Mathematics for Data Sciences **Sep 27th,2011**

Lecture 6. Diffusion Distance

Instructor: Yuan Yao, Peking University *Scribe: Lei Huang, Yue Zhao*

Introduction

Finding meaningful low-dimensional structures hidden in high-dimensional observations is an fundamental task in high-dimensional statistics. The classical techniques for dimensionality reduction, principal component analysis (PCA) and multi-dimensional scaling (MDS), guaranteed to discover the true structure of data lying on or near a linear subspace of the high-dimensional input space. PCA finds a low-dimensional embedding of the data points that best preserves their variance as measured in the high-dimensional input space. Classical MDS finds an embedding that preserves the interpoint distances, equivalent to PCA when those distances are Euclidean [1]. However, these linear techniques cannot adequately handle complex nonlinear data. Recently more emphasis is put on detecting non-linear features in the data. For example, ISOMAP [1] *etc.* extends MDS by incorporating the geodesic distances imposed by a weighted graph. It defines the geodesic distance to be the sum of edge weights along the shortest path between two nodes. The top n eigenvectors of the geodesic distance matrix are used to represent the coordinates in the new n-dimensional Euclidean space. Nevertheless, as mention in [2], in practice robust estimation of geodesic distance on a manifold is an awkward problem that require rather restrictive assumptions on the sampling. Moreover, since the MDS step in the ISOMAP algorithm intends to preserve the geodesic distance between points, it provides a correct embedding if submanifold is isometric to a convex open set of the subspace. If the submanifold is not convex, then there exist a pair of points that can not be joined by a straight line contained in the submanifold. Therefore,their geodesic distance can not be equal to the Euclidean distance. Diffusion maps [3] leverages the relationship between heat diffusion and a random walk (Markov Chain); an analogy is drawn between the diffusion operator on a manifold and a Markov transition matrix operating on functions defined on a weighted graph whose nodes were sampled from the manifold. A diffusion map, which maps coordinates between data and diffusion space, aims to re-organize data according to a new metric. In this class, we will discuss this very metric-diffusion distance and it's related properties.

1 Diffusion map, Diffusion distance

Viewing the data points x_1,x_2,...,x_n as the nodes of a weighted undirected graph $G = (V, E_W)(W = (W_{ij}))$, where the weight W_{ij} is a measure of the similarity between x_i and x_j. There are many ways to define W_{ij}, such as:

1. **Heat kernel**. If x_i and x_j are connected, put:

$$W_{ij} = e^{\frac{-\|x_i - x_j\|^2}{t}} \tag{1}$$

with some positive parameter $t \in \mathbb{R}_0^+$.

2. **Cosine Similarity**

$$W_{ij} = \cos(\angle(x_i, x_j)) = \frac{x_i}{\|x_i\|} \cdot \frac{x_j}{\|x_j\|)} \tag{2}$$

3. **Kullback-Leibler divergence**. Assume x_i and x_j are two nonvanishing probability distribution, *i.e.* $\sum_k x_i^k = 1$ and $x_i^k > 0$. Define *Kullback-Leibler divergence*

$$D^{(KL)}(x_i||x_j) = \sum_k x_i^{(k)} \log \frac{x_i^{(k)}}{x_j^{(k)}}$$

and its symmetrization $\bar{D} = D^{(KL)}(x_i||x_j) + D^{KL}(x_j||x_i)$, which measure a kind of 'distance' between distributions; *Jensen-Shannon divergence* as the symmetrization of KL-divergence between one distribution and their average,

$$D^{(JS)}(x_i, x_j) = D^{(KL)}(x_i||(x_i + x_j)/2) + D^{(KL)}(x_j||(x_i + x_j)/2)$$

A similarity kernel can be

$$W_{ij} = -D^{(KL)}(x_i||x_j) \tag{3}$$

or

$$W_{ij} = -D^{(JS)}(x_i, x_j) \tag{4}$$

The similarity functions are widely used in various applications. Sometimes the matrix W is positive semi-definite (psd), that for any vector $x \in \mathbb{R}^n$,

$$x^T W x \geq 0. \tag{5}$$

PSD kernels includes heat kernels, cosine similarity kernels, and JS-divergence kernels. But in many other cases (e.g. KL-divergence kernels), similarity kernels are not necessarily PSD. For a PSD kernel, it can be understood as a generalized covariance function; otherwise, diffusions as random walks on similarity graphs will be helpful to disclose their structures.

Define a Markov probability transition matrix A as $A = D^{-1}W$, $D = diag(\sum_{j=1}^{n} W_{ij}) \triangleq diag(d_1, d_2, \cdots, d_n)$, We used the right eigenvectors of A and corresponding eigenvalues to define the diffusion map in the following way: suppose the top right eigenvector of A is $\phi_1, \phi_2, \cdots, \phi_n$, i,e,

$$A\phi_i = \lambda_i \phi_i, \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \tag{6}$$

Then, the diffusion map $\Phi_t : V \mapsto \mathbb{R}^n$ is defined as

$$\Phi_t(x_i) = \begin{pmatrix} \lambda_1^t \phi_1(i) \\ \lambda_2^t \phi_2(i) \\ \vdots \\ \lambda_n^t \phi_n(i) \end{pmatrix} \tag{7}$$

The Euclidean distances $\|\Phi_t(x_i) - \Phi_t(x_j)\|_{\mathbb{R}^n}$ to which we refer as the diffusion distance denoted $d_t(x_i, x_j)$:

$$d_t(x_i, x_j) = \|\Phi_t(x_i) - \Phi_t(x_j)\|_{\mathbb{R}^n} \tag{8}$$

From the interpretation of the matrix A as a Markov transition probability matrix

$$A_{ij} = Pr\{s(t+1) = x_j | s(t) = x_i\} \tag{9}$$

it follows that

$$A_{ij}^t = Pr\{s(t+1) = x_j | s(0) = x_i\} \tag{10}$$

We refer to the i'th row of the matrix A^t, denoted $A_{i,*}^t$, as the probability cloud of a random walk that starts at x_i. We can express A^t using the decomposition of A. Indeed, from

$$A = \Phi \Lambda \Psi^T \tag{11}$$

we get $\Psi^T \Phi = I$, since $W = W^T$, therefore $D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$is a symmetric matrix, assume it's SVD decomposition:

$$D^{-\frac{1}{2}} W D^{-\frac{1}{2}} = V \Lambda V^T, V V^T = I \tag{12}$$

let $\Phi = D^{-\frac{1}{2}} V$,$\Psi = D^{\frac{1}{2}} V$,thus

$$A = D^{-\frac{1}{2}} D^{-\frac{1}{2}} W D^{-\frac{1}{2}} D^{\frac{1}{2}} = \Phi \Lambda \Psi^T \tag{13}$$

and $\Psi^T \Phi = V^T D^{\frac{1}{2}} D^{-\frac{1}{2}} V = I$. Then,we get

$$A^2 = \Phi \Lambda \Psi^T \Phi \Lambda \Psi^T = \Phi \Lambda^2 \Psi^T \tag{14}$$

and generally,

$$A^t = \Phi \Lambda^t \Psi^T \tag{15}$$

Written componentwise, this is equivalent to

$$A_{ij}^t = \sum_{k=1}^{n} \lambda_k^t \phi_k(i) \psi_k(j) \tag{16}$$

Lemma 1 The diffusion distance is equal to a ℓ^2 distance between the probability clouds $A_{i,*}^t$ and $A_{j,*}^t$ with weights $1/d_l$,i.e.,

$$d_t(x_i, x_j) = \|A_{i,*}^t - A_{j,*}^t\|_{\ell^2(\mathbb{R}^n, 1/d)} \tag{17}$$

Proof

$$\begin{aligned}
\|A_{i,*}^t - A_{j,*}^t\|_{\ell^2(\mathbb{R}^n,1/d)}^2 &= \sum_{l=1}^{n} (A_{il}^t - A_{jl}^t)^2 \frac{1}{d_l} \\
&= \sum_{l=1}^{n} [\sum_{k=1}^{n} \lambda_k^t \phi_k(i) \psi_k(l) - \lambda_k^t \phi_k(j) \psi_k(l)]^2 \frac{1}{d_l} \\
&= \sum_{l=1}^{n} \sum_{k,k'}^{n} \lambda_k^t (\phi_k(i) - \phi_k(j)) \psi_k(l) \lambda_{k'}^t (\phi_{k'}(i) - \phi_{k'}(j)) \psi_{k'}(l) \frac{1}{d_l} \\
&= \sum_{k,k'}^{n} \lambda_k^t \lambda_{k'}^t (\phi_k(i) - \phi_k(j))(\phi_{k'}(i) - \phi_{k'}(j)) \sum_{l=1}^{n} \frac{\psi_k(l) \psi_{k'}(l)}{d_l} \\
&= \sum_{k,k'}^{n} \lambda_k^t \lambda_{k'}^t (\phi_k(i) - \phi_k(j))(\phi_{k'}(i) - \phi_{k'}(j)) \delta_{kk'} \\
&= \sum_{k=1}^{n} \lambda_k^{2t} (\phi_k(i) - \phi_k(j))^2 \\
&= d_t^2(x_i, x_j)
\end{aligned}$$

In practice we usually do not use the mapping Φ_t but rather the truncate diffusion map Φ_t^δ that makes use of fewer than n coordinates. Specifically, Φ_t^δ uses only the eigenvectors for which the eigenvalues satisfy $|\lambda_k|^t > \delta$. When t is enough large, we can use the truncated diffusion distance:

$$d_t^\delta(x_i, x_j) = \|\Phi_t^\delta(x_i) - \Phi_t^\delta(x_j)\| = [\sum_{k:|\lambda_k|^t>\delta} \lambda_k^{2t}(\phi_k(i) - \phi_k(j))^2]^{\frac{1}{2}} \tag{18}$$

as an approximation of the weighted ℓ^2 distance of the probability clouds. We now derive a simple error bound for this approximation.

Lemma 2 $d_t^2(x_i, x_j) - \frac{2\delta^2}{d_{min}}(1-\delta_{ij}) \leq [d_t^\delta(x_i, x_j)]^2 \leq d_t^2(x_i, x_j)$, $d_{min} = \min_{1\leq i\leq n} d_i$ where $d_i = \sum_j W_{ij}$

Proof Since, $\Phi = D^{-\frac{1}{2}}V$, where V is an orthonormal matrix ($VV^T = V^TV = I$), it follows that

$$\Phi\Phi^T = D^{-\frac{1}{2}}VV^TD^{-\frac{1}{2}} = D^{-1} \tag{19}$$

Therefore,

$$\sum_{k=1}^{n} \phi_k(i)\phi_k(j) = (\Phi\Phi^T)_{ij} = \frac{\delta_{ij}}{d_i} \tag{20}$$

and

$$\sum_{k=1}^{n}(\phi_k(i) - \phi_k(j))^2 = \frac{1}{d_i} + \frac{1}{d_j} - \frac{2\delta_{ij}}{d_i} \tag{21}$$

clearly,

$$\sum_{k=1}^{n}(\phi_k(i) - \phi_k(j))^2 \leq \frac{2}{d_{min}}(1-\delta_{ij}), \quad forall \quad i,j = 1,2,\cdots,n \tag{22}$$

As a result,

$$\begin{aligned}
[d_t^\delta(x_i, x_j)]^2 &= d_t^2(x_i, x_j) - \sum_{k:|\lambda_k|^t<\delta} \lambda_k^{2t}(\phi_k(i) - \phi_k(j))^2 \\
&\geq d_t^2(x_i, x_j) - \delta^2 \sum_{k:|\lambda_k|^t<\delta} (\phi_k(i) - \phi_k(j))^2 \\
&\geq d_t^2(x_i, x_j) - \delta^2 \sum_{k=1}^{n}(\phi_k(i) - \phi_k(j))^2 \\
&\geq d_t^2(x_i, x_j) - \frac{2\delta^2}{d_{min}}(1-\delta_{ij})
\end{aligned}$$

on the other hand, it is clear that

$$[d_t^\delta(x_i, x_j)]^2 \leq d_t^2(x_i, x_j) \tag{23}$$

We conclude that

$$d_t^2(x_i, x_j) - \frac{2\delta^2}{d_{min}}(1-\delta_{ij}) \leq [d_t^\delta(x_i, x_j)]^2 \leq d_t^2(x_i, x_j) \tag{24}$$

Therefore, for small δ the truncated diffusion distance provides a very good approximation to the diffusion distance. Due to the falloff of the eigenvalues, the number of coordinates used for the truncated diffusion map is usually much smaller than n, especially when t is large.

2 Is the diffusion distance really a distance?

A distance function $d : X \times X \to \mathbb{R}$ must satisfy the following properties:

1. Symmetry: $d(x, y) = d(y, x)$
2. Non-negativity: $d(x, y) \geq 0$
3. Identity of indiscernibles: $d(x, y) = 0 \Leftrightarrow x = y$
4. Triangle inequality: $d(x, z) + d(z, y) \geq d(x, y)$

Since the diffusion map is an embedding into the Euclidean space $\mathbb{R}^n$, the diffusion distance inherits all the metric properties of $\mathbb{R}^n$ such as symmetry, non-negativity and the triangle inequality. The only condition that is not immediately implied is $d_t(x, y) = 0 \Leftrightarrow x = y$. Cleary, $x_i = x_j$ implies that $d_t(x_i, x_j) = 0$. But is it true that $d_t(x_i, x_j) = 0$ implies $x_i = x_j$? Suppose $d_t(x_i, x_j) = 0$, Then,

$$0 = d_t^2(x_i, x_j) = \sum_{k=1}^{n} \lambda_k^{2t} (\phi_k(i) - \phi_k(j))^2 \tag{25}$$

It follows that $\phi_k(i) = \phi_k(j)$ for all k with $\lambda_k \neq 0$. But there is still the possibility that $\phi_k(i) \neq \phi_k(j)$ for k with $\lambda_k = 0$. We claim that this can happen only whenever i and j have the exact same neighbors and proportional weights, that is:

$$W_{ik} = \alpha W_{jk}, \alpha > 0, \textit{for all } k = 1, 2, \cdots, n \Leftrightarrow d_t(x_i, x_j) = 0, x_i \neq x_j$$

Proof Indeed, if $d_t(x_i, x_j) = 0$, then $\sum_{k=1}^{n} \lambda_k^{2t} (\phi_k(i) - \phi_k(j))^2 = 0$ and $\phi_k(i) = \phi_k(j)$ for k with $\lambda_k \neq 0$ This implies that $d_{t'}(x_i, x_j) = 0$ for all t', because

$$d_{t'}(x_i, x_j) = \sum_{k=1}^{n} \lambda_k^{2t'} (\phi_k(i) - \phi_k(j)^2 = 0 \tag{26}$$

In particular, for $t' = 1$, we get $d_1(x_i, x_j) = 0$. But $d_1(x_i, x_j) = \|A_{i,*} - A_{j,*}\|_{\ell^2(\mathbb{R}^n, 1/d)}$, and since $\|\cdot\|_{\ell^2(\mathbb{R}^n, 1/d)}$ is a norm, we must have $A_{i,*} = A_{j,*}$, which implies $W_{ik} = W_{jk}$. In other direction, if $A_{i,*} = A_{j,*}$, then $d_1(x_i, x_j) = \sum_{k=1}^{n} \lambda_k^2 (\phi_k(i) - \phi_k(j))^2 = 0$ and therefore $\phi_k(i) = \phi_k(j)$ for k with $\lambda_k \neq 0$, from which it follows that $d_t(x_i, x_j) = 0$ for all t. □

Example In a graph with three nodes $V = \{1, 2, 3\}$ and two edges, say $E = \{(1, 2), (2, 3)\}$, the diffusion distance between nodes 1 and 3 is 0.

Summary

The mapping of points from the feature space to the diffusion map space of eigenvectors of the normalized graph Laplacian has a well defined probabilistic meaning in terms of the diffusion distance. This distance, in turn, depends on both the geometry and density of the dataset. The key concepts in the analysis of these methods, that incorporates the density and geometry of a dataset, are the characteristic relaxation times and processes of the random walk on the graph. Constructing on the finite Markov process and spectral kernel method that reflect the geometry structure of dataset, diffusion map gained a great success.

References

[1] Tenenbaum, deSilva, and Langford, A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science 290:2319-2323,22 Dec.2000.

[2] Ioannis Z. Emiris, Frank J. Sottile, Thorsten Theobald. Nonlinear computational geometry, Springer, New York, 2009.

[3] R.R.Coifman, S.Lafon, A.B.Lee, M.Maggioni, B.Nadler, F.Warner, and S.W.Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data:Diffusion maps.PNAS 102(21):7426-7431, 2005.