t-SNE

(t-distributed Stochastic Neighbor Embedding)

Introduction:

t-SNE (t-Distributed Stochastic Neighbour Embedding) is a powerful machine learning algorithm designed for dimensionality reduction, particularly suited for visualizing high-dimensional data in lower dimensions, typically two or three. Unlike traditional methods like Principal Component Analysis (PCA), t-SNE focuses on preserving the local structure of the data by converting similarities between data points in the high-dimensional space into probabilities and then mapping them to a lower-dimensional space. This method is particularly effective in revealing complex patterns, clusters, and relationships within the data that might not be easily discernible in higher dimensions. Its ability to maintain local relationships makes t-SNE a popular tool in exploratory data analysis, especially in fields like bioinformatics, natural language processing, and computer vision, where data often reside in complex, high-dimensional spaces.

Principles:

Stochastic measures for deciding Neighbourhood:

For points in higher dimension

Suppose we have high dimensional Datapoints $x_1, x_2, x_3, \cdots x_n$ and our goal is to project them onto a 2D or 3D space as Datapoints $y_1, y_2, y_3, \cdots y_n$ .The neighbourhood of a particular data point i in the higher dimensional space is decided by the probability measure $p_{j|i}$ which considers both the facts that the closer the point j is to the point i, the more is the chance of considering it as a neighbour and when the points other than j are spread even distant from i then j becomes the more obvious neighbour. Such $p_{j|i}$ would look as written below

$$p_{j|i} = \frac{\exp\left[-\left\|X_i - X_j\right\|^2 / \sigma_i\right]}{\sum_{k\neq i} \exp\left[-\|X_i - X_k\|^2 / \sigma_i\right]}$$

The choice of the standard deviation σ_i depends on the number of neighbours you would like to fit in terms of their distance from the point i in a normal probability distribution within a preselected multiple of this standard deviation

We can see that the probability is assigned proportional to the density of the normal distribution of the distance. Now, the probability that the points i and j are grouped in the higher dimensional space is given by p_{ij} which is proportional to the average of the probabilities for each of the points i, j being seen as a neighbour by the point of consideration,i.e $p_{i|j}$ and $p_{j|i}$

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

For points in lower dimension

The same neighbourhood measure can work for the points projected to the lower(2 or 3) dimensions but with a small change. The probability that point j is selected as neighbour of point i in the lower dimensional space is proportional to t-distribution with symmetric probability density curve similar to normal density curve but with heavier tails.These heavier tails can accommodate points in the lower dimensional space which went through the curse of dimensionality during projection. When we want to preserve the similarity between points in very high dimensional space while projecting them onto a lower dimension, the projections becomes sparse to compensate the lack of room as we lose on the dimensions which had accommodated them in the higher dimensional space. So the t distribution with degree of freedom '1' ,also known as Cauchy distribution can be used to consider the neighbourhood points which got spread out after projection, by giving a greater probability to them compared to normal distribution. So the probability that the point j is selected as neighbour to i in the lower dimensional space is given as

$$q_{j|i} = \frac{\left[1 + \left\|Y_i - Y_j\right\|^2\right]^{-1}}{\sum_{k \neq l} [1 + \|Y_k - Y_l\|^2]^{-1}}$$

and the probability that the point i and j are considered to be of one cluster is given as

$$q_{ij} = \frac{1}{2N} q_{i|j} + \frac{1}{2N} q_{j|i}$$

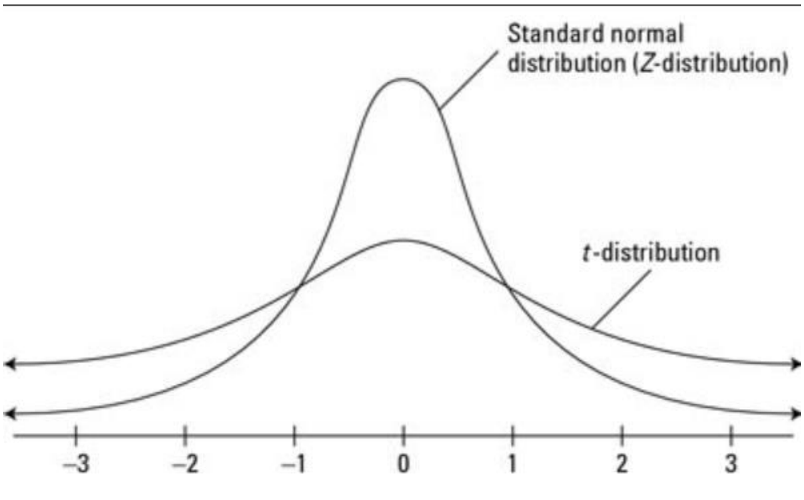

Figure 1:Comparision of Normal and Cauchy distribution

Objective function:

As similarity among the points expected to be preserved during projection, the probabilities of the points i and j belonging to the same cluster in both higher and lower dimensional spaces must be same.This needs to be achieved by minimizing the difference between $p_{ij}$and q_{ij}.KL divergence shown below, is a measure representing this difference, so minimizing KL divergence using Gradient Descent with respect to the points Y_i and Y_j would give us final projection of higher dimensional points onto a lower dimensional space.

OBJECTIVE FUNCTION

$$min(KL(P|Q))$$

where the function *KL* represents KL Divergence for given i and j, which is given by

KL DIVERGENCE

$$KL(P|Q) = \sum_{i \neq j} p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right)$$