t-Distributed Stochastic Neighbor Embedding

Oliver Zhao

1 Description

t-SNE was developed recently as a nonlinear method to visualize high-dimension data sets [van der Maaten and Hinton, 2008]. Unlike the mathematical technique PCA, t-SNE is rooted in probability. As the name suggests, it uses a Student-t distribution, rather than a Gaussian distribution used in the more outdated SNE method.

Consider a set of N high-dimensional objects $\mathbf{x}_1, \ldots, \mathbf{x}_N$. t-SNE works by trying to minimize a cost function, which is the Kullback-Leibler divergence between a joint probability distribution P in a high-dimensional space and a joint probability distribution Q in a low-dimensional space

$$KL(P||Q) = \sum_{i \neq j} p_{ij} log\left(\frac{p_{ij}}{q_{ij}}\right). \tag{1.1}$$

p_{ij} is a probability proportional to the similarity of two objectives $\mathbf{x}_i$ and $\mathbf{x}_j$, where

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}, \tag{1.2}$$

And we set the additional constraint that probabilities where $i = j$ is zero, or $p_{ii} = 0$. The conditional probability $p_{j|i}$ represents the similarity of data point $\mathbf{x}_j$ to data point $\mathbf{x}_i$, defined as

$$p_{j|i} = \frac{exp(-||\mathbf{x}_i - \mathbf{x}_j||^2/2\sigma_i^2)}{\sum_{k \neq i} exp(-||\mathbf{x}_i - \mathbf{x}_k||^2/2\sigma_i^2)}. \tag{1.3}$$

The parameter σ_i is the variance of the Gaussian centered over each high-dimensional data point $\mathbf{x}_i$. In dense regions, a smaller value is more appropriate, whereas a larger value is more appropriate in more spread out regions. Any particular value of σ_i creates a probability distribution P_i over the other data points. t-SNE searches for the value of σ_i that produces a P_i with fixed perplexity, set by the user (typically 5-40), where perplexity is defined as

$$Perp(P_i) = 2^{H(P_i)}, \tag{1.4}$$

Where $H(P_i)$ is the Shannon entropy of P_i, defined as

$$H(P_i) = -\sum_j p_{j|i} log_2(p_{j|i}). \tag{1.5}$$

Because the original high-dimension objects $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are defined by the data set, the values of p_{ij} are consequently fixed values for a given perplexity value. The goal of t-SNE is to learn a d-dimensional map $\mathbf{y}_1, \ldots, \mathbf{y}_N$, where $\mathbf{y}_i \epsilon \mathbb{R}^d$, that reflects the similarities p_{ij} closely. q_{ij} measures the similarities between two points $\mathbf{y}_i$ and $\mathbf{y}_j$, where

$$q_{ij} = \frac{(1 + ||\mathbf{y}_i - \mathbf{y}_j||^2)^{-1}}{\sum_{k \neq i}(1 + ||\mathbf{y}_i - \mathbf{y}_k||^2)^{-1}}. \tag{1.6}$$

To minimize our cost function, the Kullback-Liebler divergence with respect to the dimension-reduced points $\mathbf{y}_i$, gradient descent is applied. The gradient can be derived into the following convenient expression

$$\frac{\delta KL(P||Q)}{\delta \mathbf{y}_i} = 4 \sum_j (p_{ij} - q_{ij})(1 + ||\mathbf{y}_i - \mathbf{y}_j||^2)^{-1}(\mathbf{y}_i - \mathbf{y}_j). \tag{1.7}$$

2 Algorithm

Description: Consider a high-dimensional data set with N data points $\mathcal{X} = \{\mathbf{x}_1, \ldots \mathbf{x}_N\}$. The low-dimensional data set we are trying to acquire is $\mathcal{Y}^{(\mathcal{T})} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$, where T is the number of iterations. The learning rate is η and the momentum is $\alpha(t)$ for gradient descent.

---

1. Calculate the pairwise affinities $p_{j|i}$ with the perplexity parameter, where

$$p_{j|i} = \frac{exp(-||\mathbf{x}_i - \mathbf{x}_j||^2/2\sigma_i^2)}{\sum_{k \neq i} exp(-||\mathbf{x}_i - \mathbf{x}_k||^2/2\sigma_i^2)}. \tag{2.1}$$

2. Calculate p_{ij}, where

$$p_{ij} = \frac{p_{j|i} - p_{i|j}}{2N}. \tag{2.2}$$

3. Sample an initial solution $\mathcal{Y}^0 = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ from $\mathcal{N}(0, 10^{-4}I)$.

4. For $t = 1$ to T:

   (a) Compute the low-dimensional afinities q_{ij}, where

$$q_{ij} = \frac{(1 + ||\mathbf{y}_i - \mathbf{y}_j||^2)^{-1}}{\sum_{k \neq i}(1 + ||\mathbf{y}_i - \mathbf{y}_k||^2)^{-1}}. \tag{2.3}$$

(b) Compute the gradient $\frac{\delta KL(P||Q)}{\delta \mathcal{Y}}$, where

$$\frac{\delta KL(P||Q)}{\delta \mathbf{y}_i} = 4 \sum_j (p_{ij} - q_{ij})(1 + ||\mathbf{y}_i - \mathbf{y}_j||^2)^{-1}(\mathbf{y}_i - \mathbf{y}_j). \quad (2.4)$$

(c) Compute $\mathcal{Y}^{(t)}$, where

$$\mathcal{Y}^{(t)} = \mathcal{Y}^{(t-1)} + \eta \frac{\delta KL(P||Q)}{\delta \mathcal{Y}} + \alpha(t) \left(\mathcal{Y}^{(t-1)} - \mathcal{Y}^{(t-2)} \right). \quad (2.5)$$