# T-SNE

陳振峰, 鄭家豪, 陳煒傑

June 7, 2023

Table of content

# 1 Introduction

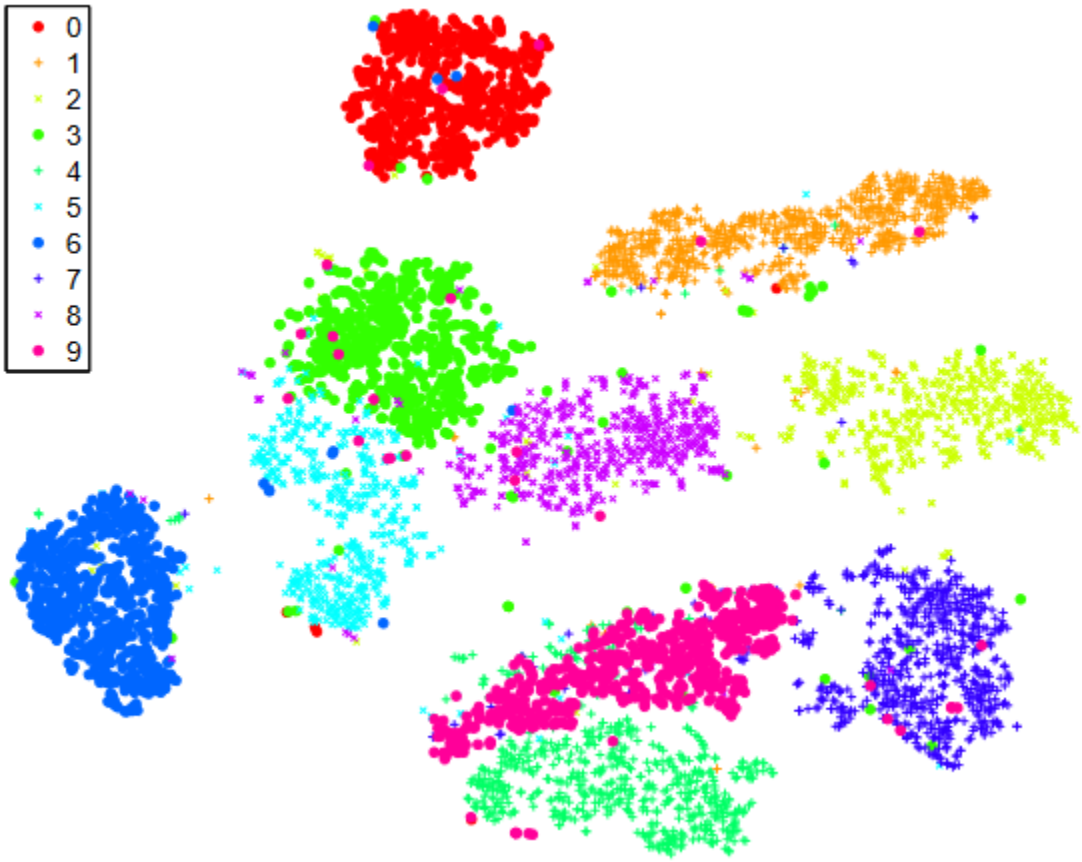


Figure 1.1: Visualizing MNIST data set using t-SNE

t-SNE, introduced by Laurens van der Maaten and Geoffrey Hinton in 2008, is a nonlinear dimensionality reduction technique. Unlike traditional dimensionality reduction algorithms that focus solely on preserving global structures, t-SNE excels at capturing both local and global relationships within the data. It accomplishes this by modelling each high-dimensional data point as a probability distribution in a low-dimensional space. By minimizing the divergence between the high-dimensional and low-dimensional distributions, t-SNE maps the data points onto a lower-dimensional plane, where their similarities and

dissimilarities are represented as distances between points. The advantages of t-SNE lie not only in its ability to reduce the dimensionality of data, but also in its capacity to uncover underlying patterns and structures that may otherwise remain hidden.

A great example of t-SNE's ability is by looking at figure 1.1 of MNIST data set visualization using t-SNE. t-SNE can effectively reveal the underlying structure and relationships between the different digit classes. It can separate the handwritten digits into distinct clusters in the low-dimensional space, where similar digits are grouped together. The visualizations produced by t-SNE often exhibit clear boundaries between clusters, allowing for easy differentiation between different digits.

In conclusion, t-SNE has emerged as a powerful tool for unraveling complex patterns and relationships within high-dimensional data. By providing a unique approach to dimensionality reduction and visualization, t-SNE enables researchers and analysts to gain valuable insights that may not be readily apparent using traditional methods. In the following sections of this essay, we will delve into the principles of t-SNE, its mathematical foundations, and its practical applications, further exploring the advantages and disadvantages of using t-SNE.

# 2 t-SNE

Refer to **Stochastic Neighbor Embedding (Hinton and Roweis (2002)),** in the high dimensional data set $X = \{x_1, x_2, ..., x_n\}$,the similarity between two point in high dimension space is defined by

$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)}$$ which means $x_i$ picks $x_j$ as its neighbor based on Gaussian distribution with $\sigma_i$ variance centered on $x_i$.

In the low dimensional data set $r=\{y_1, y_2, ...., y_n\}$(usually,2 or 3-dimensional),

$$q_{j|i} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k \neq i} \exp\left(-\|y_i - y_k\|^2\right)}$$ is the similarity in low dimension space.

Since we are only interested in modeling pairwise similarities, set 0 if j=i.

The purpose of defining these similarities in this way is the normalization to ensure the sum of similarity over j is 1.

**This is a type of semi-parameter method.**

For minimizing the mismatch between $p_{j|i}$ and $q_{j|i}$, optimizing **the Kullback-Leibler divergence** over all datapoints using a **gradient descent method**. The cost function C is given by

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

$$\frac{\delta C}{\delta y_i} = 2 \sum_j (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})(y_i - y_j)$$

$$\mathcal{Y}^{(t)} = \mathcal{Y}^{(t-1)} + \eta \frac{\delta C}{\delta \mathcal{Y}} + \alpha(t)\left(\mathcal{Y}^{(t-1)} - \mathcal{Y}^{(t-2)}\right),$$

where $\mathcal{Y}^{(t)}$ indicates the solution at iteration $t$, $\eta$ indicates the learning rate, and $\alpha(t)$ represents the momentum at iteration $t$.

The value of each $\sigma_i$ is determined by, given an initial perplexity:

$$Perp(P_i) = 2^{H(P_i)},$$

where $H(P_i)$ is the Shannon entropy of $P_i$ measured in bits

$$H(P_i) = -\sum_j p_{j|i} \log_2 p_{j|i}.$$

The perplexity can be interpreted as a smooth measure of the effective number of neighbors. The performance of SNE is a fairly robust to changes in the perplexity, and typical values between 5 and 50.

However, SNE has several problems:
1. For embedded original points, it is prone to dense crowding in the low-dimensional space, known as **the crowding problem.**
2. Outliers tend to be diluted when projected to the low-dimensional space, potentially resulting in the loss of important information.

**So, t-SNE is employed to alleviate these problems.**

To generalize the relationship between points, we first define the symmetric KL (Kullback-Leibler) divergence:

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}.$$

Where, again, we set $p_{ij}$ and $q_{ij}$ to zero. We refer to this type of SNE as symmetric SNE because it has the property that $p_{ij} = p_{ji}$ and $q_{ij} = q_{ji}$, $p_{ij}$ is defined by

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}.$$

Replacing Gaussian distribution as **t distribution(df=1)**, $q_{ij}$ is defined by

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

Similarly, for optimizing the KL divergence, using **gradient descent method** , which gradient component is different with its in SNE**:**

$$\frac{\delta C}{\delta y_i} = 4\sum_j (p_{ij} - q_{ij})(y_i - y_j)\left(1 + \|y_i - y_j\|^2\right)^{-1}.$$

The algorithm of t-SNE :

**Algorithm 1**: Simple version of t-Distributed Stochastic Neighbor Embedding.

**Data**: data set $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$,
cost function parameters: perplexity $Perp$,
optimization parameters: number of iterations $T$, learning rate $\eta$, momentum $\alpha(t)$.
**Result**: low-dimensional data representation $\mathcal{Y}^{(T)} = \{y_1, y_2, \ldots, y_n\}$.
**begin**
  compute pairwise affinities $p_{j|i}$ with perplexity $Perp$ (using Equation 1)
  set $p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$
  sample initial solution $\mathcal{Y}^{(0)} = \{y_1, y_2, \ldots, y_n\}$ from $\mathcal{N}(0, 10^{-4}I)$
  **for** $t$=1 **to** $T$ **do**
    compute low-dimensional affinities $q_{ij}$ (using Equation 4)
    compute gradient $\frac{\delta C}{\delta \mathcal{Y}}$ (using Equation 5)
    set $\mathcal{Y}^{(t)} = \mathcal{Y}^{(t-1)} + \eta \frac{\delta C}{\delta \mathcal{Y}} + \alpha(t)\left(\mathcal{Y}^{(t-1)} - \mathcal{Y}^{(t-2)}\right)$
  **end**
**end**

The use of the t-distribution is based on its property as a **heavy-tailed distribution**, which allows moderate distances in high-dimensional space to be faithfully modeled by larger distances in the map.

As a result, it eliminates the unnecessary attraction between map points representing moderately dissimilar data points.
Other heavy-tailed distributions, such as the Cauchy distribution, can also be used.

# 3 Data Analysis

## 3.1 the MNIST data set

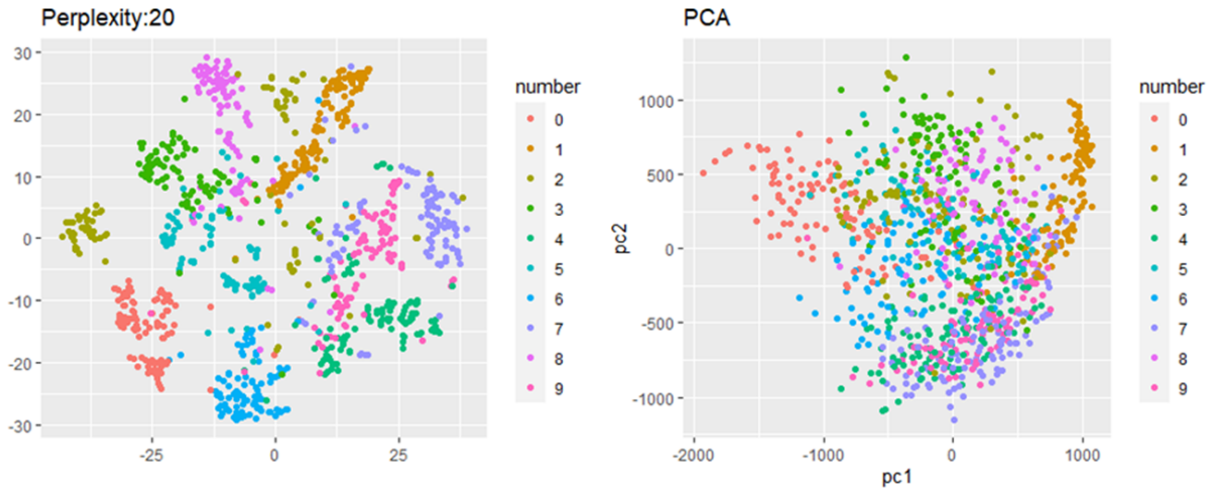


Figure 3.1: t-SNE compared to PCA on MNIST data set

The MNIST data set contains 60,000 grayscale images of handwritten digits. The digit images have 28×28 = 784 pixels (i.e., dimensions). From the graph, In contexperiments, we selected the first 1,000 of the images for computational reasons., t-SNE constructs a map in which the separation between the digit classes is almost perfect. Moreover, detailed inspection of the t-SNE map reveals that much of the local structure of the data (such as the orientation of the ones) is captured as well.

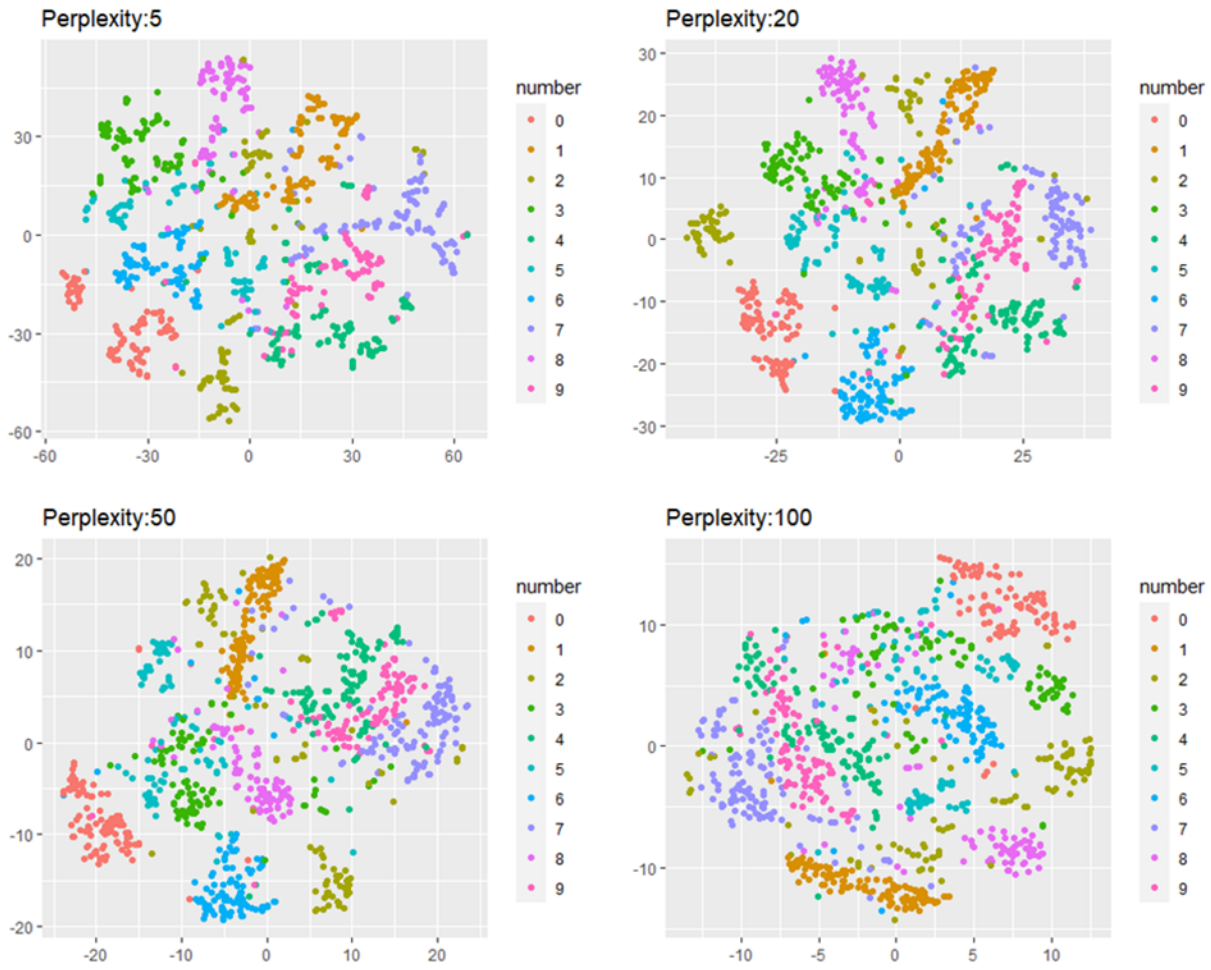


Figure 3.2: t-SNE with different perplexities parameters

Among them, the most important thing is to set the perplexity. The paper proposes that the perplexity is usually between 5 and 50, and in some cases it will be set to more than 100. Generally speaking, a large data set requires a greater perplexity. Perplexity can be interpreted as the number of effective adjacent sample points. The greater the perplexity, the more neighbors there are, and the less sensitive it is to small areas. Therefore, the following conclusions can be drawn:

- Low perplexity: Only a few neighbors are influential, and the same group may be split into multiple groups.
- High degree of perplexity: the global structure is more obvious, but it may be impossible to distinguish between groups.

Different levels of perplexity have a great influence on the results, so drawing multiple graphs for comparison can be considered when interpreting.

## 3.2 Iris data

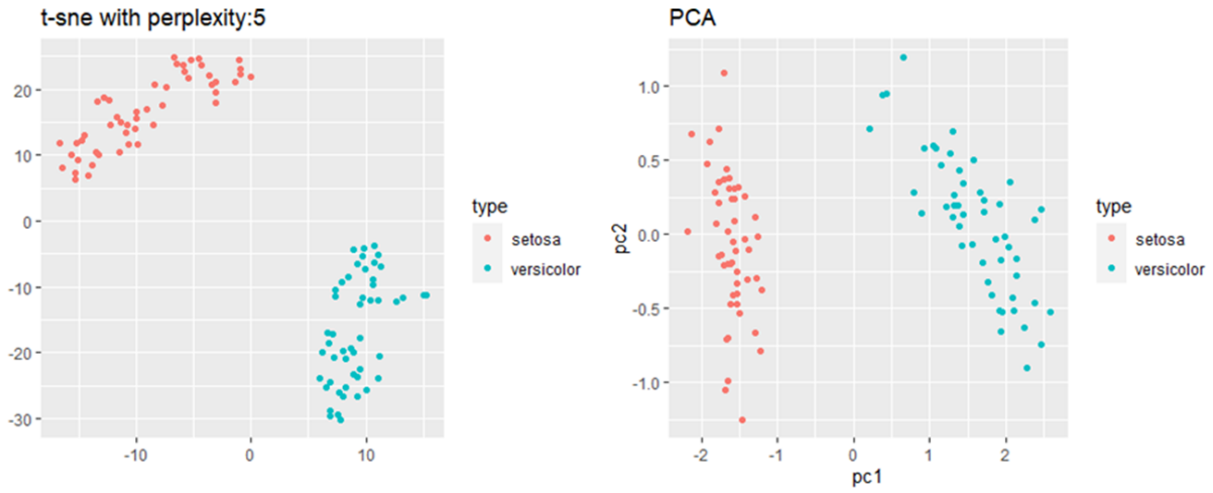


Figure 3.3: t-SNE compared to PCA on Iris dataset

The iris data contain 150 observations and 3 species(setosa, versicolor and virginica). For our experiments, we selected setosa and versicolor. It can be seen from the figure that there is not much difference between t-SNE and PCA, but PCA will be easier to interpret.

Thus, t-SNE may not offer a significant advantage over PCA when it comes to separating the Iris dataset. Since the classes in the Iris dataset are relatively well-separated, t-SNE might not reveal substantial improvements in visual separation compared to PCA.

In summary, when analyzing the Iris dataset, PCA can be a reliable method for dimensionality reduction and visualization, especially if the primary objective is to capture the global structure and retain interpretability. However, if there is a need to explore more nuanced patterns or uncover local relationships within the classes, t-SNE can offer valuable insights despite its reduced interpretability.

# 4 Conclusion

- Advantage of t-SNE
    - **Preserves local and global structure of high-dimensional data**: points which are close to one another in the high-dimensional data set will tend to be close to one another in the low-dimensional map
    - **Great for Exploratory Data Analysis** (e.g. whether we want to know if the data is separable or not)
    - **Robustness to Noise**: t-SNE is robust to noise and outliers in the data. It accomplishes this by modeling the similarities between data points using a heavy-tailed Student's t-distribution, which is less sensitive to extreme values compared to a Gaussian distribution. This robustness allows t-SNE to handle noisy datasets without significantly distorting the overall structure.
- Disadvantage of t-SNE
    - **Computationally Intensive**: t-SNE can be computationally demanding, especially for large datasets.As a result, the computational requirements increase significantly as the dataset size grows. For very large datasets, this can lead to long processing times and high memory usage.
    - **Not great for machine learning application**: mainly used for data visualization; t-SNE is not learning a function from the original space to the new (lower) dimensional one.
    - **Non-convexity of t-SNE cost function**; The results of t-SNE can vary across different runs, even with the same parameters and dataset. This non-deterministic nature arises from the algorithm's reliance on random initialization and the use of stochastic gradient descent. Consequently, researchers and analysts should be cautious when comparing different t-SNE embeddings or drawing conclusions based on a single run.

# 5 Question

1. For S={S1,S2,S3,S4} with P(S1) = 1/4,P(S2)=1/8,P(S3)=1/2,P(S4)=1/8, calculate Shannon entropy with base 2.

2. Which followings about t-SNE are true?
(A)In low-dimension space, t-SNE usually has a more severe crowding problem than SNE.
(B)We can use Gamma distribution rather than t distribution to compute similarity between two points in low-dimension space.
(C)The heavy-tailed distribution (support is real number set) is employed to alleviate crowding problem and optimization problem in low-dimension space.
(D)The cost function is symmetric.
(E)The choice for heavy-tailed distribution can be Student t , Cauchy.
(F)all of the above; (G) none of the above

# 6 Reference


1. Maaten, L. van der, & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.
2. Erdem (burnpiro), Kemal. “T-SNE Clearly Explained.” Medium, July 21, 2022. https://towardsdatascience.com/t-sne-clearly-explained-d84c537f53a.
3. G.E. Hinton and S.T. Roweis. Stochastic Neighbor Embedding. In Advances in Neural Information Processing Systems, volume 15, pages 833–840, Cambridge, MA, USA, 2002. The MIT Press.