Stochastic Neighborhood Embedding

Weekly AI pills

Fabio Brau.

2020-11-13

SSSA, Emerging Digital Technologies, Pisa.

ISTITUTO DI TECNOLOGIE DELLA COMUNICAZIONE, DELL'INFORMAZIONE E DELLA PERCEZIONE

Scuola Superiore Sant'Anna

Retis Real-Time Systems Laboratory

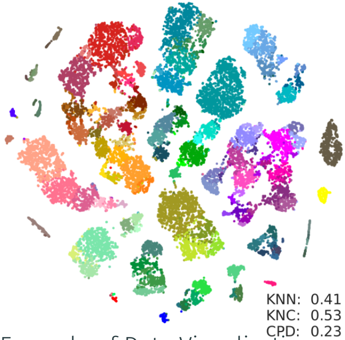
Summary

1. Entropy and Kullback–Leibler divergence
2. From SNE to t-SNE
3. Application for Visualization
4. Issues

(t)-Stochastic Neighborhood Embedding

Aims

- Dimensionality Reduction
- Data Visualization
    1. Exploration
    2. Visual Clustering



Example of Data Visualization taken from (Tasic et al., 2018)

Algorithm

Algorithm: Workflow

1. Represent each sample $x_i \in \mathcal{X}$ with $y_i \in \mathcal{Y}$ in a low-dimensional space

$$\Phi : \underset{\subseteq \mathbb{R}^n}{\mathcal{X}} \longrightarrow \underset{\subseteq \mathbb{R}^2}{\mathcal{Y}}$$

2. Convert **Geometrical Information** to **Probabilistic Distribution**

$$p_{i|j} : \text{“Probability that } x_i \text{ is } \textit{similar} \text{ to } x_j\text{”}$$
$$q_{i|j} : \text{“Probability that } y_i \text{ is } \textit{similar} \text{ to } y_j\text{”}$$

3. Compare distribution $\mathcal{P}$ of $\mathcal{X}$ to distribution $\mathcal{Q}$ of $\mathcal{X}$
4. Adjust the representation Φ to make distributions closer.

Observation
The embedding Φ is defined **point-wise**, i.e Φ is only defined over $\mathcal{X}$ through the definition $\Phi(x_i) = y_i$. Another way to say “y_i are the parameters of Φ” (Future works?).

Algorithm: Workflow

1. Represent each sample $x_i \in \mathcal{X}$ with $y_i \in \mathcal{Y}$ in a low-dimensional space

$$\Phi : \underbrace{\mathcal{X}}_{\subseteq \mathbb{R}^n} \longrightarrow \underbrace{\mathcal{Y}}_{\subseteq \mathbb{R}^2}$$

2. Convert **Geometrical Information** to **Probabilistic Distribution**

$$p_{i|j} : \text{“Probability that } x_i \text{ is } \textit{similar} \text{ to } x_j\text{”}$$
$$q_{i|j} : \text{“Probability that } y_i \text{ is } \textit{similar} \text{ to } y_j\text{”}$$

3. Compare distribution $\mathcal{P}$ of $\mathcal{X}$ to distribution $\mathcal{Q}$ of $\mathcal{X}$
4. Adjust the representation Φ to make distributions closer.

Observation
The embedding Φ is defined **point-wise**, i.e Φ is only defined over $\mathcal{X}$ through the definition $\Phi(x_i) = y_i$. Another way to say “y_i are the parameters of Φ” (Future works?).

Algorithm: Similarity and Probability Distribution

Definition (Similarity with Gaussian Kernel)

For each x_i we fix a σ_i and define

$$\forall j \neq i, \quad p_{i|j} = \frac{\exp\left(-\|x_i - x_j\|^2/(2\sigma_i)^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2/(2\sigma_i)^2\right)} \qquad (\mathcal{P}_i)$$

How σ_i impacts?

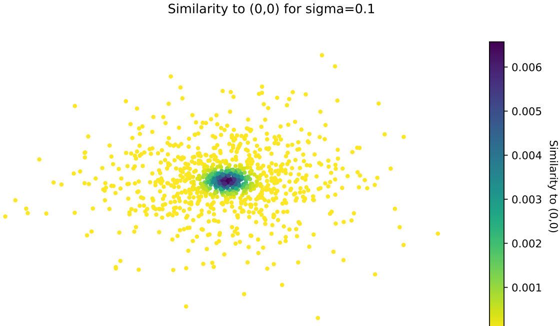
Algorithm: Similarity and Probability Distribution

Definition (Similarity with Gaussian Kernel)

For each x_i we fix a σ_i and define

$$\forall j \neq i, \quad p_{i|j} = \frac{\exp\left(-\|x_i - x_j\|^2/(2\sigma_i)^2\right)}{\sum_{k\neq i} \exp\left(-\|x_i - x_k\|^2/(2\sigma_i)^2\right)} \qquad (\mathcal{P}_i)$$

How σ_i impacts?

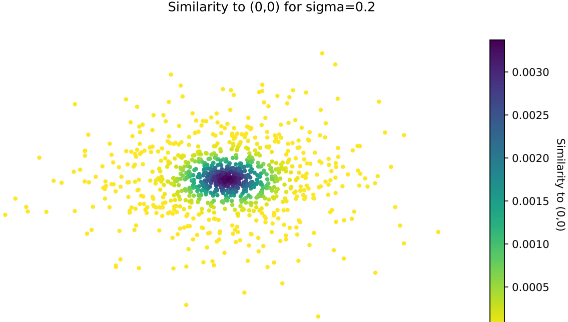


Definition (Similarity with Gaussian Kernel)

For each x_i we fix a σ_i and define

$$\forall j \neq i, \quad p_{i|j} = \frac{\exp\left(-\|x_i - x_j\|^2/(2\sigma_i)^2\right)}{\sum_{k\neq i} \exp\left(-\|x_i - x_k\|^2/(2\sigma_i)^2\right)} \qquad (\mathcal{P}_i)$$

How σ_i impacts?



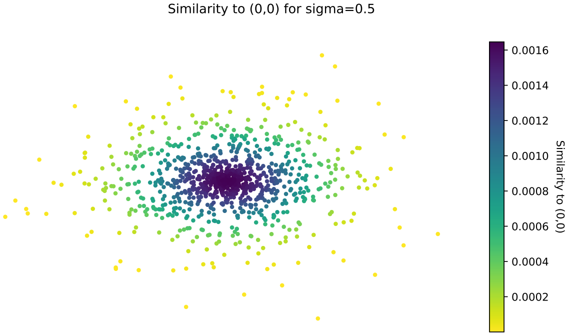
Algorithm: Similarity and Probability Distribution

Definition (Similarity with Gaussian Kernel)

For each x_i we fix a σ_i and define

$$\forall j \neq i, \quad p_{i|j} = \frac{\exp\left(-\|x_i - x_j\|^2/(2\sigma_i)^2\right)}{\sum_{k\neq i}\exp\left(-\|x_i - x_k\|^2/(2\sigma_i)^2\right)} \qquad (\mathcal{P}_i)$$

How σ_i impacts?



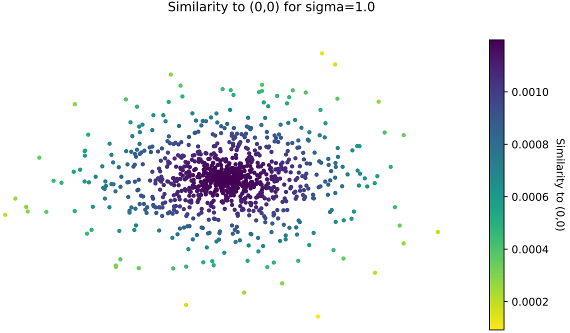
Algorithm: Similarity and Probability Distribution

Definition (Similarity with Gaussian Kernel)

For each x_i we fix a σ_i and define

$$\forall j \neq i, \quad p_{i|j} = \frac{\exp\left(-\|x_i - x_j\|^2/(2\sigma_i)^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2/(2\sigma_i)^2\right)} \qquad (\mathcal{P}_i)$$

How σ_i impacts?

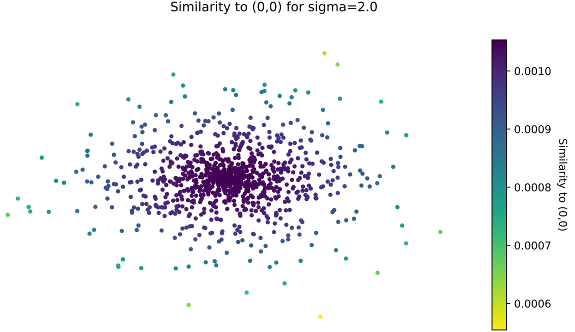


Definition (Similarity with Gaussian Kernel)

For each x_i we fix a σ_i and define

$$\forall j \neq i, \quad p_{i|j} = \frac{\exp\left(-\|x_i - x_j\|^2/(2\sigma_i)^2\right)}{\sum_{k\neq i}\exp\left(-\|x_i - x_k\|^2/(2\sigma_i)^2\right)} \qquad (\mathcal{P}_i)$$

How σ_i impacts?



Algorithm: Similarity and Probability Distribution

At the same manner we can define similarity in $\mathcal{Y}$.

$$\forall j \neq i, \quad q_{i|j} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k \neq i} \exp\left(-\|y_i - y_k\|^2\right)} \qquad (\mathcal{Q}_i)$$

Observation

$\mathcal{P}_i$ and $\mathcal{Q}_i$ are probability distributions for each i.

1. $0 \leq p_{i|j}, q_{i|j} \leq 1$ for each $j \neq i$
2. $\sum_{j \neq i} p_{i|j} = \sum_{j \neq i} q_{i|j} = 1$

Shannon Entropy, Perplexity, Kullback–Leibler Divergence

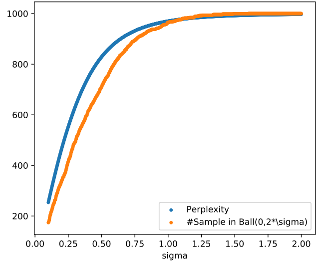
Let $\mathcal{P} = \{p_1, \cdots, p_n\}$ and $\mathcal{Q} = \{q_1, \cdots, q_n\}$ distributions

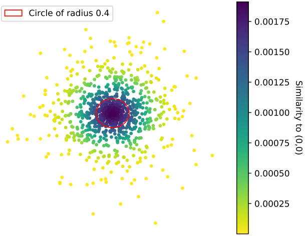
Shannon Entropy $\quad \mathbb{H}(\mathcal{P}) = -\sum_i p_i \log_2(p_i)$

Perplexity $\quad Perp(\mathcal{P}) = 2^{\mathbb{H}(\mathcal{P})}$

K.L. Divergence $\quad KL(\mathcal{P}, \mathcal{Q}) = \sum_i p_i \log_2\left(\frac{p_i}{q_i}\right)$

How σ_i and Perplexity are related?





Perplexity measures the number of samples in a neighborhood.