1 Introduction

Given a high dimensional dataset $\mathcal{X} = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$, $i = 1, \cdots, n$. How to find a meaningful dimension reduction and visualization of this dataset? The idea of dimension reduction is to preserve as much information as possible. So far, we have discussed three different methods:

1. Principal component analysis  
2. Laplacian eigenmaps  
3. Diffusion maps

PCA is a linear technique, which is to find a linear subspace which captures the maximal directions of covariance. In other words, this method will keep the dissimilar points (with high distance) far apart in the low dimensional representation. For high dimensional data, the intrinsic dimension is usually lower than the ambient dimension $d$, i.e., they may lie near a low dim object such as hypersurface or manifold. For example, a curve in $\mathbb{R}^{100}$ has its intrinsic dimension equal to 1. We hope that the low dimensional representations of similar points are still close to one another. Therefore, sometimes we prefer nonlinear dimension reduction techniques such as Laplacian eigenmaps and diffusion maps.

We will discuss t-SNE ($t$-stochastic neighboring embedding), which is able to perform dimension reduction (to $\mathbb{R}^2$ and $\mathbb{R}^3$), and capture both local and global structure of the data points.

1.1 SNE: stochastic neighboring embedding

SNE (stochastic neighboring embedding) \cite{2} starts by converting the high dimensional Euclidean distance between data points into conditional probability that represents similarities:

$$
\begin{align*}
p_{j|i} &= \frac{\exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{k=1}^n \exp(-\|x_i - x_k\|^2/2\sigma_i^2)}, \\
q_{j|i} &= \frac{\exp(-\|y_i - y_j\|^2/2)}{\sum_{k=1}^n \exp(-\|y_i - y_k\|^2/2)}.
\end{align*}
$$

Note that $p_{j|i}$ and $q_{j|i}$ define two random walks $(X_t)_t$ and $(Y_t)_t$ on $\{x_i\}_{i=1}^n \subset \mathbb{R}^d$ and $\{y_i\}_{i=1}^n \subset \mathbb{R}^2$ or $\mathbb{R}^3$ respectively:

$$
p_{j|i} = \mathbb{P}(X_{t+1} = x_j | X_t = x_i), \quad q_{j|i} = \mathbb{P}(Y_{t+1} = y_j | Y_t = y_i)
$$
If the map points $y_i$ and $y_j$ correctly models the similarity between $x_i$ and $x_j$ in the high dimensional space, then their conditional probabilities $p_{ji}$ and $q_{ji}$ will be very close. Motivated by this observation, SNE aims to find a low dimensional representation which minimizes the mismatch between $p_{ji}$ and $q_{ji}$.

We want to move those points $\{y_i\}_{i=1}^n$ in 2D so that the similarity between these two joint probability distributions is high. How to measure the similarity? The cost function is based on Kullback-Leibler divergence.

**Definition 1.1.** The Kullback-Leibler divergence from $q$ to $p$ is defined as

$$D_{KL}(p||q) = \sum_{i=1}^n p_i \log \frac{p_i}{q_i}$$

where $p = (p_1, \ldots, p_n) \geq 0$, $\sum_{i=1}^n p_i = 1$ and $q = (q_1, \ldots, q_n) \geq 0$ and $\sum_{i=1}^n q_i = 1$.

Here are a few properties about KL divergence:

1. The divergence is not symmetric, i.e.,
   $D_{KL}(p||q) \neq D_{KL}(q||p)$.

2. $D_{KL}(p||q) \geq 0$ holds and $D_{KL}(p||q) = 0$ if and only if $p = q$. The nonnegativity follows from

$$\sum_{i=1}^n p_i \log \frac{p_i}{q_i} = -\sum_{i=1}^n p_i \log \frac{q_i}{p_i} = -\mathbb{E}_{p(i)} \log \frac{q(i)}{p(i)} \geq -\log \mathbb{E}_{p(i)} \frac{q(i)}{p(i)}$$

$$= -\log \sum_{i=1}^n \frac{q_i}{p_i} p_i = 0$$

where $-\log(x)$ is a convex function. The inequality above uses Jensen’s inequality below.

**Theorem 1.1 (Jensen’s inequality).** For a convex function $f$, it holds that

$$\mathbb{E} f(X) \geq f(\mathbb{E} X)$$

where $X$ is a random variable.

SNE minimizes the sum of KL divergences over all data points by using gradient descent:

$$C(y_i) = \sum_{i=1}^n KL(P_i|Q_i) = \sum_{i=1}^n \sum_{j=1}^n p_{ji} \log \frac{p_{ji}}{q_{ji}}.$$

In other words, SNE tries to move the point clouds $\{y_i\}_{i=1}^n \in \mathbb{R}^2$ or $\mathbb{R}^3$ such that the mismatch between the conditional probabilities $\{x_i\}$ and $\{y_i\}$ is small.

The variance parameter $\sigma_i^2$ is chosen adaptively for each $i$ because the density of data points centered at any $x_i$ is unlikely to be the same, i.e., some regions are more dense/sparse than other regions. The variance $\sigma_i^2$ is chosen such that the entropy of the conditional probability $P_i$ is the same for each $i$. Here the entropy is defined as

$$H(P_i) = -\sum_{j=1}^n p_{ji} \log_2 p_{ji}$$

and the perplexity is

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

Recall that entropy is a measure of complexity of a probability distribution in $\mathbb{R}$. 

2
1.2 Crowding problem

From the construction of SNE, we can see that it aims to preserve the local structure. Suppose the dataset has intrinsic dimension 2, shown in left figure of Figure 1. If we try to reduce its dimension by preserving its local small distance, we will naturally get the right figure in which the edge length is preserved. However, the distance between the two red points are amplified.

![Crowding problem](image)

For high dimension data with intrinsic dimension greater than 2, this issue will arise very often if we want to find a 2D representation of this dataset which preserves the local structure. If the points with small pairwise distance are accurately preserved, then most points with a moderate distance have to be placed much further. If we still use Gaussian kernel in the map space, the corresponding weight will vanish due to the fast decay of Gaussian kernel.

1.3 t-SNE

While SNE uses Gaussian kernel function for the mapped data in low dimension, t-SNE \[3\] applies Student’s t-distribution to measure the local similarity in the map space. Student’s t-distribution

\[
f(t) = \frac{\Gamma((\nu + 1)/2)}{\sqrt{\nu\pi} \Gamma(\nu/2)} \left( 1 + \frac{t^2}{\nu} \right)^{-\frac{\nu+1}{2}}, \quad \nu \in \mathbb{Z}_+
\]

has a heavier tail than Gaussian, which is able to alleviate the impact of crowding problem. This allows a moderate distance in the high dimensional space to be faithfully modeled by a much larger distance in the map. In particular,

1. \(\nu = 1\),

\[
f(t) = \frac{1}{\pi(1+t^2)}
\]

is the Cauchy distribution.

2. \(\nu = \infty\),

\[
f(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}
\]

is Gaussian distribution.
The heavier tail in Student’s $t$-distribution allows us to place data points with moderate distance slightly far while preventing the weight from being too small, which compensates the loss of dimensionality.

We define a joint probability on $\mathcal{X} = \{x_1, \cdots, x_n\}$:

$$p_{ij} = \frac{\exp(-\|x_i - x_j\|^2/2\sigma^2)}{\sum_{k \neq l} \exp(-\|x_k - x_l\|^2/2\sigma^2)}; \text{ or } p_{ij} = \frac{1}{2n}(p_{ji} + p_{ij})$$

(1.1)

and $\mathcal{Y} = \{y_1, \cdots, y_n\}$:

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

It is easy to see that $p_{ij}$ defined in (1.1) is indeed a joint probability measure on $\{x_1, \cdots, x_n\} \times \{x_1, \cdots, x_n\}$ since $\sum_{i,j} p_{ij} = 1$.

We will find the low dimensional representation of $\{x_i\}$ via minimizing the KL divergence between $p$ and $q$. As a result, the cost function is

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

The training is performed by applying gradient descent to the cost function $C(y)$.

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

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

$$= C_0 + \sum_{i,j} p_{ij} \left(\log(1 + \|y_i - y_j\|^2) + \log \sum_{k \neq l}(1 + \|y_k - y_l\|^2)^{-1}\right)$$

$$= C_0 + \sum_{i,j} p_{ij} \log(1 + \|y_i - y_j\|^2) + \log \left(\sum_{k \neq l}(1 + \|y_k - y_l\|^2)^{-1}\right)$$

where $C_0 = \sum_{i,j} p_{ij} \log p_{ij}$ is the negative entropy of $p$.

By direct computation, the gradient of $C(y)$ w.r.t. $y_i$ is given by

$$\frac{\partial C}{\partial y_i} = \sum_{j=1}^{n} \frac{4p_{ij}(y_i - y_j)}{1 + \|y_i - y_j\|^2} - \sum_{j=1}^{n} \frac{4(1 + \|y_i - y_j\|^2)^{-2}(y_i - y_j)}{\sum_{k \neq l}(1 + \|y_k - y_l\|^2)^{-1}}$$

$$= \sum_{j=1}^{n} \frac{4p_{ij}(y_i - y_j)}{1 + \|y_i - y_j\|^2} - \sum_{j=1}^{n} \frac{4(y_i - y_j)q_{ij}}{1 + \|y_i - y_j\|^2}$$

$$= 4 \sum_{j=1}^{n} \frac{p_{ij} - q_{ij}}{1 + \|y_i - y_j\|^2} \cdot (y_i - y_j)$$

Then the learning process is characterized by

$$y^{(t+1)}_i = y^{(t)}_i - \eta \frac{\partial C(y^{(t)})}{\partial y_i}.$$

It is similar to a system of interacting particles.
(a) If $p_{ij} > q_{ij}$, then $\mathbf{y}_i$ is attracted by $\mathbf{y}_j$.
(b) If $p_{ij} < q_{ij}$, then $\mathbf{y}_i$ is repelled by $\mathbf{y}_j$.

All the data points $\{\mathbf{y}_j\}_{j \neq i}$ exert a force on $\mathbf{y}_i$ along the direction $\mathbf{y}_i - \mathbf{y}_j$. Whether the particles repel or attract depending on the probability in the map space and in the original space.

2 Experiments

The implementation can be found [here].

![Figure 2: Representation of MNIST dataset via t-SNE dimension reduction](image)

References

