Homework week 1

Huỳnh Lê Linh Đan - DSEB 62

Ngày 17 tháng 1 năm 2023

1 Problem 1

For SNE:

Given point $x_1, x_2, \dots x_N \in R^D$ we define the distribution P_{ij} which is the probability that point x_i chooses x_j as its neighbor:

$$P_{ij} = \frac{\exp\left(-\left\|x_i - x_j\right\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\left\|x_i - x_k\right\|^2 / 2\sigma_i^2\right)}$$

We need to find good embedding $y_1, y_2, \dots y_N \in R^d$ for $d < D$:

$$Q_{ij} = \frac{\exp\left(-\left\|y_i - y_j\right\|^2\right)}{\sum_k \sum_{k \neq l} \exp\left(-\left\|y_l - y_k\right\|^2\right)} = \frac{E}{Z}$$

let $\|y_i - y_j\| = d_{ij}$

We need to optimize Q to be close to P, we do so by minimizing KL-divergence: to find the embedding $y_1, \dots y_n \in R^d$:

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

Where P_{ij} can be inferred from the data, we treat this as constant so (4) can be rewritten as:

$$\begin{aligned} KL(P\|Q) &= -\sum_{ij} P_{ij} \log\left(Q_{ij}\right) + \text{ Constant} \\ &= -\sum_{ij} P_{ij} \log\left(\frac{E}{Z}\right) + \text{ Constant} \\ &= -\sum_{ij} P_{ij} \log(E) - P_{ij} \log(Z) + \text{ Constant} \end{aligned}$$

$$\begin{aligned} \frac{\delta L}{\delta y_i} &= \left(\frac{\delta L}{\delta d_{ij}} + \frac{\delta L}{\delta d_{ji}}\right) \frac{\delta d_{ji}}{\delta y_i} = 2 \frac{\delta L}{\delta d_{ij}} \frac{\delta d_{ij}}{\delta y_i} \\ \frac{\delta L}{\delta y_i} &= -2 \sum_{ij} P_{ij} \delta \log(E) - P_{ij} \delta \log(Z) \\ \delta E &= -2\left(y_i - y_j\right) E \end{aligned}$$

First we consider the first term:

$$\begin{aligned} \sum_{ij} P_{ij} \delta \log(E) &= \sum_{ij} P_{ij} \left(-2\left(y_i - y_j\right) E\right) \frac{1}{E} \\ &= -2 \sum_{ij} P_{ij} \left(y_i - y_j\right) \end{aligned}$$

in the second term, $\sum_{k \neq l} P_{ij} = 1$, the derivative is non-zero when k = i or l = i :

$$
\begin{aligned}
P_{ij}\delta\log(Z) &= \sum_{i\neq j}\frac{1}{Z}\delta E \\
&= 2\sum_{i\neq j}\frac{E}{Z}\left(y_i - y_j\right) \\
&= 2\sum_{i\neq j} Q_{ij}\left(y_i - y_j\right)
\end{aligned}
$$

Plug (12), (15) into (9) we have:

$$
\begin{aligned}
\frac{\delta L}{\delta y_i} &= -2\sum_{ij} -2P_{ij}\left(y_i - y_j\right) + 2Q_{ij}\left(y_i - y_j\right) \\
&= 2\sum_{ij} 2P_{ij}\left(y_i - y_j\right) - 2Q_{ij}\left(y_i - y_j\right) \\
&= 4\sum_{ij}\left(P_{ij} - Q_{ij}\right)\left(y_i - y_j\right)
\end{aligned}
$$

For T-SNE, Q_{ij} is defined as:

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

We use the same P_{ij} and our loss function is:

$$
KL(P\|Q) = -\sum_{ij} P_{ij}\log\left(E^{-1}\right) - P_{ij}\log\left(Z^{-1}\right) + \text{ Constant}
$$

$$
\begin{aligned}
\tfrac{\delta L}{\delta y_i} &= -2\textstyle\sum_{ij} P_{ij}\delta\log\left(E^{-1}\right) - P_{ij}\delta\log\left(Z^{-1}\right) \\
\delta E^{-1} &= -\left(y_i - y_j\right)E^{-2}
\end{aligned}
$$

We consider the first term:

$$
\begin{aligned}
\sum_{ij} P_{ij}\delta\log\left(E^{-1}\right) &= -\sum_{ij} P_{ij}2\left(y_i - y_j\right)\frac{E^{-2}}{E^{-1}} \\
&= -2\sum_{ij} P_{ij}\left(y_i - y_j\right)E^{-1}
\end{aligned}
$$

We consider the second term:

$$
\begin{aligned}
\sum_{ij} P_{ij}\delta\log\left(Z^{-1}\right) &= -2\sum_{ij}\left(y_i - y_j\right)\frac{1}{Z^{-1}}E^{-2} \\
&= -2\sum_{ij}\left(y_i - y_j\right)\frac{E^{-1}}{Z^{-1}}E^{-1} = -2\sum_{ij}\left(y_i - y_j\right)Q_{ij}E^{-1}
\end{aligned}
$$

Plug (23), (25) into (20) we have:

$$
\begin{aligned}
\frac{\delta L}{\delta y_i} &= 4\sum_{ij} P_{ij}\left(y_i - y_j\right)E^{-1} - \left(y_i - y_j\right)Q_{ij}E^{-1} \\
&= 4\sum_{ij}\left(P_{ij} - Q_{ij}\right)\left(y_i - y_j\right)E^{-1} \\
&= 4\sum_{ij}\left(P_{ij} - Q_{ij}\right)\left(y_i - y_j\right)\left(1 + ||y_i - y_j\ |\right)^{-1}
\end{aligned}
$$

2 Problem 4

Compare T-sne and PCA:

Similar: They are both unsupervised dimensional reduction and data visualization technique for very high dimensional data

index	PCA	T-sne
1	linear	non-linear
2	preserve global structure	preserve local structure
3	Does not involves hyper parameters	Involves hyper parameters
4	Affected by outliers	Can handle outliers
5	deterministic	randomised
6	rotating vectors for preserving variance	Minimising the distance between the point
7	preserve using eigenvalues	using hyper parameters

explanation:

(1) PCA works well when there is linear relation between features while T-sne does a decent job even in non-linear dataset.

(3) T-sne hyper parameters include perplexity, learning rate, iterations. (2), (4) PCA can lead to local inconsistencies, far away point can become nearest neighbor. For t-sne, low dimensional neighborhood should be the same as original neighborhood.

(5) PCA produces the same output each time, T-sne's intuition is based on random walk between data points, therefore may produce different result on the same data. Moreover, in Problem 2, difference in hyper parameters like perplexity can produce different result.

(6) Optimization problem for PCA is maximising the variance of the projected data. For T-sne, we minimize KL-divergence so that Q (distribution for projected data points) is close to P (distribution for given data)

(7) PCA decides how much variance to preserve using eigenvalues. T-sne decides the distance to preserve using Perplexity