t-SNE Notes

Iakovidis Ioannis
email iakoviid@auth.gr

December 25, 2019

1 Introduction

t-Distributed Stochastic Neighbor Embedding (t-SNE) is a technique for dimensionality reduction that is particularly well suited for the visualization of high-dimensional datasets.

Let $x_i \in \mathbf{R^d}, i = 1, \ldots, n$ be the data samples in high dimensional space and $y_i \in \mathbf{R^s}, i = 1, \ldots, n$ be the data samples in low dimensional space, s is usually 2 or 3. The goal of t-SNE is to preserve local structure, points that are similar (close) in the high dimensional space are mapped to similar (close) points in the low dimensional space. This is achieved by minimizing the divergence bee tween similarity weights that are defined in the high and low dimensional space.

More specifically from the distances d_{ij} of $(x_i)_{i=1}^n$ we define

$$p_{i|j} = \frac{e^{-d_{ij}^2/2\sigma_i^2}}{\sum_{l \neq i} e^{-d_{il}^2/2\sigma_i^2}} \quad \text{and} \quad p_{ij} = \frac{p_{i|j} + p_{j|i}}{2N}$$

as the similarity weights in the high dimensional space. The value $p_{j|i}$ is interpreted as the distribution of the other points given x_i or the probability that point j is a neighbor of point i. The parameters N and σ_i are chosen. For the low dimensional space we define

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

as similarity weights but this time we use the t-distribution. For notation purposes let $P = \{p_{ij}\}$, $Q = \{q_{ij}\}$ $n \times n$ matrices and $X = [x_1, \ldots, x_n]$, $Y = [y_1, \ldots, y_n]$ $d \times n$ and $s \times n$ matrices respectively.

And then we take KL divergence

$$C(Y) = KL(P \mid Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}.$$

Minimization can be done with gradient descent.

2 Computing the Gradient

First of all

$$\begin{aligned}
C(Y) &= \sum_{i\neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}} \\
&= \sum_{i\neq j} p_{ij} \log p_{ij} - \sum_{i\neq j} p_{ij} \log q_{ij} \\
&= \sum_{i\neq j} p_{ij} \log p_{ij} + \sum_{i\neq j} p_{ij} \log f_{ij} + \sum_{i\neq j} p_{ij} \log Z \\
&= \sum_{i\neq j} p_{ij} \log p_{ij} + \sum_{i\neq j} p_{ij} \log f_{ij} + \log Z.
\end{aligned}$$

Where $f_{ij} = 1 + \|y_i - y_j\|^2$ and $Z = \sum_{k\neq l}(1 + \|y_k - y_l\|^2)^{-1}$.
With some computation:

$$\begin{aligned}
\frac{\partial \sum_{i\neq j} p_{ij} \log f_{ij}}{\partial y_m} &= \sum_{i\neq m} p_{im} \frac{\partial \log f_{im}}{\partial y_m} + \sum_{i\neq m} p_{mi} \frac{\partial \log f_{mi}}{\partial y_m} \\
&= 2 \sum_{i\neq m} p_{mi} \frac{\partial \log f_{mi}}{\partial y_m} \\
&= 2 \sum_{i\neq m} p_{mi} \frac{1}{f_{mi}} \frac{\partial f_{mi}}{\partial y_m} \\
&= 2 \sum_{i\neq m} p_{mi} \frac{2(y_m - y_i)}{1 + \|y_m - y_i\|^2} \\
&= 4 \sum_{i\neq m} p_{mi} \frac{Z(y_m - y_i)}{Z(1 + \|y_m - y_i\|^2)} \\
&= 4 \sum_{i\neq m} p_{mi} q_{mi} Z(y_m - y_i).
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \log Z}{\partial y_m} &= \frac{\partial Z}{\partial y_m} \frac{1}{Z} \\
&= \frac{2}{Z} \sum_{i\neq m} \frac{\partial (1 + \|y_m - y_i\|^2)^{-1}}{\partial y_m} \\
&= -\frac{2}{Z} \sum_{k\neq m} \frac{(y_m - y_i)}{1 + \|y_m - y_i\|^2)^2} \\
&= -\frac{2}{Z} \sum_{i\neq m} \frac{2(y_m - y_i)}{1 + \|y_m - y_i\|^2} \\
&= -4 \sum_{i\neq m} q_{mi}^2 Z(y_m - y_i).
\end{aligned}$$

So

$$\frac{\partial C(Y)}{\partial y_i} = 4 \sum_{j\neq i} p_{ij} q_{ij} Z(y_i - y_j) - 4 \sum_{j\neq i} q_{ij}^2 Z(y_i - y_j)$$

3 Gradient Interpretation

The t-SNE gradient computation can be reformulated as an N-body simulation problem:

$$F_{attr,i} = \sum_{j \neq i} p_{ij} q_{ij} Z (y_i - y_j)$$

$$F_{rep,i} = \sum_{j \neq i} q_{ij}^2 Z (y_i - y_j)$$

$$\frac{1}{4} \frac{\partial C(Y)}{\partial y_i} = F_{attr,i} + F_{rep,i}$$

The forces can be also viewed with matrix computation in the following ways:

-

4 Barnes–Hut