t-SNE t-distributed stochastic neighbor embedding

a statistical method for visualizing high-dimensional data by giving each datapoint a location in a two or three-dimension map.

based on stochastic Neighbor Embedding

nonlinear dimensionality reduction technique

Specifically, it models each high-dimensional object by a two- or three-dimensional point in such a way that similiar objects are modeled by nearby points and dissimilar objects are modeled by distant points with high probability.

t-SNE algorithm comprises two main stages:

First:

t-SNE constructs a probability distribution over pairs of high-dimensional objects in such a way that similar objects are assigned a higher probability while dissimilar points are assigned a lower probability.

Second:

t-SNE defines a similar probability distribution over the points in the low-dimensional map,
& it minimizes the Kullback-Leibler divergence (KL divergence) between the two distributions with respect to the locations of the points in the map.

while the original algorithm uses the Euclidean distance between objects as the base of its similarity metric, this can be changed as appropriate. A Riemannian variant is UMAP.

Details :

stage 1

Given a set of N high-dimensional objects $x_1, \cdots, x_N$, t-SNE first computes probabilities p_{ij} that are proportional to the similarity of objects x_i and x_j, as follows

For $i \neq j$, define

$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma_i^2)}$$

and set $p_{i|i} = 0$

As van der Maaten and Hinton explained:

"The similarity of datapoint x_j to datapoint x_i is the conditional probability, $p_{j|i}$, that x_i would pick x_j as its neighbor if neighbors were picked in proportion to their probability density under a Gaussian centered at x_i"

define

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2N} \qquad (\sum_i \sum_j p_{ij} = 1)$$

stage 2 :

t-SNE aims to learn a d-dimensional map $y_1, \cdots, y_N$ (with d typically chosen as 2 or 3) that reflects p_{ij} as well as possible.

To this end, define

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

performed using gradient descent

The location of the points y_i in the map are determined by minimizing the KL-divergence of P from the distribution Q

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