ML2 - Hw2: t-SNE

Tong Thi Van Anh - 11200365

January 2023

1 Problem 1

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

Lời giải

SNE basic idea:

- "Encode" high dimensional neighborhood information as a distribution
- Find low dimensional points such that their neighborhood distribution is similar
- Intuition: Random walk between data points. High probability to jump to a close point
- Measure distance between distribution: Most common measure is KL divergence.
- Consider the neighborhood around an input data point $x_i \in \mathbb{R}^d$
- Image that we have a Gaussian distribution centered around x_i
- Then the probability that x_i chooses some other datapoint x_j as its neighbor is in proportion with the density under this Gaussian
- A point closer to x_i will be more likely one further away

The probability that point x_i chooses x_j as it neightbor:

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

with $p_{i|i} = 0$; $i, j \in \overline{1, ..., n}$ and $p_{i|j} \neq p_{j|i}$

Final distribution over pairs is symmetrized:

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

Perplexity

- The parameter σ_i sets the size of the neighborhood
  - Very low σ_i - all the probability is in the nearest neighbor
  - Very high σ_i - uniform weights
- Here we set σ_i differently for each data point

- Result depend heavily on σ_i - it defines the neighborhoods we are trying to preserve.
- For each distribution $P_{j|i}$ (depends on $sigma_i$) we define the perplexity:

$$perp(P_{j|i}) = 2^{H(P_{j|i})}$$

  with $H(P) = -\sum_i P_i \log(P_i)$ is the entropy
- If P is uniform over k elements - perplexity is k
  - Low perplexity = small σ
  - High perplexity = large σ
- Values between 5-50 usually work well
- Important parameter - different perplexity can capture different scales in the data

SNE objective:

- Given $x^1, ..., x^N \in \mathbb{R}^D$, we define the distribution P_{ij}
- Goal: Find good embedding $y^1, ..., y^N \in \mathbb{R}^d$, for some $d < D$
- For points $y^1, ..., y^N \in \mathbb{R}^d$ we can define distribution Q similarly the same (notice no σ_i and not symmetric)

$$Q_{ij} = \frac{\exp\{-||y_i - y_j||^2\}}{\sum_k \sum_{l \neq k} \exp\{-||y_l - y_k||^2\}}$$

- Optimize Q to be close to P: Minimize KL-divergence $\rightarrow$ to find the embedding (parameter) $y^1, ..., y^N \in \mathbb{R}^d$

Measure the distance between two distributions, P and Q:

$$KL(Q||P) = \sum_{ij} Q_{ij} \log \frac{Q_{ij}}{P_{ij}}$$

KL properties:

- $KL(Q||P) \geq 0$ and zero only when Q = P
- $KL(Q||P)$ is a convex function

We have P, and are looking for $y^1, ..., y^N \in \mathbb{R}^d$ such that the distribution Q we infer will minimize $L(Q) = KL(P||Q)$

$$\begin{aligned} KL(P||Q) &= \sum_{ij} P_{ij} \log \frac{P_{ij}}{Q_{ij}} \\ &= -\sum_{ij} P_{ij} \log Q_{ij} + const \end{aligned}$$

The gradients of SNE objective:

Define

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

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

$$
\begin{aligned}
L &= \sum_{k,l\neq k} p_{l|k} \log \frac{p_{l|k}}{q_{l|k}} \\
&= \sum_{k,l\neq k} \Big(p_{l|k} \log p_{l|k} - p_{l|k} \log q_{l|k} \Big) \\
&= \sum_{k,l\neq k} \Big(p_{l|k} \log p_{l|k} - p_{l|k} \log E_{kl} + p_{l|k} \log Z_k \Big)
\end{aligned}
$$

We derive w.r.t y_i. To make the derivation less cluttered, omitting the ∂y_i term at the denominator.

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

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

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

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

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

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 $i = j$ (also, in the latter case we can move Z_i inside the summation becase constant)

$$
\begin{aligned}
&= \sum_j \frac{1}{Z_j} \sum_{k\neq j} \partial E_j k \\
&= \sum_{j\neq i} \frac{E_{ji}}{Z_j} (2(y_j - y_i)) + \sum_{j\neq i} \frac{E_{ij}}{Z_j} (2(y_i - y_j)) \\
&= 2 \sum_{j\neq i} (-q_{j|i} - q_{i|j})(y_i - y_j) \qquad (3)
\end{aligned}
$$

From (1), (2) and (3), we arrive at the final result:

$$
\begin{aligned}
\frac{\partial L}{\partial y_i} &= 2 \sum_{j\neq i} (p_{j|i} - q_{j|i} + p_{i|j} - q_{j|j})(y_i - y_j) \\
&= \sum_{j\neq i} (P_{ij} - Q_{ij})(y_i - y_j)
\end{aligned}
$$

Crowding problem

- In high dimension we have more room, points can have a lot of differenct neighbors
- In 2D a point can have a few neighbors at a distance one all far from each other
- This the the "crowding problem" - we do not have enough room to accommodate all neighbors
- This is one of the biggest problems with SNE
- t-SNE solution: Change the Gaussian in Q to a heavy tailed distribution (t-dist) $\rightarrow$ if Q changes slower, we have more "wiggle room" to place points at.

t-Distributed Stochastic Neighbor Embedding

- Probability goes to zero much slower than a Gaussian
- We can now redefine

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

- We can use the same P_{ij}

The gradients of t-SNE objective: Define

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

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

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

We derive w.r.t y_i. To make the derivation less cluttered, omitting the ∂y_i term at the denominator.

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

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

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

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

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

The second term, using $\sum_{k,k\neq l} p_{kl} = 1$ and Z does not depend on k or l:

$$\begin{aligned}
\sum_{k,l\neq k} p_{lk}\partial \log Z &= \frac{1}{Z}\sum_{k',k'\neq l'} \partial E_{kl}^{-1} \\
&= 2\sum_{j\neq i} \frac{E_{ji}^{-2}}{Z}(-2(y_j - y_i)) \\
&= -4\sum_{j\neq i} q_{ij}E_{ji}^{-1}(y_i - y_j) \qquad (5)
\end{aligned}$$

From (4) and (5), we arrive at the final result:

$$\begin{aligned}
\frac{\partial L}{\partial y_i} &= 4\sum_{j\neq i}(p_{ji} - q_{ji})E_{ji}^{-1}(y_i - y_j) \\
&= 4\sum_{j\neq i}(p_{ji} - q_{ji})(1 + ||y_i - y_j||^2)^{-1}(y_i - y_j)
\end{aligned}$$

2 Problem 4

So sánh t-SNE và PCA.

Lời giải

- PCA tries to find a global structure
  - Low dimensional subspace
  - Can lead to local inconsistencies → Far away point can become nearest neighbors
- t-SNE tries to perserve local structure
  - Low dimensional neighborhood should be the same as original neighborhood