On t-SNE

Alex-Antoine Fortin

American Family Insurance

June 20, 2016

t-SNE description

t-SNE stands for: t-distributed Stochastic Neighbor Embedding

desideratum
We wish to have a representation in D-dim (R_D^*) of an input in N-dim (R_N), with N $>>$D

We wish that R_D^* preserves as much information as possible about R_N

To do that, we need that R_D^* preserves the locality of the inputs in $\mathbb{R}^N$

t-SNE description

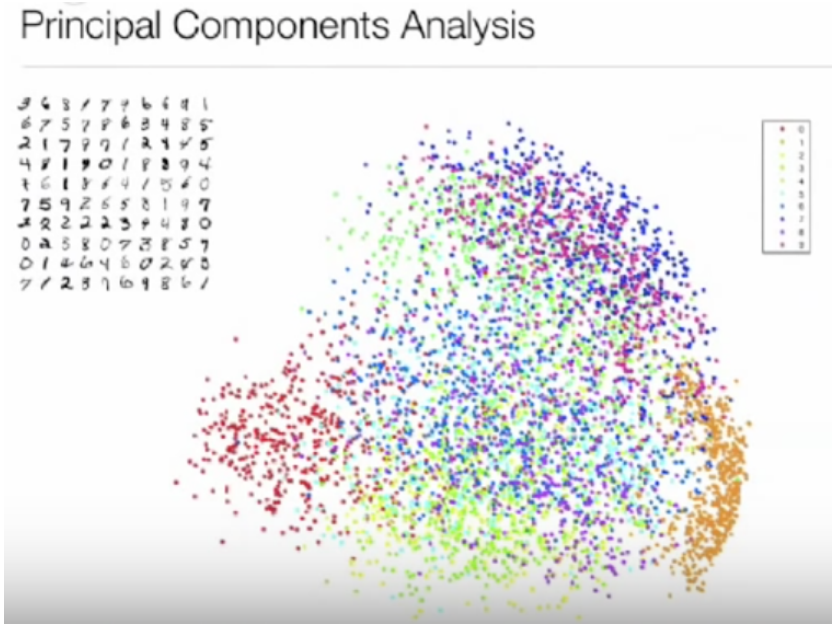
Intuition

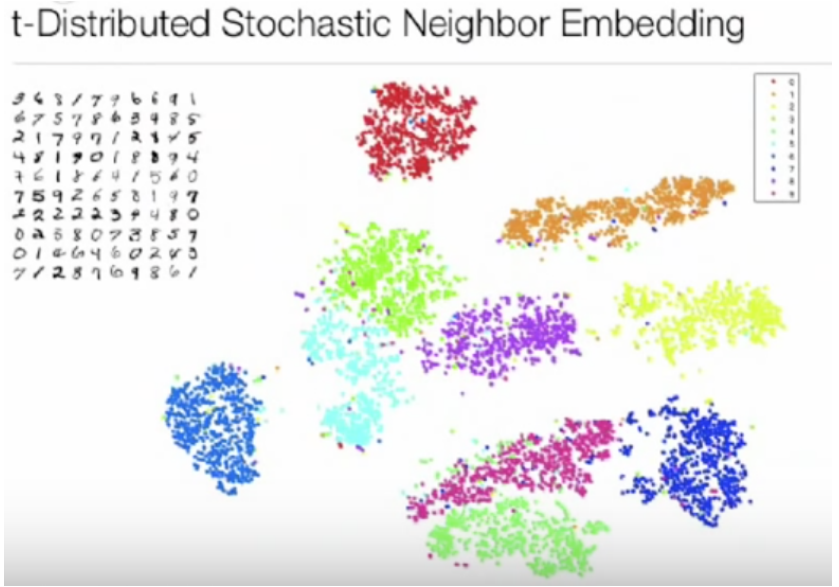
t-SNE achieves this in 3 steps:

1. fit a multivariate gaussian $\mathcal{N}(\mu, \Sigma)$ to input R_N by Maximum Likelihood Estimation (MLE)
   - preserves locality of inputs by a cute trick (see *detailed calculations*)
2. Pose a multivariate student $\boldsymbol{t}_\nu(\mu, \Sigma)$, with $\nu = 1$ as the distribution of R_D^*
3. fit R_D^* by minimizing the Kullback–Leibler divergence, $D_{KL}(P||Q)$ between $\mathcal{N}(\mu, \Sigma)$ and $\boldsymbol{t}_\nu(\mu, \Sigma)$

$D_{KL}(P||Q)$ is a measure of distance between 2 distributions and does not require the 2 distributions to be in the same dimension.

t-SNE vs PCA: MNIST





Detailed calculations

Step 1: Construct distribution of pairs of high-dim objects

$$p_{i|j} = \frac{exp\left(- \parallel \mathbf{x}_i - \mathbf{x}_j \parallel^2 / 2\sigma_i^2\right)}{\sum_{i \neq j} exp\left(- \parallel \mathbf{x}_i - \mathbf{x}_j \parallel^2 / 2\sigma_i^2\right)} \tag{1}$$

Trick:

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

Step2: Define function in low-dim ($\mathbf{y}_i \epsilon \mathbb{R}^D$, *with* $D << N$)

$$q_{ij} = \frac{\left(1 + \parallel \mathbf{y}_i - \mathbf{y}_j \parallel^2\right)^{-1}}{\sum_k \sum_{k \neq l} \left(1 + \parallel \mathbf{y}_k - \mathbf{y}_l \parallel^2\right)^{-1}} \tag{3}$$

Step3: Minimizing the distance between p_{ij} and q_{ij}
loss function:

$$\sum_{i \neq j} p_{ij} \left(log(p_{ij}) \right) - p_{ij} \left(log(q_{ij}) \right) = \sum_{i \neq j} p_{ij} \cdot log \left(\frac{p_{ij}}{q_{ij}} \right) = KL \left(p || q \right) \tag{4}$$

Using the $KL\left(p||q\right)$ is great since it penalizes a lot when p_{ij} is big and very little when p_{ij} is small.