Machine Learning 2 - Homework week 2: t-SNE

Phuong Quynh Chi
ID 11204713

Ngày 17 tháng 1 năm 2023

1 Problem 1. Biến đổi:

SNE - Stochastic Neighbor Embedding

Let go through the basic idea of SNE:

- Consider the neighborhood around an input data point $x_i \in R^D$
- Imagine that we have a Gaussian distribution centered around x_i
- Then the probability that x_i chooses some other data point x_j as its neighbor is in proportion with the density under this Gaussian.
- A point closer to x_i will be more likely than one further away

Next, the probability that point x_i chooses x_j as it neighbor:

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

with $p_{i|i} = 0$

Final distribution over pairs is symmetrized:

$$p_{ij} = \frac{1}{2N}(p_{i|j} + p_{j|i}) \tag{2}$$

The problem is that:

- Given $x^{(1)}, ..., x^{(N)} \in R^D$ we define the distribution P_{ij}
- Goal: Find good embedding $y^{(1)}, ..., y^{(N)} \in R^d$ for some $d < D$ (normally 2 or 3)
- For points $y^{(1)}, ..., y^{(N)} \in R^d$ we can define distribution Q similarly the same (notice no σ_i and not symmetric)
$$Q_{ij} = \frac{\exp\{-\|y^{(i)} - y^{(j)}\|^2\}}{\sum_k\sum_{l\neq k}\exp\{-\|y^{(l)} - y^{(k)}\|^2\}} \tag{3}$$
- Optimize Q to be close to P: Minimize KL - divergence -> to find the embedding (parameter) $y^{(1)}, ..., y^{(N)} \in R^d$

$$KL(P||Q) = \sum_{ij} P_{ij} \log \frac{P_{ij}}{Q_{ij}} = -\sum_{ij} P_{ij} \log Q_{ij} + const \tag{4}$$

We can minimize KL-divergence with gradient descent, gradient is calculated as below:

Define

$$q_{j|i} = \frac{e^{-||y_i - y_j||^2}}{\sum_{k \neq i} e^{-||y_i - y_j||^2}} = \frac{E_{ij}}{\sum_{k \neq i} E_{ij}} = \frac{E_{ij}}{Z_i} \tag{5}$$

Notice that $E_{ij} = E_{ji}$. The loss function is defined as

$$\begin{aligned} C &= \sum_{k,l \neq k} p_{l|k} log \frac{p_{l|k}}{q_{l|k}} = \sum_{k,l \neq k} p_{l|k} log\ p_{l|k} - p_{l|k} log\ q_{l|k} \\ &= \sum_{k,l \neq k} p_{l|k} log\ p_{l|k} - p_{l|k} log\ E_{kl} + p_{l|k} log\ Z_k \end{aligned} \tag{6}$$

Deriving with respect to y_i. To make the derivation less cluttered, omitting the ∂y_i, term at the denominator.

$$\frac{\partial C}{\partial y_i} = \sum_{k,l \neq k} -p_{l|k} \partial\ log E_{kl} + \sum_{k,l \neq k} p_{l|k} \partial\ log Z_k$$

Start with the first term, noting that the derivative is non-zero when $\forall j \neq i, k = i$ or $l = 1$

$$\sum_{k,l \neq k} -p_{l|k} \partial\ log E_{kl} = \sum_{j \neq i} -p_{j|i} \partial\ log E_{ij} - p_{i|j} \partial log E_{ji} \tag{7}$$

Since $\partial E_{ij} = E_{ij}(-2(y_i - y_j))$ we have

$$\begin{aligned} &\sum_{j \neq i} -p_{j|i} \frac{E_{ij}}{E_{ij}}(-2(y_i - y_j)) - p_{i|j} \frac{E_{ji}}{E_{ji}}(2(y_j - y_i)) \\ &= 2\sum_{j \neq i} (p_{j|i}) + p_{i|j}(y_i - y_j) \end{aligned} \tag{8}$$

Concluding with the second term. Since $\sum_{l \neq j} p_{l|j} = 1$ and Z_j does not depend on k, we can write (changing variable from j to j to make it more similar to the already computed terms)

$$\sum_{j,k \neq j} p_{k|j} \partial\ log Z_j = \sum_j \partial\ log Z_j$$

The derivative is non-zero when $k = i$ or $j = i$ (also, in the latter case we can move Z_i inside the summation because constant)

$$\begin{aligned} &= \sum_j \frac{1}{Z} \sum \partial E_{jk} \\ &= \sum_{j \neq i} \frac{E_{ji}}{Z_j}(2(y_j - y_i)) + \sum_{j \neq i} \frac{E_{ij}}{Z_i}(-2(y_i - y_j)) \\ &= 2\sum_{j \neq i} (-q_{j|i} - q_{i|j}(y_i - y_j) \end{aligned} \tag{9}$$

Combining equation (8) and (9) we arrive at the final result

$$\frac{\partial C}{\partial y_i} = 2\sum_{j \neq i} (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})(y_i - y_j) \tag{10}$$

t-SNE - t-Distributed Stochastic Neighbor Embedding

In high dimension we have more room, points can have a lot of different neighbors. In 2D, a point can have a few neighbors at distance one all far from each other. This is the "crowding problem" - we don't have enough room to accommodate all neighbors. Solution is that is t-SNE. Change the Gaussian in Q to a heavy tailed distribution -> if Q changes slower, we have more "wiggle room" to place points at. In t-SNE, probability goes to zero much slower then a Gaussian. We redefine Q_{ij} as:

$$Q_{ij} = \frac{(1 + \|y^{(i)} - y^{(j)}\|^2)^{-1}}{\sum_k \sum_{l \neq k} (1 + \|y^{(l)} - y^{(k)}\|^2)^{-1}} \tag{11}$$

And use the same P_{ij}

We can minimize KL-divergence with gradient descent, gradient of t-SNE is calculated as below:

Define

$$q_{ji} = q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k,l \neq k} (1 + \|y_k - y_l\|^2)^{-1}} = \frac{E_{ij}^{-1}}{\sum_{k,l \neq k} E_{kl}^{-1}} = \frac{E_{ij}^{-1}}{Z} \tag{12}$$

Notice that $E_{ij} = E_{ji}$. The loss function is defined as:

$$\begin{aligned} C &= \sum_{k,l \neq k} p_{lk} \log \frac{p_{lk}}{q_{lk}} && (13) \\ &= \sum_{k,l \neq k} p_{lk} \log p_{lk} - p_{lk} \log q_{lk} && (14) \\ &= \sum_{k,l \neq k} p_{lk} \log p_{lk} - p_{lk} \log E_{kl}^{-1} + p_{lk} \log Z && (15) \end{aligned}$$

We derive with respect to y_i:

$$\frac{\partial L}{\partial y_i} = \sum_{k,l \neq k} -p_{lk} \frac{\partial \log E_{kl}^{-1}}{\partial y_i} + \sum_{k,l \neq k} p_{lk} \frac{\partial \log Z}{\partial y_i} \tag{16}$$

We start with the first term, noting that the derivative is non-zero when $\forall j, k = i$ or $l = i$, that $p_{ij} = p_{ji}$ and $E_{ji} = E_{ij}$

$$\sum_{k,l \neq k} -p_{lk} \frac{\partial \log E_{kl}^{-1}}{\partial y_i} = -2 \sum_{j \neq i} p_{ij} \frac{\partial \log E_{ij}^{-1}}{\partial y_i} \tag{17}$$

Since $\frac{\partial E_{ij}^{-1}}{\partial y_i} = E_{ij}^{-2}(-2(y_i - y_j))$, we have:

$$-2 \sum_{j \neq i} p_{ij} \frac{E_{ij}^{-2}}{E_{ij}^{-1}} (-2(y_i - y_j)) = 4 \sum_{j \neq i} p_{ij} E_{ij}^{-1} (y_i - y_j) \tag{18}$$

We conclude with the second term. Using the fact that $\sum_{k,l \neq k} p_{kl} = 1$ and Z does not depend on k or l

$$\sum_{k,l\neq k} p_{lk}\frac{\partial \log Z}{\partial y_i} = \frac{1}{Z}\sum_{k',l'\neq k'}\frac{\partial E_{kl}^{-1}}{\partial y_i} \tag{19}$$

$$= 2\sum_{j\neq i}\frac{E_{ji}^{-2}}{Z}(-2(y_j-y_i)) \tag{20}$$

$$= -4\sum_{j\neq i} q_{ij}E_{ji}^{-1}(y_i-y_j) \tag{21}$$

Then, we arrive at the result:

$$\frac{\partial C}{\partial y_i} = 4\sum_{j\neq i}(p_{ji}-q_{ji})E_{ji}^{-1}(y_i-y_j) \tag{22}$$

$$= 4\sum_{j\neq i}(p_{ji}-q_{ji})(1+\|y_i-y_j\|^2)^{-1}(y_i-y_j) \tag{23}$$

2 Problem 4. Compare PCA and t-SNE:

No.	PCA	t-SNE
1	linear	non-linear
2	find global structure	preserve the local cluster
3	not as good as t-SNE	better than PCA
4	not involve Hyperparameters	involves Hyperparameters (perplexity, learning rate and number of steps)
5	highly affected by outliers	can handle outlier)
6	deterministic algorithm	randomised algorithm.
7	try to preserve high variance	try to find similar distribution
8	local inconsistencies (far away point can become nearest neighbors)	low dimensional neighborhood should be the same as original neighborhood