MATH 690: Topics in Data Analysis and Computation	September 26, 2017

Dimension Reduction

Lecturer: Xiuyuan Cheng *Scribe: Didong Li*

1 tSNE

Definition 1. *Let p and q be two probabilist distribution on space Ω, then the Kullback-Leibler divergence between p and q is*

$$KL(p||q) \coloneqq \int_\Omega p(x) \log \frac{p(x)}{q(x)} \,\mathrm{dx}$$

when p is continuous, and

$$KL(p||q) \coloneqq \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i}$$

when p is discrete.

Proposition 1. *$D(p||q) \geq 0$, $KL(p||q) = 0 \Longleftrightarrow p = q$.*

Proof. $\Longleftarrow$: Trivial.

$\Longrightarrow$: Since $\log x \leq x - 1$, $\log x = x - 1 \Longleftrightarrow x = 1$,

$$\begin{aligned} KL(p||q) &= -\sum_i p_i \frac{q_i}{p_i} \\ &\geq -\sum_i p_i (\frac{q_i}{p_i} - 1) \\ &= 0, \end{aligned}$$

and the inequality is equality if and only if $\frac{q_i}{p_i} - 1 = 1$, that is $p_i = q_i$, $\forall i$. □

Definition 2. $K(x_i, x_j) \coloneqq e^{-\frac{\|x_i - x_j\|^2}{2\sigma_i^2}}$, $P_{ij} \coloneqq \frac{K(x_i,x_j)}{\sum_{k\neq l} K(x_k,x_l)}$, $\bar{P}_{ij} \coloneqq \frac{P_{ij}+P_{ji}}{2}$.

The t-distribution stochastic neighborhood embedding(tSNE) algorithm is to find the optimal $Y = \{y_i\}_{i=1}^n$ to minimize the function

$$\min_Y KL(P||Q) = \sum_{i\neq j} P_{ij} \log \frac{P_{ij}}{Q_{ij}},$$

where $Q_{ij} \coloneqq \frac{K(y_i,y_j)}{\sum_{k\neq l} K(y_k,y_l)}$.

Remark 1. *The kernel has heavy tail so it decays slowly, so p could be fitted better when the dimension is high.*

2 Convergence of Graph Laplacian

Let $\{x_i\}_{i=1}^n \subset \mathbb{R}^D$ be the observed data set. Assume x_i is sampled from some d dimensional manifold M embedded in $\mathbb{R}^D$.

Definition 3. $K_\epsilon(x,y) \coloneqq (2\pi\epsilon)^{-\frac{d}{2}} \exp^{-\frac{\|x-y\|^2}{2\epsilon}}$ *is called the heat kernel parametrized by ϵ.*

Recall the following definitions in the previous lectures:

Definition 4. $W_{ij} \coloneqq K_\epsilon(x_i, x_j)$, $L_{n,\epsilon} \coloneqq L_{rw} \coloneqq= I - P \coloneqq= I - D^{-1}W$, *where* $D = \operatorname{diag}\{d_{ii}\}_{i=1}^n$, $d_{ii} = \sum_{j=1}^n W_{ij}$.

The goal of this lecture is to prove the following theorem.

Theorem 1. *Assume $x_i \sim p$, let $u(x) = -2\log p(x)$, then*

$$\frac{1}{\epsilon} L_{n,\epsilon} \xrightarrow[\epsilon \to 0]{n \to \infty} -\frac{1}{2}\Delta_M - \nabla u \cdot \nabla,$$

where Δ_M is the Beltrami-Laplacian operator on M. In particular, when x_i is sampled from uniform distribution, $\nabla u = 0$, so

$$\frac{1}{\epsilon} L_{n,\epsilon} \xrightarrow[\epsilon \to 0]{n \to \infty} -\frac{1}{2}\Delta_M.$$

Proof. We prove this theorem in two steps:
Step 1: $\frac{1}{\epsilon} L_{n,\epsilon} \xrightarrow{n \to \infty} L_\epsilon$.
Step 2: $\frac{1}{\epsilon} L_\epsilon \xrightarrow{\epsilon \to 0} L = -\frac{1}{2}\Delta_M - \nabla u \cdot \nabla$.

1. Proof of Step 1.
   For any $v \in \mathbb{R}^n$,

   $$[L_{n,\epsilon}(v)](i) = [(I - D^{-1}W)v](i) = v(i) - \frac{\sum_j W_{ij} v(j)}{D_{ii}},$$

   where $v(i)$ denotes the i-th coordinate of vector v. Rewrite $L_{n,\epsilon}$ in term of kernel function:

   $$[L_{n,\epsilon}(v)](i) = v(i) - \frac{\sum_j K_\epsilon(x_i, x_j) v(j)}{\sum_j K_\epsilon(x_i, x_j)}. \tag{1}$$

   Then we can extend $L_{n,\epsilon} : \mathbb{R}^n \to \mathbb{R}^n$ to a operator $\bar{L}_{n,\epsilon} : C^2(M) \to C(M)$, defined by

   $$[\bar{L}_{n,\epsilon}(f)](x) \coloneqq f(x) - \frac{\sum_j K_\epsilon(x, x_j) f(x_j)}{\sum_j K_\epsilon(x, x_j)} \tag{2}$$

Suppose ψ is an eigenfunction of $\bar{L}_{n,\epsilon}$, that is, $L_{n,\epsilon} = \lambda\phi$, then the discrete version of ϕ: $(\phi(x_1), \cdots, \phi(x_n))$ is an eigenvector of $L_{n,\epsilon}$ associated with eigenvalue λ.

Assume x_i is sampled from uniform distribution, then by the Law of Large Numbers, as $n \to \infty$

$$\frac{1}{n}\sum_j K_\epsilon(x, x_j) f(x_j) \to \int_M K_\epsilon(x, y) f(y)\,\mathrm{dV}(y), \tag{3}$$

$$\frac{1}{n}\sum_j K_\epsilon(x, x_j) \to \int_M K_\epsilon(x, y)\,\mathrm{dV}(y), \tag{4}$$

where dV is the volume form of M. As a result,

$$\begin{aligned}
[\bar{L}_{n,\epsilon} f](x) \xrightarrow{n \to \infty} L_\epsilon &= f(x) - \frac{\int_M K_\epsilon(x, y) f(y)\,\mathrm{dV}(y)}{\int_M K_\epsilon(x, y)\,\mathrm{dV}(y)} \\
&= f(x) - \frac{\int_M e^{-\frac{\|x-y\|^2}{2\epsilon}} f(y)\,\mathrm{dV}(y)}{\int_M e^{-\frac{\|x-y\|^2}{2\epsilon}}\,\mathrm{dV}(y)}
\end{aligned}$$

2. Proof of Step 2.

   (a) Consider the simplest case first: $M = [0, 1]$. Then (3) becomes

$$\begin{aligned}
\int_0^1 e^{-\frac{(x-y)^2}{2\epsilon}} f(y)\,\mathrm{dy} &= \int_{-\frac{x}{\sqrt{\epsilon}}}^{\frac{1-x}{\sqrt{\epsilon}}} e^{-\frac{z^2}{2}} f(x + \sqrt{\epsilon} z)\sqrt{\epsilon}\,\mathrm{dz} \\
&= \sqrt{\epsilon}\int_{-\frac{x}{\sqrt{\epsilon}}}^{\frac{1-x}{\sqrt{\epsilon}}} e^{-\frac{z^2}{2}} [f(x) + \sqrt{\epsilon} z f'(x) + \frac{\epsilon z^2}{2} f''(x) + o(\epsilon^{\frac{3}{2}})]\,\mathrm{dz} \\
&= \sqrt{\epsilon}(f(x)\int_{-\frac{x}{\sqrt{\epsilon}}}^{\frac{1-x}{\sqrt{\epsilon}}} e^{-\frac{z^2}{2}}\,\mathrm{dz} + \frac{\epsilon}{2} f''(x)\int_{-\frac{x}{\sqrt{\epsilon}}}^{\frac{1-x}{\sqrt{\epsilon}}} e^{-\frac{z^2}{2}} z^2\,\mathrm{dz} + o(\epsilon^{\frac{3}{2}})) \\
&= \sqrt{2\pi\epsilon}(f(x) + \frac{\epsilon}{2} + o(\epsilon^{\frac{3}{2}})).
\end{aligned}$$

   The last equation holds when ϵ is sufficiently small, since the odd order moments of standard normal distribution are all zero. Similarly, (4) could be written as

$$\int_0^1 e^{-\frac{(x-y)^2}{2\epsilon}}\,\mathrm{dy}(y) = \sqrt{2\pi\epsilon}.$$

   As a result,

$$L_\epsilon = f(x) - (f(x) + \frac{\epsilon}{2} f''(x) + o(\epsilon^{\frac{3}{2}})) = -\frac{\epsilon}{2} f''(x) - o(\epsilon^{\frac{3}{2}}),$$

$$\frac{1}{\epsilon} L_\epsilon f \xrightarrow{\epsilon \to 0} -\frac{1}{2} f''.$$

(b) When the manifold is $c(t) = \begin{bmatrix} t \\ at^2 + o(t^3) \end{bmatrix}$. When the density is uniform, similar computation simplifies equation (3) as:

$$\int K_\epsilon(x,y)f(y)\,\mathrm{ds(y)} = f(x) + c\frac{\epsilon}{2}(f''(x) + a^2 f(x)) + o(\epsilon^2) \tag{5}$$

When the density is p, (3)/(4) is

$$\frac{fp + c\frac{\epsilon}{2}((fp)'' + a^2 fp + o(\epsilon^2)}{p + c\frac{\epsilon}{2}(p'' + 2f'\frac{p'}{p}) + o(\epsilon^2)} = f + c\frac{\epsilon}{2}(f'' + 2f'\frac{p'}{p}) + o(\epsilon^2).$$

As a result,

$$\frac{1}{\epsilon}\bar{L}_\epsilon(f) = c(f'' + 2f'\frac{p'}{p}) + o(\epsilon) \xrightarrow[\epsilon \to 0]{} c(f'' - u'f'),$$

where $u = -2\log f$.

(c) For general case, that is $d > 1$,

$$\frac{1}{\epsilon}\bar{L}_\epsilon \xrightarrow[\epsilon \to 0]{} \Delta_M f - \nabla u \cdot \nabla f.$$

$\square$

Remark 2. *When $d > 1$, let $\{s_1, \cdots, s_d\}$ be the local orthonormal basis, then the correction term, a^2 in (5), is*

$$E(x) = \sum_{i=1}^{d} a_i(x) - \sum_{i \neq j} a_i(x)a_j(x),$$

where $a_i(x)$ is the directional curvature of s_i.