Dimensionary Reduction via Dynamical Systems

From gif, the t-sne plot at different step
Looks like particle moving.

Remind the gradient in t-SNE is

$$\nabla_{z^{(i)}} L(Z) = 4 \sum_j (z^{(i)} - z^{(j)})(P_{ij} - Q_{ij})(1 + \|z^{(i)} - z^{(j)}\|^2)^{-1}$$

Note $Q_{ij} = \dfrac{(1 + \|z^{(i)} - z^{(j)}\|^2)^{-1}}{Z}$

Z is the normalization constant.
$= \sum_{k \neq l} (1 + \|z^{(k)} - z^{(l)}\|^2)^{-1}$

$$= 4 \sum_j (z^{(i)} - z^{(j)})(P_{ij} - Q_{ij}) \cdot Q_{ij} \cdot Z$$

$$\Rightarrow -\frac{1}{4} \nabla_{z^{(i)}} L(Z)$$

attractive term

$$= \sum_j P_{ij} Q_{ij} Z (z^{(j)} - z^{(i)})$$

$$- \sum_j Q_{ij}^2 Z (z^{(j)} - z^{(i)})$$

repulsive term

attract terms to be together base on distance in original space.

① First term moves $z^{(i)}$ to a weighted average of the other $z^{(j)}$.

The weights are big if P_{ij} is big ($x^{(i)}$ and $x^{(j)}$ are close)

The weights are small if P_{ij} is small.

push points apart when they are too close in embedding space.

② second term is repulsion effect. It only relies on the closeness of points in the embedding space

---

Introduce early exaggeration.

$$\frac{1}{4}\nabla_{z^{(i)}} L(z) \qquad \alpha > 1$$

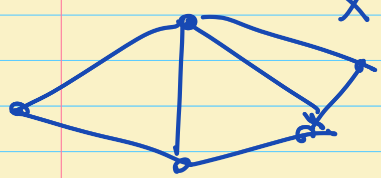
$$= \sum_j \alpha P_{ij} Q_{ij} Z (z^{(i)} - z^{(j)})$$

$$- \sum_j Q_{ij}^2 Z (z^{(i)} - z^{(j)})$$

increase the attraction term to speed up convergence

Big picture.

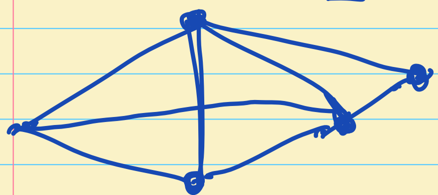
$x^{(i)}$



$\subset R^D$

$\{x^{(i)}\}_{i=1}^N$ generates the first probability matrix P_{ij}

$z^{(i)}$



$\subset R^2$

$\{z^{(i)}\}_{i=1}^N$ generates the second probability matrix Q_{ij}

$$\begin{pmatrix} P_{11} & P_{12} & \cdots & P_{1N} \\ \vdots & & & \vdots \\ P_{N1} & \cdots & \cdots & P_{NN} \end{pmatrix} \sim \begin{pmatrix} Q_{11}, & \cdots & Q_{1N} \\ \vdots & & \vdots \\ Q_{N1} & \cdots & Q_{NN} \end{pmatrix}$$

The idea of t-sne is to move the points $\{z^{(i)}\}_{i=1}^N$ s.t P and Q become as similar as possible.

We will have certain flexibility

① different ways to define similarity P_{ij}

- use $\exp(-d(x_i, x_j)^2/\sigma_i^2)$

  only few points has the significant weight

- use k-NN

  x_i is close to x_j

  if $i \in Nbr(j)$.

② measure the "distance" between probability distributions meaningfully.

- KL-divergence.
- total-variation distance
- earth-mover's distance/ Wasserstein metric.

③ the similarity q_{ij} assigned to points in some sense decreasing with euclidean distance

$Q_{ii} = 0.$

In t-SNE.

$$Q_{ij} = \frac{(1 + \|z^{(i)} - z^{(j)}\|^2)^{-1}}{\sum_{k \neq l} (1 + \|z^{(k)} - z^{(l)}\|^2)^{-1}}$$

In symmetric SNE.

$$Q_{ij} = \frac{\exp(-\|z^{(i)} - z^{(j)}\|^2)}{\sum_{k \neq l} \exp(-\|z^{(k)} - z^{(l)}\|^2)}$$

can be interpreted as normalize t-dist' with the kernel in R^2.

$$k(\|z^{(i)} - z^{(j)}\|) = \frac{1}{1 + \|z^{(i)} - z^{(j)}\|^2}$$

Def: $$K_\alpha(\|z^{(i)} - z^{(j)}\|) = \left(1 + \frac{\|z^{(i)} - z^{(j)}\|^2}{\alpha}\right)^{-\alpha}.$$

Note $\alpha = 1$, it recovers t-SNE.

$\alpha \to +\infty$, the kernel converges to a gaussian and recovers symmetric SNE.

---

different α may recover different structure

The dynamical system viewpoint.

$$y_i(t+1) = y_i(t) + \sum_{\substack{x^{(j)} \text{ close} \\ \text{to } x^{(i)}}} a_{ijt}\left(z^{(j)} - z^{(i)}\right)$$

$$- \sum_{\substack{x^{(j)} \text{ far} \\ \text{from } x^{(i)}}} a_{ijt}\left(z^{(j)} - z^{(i)}\right)$$

time dependent attraction - repulsion dynamical system.