Student Name: Tran Thi Hanh
Student ID: 11201381

Machine Learning 2
Assignment 2

Question 1:

Stochastic Neighbor Embedding (SNE):

Define

$$q_{j/i} = \frac{e^{-||y_i - y_j||^2}}{\sum_{k \neq i} e^{-||y_i - y_k||^2}} = \frac{E_{ij}}{\sum_{k \neq i} E_{ik}} = \frac{E_{ij}}{Z_i} \tag{1}$$

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

$$\begin{aligned} C &= \sum_{k, l \neq k} p_{l|k} \log \frac{p_{l|k}}{q_{l|k}} = \sum_{k, l \neq k} p_{l|k} \log p_{l|k} - p_{l|k} \log q_{l|k} \\ &= \sum_{k, l \neq k} p_{l|k} \log p_{l|k} - p_{l|k} \log E_{kl} + p_{l|k} \log Z_K \end{aligned} \tag{2}$$

We derive with respect to y_i . To make the derivation less cluttered, I will omit the ∂y_i term at the denominator

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

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

$$\sum_{k, l \neq k} -p_{l|k} \partial \log E_{kl} = \sum_{j \neq i} -p_{j|i} \partial \log E_{ij} - p_{i|j} \partial \log E_{ji} \tag{3}$$

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

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

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)

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

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)

$$= \sum_j \frac{1}{Z_j} \sum_{k \neq j} \partial E_{jk}$$
$$= \sum_{j \neq i} \frac{E_{ji}}{Z_j}(2(y_j - y_i)) + \sum_{j \neq i} \frac{E_{ij}}{Z_i}(-2(y_i - y_j))$$
$$= 2 \sum_{j \neq i} (-q_{j|i} - q_{i|j})(y_i - y_j) \qquad (5)$$

Combining eq. (4) and (5) we arrive at the final result

$$\frac{\partial C}{\partial y_i} = 2 \sum_{j \neq i} (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})(y_i - y_j) \qquad (6)$$

t-distributed Stochastic Neighbor Embedding (t-SNE):

Define

$$q_{ji} = q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k,l \neq k} (1 + ||y_k - y_l||^2)^{-1}} = \frac{E_{ij}^{-1}}{\sum_{k,l \neq k} E_{kl}^{-1}} = \frac{E_{ij}^{-1}}{Z} \qquad (7)$$

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

$$C = \sum_{k,l \neq k} p_{lk} \log \frac{p_{lk}}{q_{lk}} = \sum_{k,l \neq k} p_{lk} \log p_{lk} - p_{lk} \log q_{lk}$$
$$= \sum_{k,l \neq k} p_{lk} \log p_{lk} - p_{lk} \log E_{kl}^{-1} + p_{lk} \log Z \qquad (8)$$

We derive with respect to y_i . To make the derivation less cluttered, I will omit the ∂y_i term at the denominator.

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

We start with 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 $E_{ji} = E_{ij}$

$$= \sum_{k,l \neq k} -p_{lk} \partial \log E_{kl}^{-1} = -2 \sum_{j \neq i} -p_{ji} \partial \log E_{ij}^{-1} \qquad (9)$$

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

$$-2 \sum_{j \neq i} p_{ji} \frac{E_{ij}^{-2}}{E_{ij}^{-1}}(2(y_i - y_j)) = 4 \sum_{j \neq i} p_{ji} E_{ji}^{-1}(y_i - y_j) \qquad (10)$$

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

$$\sum_{k,l\neq k} p_{lk}\partial \log Z = \frac{1}{Z}\sum_{k',l'\neq k'} \partial E_{kl}^{-1}$$
$$= 2\sum_{j\neq i} \frac{E_{ji}^{-2}}{Z}(-2(y_j - y_i))$$
$$= -4\sum_{j\neq i} q_{ij}E_{ji}^{-1}(y_i - y_j) \qquad (11)$$

Combining eq. (10) and (11) we arrive at the final result

$$\frac{\partial C}{\partial y_i} = 4\sum_{j\neq i}(p_{ji} - q_{ji})E_{ji}^{-1}(y_i - y_j)$$
$$\frac{\partial C}{\partial y_i} = 4\sum_{j\neq i}(p_{ji} - q_{ji})(1 + ||y_i - y_j||^2)^{-1}(y_i - y_j) \qquad (12)$$

Question 4: Compare PCA and t-SNE

No	PCA	t-SNE
1	Linear dimensionality reduction technique.	Non-linear dimensionality reduction technique.
2	Preserve global structure	Preserve local structure
3	Not involve Hyperparameters	Involves Hyperparameters such as perplexity, learning rate and number of steps
4	Gets highly affected by outliers	Can handle outliers
5	Deterministic algorithm	Non-deterministic or randomised algorithm
6	Rotating the vectors for preserving variance.	Minimising the distance between the point in a gaussian
7	We can find decide on how much variance to preserve using eigen values	We cannot preserve variance instead we can preserve distance using hyperparameters.