W2.t-SNE

Trinh Phuong Anh - 11200417

January 2023

1 Ex 1

t-SNE minimizes the Kullback-Leibler divergence between the joint probabilities p_{ij} in the high-dimensional space and the joint probabilities q_{ij} in the low-dimensional space.

Supposed these probabilities are defined to be the symmetrized conditional probabilities, we form the distribution P, then have the joint probability distribution below represents our high-dimensional:

$$P = (p_i j)_{i,j=1}^n, \, p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$$

The values of q_{ij} are obtained by means of a Student-t distribution with one degree of freedom. We define distribution Q represents low-dimensional:

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

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 Ex 4

PCA	t-SNE
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 technique
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 randomised algorithm
It works by rotating the vectors for preserving variance	It works by minimising the distance between the point in a gaussian
We can find decide on how much variance to preserve using eigen values	We cannot preserve variance instead we can preserve distance using hyperparameters