A Note of Diffusion Map

Singyuan Yeh

2020 NCTS Mini-Course

July, 2020

Table of Contents

Outline

Definition

Given n data points $\{x_1, \cdots, x_n\}$ in $\mathbb{R}^d$. Consider a function $f : \mathbb{R}^d \to \mathbb{R}$. Write $f_i = f(x_i)$, $1 \leq i \leq n$. We hope to minimize

$$E(f) = \sum_{i,j=1}^{n} A_{ij} \left(f_i - f_j \right)^2 .$$

The matrix A is symmetric transition matrix, *i.e.* $A_{ij} \geq 0$ and $\sum_i A_{ij} = \sum_j A_{ij} = 1$.

Definition

Let's look at the quadratic form $A_{ij}\left(f_i - f_j\right)^2$.

$$\sum_{i,j=1}^{n} A_{ij}\left(f_i - f_j\right)^2 = \sum_{i=1}^{n} f_i^2 \left(\sum_{j=1}^{n} A_{ij}\right) - 2\sum_{i,j=1}^{n} f_i A_{ij} f_j + \sum_{j=1}^{n} f_j^2 \left(\sum_{i=1}^{n} A_{ij}\right)$$
$$= \sum_{i=1}^{n} f_i^2 + 2\sum_{i,j=1}^{n} f_i A_{ij} f_j + \sum_{j=1}^{n} f_j^2$$
$$= 2f^T(I - A)f$$

Define $L = I - A$. Hope to minimize $E(f) = 2f^T Lf$, subject to some constraints that is not mentioned here.

Relation

Now we turn our attention to Euclidean $\mathbb{R}^1$ Laplacian operator.

$$L = \begin{bmatrix} \ddots & & & & \\ & 1 & -2 & 1 & \\ & & & & \ddots \end{bmatrix}$$

Then, matrix $A = \begin{bmatrix} \ddots & & & & \\ & 1 & 0 & 1 & \\ & & & & \ddots \end{bmatrix}$ and $D = \begin{bmatrix} \ddots & & \\ & -2 & \\ & & \ddots \end{bmatrix}$, where A is adjacent matrix in Euclidean space and $D_{ii} = -\sum_j A_{ij}$ shows degree of point i. Therefore $-L = D - A$ is roughly called **unnormalized** Laplacian operator.

Outline

Affinity graph

Given a pair of graph $G = (V, E)$. If we allow the **affinity** function $\omega : E \to \mathbb{R}_+$, then we called **affinity** graph $G = (V, E, \omega)$.

Remark

A function ω can regarded as some kind of "distance function" between two vertices. We can convert a Euclidean discrete space into an affinity graph by setting a function $\omega(i, j) = 1$ for all $(i, j) \in E$.

Affinity matrix

- Given a graph $G = (V, E)$ and $|V| = n$, the adjacency matrix of G is matrix $W \in \mathbb{R}^{n \times n}$ defined by

$$W_{i,j} = \begin{cases} 1 & \text{if } (i,j) \in E \\ 0 & \text{otherwise} \end{cases}.$$

- Given a **affinity** graph $G = (V, E, W)$ and $|V| = n$, the adjacency matrix of G is matrix $W \in \mathbb{R}^{n \times n}$ defined by

$$W_{i,j} = \begin{cases} \omega_{ij} & \text{if } (i,j) \in E \\ 0 & \text{otherwise} \end{cases}.$$

Degree matrix

Let G is affinity graph. The degree function $d : V \to \mathbb{R}_+$ is defined by

$$d(i) = \sum_{(i,j)\in E} W_{ij} .$$

The degree matrix $D \in \mathbb{R}^{n\times n}$ is defined by a diagonal matrix

$$D = \begin{bmatrix} d(1) & & \\ & \ddots & \\ & & d(n) \end{bmatrix}.$$

Outline

Graph Laplacian

Definition

We focus on **undirected** graph $G = (V, E)$ with n vertices. Let $G = (V, E, \omega)$ be an undirected affinity graph.

- The unnormalized graph Laplacian (GL) is defined as $\tilde{L} = D - W$.
- If there is no isolated vertex, the normalized graph Laplacian (NGL) is defined as $L = I_n - D^{-1}W$. (NOT necessary symmetric)
- The **symmetrized** normalized graph Laplacian is defined as $\mathcal{L} = I_n - D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$.

Some relation between definition:

- $L = D^{-1}\tilde{L}$. *i.e.* Normalized $\tilde{L}$.
- $\mathcal{L} = D^{\frac{1}{2}}(I - D^{-1}W)D^{-\frac{1}{2}} = D^{\frac{1}{2}}LD^{-\frac{1}{2}}$. *i.e.* L is similar to $\mathcal{L}$.

Transition matrix

Definition

Define the transition matrix of the random walk on the graph as $A = D^{-1}W$. It is **NOT** necessary symmetric.

Proposition

- $\sum_j A_{ij} = 1$
- $\sum_{j=1}^{n} \left(A^k\right)_{ij} = 1$

Remark: . A entry A_{ij} can be thought of as the probability of moving from i to j in one step of a random walk on G.

Proof of proposition

Proposition

- $\sum_j A_{ij} = 1$
- $\sum_{j=1}^{n} \left(A^k\right)_{ij} = 1$

Proof:

- $\sum_j A_{ij} = \sum_j \frac{1}{d_i} \delta_{ik} W_{kj} = \sum_j \frac{1}{d_i} W_{ij} = 1$
-

$$
\begin{aligned}
\sum_{j=1}^{n} \left(A^k\right)_{ij} &= \sum_{j_1, \cdots, j_k, j=1}^{n} A_{ij_1} A_{j_1 j_2} \cdots A_{j_{k-1} j_k} A_{j_k j} \\
&= \sum_{j_1=1}^{n} A_{ij_1} \sum_{j_2=1}^{n} A_{j_1 j_2} \cdots \sum_{j_k=1}^{n} A_{j_{k-1} j_k} \sum_{j=1}^{n} A_{j_k j} \\
&= 1
\end{aligned}
$$

Notation

Before some basic spectral properties of the GL are provided, we introduced some notation.

- Denote $\sigma(M)$ to be the spectrum of a given matrix M.
- Denote $\rho(M)$ to be the associated spectral radius $\rho(M) = \max_{f \neq 0} \frac{\|f^T M f\|}{f^T f} = \max\{|\lambda| : \lambda \in \sigma(M)\}$

Definition

The Rayleigh quotient of a matrix $M \in \mathbb{R}^{n \times n}$ is defined as

$$RM(v) = \frac{\langle v, Mv \rangle}{\langle v, v \rangle}$$

Nonnegative definite GL (1)

Proposition of GL

The unnormalized graph Laplacian $\tilde{L} = D - W$ is nonnegative definite and $\sigma(\tilde{L}) \subset [0, 2\rho(D)]$.

Proof: Let $f \in \mathbb{R}^n$ and $d_i = \sum_j W_{ij}$. Since $W_{ij} \geq 0$ and $W_{ij} = W_{ji}$,

$$
\begin{aligned}
f^T \tilde{L} f = f^T (D - W) f &= \sum_{i,j=1}^{n} f_i \left(d_i \delta_{ij} - W_{ij} \right) f_j = \sum_{i=1}^{n} d_i f_i^2 - \sum_{i,j=1}^{n} f_i W_{ij} f_j \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij} f_i^2 - \sum_{i,j=1}^{n} f_i W_{ij} f_j \\
&= \frac{1}{2} \left(\sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij} f_i^2 + \sum_{j=1}^{n} \sum_{i=1}^{n} W_{ji} f_j^2 - \sum_{i,j=1}^{n} 2 f_i W_{ij} f_j \right) \\
&= \frac{1}{2} \sum_{i,j=1}^{n} W_{ij} (f_i - f_j)^2 \geq 0
\end{aligned}
$$

Nonnegative definite GL (2)

Proof: Since above equality holds if $f_1 = \cdots f_n$, which implies the smallest eigenvalue of $\tilde{L}$ is 0 w.r.t. eigenvector $\mathbf{1}$.
On the other hands, since $(f_i - f_j)^2 \le 2(f_i^2 + f_j^2)$ and $\rho(D) = \max_i d_i$,

$$\begin{aligned} f^T \widetilde{L} f &= \frac{1}{2} \sum_{i=1,j=1}^{n} W_{ij} \left(f_i - f_j\right)^2 \le \sum_{i=1,j=1}^{n} W_{ij} \left(f_i^2 + f_j^2\right) \\ &= \sum_{i=1,j=1}^{n} W_{ij} f_i^2 + \sum_{i=1,j=1}^{n} W_{ij} f_j^2 \\ &= 2 \sum_i d_i f_i^2 \le 2\rho(D) f^T f \end{aligned}$$

Some remarks of eigenvalue of NG

- From above, we know $\mathbf{1}^T \tilde{L}\mathbf{1} = 0$. Furthermore, $\tilde{L}\mathbf{1} = 0\mathbf{1} = 0$
- Since $\tilde{L} = DL$ and D is invertible, $0 = L\mathbf{1} = (I - D^{-1}W)\mathbf{1}$. Hence, $D^{-1}W\mathbf{1} = \mathbf{1}$. Thus, 1 is eigenvalue of A.
- Note that A is not necessary symmetric. Since A is similar to $D^{\frac{1}{2}}AD^{-\frac{1}{2}} = D^{\frac{1}{2}}(D^{-1}W)D^{-\frac{1}{2}} = D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$, which is symmetric.
- As mentioned above, $\mathcal{L}$ is similar to $L = D^{-1}W$. If λ is eigenvalue of A, then $1 - \lambda$ is eigenvalue of L and $\mathcal{L}$.

Some spectrum properties

Lemma

$\rho(A) = 1$, $\sigma(A) \subset [-1, 1]$ and $\sigma(L) = \sigma(\tilde{L}) \subset [0, 2]$.

Diagonalization of GL (1)

1. As mentioned above, A is similar to symmetric matrix $D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$.
2. Since $D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$ is symmetric, exist orthonormal matrix O (*i.e.* $O^TO = 1$) such that

$$D^{-\frac{1}{2}}WD^{-\frac{1}{2}} = O\Lambda O^T$$

where diagonal matrix $\Lambda = \begin{bmatrix} \mu_1 & & \\ & \ddots & \\ & & \mu_n \end{bmatrix}$ and $1 = |\mu_1| \geq |\mu_2| \geq \cdots \geq |\mu_n|$.

3. The eigenvalue of L and $\mathcal{L}$ is $\{\lambda_i = 1 - \mu_i\}$

Diagonalization of GL (2)

4. Since $D^{-\frac{1}{2}}WD^{-\frac{1}{2}} = O\Lambda O^T$, we can build a relation to A

$$\begin{aligned} A &= D^{-1}W = D^{-\frac{1}{2}}D^{-\frac{1}{2}}WD^{-\frac{1}{2}}D^{\frac{1}{2}} = D^{-\frac{1}{2}}O\Lambda O^T D^{\frac{1}{2}} \\ &= U\Lambda V^T, \end{aligned}$$

where $U = D^{-\frac{1}{2}}O$ and $V = D^{\frac{1}{2}}O$.

Properties of $U\Lambda V^T$ (1)

Proposition of $U\Lambda V^T$

1. $UV^T = VU^T = U^TV = V^TU = I$
2. $AU = U\Lambda$ and $V^TA = \Lambda V^T$
3. Denote two vectors $u = \frac{1}{n}\mathbf{1}$ and $v = \frac{1}{\sum d_i}[d_1, \cdots, d_n]^T$, which are normalized by 1-norm, *i.e.* $\|\cdot\|_1$. Then, $Au = u$ and $v^TA = v^T$.

Proof:

1. Plug $U = D^{-\frac{1}{2}}O$ and $V = D^{\frac{1}{2}}O$ into equation.
2. By $A = U\Lambda V^T$, the following can computed directly

$$AU = U\Lambda V^TU = U\Lambda.$$

Properties of $U \Lambda V^T$ (2)

3. First, we know $A\mathbf{1} = \mathbf{1}$, so $Au = u$ done!.
   Second, since $A = D^{-1}W$, we can get

$$\left(v^T A\right)_j = \sum_i v_i A_{ij} = \frac{\sum_i d_i \frac{w_{ij}}{d_i}}{\sum_l d_l} = \frac{\sum_i W_{ij}}{\sum_l d_l} = \frac{d_j}{\sum_l d_l} = v_j$$

Now, it's sufficient to introduce **diffusion map**.

Eigenmap (1)

As setting above, let $A = D^{-1} W = U \Lambda V^T$ where $\Lambda = \text{diag}(\mu_1, \cdots, \mu_n)$ with $1 = \mu_1 \geq \mu_2 \geq \cdots \geq \mu_n$. Take m with $m + 1 \leq n$. The m-dimension eigenmap for ith-vertex is defined as

$$\text{Eig}_m(i) = [u_2(i), \cdots, u_{m+1}(i)]^T$$

Remark

It map vertex i in $\mathbb{R}^n$ to $\mathbb{R}^m$, where $m \leq n$.

Eigenmap (2)

Note that the $n \times m$ (n data reduced in m dimension) matrix

$$\begin{bmatrix} \mathrm{Eig}_m(1)^T \\ \mathrm{Eig}_m(2)^T \\ \vdots \\ \mathrm{Eig}_m(n)^T \end{bmatrix} = \begin{bmatrix} u_2(1) & u_3(1) & \cdots & u_{m+1}(1) \\ u_2(2) & u_3(2) & \cdots & u_{m+1}(2) \\ \vdots & \vdots & & \vdots \\ u_2(n) & u_3(n) & \cdots & u_{m+1}(n) \end{bmatrix} = \begin{bmatrix} | & | & & | \\ u_2 & u_3 & \cdots & u_{m+1} \\ | & | & & | \end{bmatrix}$$

Diffusion map

Definition

As setting above, let $A = D^{-1}W = U\Lambda V^T$ where $\Lambda = \mathrm{diag}(\mu_1, \cdots, \mu_n)$ with $1 = \mu_1 \geq \mu_2 \geq \cdots \geq \mu_n$. Take diffusion time $t > 0$. The diffusion map (DM) $\Phi_t : V \to \mathbb{C}^{n-1}$ is defined by

$$\Phi_t(i) = \left[\mu_2^t u_2(i), \mu_3^t u_3(i), \cdots, \mu_n^t u_n(i)\right]^T .$$

Furthermore, if all eigenvalue are nonnegative, then the embedding is into $\mathbb{R}^n$.

Truncated diffusion map (tDM)

Definition

As setting above, let $A = D^{-1}W = U\Lambda V^T$ where $\Lambda = \mathrm{diag}(\mu_1, \cdots, \mu_n)$ with $1 = \mu_1 \geq \mu_2 \geq \cdots \geq \mu_n$. Take diffusion time $t > 0$. The truncated diffusion map with time t and threshold δ is a map $\Phi_t^\delta : V \to \mathbb{C}^{m(t,\delta)-1}$ is defined by

$$\Phi_t^\delta(i) = \left[\mu_2^t u_2(i), \cdots, \mu_{m(t,\delta)}^t u_{m(t,\delta)}(i)\right]^T,$$

where $m(t,\delta) := \max\left\{i : |\mu_i|^t > \delta\, |\mu_2|^t\right\}$.

Outline

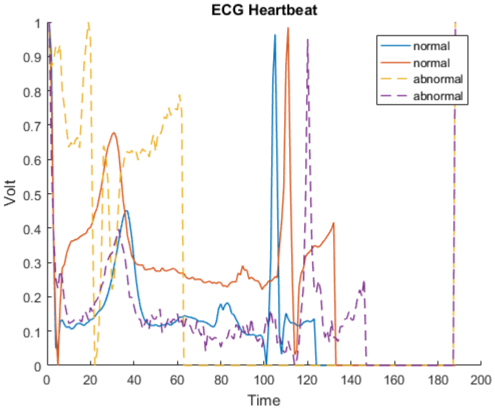
Introduction to dataset

This dataset is from website kaggle, called The PTB Diagnostic ECG Database. This dataset has been used in exploring heartbeat classification. The signals correspond to electrocardiogram (ECG) shapes of heartbeats for the normal case and the cases affected by PTB diagnostic (lung problem).

Because it is a large dataset, I just use 1000 normal and 500 abnormal to train the model.

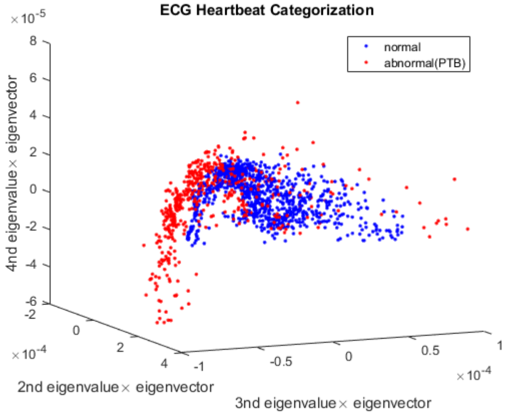
Plot the data

The following figure is about ECG heartbeats with PTB diagnostic or without PTB diagnostic.

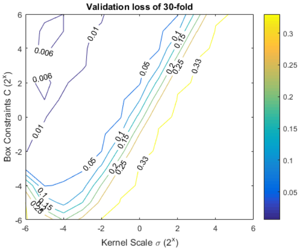


Diffusion map

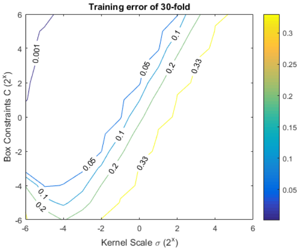
Choose kernel scale $\sigma = 0.005$. Then,



Tuning procedure between box constraints C and kernel scale σ by 30-fold.



(a) Validation loss of SVM (σ, C)



(b) Training error of SVM (σ, C)

The best cross validation loss is 0.0047 with $(\sigma, C) = (0.03125, 4)$.

Outline

Reference

[1] Maopei Tsui, *2020 NCTS Mini-Course on Manifold Learning Lecture Notes*.

[2] Hautieng Wu, *Mathematics of Massive Data Analysis*.