Visualizing Data using t-SNE

Taran Lynn, Xiaoli Yang, Xiaoxin Chen

October 26, 2020

Data Visualization and Manifold Learning

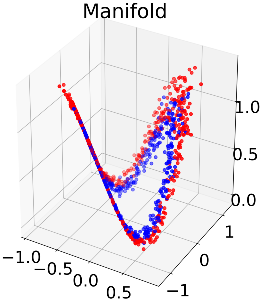
1. What is the goal of data visualization?

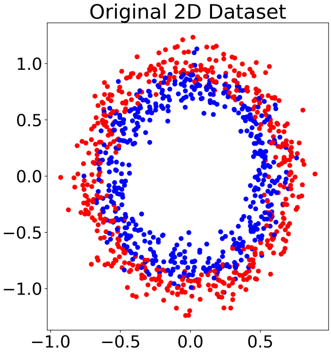
Data Visualization and Manifold Learning

1. What is the goal of data visualization?
2. We want to preserve as much structure between data points when mapping to a lower dimension
3. What is the ideal case when reducing to 2D from a higher dimension?

Data Visualization and Manifold Learning

1. What is the goal of data visualization?
2. We want to preserve as much structure between data points when mapping to a lower dimension
3. What is the ideal case when reducing to 2D from a higher dimension?

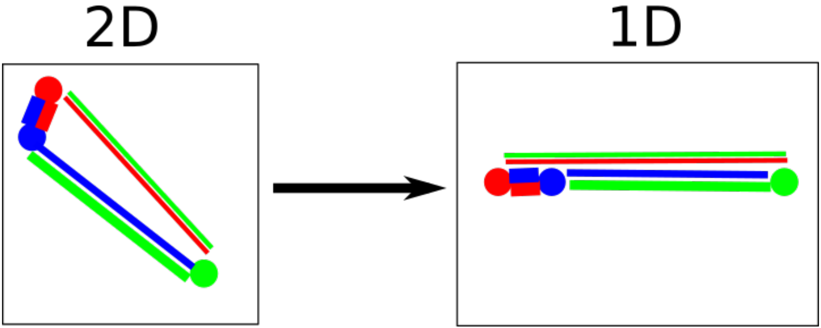




Stochastic Neighbor Embedding (SNE)

High-Level Overview

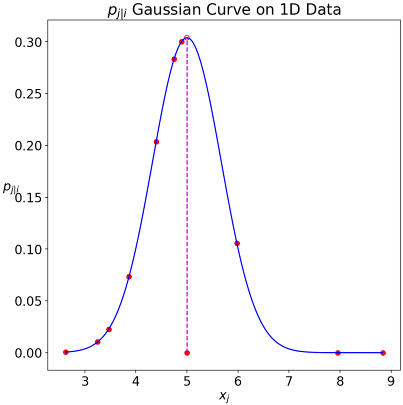
1. Each pair of data points has a measurable neighbor relationship with every other point
2. This relationship is non-symmetric
3. Stochastic neighbor embedding tries to preserve this relationship



Stochastic Neighbor Embedding (SNE)

What is $p_{j|i}$?

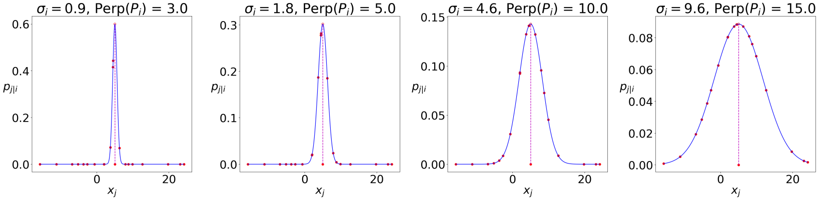
1. $p_{j|i} = \frac{\exp\left(-||x_i - x_j||^2/2\sigma_i^2\right)}{\sum_{k\neq i}\exp\left(-||x_i - x_k||^2/2\sigma_i^2\right)}$
2. Probability datapoint x_i views x_j as its neighbor given Gaussian with variance σ_i



Stochastic Neighbor Embedding (SNE)

Determining σ_i

σ_i	Determined via binary search on perplexity
Perplexity	Smooth approximation of number of neighbors $Perp(P_i) = 2^{H(P_i)}$
Shannon Entropy	Information present in probability space $H(P_i) = -\sum_j p_{j\|i} \log_2 p_{j\|i}$



Stochastic Neighbor Embedding (SNE)

Cost Function

1. $q_{j|i}$ is analogous to $p_{j|i}$ for lower dimension points y_i and y_j. $\sigma = 1/\sqrt{2}$ for all $q_{j|i}$
2. Cost function is $C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$
3. Cost function uses Kullbeck-Leibler divergence to measure the difference between the set of $q_{j|i}$ and $p_{j|i}$

Stochastic Neighbor Embedding (SNE)

Gradient Descent Optimization

1. $\frac{\partial C}{\partial y_i} = 2 \sum_j \left(p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j} \right) (y_i - y_j)$
2. $Y^{(t)} = Y^{(t-1)} + \nu \frac{\partial C}{\partial Y} + \alpha(t) \left(Y^{(t-1)} - Y^{(t-2)} \right)$
3. Simulated annealing through decaying Gaussian noise

Problems with SNE

- Cost function is not convex
  - Possible to get stuck on local minima
- Computational inefficiency
  - Must compute $p_{j|i}$ many times, especially when estimating σ_i
  - Simulated annealing is slow. Suffers compared to convex optimization
- Overcrowding