection between data. The eigenvectors of this kernel represent the data in lower dimension. The diffusion map framework consists of the following steps [TCGC13]:

1. Constructing a weighted graph $(V, E, \mathbf{W})$ on the data. The pairwise weights measure the closeness between data points.

2. Defining a random walk on the graph determined by a transition matrix constructed from the weights $\mathbf{W}$

3. Non-linear embedding of points in a lower dimensional space based on the parameters of the graph and the respective transition matrix

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ be n samples. We start from constructing a weighted graph $(V, E, \mathbf{W})$. Nodes which are connected by an edge with large weight are considered to be close. Each sample $\mathbf{x}_i$ is associated with a vertex v_i. The weight of an edge between $\mathbf{x}_i$ and $\mathbf{x}_j$ is given by the weight function or kernel $w_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$. The kernel should satisfy three properties:

- Symmetry: $K(\mathbf{x}_i, \mathbf{x}_j) = K(\mathbf{x}_j, \mathbf{x}_i)$

- Non-negativity: $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) \geq 0$

- Locality: there is a scale parameter ϵ such that if $\|\mathbf{x}_i - \mathbf{x}_j\| \ll \epsilon$ then $K(\mathbf{x}_i, \mathbf{x}_j) \to 1$, and if $\|\mathbf{x}_i - \mathbf{x}_j\| \gg \epsilon$ then $K(\mathbf{x}_i, \mathbf{x}_j) \to 0$.

Note that the kernel function encapsulates the notion of closeness between the points. Setting the scale parameter ϵ, similar to the choice of ϵ in ISOMAP, is important. Small ϵ may lead to a disconnected graph and large ϵ may miss the underlying geometry. Gaussian kernel is one of the well known weight functions and is defined as:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\epsilon^2}\right).$$

Using kernel functions, the weight matrix is constructed.

Next, we construct a random walk $X_t, t = 0, 1, 2, \ldots$ on the vertices of the graph $V = \{v_1, \ldots, v_n\}$ with transition matrix:

$$\mathbf{M} = (M_{ij})_{i,j=1,\ldots,n} \text{ with } M_{ij} = \frac{w_{ij}}{\deg(i)}, 1 \leq i, j \leq n.$$

with $\mathbf{W} = (w_{ij})_{1\leq i,j\leq n}$ and $\deg(i) = \sum_j w_{ij}$. The transition matrix represents the probability of moving from the node v_i at time t to v_j at time $t+1$, namely:

$$\mathbb{P}(X_{t+1} = j|X_t = i) = M_{ij}.$$

The transition matrix $\mathbf{M}$ can be written as $\mathbf{D}^{-1}\mathbf{W}$ where $\mathbf{D} = \mathrm{diag}(\deg(1), \ldots, \deg(n))$. The conditional distribution of being at the vertex v_j having started at the vertex v_i is given by:

$$\mathbb{P}(X_t = j|X_0 = i) = (\mathbf{M}^t)_{i,j}, j = 1, \ldots, n.$$

The probability of being at each vertex after step time t starting from v_i is given by i^{th} row of $\mathbf{M}^t = (M_{ij}^{(t)})_{1\leq i,j\leq n}$. This distribution is given by:

$$v_i \to \mathbf{e}_i^T\mathbf{M}^t = (M_{i1}^{(t)}, \ldots, M_{in}^{(n)}).$$

Therefore to each vertex v_i, a vector of probabilities is assigned. This vector contains information about underlying geometry. If v_i and v_j are close - strongly connected in the graph - then $\mathbf{e}_i^T\mathbf{M}^t$ and $\mathbf{e}_j^T\mathbf{M}^t$ will be similar.

However it is still not clear how this representation can be embedded in a low-dimensional space. To do this, we focus on the spectrum of $\mathbf{M}^t$. The transition $\mathbf{M} = \mathbf{D}^{-1}\mathbf{W}$ is not symmetric, however the normalized matrix $\mathbf{S} = \mathbf{D}^{1/2}\mathbf{M}\mathbf{D}^{-1/2}$ is symmetric because $\mathbf{S} = \mathbf{D}^{1/2}\mathbf{D}^{-1}\mathbf{W}\mathbf{D}^{-1/2} = \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$ and $\mathbf{W}$ is symmetric. Spectral decomposition of $\mathbf{S}$ is then given by $\mathbf{S} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T$, with $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ eigenvalue matrix such that $\lambda_1 \geq \ldots \lambda_n$. Therefore $\mathbf{M}$ can be written as:

$$\mathbf{M} = \mathbf{D}^{-1/2}\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T\mathbf{D}^{1/2} = \boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Psi}$$

where $\boldsymbol{\Phi} = \mathbf{D}^{-1/2}\mathbf{V} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_n)$ and $\boldsymbol{\Psi} = \mathbf{D}^{1/2}\mathbf{V} = (\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_n)$.

$\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ are bi-orthogonal, i.e., $\boldsymbol{\Phi}^T\boldsymbol{\Psi} = \mathbf{I}_n$, or equivalently $\boldsymbol{\phi}_i^T\boldsymbol{\psi}_j = \delta_{ij}$. λ_k's are the eigenvalues of $\mathbf{M}$ with right and left eigenvectors $\boldsymbol{\phi}_k$ and $\boldsymbol{\psi}_k$:

$$\mathbf{M}\boldsymbol{\phi}_k = \lambda_k\boldsymbol{\phi}_k, \boldsymbol{\psi}_k^T\mathbf{M} = \lambda_k\boldsymbol{\psi}_k^T$$

In summary:

$$\mathbf{M} = \sum_{k=1}^{n} \lambda_k\boldsymbol{\phi}_k\boldsymbol{\psi}_k^T$$

and hence:

$$\mathbf{M}^t = \sum_{k=1}^{n} \lambda_k^t\boldsymbol{\phi}_k\boldsymbol{\psi}_k^T.$$

$$\mathbf{e}_i^T\mathbf{M}^t = \sum_{k=1}^{n} \lambda_k^t\mathbf{e}_i^T\boldsymbol{\phi}_k\boldsymbol{\psi}_k^T = \sum_{i=1}^{n} \lambda_k^t\phi_{k,i}\boldsymbol{\psi}_k^T,$$

Therefore the distribution $\mathbf{e}_i^T\mathbf{M}^t$ can be represented in terms of basis vectors $\boldsymbol{\psi}_k$ with coefficients $\lambda_k^t\phi_{k,i}$ for $k = 1, \ldots, n$ with $\boldsymbol{\phi}_k = (\phi_{k,1}, \ldots, \phi_{k,n})^T$. These coefficients are used to define the diffusion map.

Definition 4.5. The diffusion map at step time t is defined as:

$$\boldsymbol{\phi}_t(v_i) = \begin{bmatrix} \lambda_1^t \phi_{1,i} \\ \vdots \\ \lambda_n^t \phi_{n,i} \end{bmatrix}, i = 1, \ldots, n$$

In the diffusion map, $\phi_{k,i}$ does not vary with t but each element is dependent on t via λ_i^t. The eigenvalues of transition matrix therefore capture the main components of the data. Following theorem provides some information about the eigenvalues of $\mathbf{M}$.

Theorem 4.6. *The eigenvalues $\lambda_1, \ldots, \lambda_n$ of $\mathbf{M}$ satisfy $|\lambda_k| \leq 1$. It also holds that $\mathbf{M}\mathbf{1}_n = \mathbf{1}_n$ and 1 is an eigenvalue of $\mathbf{M}$.*

Proof. Since $\mathbf{M}$ is a stochastic matrix, then sum of each row elements is one which implies $\mathbf{M}\mathbf{1}_n = \mathbf{1}_n$. Let $\mathbf{m}_k = (m_{k,1}, \ldots, m_{k,n})^T$ be the eigenvector corresponding to λ_k. Suppose that $|m_{k,l}| = \max_{1\leq j\leq n} |m_{k,j}|$, which means that $|m_{k,j}| \leq |m_{k,l}|$. It can be seen that:

$$\sum_{j=1}^{n} M_{lj} m_{k,j} = \lambda_k m_{k,l} \implies |\lambda_k| \leq \sum_{j=1}^{n} M_{lj} \frac{|m_{k,j}|}{|m_{k,l}|} \leq \sum_{j=1}^{n} M_{lj} = 1.$$

□

An interesting point is that $\lambda_1 = 1$ and $\boldsymbol{\phi}_1 = \mathbf{1}_n$. Therefore the first element of the diffusion map in above definition is always one for all points. Therefore we simply drop this from the diffusion map and rewrite it as:

$$\boldsymbol{\phi}_t(v_i) = \begin{bmatrix} \lambda_2^t \phi_{2,i} \\ \vdots \\ \lambda_n^t \phi_{n,i} \end{bmatrix}, i = 1, \ldots, n.$$

It is possible to have more than one eigenvalues with absolute value equal to one. In this case, the underlying graph is either disconnected or bipartite.

If λ_k is small, λ_k^t is rather small for moderate t. This motivates truncating the diffusion maps to d dimensions.

Definition 4.7. The diffusion map truncated to d dimensions is defined as:

$$\boldsymbol{\phi}_t^{(d)}(v_i) = \begin{bmatrix} \lambda_2^t \phi_{1,i} \\ \vdots \\ \lambda_{d+1}^t \phi_{n,i} \end{bmatrix}, i = 1, \ldots, n$$

$\boldsymbol{\phi}_t^{(d)}(v_i)$ is an approximate embedding of $v_1, \ldots, v_n$ in a $d-$dimensional Euclidean space. If the graph structure $(V, E, \mathbf{W})$ is appropriately chosen, non-linear geometries can also be recovered using diffusion maps.

The connection between the Euclidean distance in the diffusion map coordinates (diffusion distance) and the distance between the probability distributions is described in the following [Ban08, Theorem 2.11].

Theorem 4.8. *For any pair of nodes v_i and v_j it holds that:*

$$\|\boldsymbol{\phi}_t(v_i) - \boldsymbol{\phi}_t(v_j)\|^2 = \sum_{l=1}^{n} \frac{1}{\deg(l)} \big(\mathbb{P}(X_t = l | X_0 = i) - \mathbb{P}(X_t = l | X_0 = j)\big)^2.$$

Proof. Exercise. □