Problem. Set up t-SNE problem and optimize (calculate derivatives of Cost function with respect to parameters)

Solution

SNE converts euclidean distances to similarities, that can be interpreted as probabilities (converting the high-dimensional Euclidean distances between datapoints into conditional probabilities that represent similarities). The similarity of datapoint x_j to datapoint x_i is the conditional probability, $p_{j|i}$, that x_i would pick x_j as its neighbor if neighbors were picked in proportion to their probability density under a Gaussian centered at x_i. The conditional probability $p_{j|i}$ is computed as:

$$p_{j|i} = \frac{exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k\neq i} exp(-||x_i - x_k||^2/2\sigma_i^2)} \tag{1}$$

For the low-dimensional counterparts y_i and y_j of the high-dimensional datapoints x_i and x_j, it is possible to compute a similar conditional probability:

$$q_{j|i} = \frac{exp(-||y_i - y_j||^2)}{\sum_{k\neq i} exp(-||y_i - y_k||^2)} \tag{2}$$

$$p_{i|i} = 0 \text{ and } q_{i|i} = 0 \ \forall i$$

If the points y_i, $y_j \in$ Y correctly model the similarity between the high-dimensional datapoints x_i, $x_j \in$ X, the conditional probabilities $p_{j|i}$ and $q_{j|i}$ will be equal. SNE aims to find an embedding that minimizes the mismatch between $p_{j|i}$ and $q_{j|i}$.

Kullback-Leibler divergence from Q to P is a natural measure of the faithfulness with which $q_{j|i}$ models $p_{j|i}$, SNE minimizes the sum of KL divergences over all datapoints using gradient descent.

$P_i = \{p_{1|i},\ p_{2|i},\ ...,\ p_{n|i}\}$ and $Q_i = \{q_{1|i},\ q_{2|i},\ ...,\ q_{n|i}\}$ are the distributions on the neighbors of datapoint i.

KL Divergence compares 2 distributions. The cost function C is given by

$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}} \tag{3}$$

where Pi represents the conditional probability distribution over all other datapoints given datapoint xi, and Qi represents the conditional probability distribution over all other map points given point yi $\in$ Y.

Symmetric SNE has the property that $p_{i|j} = p_{j|i}$ and $q_{i|j} = q_{j|i}$ which allows for a simpler gradient descent to the cost function C, effectively making the calculations faster (and even give slightly better result). The high- and low-dimensional are now defined as

$$p_{i|j} = \frac{p_{i|j} + p_{j|i}}{2N} q_{ij} = \frac{exp(-||y_i - y_j||^2)}{\sum_{k\neq l} exp(-||y_k - y_l||^2)} \tag{4}$$

respectively, while the cost function C is calculated as a single KL divergence between 2 joint probability distributions P and Q:

$$C = \sum_i KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{5}$$

A problem know as the “crowding problem", which is characteristic of many multidimensional scaling techniques, including SNE, is being alleviated in t-SNE by using a heavy-tailed Student t-distribution with one degree of freedom for low-dimensional q_{ij}:

$$q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k\neq l}(1 + ||y_l - y_k||^2)^{-1}} \tag{6}$$

This also speeds up the calculations.
Given a high-dimensional dataset X, t-SNE first computes the pairwise affinities p_{ij} in the same way as Symmetric SNE. The points in the low-dimensional space Y are initialized randomly from a Gaussian distribution. The objective of t-SNE is to minimize the cost function C(Y) , using gradient descent.

1 Lemma Gradient to the cost function C(Y) defined as the Kullback-Leibler divergence

$$C(Y) = KL(P||Q) = \sum_{i\neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{7}$$

is given by

$$\frac{\partial C}{\partial y_i} = 4 \sum_j (p_{ij} - q_{ij})(1 + ||y_i - y_j||^2)^{-1}(y_i - y_j) \tag{8}$$

Proof
Put $d_{jk} := ||y_i - y_k||, f_{jk} := (1 + d_{jk}^2)^{-1}, Z := \sum_{l\neq m} f_{lm}$
Note that $\frac{\partial f_{ij}}{\partial d_{kl}} = 0$ unless i = k, j = l. By the chain rule:

$$\frac{\partial C}{\partial y_i} = \sum_{j,k} \frac{\partial C}{\partial q_{jk}} \sum_{l,m} \frac{\partial q_{jk}}{\partial f_{lm}} \frac{\partial f_{lm}}{\partial d_{lm}} \frac{\partial d_{lm}}{\partial y_i} \tag{9}$$

Definition of the KL divergence:

$$C = \sum_{j,k} p_{jk} \log \frac{p_{jk}}{q_{jk}} \tag{10}$$

Thus, we have

$$\frac{\partial C}{\partial q_{jk}} = -\frac{p_{jk}}{q_{jk}} \tag{11}$$

So,

$$\frac{\partial C}{\partial y_i} = \sum_{j,k} -\frac{p_{jk}}{q_{jk}} \sum_{l,m} \frac{\partial q_{jk}}{\partial f_{lm}} \frac{\partial f_{lm}}{\partial d_{lm}} \frac{\partial d_{lm}}{\partial y_i} \tag{12}$$

Note that $\frac{\partial d_{kl}}{\partial y_i} = 0$ unless l = i or k = i. Thus, we obtain:

$$\frac{\partial C}{\partial y_i} = -(\sum_{j,k} \frac{p_{jk}}{q_{jk}} \sum_{l} \frac{\partial q_{jk}}{\partial f_{il}} \frac{\partial f_{il}}{\partial d_{il}} \frac{\partial d_{il}}{\partial y_i} + \sum_{j,k} \frac{p_{jk}}{q_{jk}} \sum_{m} \frac{\partial q_{jk}}{\partial f_{mi}} \frac{\partial f_{mi}}{\partial d_{mi}} \frac{\partial d_{mi}}{\partial y_i}) \tag{13}$$

Moreover, since the arguments of d and f commute, we obtain

$$\frac{\partial C}{\partial y_i} = -2 \sum_{j,k} \frac{p_{jk}}{q_{jk}} \sum_{l} \frac{\partial q_{jk}}{\partial f_{il}} \frac{\partial f_{il}}{\partial d_{il}} \frac{\partial d_{il}}{\partial y_i} \tag{14}$$

$$\frac{\partial C}{\partial y_i} = -2 \sum_{l} (\sum_{j,k} \frac{p_{jk}}{q_{jk}} \frac{\partial q_{jk}}{\partial f_{il}}) \frac{\partial f_{il}}{\partial d_{il}} \frac{\partial d_{il}}{\partial y_i} \tag{15}$$

We have

$$\frac{\partial f_{il}}{\partial d_{il}} = -\frac{2d_{il}}{(1+d_{il})^2} = -2d_{il} f_{il}^2 = -2d_{il} Z^2 q_{il}^2 (1) \tag{16}$$

and

$$\frac{\partial d_{il}}{\partial y_i} = \frac{1}{d_{il}} (y_i - y_l)(2) \tag{17}$$

Plug (1) & (2) in $\frac{\partial C}{\partial y_i}$:

$$\frac{\partial C}{\partial y_i} = -4 \sum_l (\sum_{j,k} \frac{p_{jk}}{q_{jk}} \frac{\partial q_{jk}}{\partial f_{il}}) Z^2 q_{il}^2 (y_i - y_l) \tag{18}$$

Due to the definition of q_{jk} including both the factor f_{jk}, and the sum of all terms $Z = \sum_{l \neq m} f_{lm}$ in the denominator, we obtain the partial derivatives

$$\frac{\partial q_{jk}}{\partial f_{jk}} = \frac{Z - f_{jk}}{Z^2} = \frac{1}{Z}(1 - q_{jk}) and \frac{\partial f_{lm}}{\partial f_{jk}} = -\frac{f_{lm}}{Z^2} = -\frac{q_{lm}}{Z} \tag{19}$$

Therefore,

$$\frac{\partial C}{\partial y_i} = -4 \sum_l \frac{1}{Z} (\frac{p_{jk}}{q_{jl}} - \sum_{j,k} \frac{p_{jk}}{q_{jk}} q_{jk}) Z^2 q_{il}^2 (y_i - y_l) \tag{20}$$

Together with $\sum_{j,k} p_{jk} = 1$, yields

$$\frac{\partial C}{\partial y_i} = -4 \sum_l \frac{1}{Z} (-\frac{p_{il}}{q_{il}} + \sum_{j,k} p_{jk}) Z^2 q_{il}^2 (y_i - y_l) \tag{21}$$

$$\frac{\partial C}{\partial y_i} = -4 \sum_l \frac{1}{Z} (-\frac{p_{il}}{q_{il}} + 1) Z^2 q_{il}^2 (y_i - y_l) \tag{22}$$

$$\frac{\partial C}{\partial y_i} = -4 \sum_l (-p_{il} + q_{il}) Z q_{il} (y_i - y_l) \tag{23}$$

$$\frac{\partial C}{\partial y_i} = 4 \sum_l (p_{il} - q_{il}) Z q_{il} (y_i - y_l) \tag{24}$$

$$\frac{\partial C}{\partial y_i} = 4 \sum_j (p_{ij} - q_{ij}) (1 + ||y_i - y_j||^2)^{-1} (y_i - y_j) \tag{25}$$