Demystifying the high dimensional data: tSNE
t-Distribution stochastic neighbour embedding

Shourya[1]

CS460: Machine Learning
Supervisor: Dr. Subhankar Mishra

National Institute of Science Education and Research, Bhubaneshwar

Background

After the advancements in technology, the world has been brimmed with data which has brought a wealth of information, but with it comes a challenge: "**the curse of dimensionality**". As the number of dimensions (features) in our data increases, visualizing and processing it becomes increasingly difficult. This is where dimension reduction techniques, like t-SNE, come into play. These techniques allow us to uncover hidden patterns within the data, making complex information easier to explore and analyze. Ultimately, this leads to increased computational efficiency.

What is tSNE?

Non-linear form dimension reduction, its objective is to embed data from high dimension to lower dimension so as to optimally preserve neighbourhood identity.

Algorithm of t-SNE

Each data point in the high-dimensional space is captured as a probability distribution over its neighbors using a Gaussian kernel, $\boldsymbol{p}_{ij}$ is calculated for $\boldsymbol{d}_{ij}$

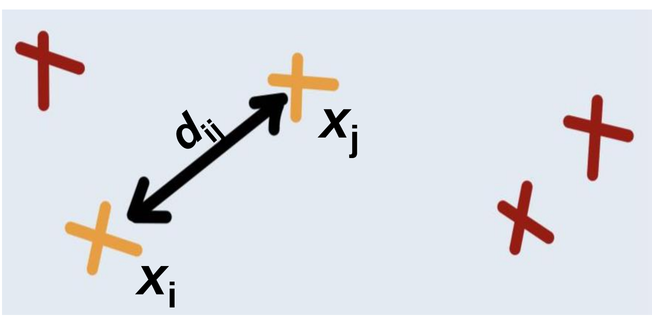


Figure 1: *Data points in 2D*

$$p_{j|i} = \frac{\exp\left(-\frac{d_{ik}^2}{2\sigma_i^2}\right)}{\sum_{k \neq i} \exp\left(-\frac{d_{ik}^2}{2\sigma_i^2}\right)}$$

Equation 1: *Gaussian Kernel*

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

It is then mapped to a corresponding point in the lower-dimensional space, Here also probability distribution is calculated for embedded data using t-distribution, q_{ij} is calculated for $\mathbf{e}_{ij}$

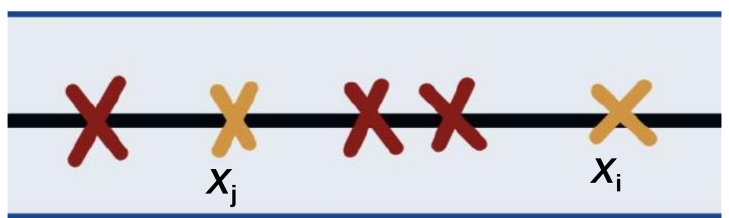


Figure 2: *Data points in 1D*

$$q_{ij} = \frac{\left(1+e_{ij}^2\right)^{-1}}{\sum_{k \neq l}\left(1+e_{kl}^2\right)^{-1}}$$

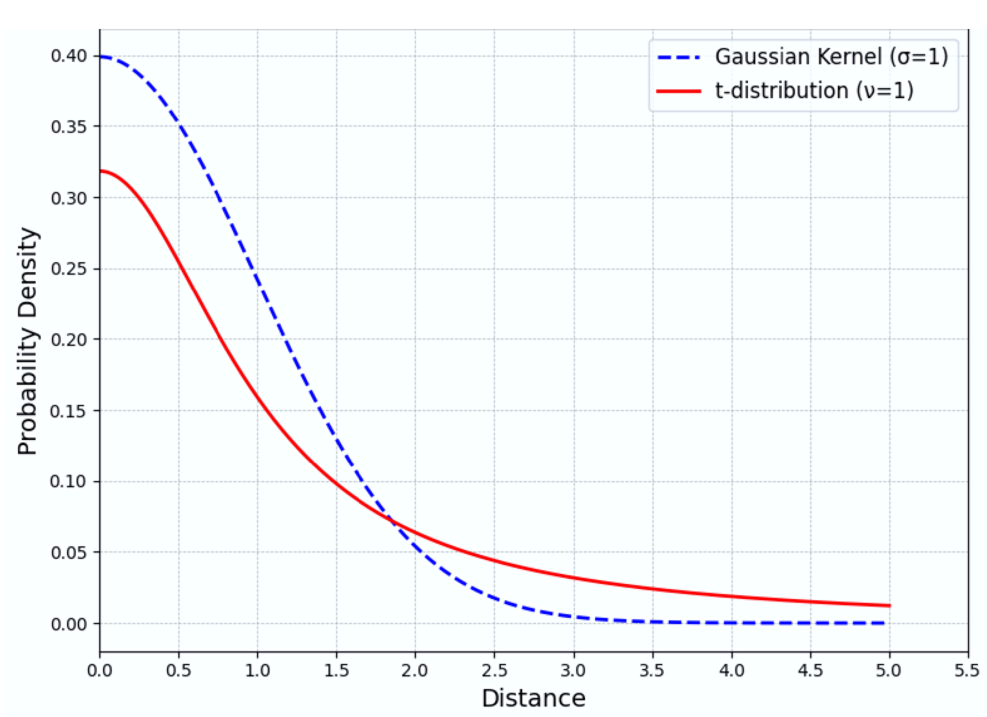
Equation 2: *Heavier tail t-Distribution*

To achieve its goal, an iterative optimization process is employed. A cost function is calculated, measuring the difference between the probability distributions of neighbors in the high-dimensional space and the corresponding low-dimensional representation.

Figure 3: *Data points in 1D, aligned*

$$L = \sum_{i,j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

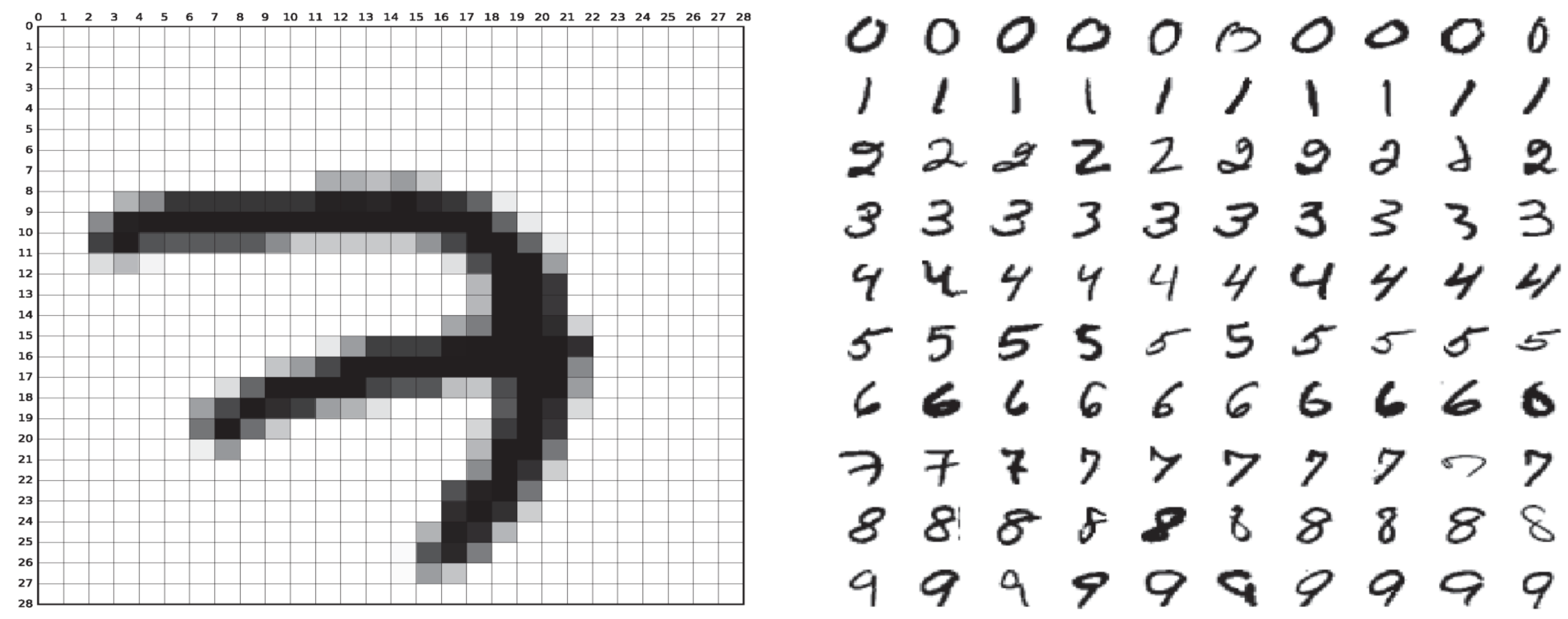
Equation 4: *Cost function (Kullback–Leibler divergence)*



Graph 1: *Gaussian distribution in High-D , t-Distribution in Low-D*

Why tSNE?

To illustrate the benefits of t-SNE and the types of problems it excels at, let's consider the MNIST handwritten digit dataset, a popular dataset in machine learning



(**a**) MNIST sample belonging to the digit '7'. (**b**) 100 samples from the MNIST training set.

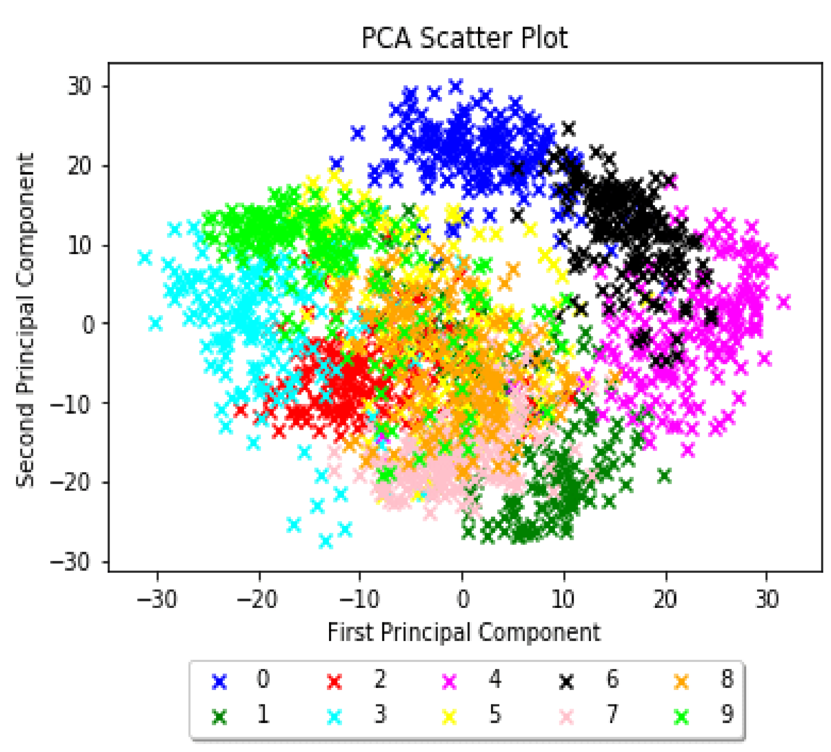


Figure 4: *PCA for MNIST*

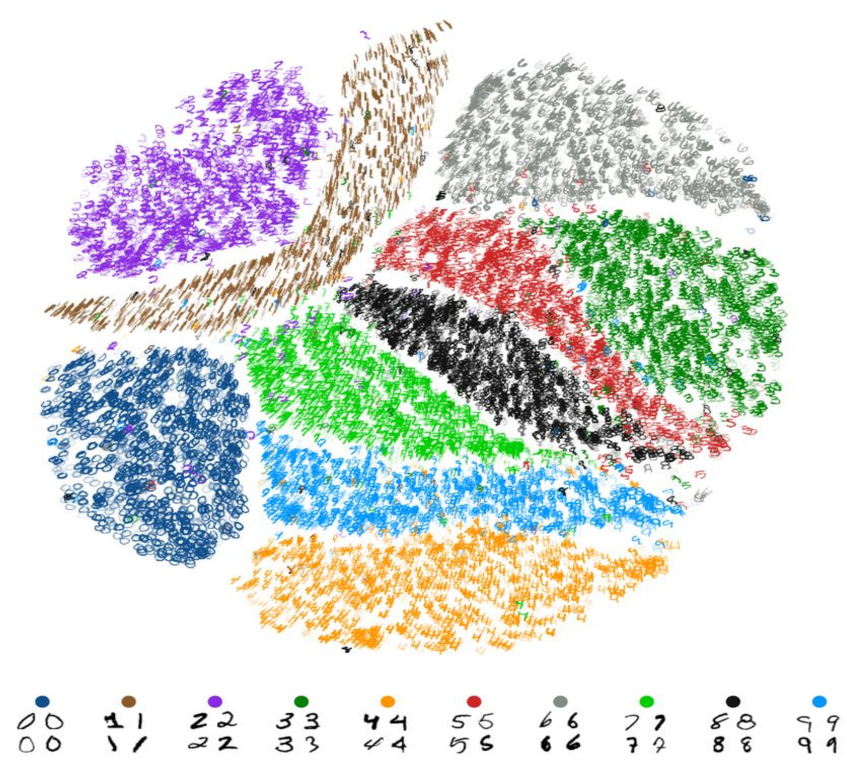


Figure 5: *t-SNE for MNIST*

Applications

- Bioinformatics and Genomics
- Natural language processing
- Scientific Visualization
- Biomedical Signal processing
- Geological Domain interpretation

References

1. https://en.wikipedia.org/wiki/T-distributed_stochastic_neighbor_embedding
2. *van der Maaten, L.J.P.; Hinton, G.E. (Nov 2008). "Visualizing Data Using t-SNE" (PDF). Journal of Machine Learning Research. **9**: 2579–2605.*
3. https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2020.620143/full
4. Baldominos, A.; Saez, Y.; Isasi, P. A Survey of Handwritten Character Recognition with MNIST and EMNIST. *Appl. Sci.* **2019**, *9*, 3169. https://doi.org/10.3390/app9153169
5. Pezzotti, Nicola. (2019). Dimensionality-Reduction Algorithms for Progressive Visual Analytics. 10.13140/RG.2.2.35141.50403.
6. https://www.scikit-yb.org/en/latest/api/text/tsne.html
7. https://doi.org/10.3389/fgene.2020.620143

shourya.2021@niser.ac.in
2111118