1 Method

Given embedding $Z \in \mathbb{R}^{n\times q} \mapsto \mathbb{R}^{n\times p}$, we want to embed the out-of-sample point $w \in \mathbb{R}^s q$. For $i = 1, \ldots, n$, define

$$P_i = \frac{\exp\left(-\frac{||w-z_i||^2}{2\sigma^2}\right)}{\sum_j \exp\left(-\frac{||w-z_j||^2}{2\sigma^2}\right)}.$$

Given a potential solution $y \in \mathbb{R}^p$, define

$$Q_i = \frac{(1+||y-x_i||^2)^{-1}}{\sum_j (1+||y-x_j||^2)^{-1}}.$$

We want to find the vector y that minimizes

$$D_{KL}(P||Q) = \sum_i P_i \log \frac{P_i}{Q_i}.$$

2 Gradient

$$\begin{aligned}
\frac{\partial}{\partial y}\sum_i P_i \log\frac{P_i}{Q_i} &= \sum_i P_i \frac{Q_i}{P_i}\frac{\partial}{\partial y}\frac{P_i}{Q_i} \\
&= \sum_i P_i Q_i \frac{\partial}{\partial y}\frac{1}{Q_i} \\
&= -\sum_i P_i Q_i \frac{1}{Q_i^2}\frac{\partial}{\partial y} Q_i \\
&= -\sum_i \frac{P_i}{Q_i}\frac{\partial}{\partial y} Q_i
\end{aligned}$$

$$\frac{\partial}{\partial y}(1+||y-x_i||^2)^{-1} = -\frac{2(y-x_i)}{(1+||y-x_i||^2)^2}$$

$$\begin{aligned}
\frac{\partial}{\partial y}Q_i &= \frac{\partial}{\partial y}\frac{(1+||y-x_i||^2)^{-1}}{\sum_j(1+||y-x_j||^2)^{-1}} \\
&= \frac{\left[\sum_j(1+||y-x_j||^2)^{-1}\right]\frac{\partial}{\partial y}(1+||y-x_i||^2)^{-1} - (1+||y-x_i||^2)^{-1}\frac{\partial}{\partial y}\left[\sum_j(1+||y-x_j||^2)^{-1}\right]}{\left[\sum_j(1+||y-x_j||^2)^{-1}\right]^2} \\
&= \frac{-\left[\sum_j(1+||y-x_j||^2)^{-1}\right]\frac{2(y-x_i)}{(1+||y-x_i||^2)^2} + (1+||y-x_i||^2)^{-1}\left[\sum_j \frac{2(y-x_j)}{(1+||y-x_j||^2)^2}\right]}{\left[\sum_j(1+||y-x_j||^2)^{-1}\right]^2}
\end{aligned}$$

If we define

$$a = \begin{bmatrix} (1+||y-x_1||^2)^{-1} \\ \vdots \\ (1+||y-x_n||^2)^{-1} \end{bmatrix} \text{ and } b = \begin{bmatrix} | & & | \\ \frac{2(y-x_1)}{(1+||y-x_1||^2)^2} & \cdots & \frac{2(y-x_n)}{(1+||y-x_n||^2)^2} \\ | & & | \end{bmatrix},$$

then

$$\frac{\partial}{\partial y}Q_i = \frac{-\mathrm{sum}(a) * b[,i] + a_i * \mathrm{rowSums}(b)}{\mathrm{sum}(a)^2}.$$

Using vectorization in R,

$$\text{grad}_Q := \begin{bmatrix} — & \frac{\partial}{\partial y}Q_1 & — \\ & \vdots & \\ — & \frac{\partial}{\partial y}Q_n & — \end{bmatrix} = \left(-\text{sum}(a) * b^T + a * \begin{bmatrix} — & \text{rowSums}(b) & — \\ & \vdots & \\ — & \text{rowSums}(b) & — \end{bmatrix} \right) / \text{sum}(a)^2$$

$$\frac{\partial}{\partial y} D_{KL}(P||Q) = -\sum_i \frac{P_i}{Q_i} \frac{\partial}{\partial y} Q_i = -\text{colSums}\left(\frac{P}{Q} * \text{grad}_Q \right).$$

3 Choosing σ

σ is chosen so that

$$\text{perplexity} = 2^{-\sum P_i \log_2 P_i}$$

is equal to some pre-specified value. The creators of t-SNE suggested the perplexity should range form 5 to 50 based on sample size.