Homework week 2

Nguyễn Minh Đức
Student code: 11204838

January 2023

1 Question 1

The idea of SNE is, we want to encode high dimensional neighborhood information as a distribution and find low dimensional points that their neighborhood distribution is similar

The probability that point x_i choose x_j as it neighbor

$$P(j|i) = \frac{\exp(-||x^{(i)} - x^{(j)}||^2/2\sigma_i^2)}{\sum_{k \neq j} \exp(-||x^{(i)} - x^{(k)}||^2/2\sigma_i^2)} \text{ with } P_{(i|i)} = 0$$

The parameters 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

Final distribtution over pairs is symmetrized:

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

For each distribution $P_{j|i}$ (depends on σ_i) we define the perplexity

$$prep(P_{j|i}) = 2^{H(P_{j|i})} \text{ where } H(P_{j|i}) = -\sum_i P_i log(P_i)$$

Given $x^{(1)}, x^{(2)}, ..., x^{(N)} \in R^D$ we define in the distribution P_{ij}. We will have $y^{(1)}, y^{(2)}, ..., y^{(N)} \in R^d$ where $d < D$ we can define distribution Q similar the same (notices no σ_i and not symmetric)

$$Q_{ij} = \frac{\exp(-||y^{(i)} - y^{(j)}||^2)}{\sum_k \sum_{l \neq k} \exp(-||y^{(l)} - y^{(k)}||^2)}$$

We optimize Q to be close to P we use KL divergence. To measure distance between two distributions, P and Q:

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

$$\text{minimize } KL(Q||P)$$

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

The biggest problem with SNE is crowding problem. The solution for this is using another distribution. Change the Gaussian in Q to a heavy tailed distribution (t-distribution) Student-t Probability density:

$$P(x) = (1 + \frac{x^2}{v})^{-(v+1)/2} \text{ for v} = 1 \text{ we get } P(x) = \frac{1}{1+x^2}$$

Now we can define Q as

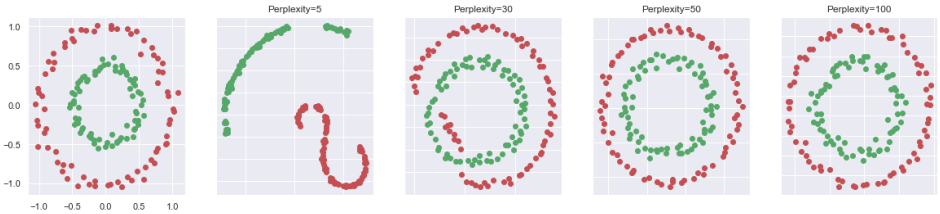
$$Q_{ij} = \frac{(1 + ||y^{(i)} - y^{(j)}||^2)^{-1}}{\sum_k \sum_{l \neq k} (1 + ||y^k - y^{(l)}||^2)^{-1}}$$

Then we have optimization problem

$$\text{minimize } KL(Q||P)$$

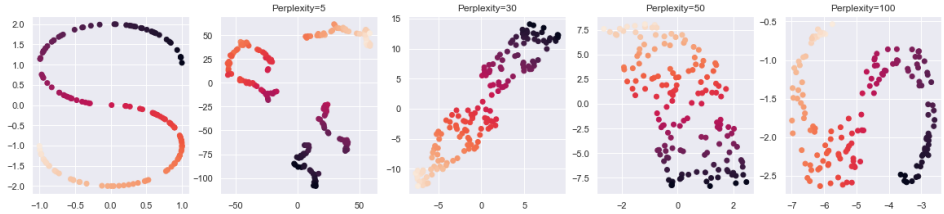
$$\begin{aligned} KL(Q||P) &= \sum_{ij} Q_{ij} log(\frac{Q_{ij}}{P_{ij}}) \\ &= \sum_{ij} Q_{ij} log(Q_{ij}) + const \\ \frac{\partial L}{\partial y^{(i)}} &= \sum_{j} (P_{ij} - Q_{ij})(y^{(i)} - y^{(j)})(1 + ||y^{(i)} - y^{(j)}||^2)^{-1} \end{aligned}$$

2 Question 2

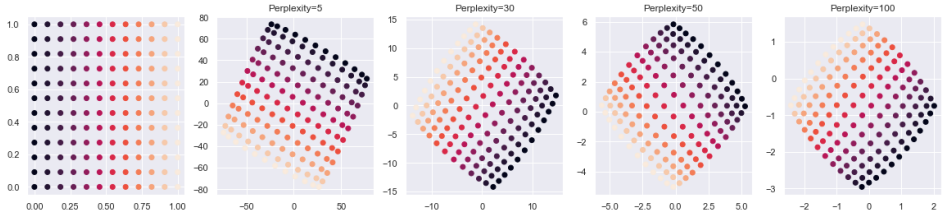


With circle-shape dataset, if we choose perplexity = 5, we will have a well-separated space between points after 5000 iter. Increasing perplexity don't

create better well-separated space (new space even similar to the initial space).



With S-shape dataset, using perplexity = 5 will create well-separated space. With perplexity = 30, the new space is the S-shape stretched to line . Keep increasing perplexity don't create better sub-space.



With uniform grid dataset, TSNE algorithm basic just rotate the dataset in new space.

3 Question 3

Words with close embedding is either synonymous or in one specific domain knowledge .Some embedding words are information related to the initial word (eg: Biologist Anatoly Kochnev in Russia). If we choose 2 or more synonymous, result with create words with similar embedding (eg: university is embedding of college, and vice versa)

words with embedding độc lập về nghĩa khi project xuống 2D dimension thường tách biệt độc lập với nhau .trong khi đó các từ có từ gốc gần nghĩa nhau như university hay college có các từ embedding nằm gần nhau trong 2D dimension mới.

4 Question 4

PCA
- It is a linear Dimensionality reduction technique.
- It tries to preserve the global structure of the data.
- PCA is a deterministic algorithm.
- We can find decide on how much variance to preserve using eigen values.
- New/unseen data can be direct use.

t-SNE
- It is a non-linear Dimensionality reduction technique.
- It tries to preserve the local structure(cluster) of data.
- It is a non-deterministic or randomised algorithm.
- We cannot preserve variance instead we can preserve distance using hyperparameters.
- New/unseen data can't be direct use (have to recompute).