DSEB K61, MFE, NEU

Machine Learning 2 Assignment #2

Tô Đức Anh | 11196328

January 2022

1 Problems

1. Xây dựng lại bài toán t-SNE. Tính đạo hàm loss với các parameter (y) trong bài toán t-SNE.
2. Dùng t-SNE (sklearn) giảm chiều dữ liệu MNIST về 2 chiều, so sánh chuyện giảm chiều PCA và t-SNE.
3. (Optional, cộng điểm) tự implement thuật toán t-SNE để giảm chiều dữ liệu IRIS về 2 chiều.

2 Answers

2.1 About SNE...

SNE converts euclidean distances to similarities, that can be interpreted as probabilities.

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

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

with

$$p_{i|i} = 0, q_{i|i} = 0$$

We are reducing the dataset dimension, so that the pair-wise similarity (or distribution) should stay the same.
In other word, our target is to find y so that:

$$p_{i|j} = q_{i|j}$$

2.2 The breaking point where "SNE" turned into "t-SNE"

We have that:

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

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

But, the thing is that, let's say x_j is a point in the dataset, with $j \neq i$, the further x_j is, the lower the $p_{j|i}$ and, at the extreme point, this value will approach and will be 0.
Moving on, notice that, we have the σ^2, the higher the σ the more spread out the Gaussian distribution is, so that, with lower standard deviation, the further point from x_i will have a higher probability.
Let's remind our self back to the purpose of the algorithm. The goal is to find similar probability distribution in lower-dimensional space. The most obvious

choice for new distribution would be Gaussian distribution, but that's not the case here. One of the properties of Gaussian is that it has a "short tail" and because of that, it creates a problem called: "the crowding problem". If we use Gaussian again, the data will be crowed (aka stick too close with each other), so we need to use a distribution that has a heavier tail, or more spread out. So the ideal solution is to use t-Student distribution with a single degree of freedom.
Using t-Student distribution has exactly what we need. The distribution falls quickly and has a "long tail" so points won't get squashed into a single point. Welp, by using t-Student, we don't have to care about the σ anymore.

Hence, the name "t"-SNE.

2.3 Calculate the derivative

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 p_{ij} 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

$$\begin{aligned} p_{ij} &= \tfrac{p_{j|i}+p_{i|j}}{2n} \\ q_{ij} &= \tfrac{\left(1+\|y_i-y_j\|^2\right)^{-1}}{\sum_{k\neq l}\left(1+\|y_k-y_l\|^2\right)^{-1}} \end{aligned}$$

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

$$\begin{aligned} d_{ij} &= \|y_i - y_j\|, \\ Z &= \textstyle\sum_{k\neq l}\left(1+d_{kl}^2\right)^{-1}. \end{aligned}$$

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 to y_i is given by

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

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

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

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

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

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

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

We obtain the gradient:

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