Machine Learning 2: t-SNE

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

Trả lời:
* Biến đổi lại:
1. t-SNE Definition
- Popular dimension reduction technique
- Ability to preserve local structure
- Focuses on maintaining the nearest neighbours in lower dimensional map
- Converts the pair-wise Euclidean distances between points into a probability density, $P_{(j|i)}$, and we model probability that a point j is a neighbor of point i

2. Goal
Find a low dimensional representation that make $P_{(i)}$ match $Q_{(i)}$

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

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

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

Different errors have a different "cost":
- Close points mapped to far points ($p_{(j|i)}$ high, $q_{j|i}$ low)
- Far points mapped to close points ($p_{(j|i)}$ low, $q_{j|i}$ high)

3. Perplexity:
- Global parameter: effective number of neighbors
- Optimal choice local density: small value for dense region, large value for sparse region
- Entropy of $P(i)$ increases with larger values

4. Limitations:
- Fails when data is on a highly varying maniford
- The local linearity assumption on the maniford may be violated if data is high intrinsic dimensional

* Tính đạo hàm loss với các parameter:
1. Stochastic Neighbor Embedding (SNE):
Define

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

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

$$C = \sum_{k,l \neq k} log \frac{p_{l|k}}{q_{l|k}} = \sum_{k,l \neq k} log p_{l|k} - log q_{l|k} \quad (1)$$

$$= \sum_{k,l \neq k} p_{l|k} log p_{l|k} - p_{l|k} log E_{kl} + p_{l|k} log Z_k \quad (2)$$

Derive with respect to y_i

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

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

$$\frac{\partial(\sum_{k,l \neq k} -p_{l|k} log E_{kl})}{\partial y_i} = \sum_{j \neq i} -p_{j|i} \frac{\partial log E_{ij}}{\partial y_i} - p_{i|j} \frac{\partial log E_{ji}}{\partial y_i}$$

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

$$= \sum_{j \neq i} -p_{j|i} \frac{E_{ij}}{E_{ij}}(-2(y_i - y_j)) - p_{i|j} \frac{E_{ji}}{E_{ji}}(-2(y_j - y_i)) = 2 \sum_{j \neq i} (p_{j|i} + p_{i|j})(y_i - y_j)$$

We conclude 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 l to j to make it more similar to the already computed terms)

$$\frac{\partial(\sum_{l,k \neq j} p_{k|j} log Z_j)}{\partial y_i} = \frac{\partial(\sum_j log Z_j)}{\partial y_i}$$

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)

$$= \sum_j \frac{1}{Z_j} \sum_{k \neq j} \frac{\partial E_{jk}}{\partial y_i} \quad (3)$$

$$= \sum_{j \neq i} \frac{E_{ji}}{Z_j}(2(y_j - y_i)) + \sum_{j \neq i} \frac{E_{ij}}{Z_i}(2(y_i - y_j)) \quad (4)$$

$$= 2 \sum_{j \neq i} (q_{j|i} - q_{i|j})(y_i - y_j) \quad (5)$$

Combine and we have the final result

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

2. t-distributed Stochastic Neighbor Embedding (t-SNE)
Define

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

The loss function is defined as

$$\begin{aligned} C &= \sum_{k,l \neq k} p_{lk} log \frac{p_{lk}}{q_{lk}} = \sum_{k,l \neq k} logp_{lk} - p_{lk} logq_{lk} && (6) \\ &= \sum_{k,l \neq k} p_{lk} logp_{lk} - p_{lk} logE_{kl}^{-1} + p_{lk} logZ && (7) \end{aligned}$$

Derive with respect to y_i

$$\frac{\partial C}{\partial y_i} = \sum_{k,l \neq k} -p_{lk} \frac{\partial logE_{kl}^{-1}}{\partial y_i} + \sum_{k,l \neq k} p_{lk} \frac{\partial logZ}{\partial y_i}$$

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

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

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

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

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

$$\begin{aligned} sum_{k,l \neq k} p_{lk} \frac{\partial logZ}{\partial y_i} &= \frac{1}{Z} sum_{k',l' \neq k'} \frac{E_{kl}^{-1}}{\partial y_i} && (8) \\ &= 2 \sum_{j \neq i} \frac{E_{ij}^{-2}}{Z} (-2(y_j - y_i)) && (9) \\ &= -4 \sum_{j \neq i} q_{ij} E_{ij}^{-1} (y_i - y_j) && (10) \end{aligned}$$

Combine and we have the final result

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

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

Câu 4: So sánh t-SNE và PCA
The most significant difference is that PCA is a linear dimension reduction technique. At the same

time, t-SNE is a non-linear dimensionality reduction method, and it has an advantage in this case. Similar data points must be represented close together (Points with similar characteristics are together whether in a higher dimension or a lower dimension), which is not what linear dimensionality reduction algorithms do. Aim of PCA algorithm is not to keep the neighbors intact but to keep the maximum variance dimensions for more information.