t-SNE

Đạt Nguyễn Ngọc

January 2023

1 Biến đổi lại công thức SNE, t-SNE, có tính đạo hàm loss các parameter

t-SNE minimizes the Kullback-Leibler divergence between the joint probabilities p_{ij} in the highdimensional space and the joint probabilities q_{ij} in the low-dimensional space. The values of are defined to be the symmetrized conditional probabilities, whereas the values of q_{ij} are obtained by means of a Student-t distribution with one degree of freedom

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

The values of p_{ii} and q_{ii} are set to zero. The Kullback-Leibler divergence between the two joint probability distributions P and Q is given by

$$\begin{aligned} C = KL(P||Q) &= \sum_i \sum_j p_{ij} log \frac{p_{ij}}{q_{ij}} \\ &= \sum_i \sum_j p_{ij} log p_{ij} - p_{ij} log q_{ij} \end{aligned}$$

In order to make the derivation less cluttered, we define two auxiliary variables d_{ij} and Z as follows

$$d_{ij} = ||y_i - y_j||,$$
$$Z = \sum_{k\neq 1}(1 + d_{kl}^2)^{-1}$$

Note that if y_i changes, the only pairwise distances that change are d_{ij} and d_{ji} for $\forall j$.Hence, the gradient of the cost function C with respect y_i to is given

by

$$\begin{aligned}\frac{\partial C}{\partial y_i} &= \sum_j (\frac{\partial C}{\partial d_{ij}} + \frac{\partial C}{\partial d_{ji}})(y_i - y_j) \\ &= 2\sum_j \frac{\partial C}{\partial d_{ij}}(y_i - y_j)\end{aligned}$$

The gradient $\frac{\partial C}{\partial d_i}$ is computed from the definition of the Kullback-Leibler divergence in Equation 6 (note that the first part of this equation is a constant).

$$\begin{aligned}\frac{\partial C}{\partial d_{ij}} &= -\sum_{k \neq l} p_{kl} \frac{\partial (log q_{kl})}{\partial d_{ij}} \\ &= -\sum_{k \neq l} p_{kl} \frac{\partial (log q_{kl} Z - log Z)}{\partial d_{ij}} \\ &= -\sum_{k \neq l} p_{kl} (\frac{1}{q_{kl} Z} \frac{\partial ((1 + d_{kl}^2)^{-1})}{\partial d_{ij}} - \frac{1}{Z} \frac{\partial Z}{\partial d_{ij}})\end{aligned}$$

The gradient $\frac{\partial ((1+d_j^2)^{-1})}{\partial d_{ij}}$ is only nonzero when $k = i$ and $l = j$. Hence, the gradient $\frac{\partial C}{d_{ij}}$ is given by

$$\frac{\partial C}{d_{ij}} = 2\frac{p_{ij}}{q_{ij} Z}(1 + d_{ij}^2)^{-2} - 2\sum_{k \neq l} p_{kl} \frac{(1+d_{ij}^2)^{-2}}{Z}$$

Noting that $\sum_{k \neq l} p_{kl} = 1$, we see that the gradient simplifies to

$$\begin{aligned}\frac{\partial C}{d_{ij}} &= 2p_{ij}(1 + d_{ij}^2)^{-1} - 2q_{ij}(1 + d_{ij}^2)^{-1} \\ &= 2(p_{ij} - q_{ij})(1 + d_{ij}^2)^{-1}\end{aligned}$$

We obtain the gradient:

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

2 So sánh PCA và t-SNE

PCA are used for visualization and dimensionality reduction but T-SNE is specifically used for visualization purposes only. It is well suited for the visualization of high-dimensional datasets.

T-SNE is a non-linear data visualizer. It means it doesn't form a linear line to separate the classes or to calculate the variance and it doesn't use any norm or distance metric to calculate the distance between points.

	PCA	t-SNE
Similarities	Unsupervised machine learning method	
Difference	It is a linear Dimensionality reduction technique.	It is a non-linear Dimensionality reduction technique.
	It tries to preserve the global structure of the data.	It tries to preserve the local structure(cluster) of data.
	It does not work well as compared to t-SNE.	It is one of the best dimensionality reduction techniques.
	It does not involve Hyperparameters.	It involves Hyperparameters such as perplexity, learning rate and number of steps.
	It gets highly affected by outliers.	It can handle outliers
	PCA is a deterministic algorithm.	It is a non-deterministic or randomized algorithm.
	It works by rotating the vectors for preserving variance.	It works by minimizing the distance between the points in a gaussian.
	We can decide on how much variance to preserve using eigenvalues.	We cannot preserve variance instead we can preserve distance using hyperparameters.