01.12.2017

4.3. Diffusion maps

$x_1, \dots, x_n \in \mathbb{R}^p$

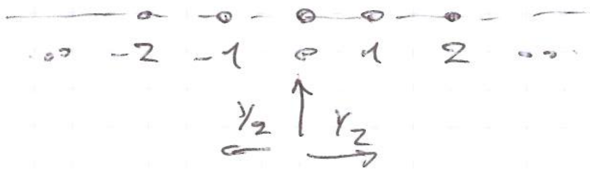
$G(\underline{\underline{V}}, \underline{\underline{E}}, \underline{\underline{W}})$

$v_i \longleftrightarrow x_i$

First construct a homogenous random walk $X_t, t = 0, 1, \dots,$ on $V = \{v_1, \dots, v_n\}$.

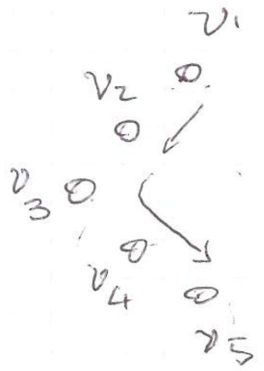
Random Walk:

Path on a space from random steps.



$$P(X_t = j \mid X_{t-1} = i) = \begin{cases} 1/2 & j = i-1, i+1 \\ 0 & \text{o.w.} \end{cases}$$

transition probability.



(equilateral triangle)

$$P(X_t = v_j \mid X_{t-1} = v_i) \propto d(v_i, v_j)$$

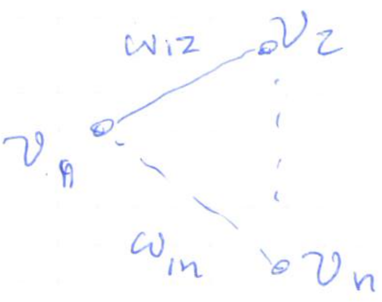
$t = 1 \quad P(X_1 = v_5 \mid X_0 = v_1) = P(X_1 = v_3 \mid X_0 = v_1)$

$t = 2 \quad P(X_2 = v_5 \mid X_0 = v_1) < P(X_2 = v_3 \mid X_0 = v_1)$

Transition matrix $\to$ Stochastic matrix

$$M = (m_{ij})_{i,j=1,\dots,n} \qquad 0 \le m_{ij} \le 1$$

$$\sum_{j=1}^{n} m_{ij} = 1$$



$$m_{12} = \frac{w_{12}}{\sum_i w_{1i}}$$

$$\deg(i) = \sum_j w_{ij}$$

$$W = (w_{ij})_{1 \le i,j \le n} \qquad M = (m_{ij})_{1 \le i,j \le n}$$

- $P(X_{t+1} = j \mid X_t = i) = m_{ij}$
- $m_{ij} = \dfrac{w_{ij}}{\deg(i)} \Rightarrow M = D^{-1} W$

$$D = \mathrm{diag}(\deg(1), \dots, \deg(n))$$

$$P(X_t = j \mid X_0 = i) = ?$$

$$P(X_0 = i) = p_i \to P(X_1 = j) = \sum P(X_0 = i) P(X_1 = j \mid X_0 = i)$$

$$\to P(X_1 = j) = \sum p_i m_{ij}$$

$$p = (p_1, \dots, p_n)$$

$$p^T M = \underbrace{(P(X_1 = 1), \dots, P(X_1 = n))}_{p^{(1)T}}$$

$$\longrightarrow \cancel{p^T} \, \underline{\underline{M}}^{t} = p^{(t)\,\cancel{T}}$$

state proba at time t.

$M^t \longrightarrow$ characterizes conditional probabilities

of $P(X_t = j \mid X_0 = i)$

$M^t = (m_{ij}^{(t)})_{1 \le i,j \le n}$.

$(m_{i1}^{(t)}, \dots, m_{in}^{(t)})$

$v_i \longrightarrow e_i^T M^t$

$e_i = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \longrightarrow i\text{th}.$

If v_i and v_j are close - strongly connected in the graph - then $e_i^T M^t$ and $e_j^T M^t$ will be similar.

$\underline{\underline{M}}^t$

$M = D^{-1} W$ is not symmetric, the normalized matrix $S = D^{1/2} M D^{-1/2} = D^{-1/2} W D^{-1/2}$;

S is symmetric and has the spectral decomposition

$$S = V \Lambda V^T \qquad \Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$$

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n.$$

RWTH AACHEN UNIVERSITY

$$M = D^{-1/2} S D^{1/2} = D^{-1/2} \cancel{A} V \Lambda V^T D^{1/2}$$

$$= \underbrace{(D^{-1/2} V)}_{\Phi} \Lambda \underbrace{(D^{1/2} V)}_{\Psi}^T$$

$$= \Phi \Lambda \Psi^T$$

$\Phi = (\phi_1, \dots, \phi_n)$ and $\Psi = (\psi_1, \dots, \psi_n)$

then:

$$\Phi^T \Psi = V^T D^{-1/2} D^{1/2} V = V^T V = I$$

Φ and Ψ are bi-orthogonal.

equivalently $\phi_i^T \psi_j = \delta_{ij}$

and λ_k's are the eigenvalues of M with left and right eigenvalues ϕ_i, ψ_j.

$$M \psi_k = \Phi \Lambda \Psi^T \psi_k = \Phi \Lambda e_k = \lambda_k \Phi e_k$$

$$= \lambda_k \phi_k$$

$$\Rightarrow \psi_k^T M = \lambda_k \psi_k^T.$$

$$M = \sum_{k=1}^{n} \lambda_k \varphi_k \psi_k^T$$

$$M^t = \sum_{k=1}^{n} \lambda_k^t \varphi_k \psi_k^T.$$

$$e_i^T M^t = \sum_{k=1}^{n} (\lambda_k^t \, e_i^T \phi_k) \psi_k^T$$
$$= \sum_{k=1}^{n} \underbrace{\lambda_k^t \phi_{ki}}\, \psi_k^T$$

$e_i^T M^t$ are represented in terms of the basis

$\psi_1, \dots, \psi_n$ with coefficients $\lambda_k^t \phi_{ki}$, $k = 1, \dots, n$.

with $\phi_k = (\phi_{k,1}, \dots, \phi_{k,n})^T$.

These coefficients define the diffusion map.

Def. 4.5. The diffusion map at step time t is defined as
$$\Phi_t(v_i) = \begin{bmatrix} \lambda_1^t \phi_{1i} \\ \vdots \\ \lambda_n^t \phi_{ni} \end{bmatrix} \quad i = 1, \dots, n.$$

Thm. 4.6 The eigenvalues of M, given by $\lambda_1, \dots, \lambda_n$, satisfy $|\lambda_k| \leq 1$. It holds that
$$M 1_n = 1_n,$$ i.e., 1_n is an eigenvector of M and 1 an eigenvalue.

Proof:

$$M = (m_{ij})_{1 \le i,j \le n}$$

$$\sum_j m_{ij} = 1 \longleftrightarrow \sum_j P(X_t = j \mid X_{t-1} = i) = 1$$

$$M \times \underbrace{\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}}_{1_n} = 1_n \Rightarrow \lambda = 1 \text{ and } 1_n \text{ eigenvector}$$

$$|\lambda_k| \le 1, \qquad (\lambda_k, u_k) \text{ eigenvalue / eigenvector}$$

$$M u_k = \lambda_k u_k \Rightarrow \sum_j m_{ij} u_{kj} = \lambda_k u_{ki}$$

$$|\lambda_k u_{ki}| = \left|\sum_j m_{ij} u_{kj}\right| \le \sum_j m_{ij} |u_{kj}|$$

$$u_k = (u_{k1}, \dots, u_{kn})^T.$$

$$u_{k,i} \ne 0 \Rightarrow |\lambda_k| \le \sum_j m_{ij} \frac{|u_{kj}|}{|u_{k,i}|}$$

$$i : |u_{k,i}| = \max_{1 \le j \le n} |u_{kj}| \Rightarrow \frac{|u_{kj}|}{|u_{k,i}|} \le 1$$

$$\Rightarrow |\lambda_k| \le \sum_j m_{ij} = 1. \qquad \square$$

$$\phi_t(v_i) = \begin{bmatrix} 1 \\ \lambda_2^t \phi_{2,i} \\ \lambda_n^t \phi_{ni} \end{bmatrix}$$

RWTH AACHEN UNIVERSITY

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

Def 4.7 The diffusion map truncated to d dimensions is defined as

$$\phi_t^{(d)}(v_i) = \begin{pmatrix} \lambda_2^t \phi_{2,i} \\ \vdots \\ \lambda_{d+1}^t \phi_{d+1,i} \end{pmatrix}$$

$\phi_t^{(d)}$ is an approximate embedding of $v_1, \dots, v_n$ in a d-dimensional Euclidean space.

The connection between the Euclidean distance in the diffusion map coordinates and the distance between the probabilities is given in the following theorem.

Thm. 4.8 For any pair of nodes v_i and v_j it holds:

$$\| \phi_t(v_i) - \phi_t(v_j) \|^2 = \sum_{l=1}^{n} \frac{1}{\deg(l)} \left(P(X_t = l \mid X_0 = i) - P(X_t = l \mid X_0 = j) \right)^2$$

Thinking the Future
Zukunft denken

5. Classification and Clustering

Classification and clustering is one of the central tasks, in machine learning.

Given a set of data points, the purpose is to classify the points into sub-groups, which express closeness or similarity of the points and which are represented by a cluster head.

5.1 Discriminant analysis.

Consider g populations/groups/classes $C_1, \dots, C_g$ represented by a p.d.f $f_i(x)$ on $\mathbb{R}^p$, $i = 1, \dots, g$.

A discriminant rule divides $\mathbb{R}^p$ into disjoint regions $R_1, \dots, R_g$, $\bigcup_{i=1}^{g} R_i = \mathbb{R}^p$.

The rule is defined by:

allocate some observation x to C_i, if $x \in R_i$

Often the p.d.f. is completely unknown or the parameters must be estimated from a training set $x_1, \dots, x_n \in \mathbb{R}^p$. with known class allocation.

$\rightarrow$ Fisher's linear discriminant analysis