Mathematics for Data Sciences **2011.10.31**

Lecture 10. Diffusion Distance and Commute Time Distance

Instructor: Yuan Yao, Peking University *Scribe: Longlong Jiang, Tangjie Lv*

1 Diffusion Distance

Recall $x_i \in \mathbb{R}^p, i = 1, 2 \cdots, n$,

$$W_{ij} = \exp(-\frac{||x_i - x_j||^2}{2\varepsilon})$$

W is a symmetric $n \times n$ matrix.

Let $d_i = \Sigma_{j=1}^n W_{ij}$, $D = \text{diag}(d_i)$, and $P = D^{-1}W$, then P is a row-Markov matrix.

1) W is symmetric$\Rightarrow$ P is reversible, *i.e.* there is a probability distribution π, such that

$$\pi_i P_{ij} = \pi_j P_{ji},$$

π is a stationary distribution. P has n real eigenvalues $\{\lambda_i\}_{i=1}^n$and n independent eigenvectors $\{\phi_i\}_{i=1}^n$ such that

$$P\phi_i = \lambda_i \phi_i.$$

2) $W_{ij} > 0 \Rightarrow P$ is primitive, *i.e.*

$$1 = \lambda_1 > |\lambda_2| \geq \cdots \geq |\lambda_n|.$$

Define diffusion map

$$\Phi_t(x_i) = (\lambda_2^t \phi_2(i), \cdots, \lambda_n^t \phi_n(i))^T \in \mathbb{R}^{n-1}$$

and diffusion distance

$$\begin{aligned} d_t^2(x_i, x_j) &= ||\Phi_t(x_i) - \Phi_t(x_j)||_{l^2}^2 \\ &= ||P_{i,*}^t - P_{j,*}^t||_{l^2(\frac{1}{d_i})}^2 \\ &= \sum_k \frac{(P_{i,k}^t - P_{j,k}^t)^2}{d_k}. \end{aligned}$$

Diffusion distance depends on time scale parameter t which is hard to select in applications. In this section we introduce another closely related distance, namely *commute time distance*, derived from mean first passage time. For such distances we do not need to choose the time scale t.

2 Commute Time Distance

Definition.

1. *First passage time (or hitting time)*: $\tau_{ij} := \inf(t \geq 0 | x_t = j, x_0 = i)$;
2. *Mean First Passage Time*: $T_{ij} = \mathbb{E}_i \tau_{ij}$;
3. $\tau_{ij}^+ := \inf(t > 0 | x_t = j, x_0 = i)$, where τ_{ii}^+ is also called *first return time*;
4. $T_{ij}^+ = \mathbb{E}_i \tau_{ij}^+$, where T_{ii}^+ is also called *mean first return time.*

Here $\mathbb{E}_i$ denotes the conditional expectation for fixing initial condition $x_0 = i$.

All the below will show that the (average) commute time between x_i and x_j, *i.e.* $T_{ij} + T_{ji}$, in fact leads to an Euclidean distance metric which can be used for embedding.

Theorem 2.1. $d_c(x_i, x_j) := \sqrt{T_{ij} + T_{ji}}$ is an Euclidean distance metric, called *commute time distance.*

Proof. We will give a constructive proof that $T_{ij} + T_{ji}$ is a squared distance of some Euclidean coordinates for x_i and x_j.

By definition, we have

$$T_{ij}^+ = P_{ij} \cdot 1 + \sum_{k \neq j} P_{ik} (T_{kj}^+ + 1) \tag{1}$$

Let $E = 1 \cdot 1^T$ where $1 \in \mathbb{R}^n$ is a vector with all elements one, $T_d^+ = \text{diag}(T_{ii}^+)$. Then 1 becomes

$$T^+ = E + P(T^+ - T_d^+). \tag{2}$$

For stationary distribution π, $\pi^T P = P$, whence we have

$$\begin{aligned}
\pi^T T^+ &= \pi^T 1 \cdot 1^T + \pi^T P (T^+ - T_d^+) \\
\pi^T T^+ &= 1^T + \pi^T T^+ - \pi^T T_d^+ \\
1 &= T_d^+ \pi \\
T_{ii}^+ &= \frac{1}{\pi_i}
\end{aligned}$$

Since $T = T^+ - T_d^+$, then (1) becomes

$$\begin{aligned}
T &= E + PT - T_d^+ \\
(I - P)T &= E - T_d^+ \\
(I - D^{-1}W)T &= F \\
(D - W)T &= DF \\
LT &= DF
\end{aligned}$$

where $F = E - T_d^+$ and $L = D - W$ is the (unnormalized) graph Laplacian. Since L is symmetric and irreducible, we have $L = \sum_{k=1}^n \mu_k \nu_k \nu_k^T$, where $0 = \mu_1 < \mu_2 \leq \cdots \leq \mu_n, \nu_1 = 1/||1||, \nu_k^T \nu_l = \delta_{kl}$. Let $L^+ = \sum_{k=2}^n \frac{1}{\mu_k} \nu_k \nu_k^T$, L^+ is called the *pseudo-inverse* (or *Moore-Penrose inverse*) of L. We can test and verify L^+ satisfies the following four conditions

$$\begin{cases}
L^+ L L^+ = L^+ \\
L L^+ L = L \\
(L L^+)^T = L L^+ \\
(L^+ L)^T = L^+ L
\end{cases}$$

From $LT = D(E - T_d^+)$, multiplying both sides by L^+ leads to

$$T = L^+DE - L^+DT_d^+ + 1 \cdot u^T,$$

as $1 \cdot u^T \in \ker(L)$, whence

$$\begin{aligned} T_{ij} &= \sum_{k=1}^{n} L_{ik}^+ d_k - L_{ij}^+ d_j \cdot \frac{1}{\pi_j} + u_j \\ u_i &= -\sum_{k=1}^{n} L_{ik}^+ d_k + L_{ii}^+ vol(G), \quad j = i \\ T_{ij} &= \sum_k L_{ik}^+ d_k - L_{ij}^+ vol(G) + L_{jj}^+ vol(G) - \sum_k L_{jk}^+ d_k \end{aligned}$$

Note that $vol(G) = \sum_i d_i$ and $\pi_i = d_i / vol(G)$ for all i.

Then

$$T_{ij} + T_{ji} = vol(G)(L_{ii}^+ + L_{jj}^+ - 2L_{ij}^+). \tag{3}$$

To see it is a squared Euclidean distance, we need the following lemma.

Lemma 2.2. If K is a symmetric and positive semidefinite matrix, then

$$K(x,x) + K(y,y) - 2K(x,y) = d^2(\Phi(x), \Phi(y)) = \langle \Phi(x), \Phi(x) \rangle + \langle \Phi(y), \Phi(y) \rangle > -2\langle \Phi(x), \Phi(y) \rangle$$

where $\Phi = (\phi_i : i = 1, \ldots, n)$ are orthonormal eigenvectors with eigenvalues $\mu_i \geq 0$, such that $K(x,y) = \sum_i \mu_i \phi_i(x) \phi_i(y)$.

Clearly L^+ is a positive semidefinite matrix and we define the *commute time map* by its eigenvectors,

$$\Psi(x_i) = \left(\frac{1}{\sqrt{\mu_2}} \nu_2(i), \cdots, \frac{1}{\sqrt{\mu_n}} \nu_n(i) \right)^T \in \mathbb{R}^{n-1}.$$

then $L_{ii}^+ + L_{jj}^+ - 2L_{ij}^+ = ||\Psi(x_i) - \Psi(x_j))||_{l^2}^2$, and we call $d_r(x_i, x_j) = \sqrt{L_{ii}^+ + L_{jj}^+ - 2L_{ij}^+}$ the *resistance distance*.

So we have $d_c(x_i, x_j) = \sqrt{T_{ij} + T_{ji}} = \sqrt{vol(G)} d_r(x_i, x_j)$. □

3 Comparisons between diffusion map and commute time map

Table 1: Comparisons between diffusion map and commute time map. Here $x \sim y$ means that x and y are in the same cluster and $x \nsim y$ for different clusters.

Diffusion Map	Commute Time Map
P's right eigenvectors	L^+'s eigenvectors
scale parameters: t and ε	scale: ε
$\exists t$, s.t. $x \sim y$, $d_t(x,y) \to 0$ and $x \nsim y$, $d_t(x,y) \to \infty$	$x \sim y$, $d_c(x,y)$ small and $x \nsim y$, $d_c(x,y)$ large?

However, recently Radl, von Luxburg, and Hein give a *negative* answer for the last desired property of $d_c(x, y)$ in geometric random graphs. Their result is as follows. Let $\mathcal{X} \subseteq \mathbb{R}^p$ be a compact set and let $k : \mathcal{X} \times \mathcal{X} \to (0, +\infty)$ be a symmetric and continuous function. Suppose that $(x_i)_{i\in\mathbb{N}}$ is a sequence of data points drawn i.i.d. from $\mathcal{X}$ according to a density function $p > 0$ on $\mathcal{X}$. Define $W_{ij} = k(x_i, x_j)$, $P = D^{-1}W$, and $L = D - W$. Then Radl et al. shows

$$\lim_{n\to\infty} n d_r(x_i, x_j) = \frac{1}{d(x_i)} + \frac{1}{d(x_j)}$$

where $d(x) = \int_X k(x, y)dp(y)$ is a smoothed density at x, $d_r(x_i, x_j) = \frac{d_c(x_i,x_j)}{\sqrt{vol(G)}}$ is the resistance distance. This result shows that in this setting commute time distance has no information about cluster information about point cloud data, instead it simply reflects density information around the two points.