MATH466/MATH766
Math of machine learning
03/31 Lecture 19 tSNE

References:
- SNE: Stochastic Neighbor Embedding by Hinton and Roweis, 2002, Neural Information Processing systems
- t-SNE: Visualizing Data Using t-SNE by van der Maaten and Hinton, 2008, Journal of Machine Learning Research
- https://www.cs.toronto.edu/~jlucas/teaching/csc411/lectures/lec13_handout.pdf

Todays contents:
- stochastic neighbor embedding (SNE)
- t-distribution SNE (t-SNE)

Important concepts:
-

Recommend reading:
- https://distill.pub/2016/misread-tsne/

Recall

1. $p, q \in \mathbb{R}^n$, $p_i, q_i \geq 0$, $\sum_i p_i = \sum_i q_i = 1$

$KL(p \| q) := \sum_i p_i \log \frac{p_i}{q_i}$ $\quad \frac{0}{0} = 0, \ 0 \log 0 = 0$.

- $KL(p \| q) \geq 0$
- $KL(p \| q) = 0$ iff $p = q$

2. Gaussian distribution

$$p(x) \sim \exp\left(-\|x - x_0\|^2 / 2\sigma^2\right)$$

3. t-distribution

$$p(x) \sim \frac{1}{1 + \|x - x_0\|^2}$$ heavy tail

Stochastic Neighbor Embedding

$\{x_i\}_{i=1}^n \subseteq \mathbb{R}^D \quad \to \quad \{y_i\}_{i=1}^n \subseteq \mathbb{R}^d$

Euclidean distance $\to$ conditional probabilities ~ similarity

$$d_{ij} := \| x_i - x_j \|$$

$$p_{j|i} := \frac{\exp(-d_{ij}^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-d_{ik}^2 / 2\sigma_i^2)} \quad \sim \text{similarity}$$

" given x_i, x_j is picked in proportion to their prob. density under a Gaussian centered at x_i. "

$P_i := p_{\cdot|i}$ is a probability distribution associated to x_i

$$q_{j|i} := \frac{\exp(-\|y_i - y_j\|^2)}{\sum_{k \neq i} \exp(-\|y_i - y_k\|^2)}, \quad Q_i := q_{\cdot|i}$$

Goal: $\min\limits_{\{y_i\}_{i=1}^n} \quad C := \sum_i KL(P_i \| Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$

① small $\|x_i - x_j\| \to$ large $p_{j|i}$
large $\|y_i - y_j\| \to$ small $q_{j|i}$ $\quad \to$ high cost

② $$\frac{\partial C}{\partial y_i} = 2 \sum_j \underbrace{\underbrace{(p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})}_{\text{mismatch} \sim \text{stiffness}} \underbrace{(y_i - y_j)}_{\text{length}}}_{f_{ij} \text{ force}}$$

imagine there are a set of springs b/t y_i and other y_j
aims at an balanced state ($\frac{\partial C}{\partial y_i} = 0$)

Remark: the choose σ_i depends on the size of the nbhd you want to preserve
dense near x_i ~ small σ_i, sparse near x_i ~ large σ_i

Symmetric SNE

$$P. \quad p_{ij} := \frac{p_{j|i} + p_{i|j}}{2n}$$

$$Q. \quad q_{ij} := \frac{\exp(-\|y_i - y_j\|^2)}{\sum_{k \neq l} \exp(-\|y_k - y_l\|^2)}$$

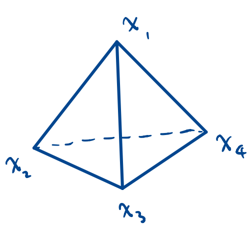
why different formulations ?

$$\min_{\{y_i\}} \quad C := KL(P \| Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

simpler gradient. $\frac{\partial C}{\partial y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)$

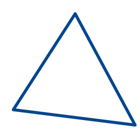
Crowding Problem

$\mathbb{R}^3$



equidistant $\longrightarrow$

$\mathbb{R}^2$



impossible to embed $\{x_1, \ldots, x_4\}$ to $\mathbb{R}^2$ while preserving all pairwise distance

no enough room to accommodate all nbhds.

t-distribution Stochastic Neighbor Embedding

SNE : match "joint prob" instead of "distance"

t-SNE : in low-dim, allocate more space for moderately dissimilar data points

$$q_{ij} := \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}}$$

$$\frac{\partial C}{\partial y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j) \frac{1}{1 + \|y_i - y_j\|^2}$$

Good for visualizing data with clusters

potential issue : create artificial clusters