LECTURE 3 14/10/2024

MORE ON DIM. RED.

① PCA: Eigenvalue - Eigenvector
Covariance matrix

- Rotation
- Linear transformation

② MDS → Reproduce the distances from the embedding space as euclidean distances in the projected space

ⓐ Classical MDS:

$S = \| \Theta - \Delta \|^2$

diagonal

$\Delta \rightarrow G \longrightarrow G = V \Lambda V^1$

$G = Y Y^T$

$Y = V \Lambda^{1/2}$

If we use Euclidean distances <u>also</u> in the embedding space Classical MDS is equiv. to PCA

(b) Metric MDS:

$$S = \sum_{ij} \left(\frac{\Delta_{ij} - \theta_{ij}}{\Delta_{ij}} \right)^2$$

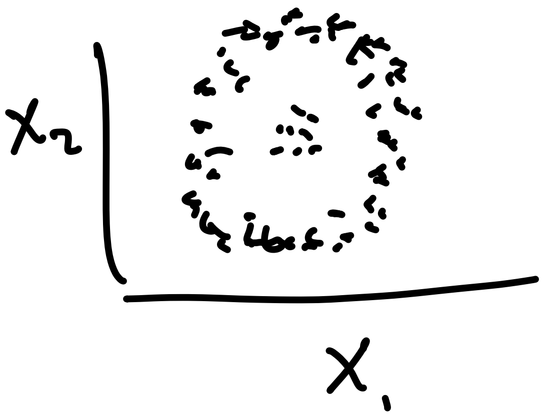
Minimize S with GD

(c) Non-metric MDS

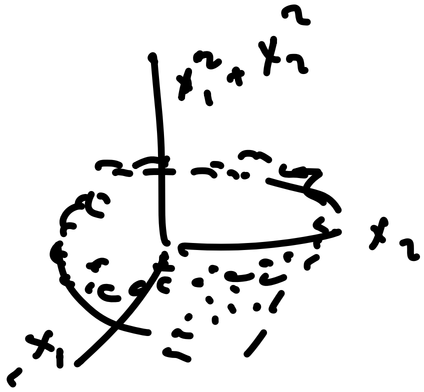
$$S = \sum_{ij} \left(\frac{g(\Delta_{ij}) - f(\theta_{ij})}{g(\Delta_{ij})} \right)^2$$

Kernel PCA:

The Kernel Trick



$$(x_1, x_2) \xrightarrow[\phi]{} (x_1, x_2, \underline{x_1^2 + x_2^2})$$



ϕ

$$K_{ij} = K(x^i, x^j) = \langle \phi^i, \phi^j \rangle$$

$$x^i \to \phi^i$$

$$k_{ij} = \langle \phi^i, \phi^j \rangle$$

$$K(x^i, x^j) = x^i \, x^{j^T} \quad \text{linear Kernel}$$

$$K(x^i, x^j) = (x^i \cdot x^{j^T})^\delta \quad V_1 \quad \text{Polynomial Kernel}$$

$$(x^i x^{j^T} + 1)^\delta \quad V_2$$

$$\delta = 2 \quad D = 2$$

$$K_{il} = \Big(\underbrace{1 + (x_1^i, x_2^i) \cdot (x_1^l, x_2^l)} \Big)^2 =$$

$$1 + (x_1^i, x_1^l)^2 + (x_2^i, x_2^l)^2 + 2(x_1^i x_1^l)$$

$$+ 2(x_2^i x_2^l) + 2 x_1^i x_1^l x_2^i x_2^l =$$

$$\to \left(1, \sqrt{2} x_1^i, \sqrt{2} x_2^i, (x_1^i)^2, \sqrt{2}(x_1^i x_2^i), (x_2^i)^2\right) \cdot$$

$$\left(1, \sqrt{2}, x_1^l, \sqrt{2} x_2^l, (x_1^l)^2, \sqrt{2}(x_1^l x_2^l), (x_2^l)^2\right)$$

$$= \underline{\phi^i \cdot \phi^l} \to \qquad D(\phi) = \frac{(\delta \cdot D)^2 - (\delta D)}{2}$$

Made with Goodnotes

Gaussian Kernel

$$k_{il} = \exp\left(\frac{-\|x^i - x^l\|^2}{2\cancel{\sigma^2}}\right)$$

$D = 2$

$$k_{il} = \exp\left(-(x_1^i - x_1^l)^2 - (x_2^i - x_2^l)^2\right) =$$

$$\exp\left(-\underline{(x_1^i)^2} + 2x_1^i x_1^l - \underline{(x_1^l)^2} - \underline{(x_2^i)^2} + 2x_2^i x_2^l - \underline{(x_2^l)^2}\right) = \underline{\exp\left(-\|\underline{x^i}\|^2\right)} \cdot \underline{\exp\left(-\|\underline{x^l}\|\right)}$$

$$\exp\left(\underline{2x^i \cdot x^l}\right) = \underline{\exp\left(-\|x^i\|^2\right)} \cdot \underline{\exp\left(-\|x^l\|^2\right)}$$

$$\underbrace{\cdot \sum_{n=0}^{\infty} \left(\frac{(2x^i x^l)^n}{n!}\right)}_{\underline{\phi_i} \cdot \phi_l} \qquad D(\phi) = \infty$$

$$\phi_i = \frac{(\sqrt{2}\,x^i)^n}{\sqrt{n!}} \; \dots, \quad n = 0, \dots, \infty$$

$$\phi^l =$$

Linear kernel $\rightarrow$ kernel PCA $\sim$ PCA

Non-linear Kernel $\rightarrow$ Non-linear Dim. Red.

(1) Mercer theorem $\quad K \rightarrow$ positive semi def.

sym.

$$k(x^i, x^l) = \langle \phi^i, \phi^l \rangle$$

$$\tilde{\phi}^i = \phi^i - \frac{1}{N} \sum_k \phi^k$$

$$\tilde{K} \quad \langle \tilde{\phi}^i, \tilde{\phi}^l \rangle = \langle \phi^i, \phi^l \rangle - \frac{1}{N} \sum_k \langle \phi^i, \phi^k \rangle$$

$$- \frac{1}{N} \sum_k \langle \phi^l, \phi^k \rangle - \frac{1}{2N^2} \sum_{k,m} \langle \phi^k, \phi^m \rangle$$

Kernel PCA in a nutshell

(a) Pick a kernel

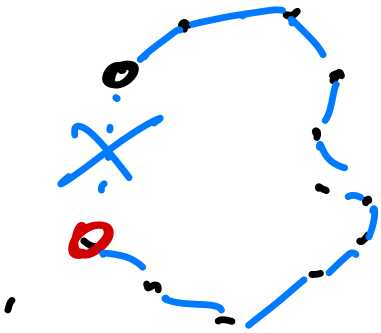
(b) Build the kernel matrix K

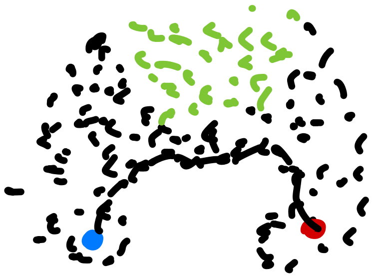
(c) Recover G in the feature space by double centering $\tilde{K}$

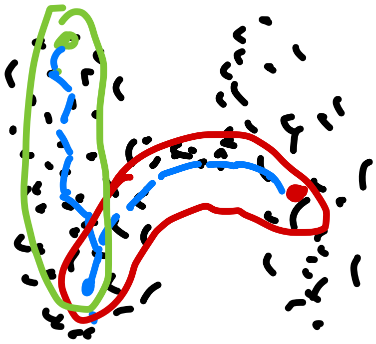
(d) $\tilde{K} \alpha_i = \lambda_i \alpha_i$

(e) $Y = \Lambda^{1/2} \alpha$

* Diffusion Map







Geodesic distance → Diffusion distance
(ISOMAP)

$$k_{ij} = \exp\left(\frac{-\|x^i - x^j\|^2}{2\sigma^2}\right)$$

σ max. distance with a single jump

$$\tilde{K}_{il} = \frac{K_{il}}{\sqrt{\sum_k K_{ik} \cdot \sum_m K_{lm}}}$$

$$p_{il} = \frac{\tilde{K}_{il}}{\underbrace{\sum_n \tilde{K}_{in}}}$$

$\underline{P} v = \lambda v$

we choose as new coordinates the d eigenvectors corresponding to the highest λ

Sketch Map : Non metric MDS

$$S = \sum_{ie} \frac{f(\theta_{ie}) - g(\Delta_{ie})}{g(\Delta_{ie})}$$

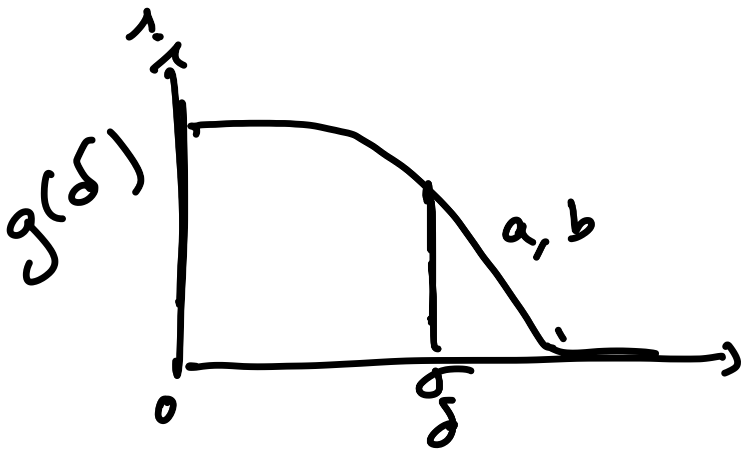
$$S = \frac{\sum_{ie} \omega_i \omega_e \left[g(\Delta_{ie}) - g(\theta_{ie}) \right]}{\sum_{ie} \omega_i \omega_e}$$

$$g(\delta) = 1 - \left[\left(2^{a/b} - 1\right) \left(\delta / \sigma \right)^a \right]^{-b/a}$$

$\omega_i \Rightarrow$ relative importance of point i

σ = distance scale that should be reproduced

$a, b \rightarrow$ rate of changing



Reproducing distances of the embedding space into the projected space

$\Big\downarrow$ probabilistic view

Probabilities of being a neighbor

t-SNE

t-(distributed) Stochastic Neighbor Embedding

Y maximizes the similarities between the probabilities distributions of being neighbors in the original and projected spaces

Probability that point ℓ is a neighbor of the point i

$$p_{\ell|i} = \frac{\exp\left(-\frac{1}{2}\left(\frac{\Delta_{i\ell}}{\sigma_i}\right)^2\right)}{\sum_{k \neq i} \exp\left(-\frac{1}{2}\left(\frac{\Delta_{ik}}{\sigma_i}\right)^2\right)}$$

Sym: $p_{i\ell} = \dfrac{p_{i|\ell} + p_{\ell|i}}{2\underset{\downarrow}{N}}$

total number of data points

$\begin{cases} p_{i\ell} = p_{\ell i} \\ \sum_{i,\ell} p_{i\ell} = 1 \end{cases}$

σ_i

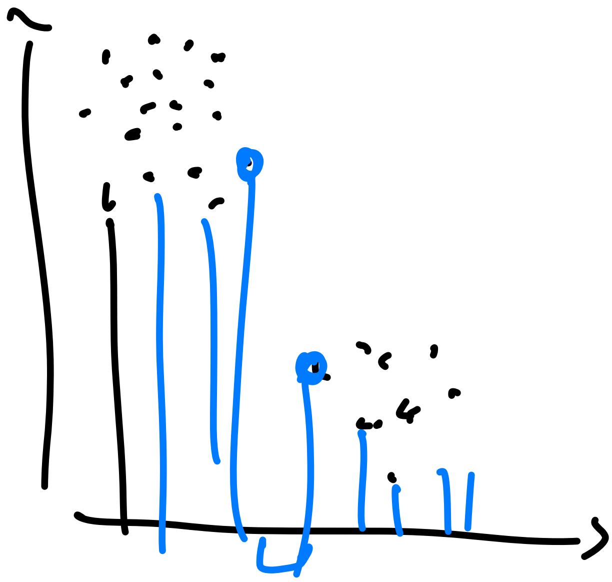
Perplexity $= 2^{-\sum_\ell p_{\ell|i} \log_2 p_{\ell|i}}$

~ Preserve the distances up to the perp NN

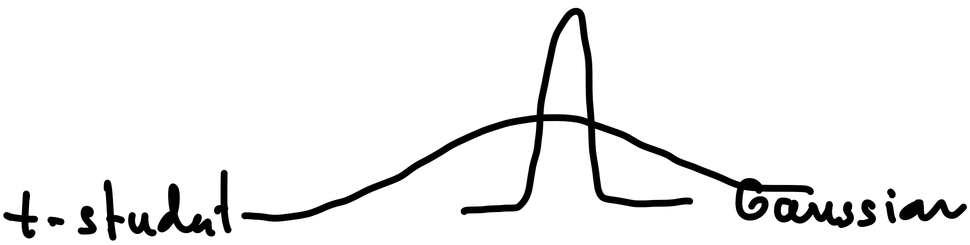
$P \to$ prob. dist. in the embedding space

$Q \to$ prob. dist. in the proj. space

minimize the KL divergence between P & Q $\quad KL(P\|Q) = \sum_{\ell \neq i} p_{i\ell} \log \dfrac{p_{i\ell}}{q_{i\ell}}$



Over crowding



$$q_{il} = \frac{(1+\theta_{il}^2)^{-1}}{\sum_{k\neq i} (1+\theta_{ik}^2)^{-1}}$$

θ: Euclidean distance in Y

$$KL(P \| Q) = \sum_{l \neq i} \underline{p_{il}} \log \frac{p_{il}}{\underline{q_{il}}(Y)}$$

Gradient Descent $\rightarrow$ obtain

$$Y = \min_{Y} \arg \; KL(P \| Q)$$

Problems:

1. Projection is different at each run
2. $d > 2$ convergence problems
3. Loss the original structure

UMAP Uniform Manifold Approximation & Projection

variation t-SNE

1. $$p_{i\ell} = \exp\left(\frac{-\Delta_{i\ell} - \rho}{\sigma_i}\right)$$

   a. we do not normalize

   b. $\rho \rightarrow$ min_dist
      if $\Delta_{i\ell} < \rho$ $\quad y_i \sim y_\ell$

2. σ_i

$$K = 2^{\sum_\ell p_{i\ell}} \rightarrow \sigma_i$$

K: number of neighbors of interest

3. Symmetrizing scheme

t-SNE = $p_{il} = \frac{p_{i|l} + p_{l|i}}{2N}$

UMAP = $p_{il} = p_{i|l} + p_{l|i} - p_{i|l} \cdot p_{l|i}$

4. $q_{il} = \left(1 + a\,\Delta_{il}^{2b}\right)^{-1}$

$a, b \to$ fit

$$\left(1 + a\left(\Delta_{il}\right)^{2b}\right)^{-1} \begin{cases} 1 & \text{if } \Delta_{il} \leq \rho \\ e^{-\Delta_{il} - \rho} & \text{if } \Delta_{il} > \rho \end{cases}$$

5. Instead of KL divergence binary cross entropy

$$L = \sum_{i,l} \left[\underbrace{p_{il} \log \frac{p_{il}}{q_{il}}}_{KL} + (1 - p_{il}) \log \frac{(1 - p_{il})}{(1 - q_{il})} \right]$$

6. Init Y from graph Laplacian

Why UMAP performs better than t-SNE?

(A) ρ hyperparameter

→ (1) & (4)

(B) Faster & Not changes from run to run

(1), (2), (3)   (6)

(3) Preserves better the global structure

(5)