t-Stochastic Neighbor Embedding

Department of Electrical & Computer Engineering,
University of Waterloo, ON, Canada

Data and Knowledge Modeling and Analysis (ECE 657A)
Course Instructor: Prof. Mark Crowley
TA and Presenter of Slides: Benyamin Ghojogh

Dataset Notations

$$\text{training dataset: } \{\boldsymbol{x}_i \in \mathbb{R}^d\}_{i=1}^n, \boldsymbol{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_n] \in \mathbb{R}^{d \times n} \tag{1}$$

$$\text{embedding of training data: } \{\boldsymbol{y}_i \in \mathbb{R}^h\}_{i=1}^n, \boldsymbol{Y} = [\boldsymbol{y}_1, \dots, \boldsymbol{y}_n] \in \mathbb{R}^{h \times n} \tag{2}$$

$$h \leq d, \quad \text{usually: } h \ll d \tag{3}$$

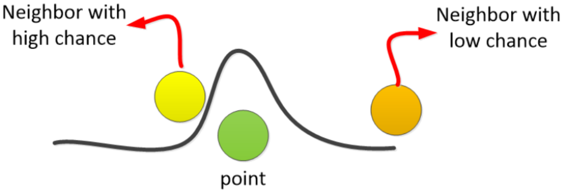
SNE and t-SNE are usually used for data visualization so we usually have $h = 2$ or $h = 3$.

Lecture Outline

1. Probabilistic Approach of Embedding
2. Stochastic Neighbor Embedding
3. Symmetric Stochastic Neighbor Embedding
4. The Crowding Problem
5. t-Stochastic Neighbor Embedding
6. Early Exaggeration
7. General Degrees of Freedom in t-SNE
8. Examples

Probabilistic Approach of Embedding

- Rather than saying that this point is neighbor of that point but the other point is not a neighbor, we can have a **probabilistic** approach.
- We say, all points are neighbors of a point with some probability. A very **similar/dissimilar** point to some other point is its neighbor with **high/low probability**.
- SNE uses **Gaussian** distribution for probability of neighborhood.



Lecture Outline

1. Probabilistic Approach of Embedding
2. Stochastic Neighbor Embedding
3. Symmetric Stochastic Neighbor Embedding
4. The Crowding Problem
5. t-Stochastic Neighbor Embedding
6. Early Exaggeration
7. General Degrees of Freedom in t-SNE
8. Examples

Stochastic Neighbor Embedding

$$f(d) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{d^2}{2\sigma^2}) \propto \exp(-\frac{d^2}{2\sigma^2}) \tag{4}$$

$$\mathbb{R} \ni p_{ij} := \frac{\exp(-d_{ij}^2)}{\sum_{k \neq i} \exp(-d_{ik}^2)} \tag{5}$$

$$\mathbb{R} \ni d_{ij}^2 := \frac{||\boldsymbol{x}_i - \boldsymbol{x}_j||_2^2}{2\sigma_i^2} \tag{6}$$

The σ_i^2 is the variance which we consider for the Gaussian distribution used for the $\boldsymbol{x}_i$. It can be set to a fixed number or by a binary search to make the entropy of distribution some specific value [1].

$$\mathbb{R} \ni q_{ij} := \frac{\exp(-z_{ij}^2)}{\sum_{k \neq i} \exp(-z_{ik}^2)} \tag{7}$$

$$\mathbb{R} \ni z_{ij}^2 := ||\boldsymbol{y}_i - \boldsymbol{y}_j||_2^2 \tag{8}$$

Stochastic Neighbor Embedding

$$\mathbb{R} \ni c_1 := \sum_{i=1}^{n} \mathrm{KL}(P_i || Q_i) = \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} p_{ij} \log(\frac{p_{ij}}{q_{ij}}) \tag{9}$$

$$\mathbb{R}^h \ni \frac{\partial c_1}{\partial \boldsymbol{y}_i} = 2 \sum_{j=1}^{n} (p_{ij} - q_{ij} + p_{ji} - q_{ji})(\boldsymbol{y}_i - \boldsymbol{y}_j) \tag{10}$$

For derivation of gradient, see [2].

Optimization using gradient descent with momentum:

$$\Delta \boldsymbol{y}_i^{(t)} := -\eta \frac{\partial c_1}{\partial \boldsymbol{y}_i} + \alpha(t)\, \Delta \boldsymbol{y}_i^{(t-1)} \tag{11}$$

$$\boldsymbol{y}_i^{(t)} := \boldsymbol{y}_i^{(t-1)} + \Delta \boldsymbol{y}_i^{(t)} \tag{12}$$

The momentum parameter:

$$\alpha(t) := \begin{cases} 0.5 & t < 250, \\ 0.8 & t \geq 250. \end{cases} \tag{13}$$

In SNE, we add jitter to the solution of initial iterations for better convergence of embedding.

Lecture Outline

1. Probabilistic Approach of Embedding
2. Stochastic Neighbor Embedding
3. Symmetric Stochastic Neighbor Embedding
4. The Crowding Problem
5. t-Stochastic Neighbor Embedding
6. Early Exaggeration
7. General Degrees of Freedom in t-SNE
8. Examples

Symmetric Stochastic Neighbor Embedding

$$\boxed{\mathbb{R} \ni p_{ij} := \frac{\exp(-d_{ij}^2)}{\sum_{k \neq l} \exp(-d_{kl}^2)}} \quad \text{vs.} \quad \mathbb{R} \ni p_{ij} := \frac{\exp(-d_{ij}^2)}{\sum_{k \neq i} \exp(-d_{ik}^2)} \tag{14}$$

$$\mathbb{R} \ni d_{ij}^2 := \frac{||\boldsymbol{x}_i - \boldsymbol{x}_j||_2^2}{2\sigma_i^2} \tag{15}$$

The first p_{ij} has a problem with outliers. If the point $\boldsymbol{x}_i$ is an outlier, its p_{ij} will be extremely small because the denominator is fixed for every point and numerator will be small for the outlier. However, If we use the second p_{ij}, the denominator for all the points is not the same and therefore, the effect of denominator waives out the effect of numerator. So we use instead:

$$\mathbb{R} \ni p_{j|i} := \frac{\exp(-d_{ij}^2)}{\sum_{k \neq i} \exp(-d_{ik}^2)} \tag{16}$$

$$\mathbb{R} \ni p_{ij} := \frac{p_{i|j} + p_{j|i}}{2n} \tag{17}$$

Symmetric Stochastic Neighbor Embedding

$$\mathbb{R} \ni p_{j|i} := \frac{\exp(-d_{ij}^2)}{\sum_{k \neq i} \exp(-d_{ik}^2)} \tag{18}$$

$$\mathbb{R} \ni p_{ij} := \frac{p_{i|j} + p_{j|i}}{2n} \tag{19}$$

But that problem does not exist in the embedding space because even for an outlier, the embedded points are initialized close together and not far.

$$\boxed{\mathbb{R} \ni q_{ij} := \frac{\exp(-z_{ij}^2)}{\sum_{k \neq l} \exp(-z_{kl}^2)},} \quad \text{vs.} \quad \mathbb{R} \ni q_{ij} := \frac{\exp(-z_{ij}^2)}{\sum_{k \neq i} \exp(-z_{ik}^2)} \tag{20}$$

$$\mathbb{R} \ni z_{ij}^2 := ||\boldsymbol{y}_i - \boldsymbol{y}_j||_2^2 \tag{21}$$

Symmetric Stochastic Neighbor Embedding

$$\mathbb{R} \ni c_2 := \sum_{i=1}^{n} \text{KL}(P_i || Q_i) = \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} p_{ij} \log(\frac{p_{ij}}{q_{ij}}) \tag{22}$$

$$\mathbb{R}^h \ni \frac{\partial c_2}{\partial \boldsymbol{y}_i} = 4 \sum_{j=1}^{n} (p_{ij} - q_{ij})(\boldsymbol{y}_i - \boldsymbol{y}_j) \tag{23}$$

For derivation of gradient, see [2].

As before, optimization is done using gradient descent with momentum.

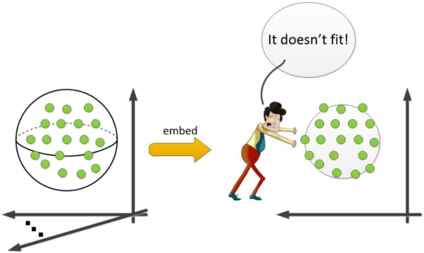
In symmetric SNE, we add jitter to the solution of initial iterations for better convergence of embedding.

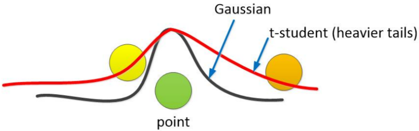
Lecture Outline

1. Probabilistic Approach of Embedding
2. Stochastic Neighbor Embedding
3. Symmetric Stochastic Neighbor Embedding
4. The Crowding Problem
5. t-Stochastic Neighbor Embedding
6. Early Exaggeration
7. General Degrees of Freedom in t-SNE
8. Examples

The Crowding Problem

- If we want to fit one million people in a room, they don't fit. So, let's enlarge the room!
- **Crowding problem**: If we want to fit the large information of high dimensional data into low dimensional subspace, we should enlarge the distribution!
- **Student-t distribution** has heavier tails than the Gaussian distribution.





Lecture Outline

1. Probabilistic Approach of Embedding
2. Stochastic Neighbor Embedding
3. Symmetric Stochastic Neighbor Embedding
4. The Crowding Problem
5. t-Stochastic Neighbor Embedding
6. Early Exaggeration
7. General Degrees of Freedom in t-SNE
8. Examples

t-Stochastic Neighbor Embedding

$$\mathbb{R} \ni p_{j|i} := \frac{\exp(-d_{ij}^2)}{\sum_{k \neq i} \exp(-d_{ik}^2)}, \quad \mathbb{R} \ni p_{ij} := \frac{p_{i|j} + p_{j|i}}{2n} \tag{24}$$

$$\mathbb{R} \ni d_{ij}^2 := \frac{||\boldsymbol{x}_i - \boldsymbol{x}_j||_2^2}{2\sigma_i^2} \tag{25}$$

$$q_{ij} = \frac{(1 + z_{ij}^2)^{-1}}{\sum_{k \neq l}(1 + z_{kl}^2)^{-1}} \tag{26}$$

$$\mathbb{R} \ni z_{ij}^2 := ||\boldsymbol{y}_i - \boldsymbol{y}_j||_2^2 \tag{27}$$

In the embedding space, we use Student-t distribution (with one degree of freedom) which is the standard Cauchy distribution:

$$f(z) = \frac{1}{\pi(1 + z^2)} \tag{28}$$

t-Stochastic Neighbor Embedding

$$\mathbb{R} \ni c_3 := \sum_{i=1}^{n} \text{KL}(P_i || Q_i) = \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} p_{ij} \log(\frac{p_{ij}}{q_{ij}}) \tag{29}$$

$$\frac{\partial c_3}{\partial \boldsymbol{y}_i} = 4 \sum_{j=1}^{n} (p_{ij} - q_{ij})(1 + ||\boldsymbol{y}_i - \boldsymbol{y}_j||_2^2)^{-1}(\boldsymbol{y}_i - \boldsymbol{y}_j) \tag{30}$$

For derivation of gradient, see [2].

As before, optimization is done using gradient descent with momentum.

For t-SNE, there is no need to add jitter to the solution of initial iterations because it is more robust than SNE.

Lecture Outline

1. Probabilistic Approach of Embedding
2. Stochastic Neighbor Embedding
3. Symmetric Stochastic Neighbor Embedding
4. The Crowding Problem
5. t-Stochastic Neighbor Embedding
6. Early Exaggeration
7. General Degrees of Freedom in t-SNE
8. Examples

Early Exaggeration

In t-SNE, it is better to multiply all p_{ij}'s by a constant (e.g., 4) in the initial iterations:

$$p_{ij} := p_{ij} \times 4, \tag{31}$$

which is called **early exaggeration**.

- This heuristic helps the optimization to focus on the large p_{ij}'s (close neighbors) more in the early iterations.
- This is because large p_{ij}'s are affected more by multiplying by 4 than the small p_{ij}'s.

After the neighbors are embedded close to one another, we are free not to do it and let far away points be handled as well. Note that the early exaggeration is optional and not mandatory.

Lecture Outline

1. Probabilistic Approach of Embedding
2. Stochastic Neighbor Embedding
3. Symmetric Stochastic Neighbor Embedding
4. The Crowding Problem
5. t-Stochastic Neighbor Embedding
6. Early Exaggeration
7. General Degrees of Freedom in t-SNE
8. Examples

General Degrees of Freedom in t-SNE

$$\text{Cauchy distribution: } f(z) = \frac{1}{\pi(1+z^2)} \tag{32}$$

$$\text{Student-t distribution: } f(z) = \frac{\Gamma(\frac{\delta+1}{2})}{\sqrt{\delta \times \pi}\ \ \Gamma(\frac{\delta}{2})}(1+\frac{z^2}{\delta})^{-\frac{\delta+1}{2}} \tag{33}$$

$$q_{ij} = \frac{(1+z_{ij}^2/\delta)^{-(\delta+1)/2}}{\sum_{k \neq l}(1+z_{kl}^2/\delta)^{-(\delta+1)/2}} \tag{34}$$

$$\mathbb{R} \ni p_{j|i} := \frac{\exp(-d_{ij}^2)}{\sum_{k \neq i}\exp(-d_{ik}^2)}, \quad \mathbb{R} \ni p_{ij} := \frac{p_{i|j}+p_{j|i}}{2n} \tag{35}$$

$$\mathbb{R} \ni d_{ij}^2 := \frac{||\boldsymbol{x}_i - \boldsymbol{x}_j||_2^2}{2\sigma_i^2} \tag{36}$$

General Degrees of Freedom in t-SNE

For determining the degrees of freedom, we can do one the following items:

1. Set δ to be fixed
2. $\delta = h - 1$
    - Volume $\propto \exp(h)$ [example: πr^2 and $(4/3)\pi r^3$]
    - In Eq. (33): $f(z) \propto \exp(-\delta)$
    - $\delta \propto h$, special case: $\delta = 1, h = 2 \implies \delta = h - 1$
3. Updating both variables δ and $\{\boldsymbol{y}_i\}_{i-1}^{n}$ by restricted Boltzmann machine [3] or backpropagation
4. Alternating optimization approach [4]:
   alternate between δ and $\{\boldsymbol{y}_i\}_{i-1}^{n}$ in optimization

General Degrees of Freedom in t-SNE

$$\mathbb{R} \ni c_3 := \sum_i \mathrm{KL}(P_i || Q_i) = \sum_i \sum_{j \neq i} p_{ij} \log(\frac{p_{ij}}{q_{ij}}) \tag{37}$$

$$\frac{\partial c_3}{\partial \delta} = \sum_{i \neq j} \Big(\frac{-(1+\delta) z_{ij}^2}{2\delta^2 (1 + \frac{z_{ij}^2}{\delta})} + \frac{1}{2} \log(1 + \frac{z_{ij}^2}{\delta}) \Big) (p_{ij} - q_{ij}) \tag{38}$$

$$\delta := \delta - \mathrm{sign}(\frac{\partial c_3}{\partial \delta}) \tag{39}$$

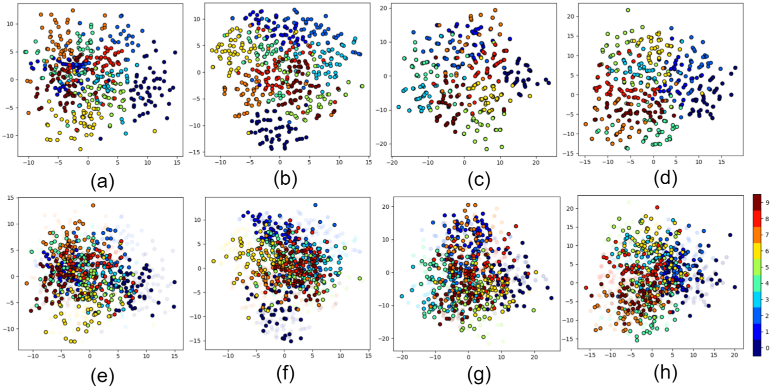
$$\frac{\partial c_3}{\partial \boldsymbol{y}_i} = \frac{2\delta + 2}{\delta} \times \sum_j (p_{ij} - q_{ij})(1 + \frac{||\boldsymbol{y}_i - \boldsymbol{y}_j||_2^2}{\delta})^{-1} (\boldsymbol{y}_i - \boldsymbol{y}_j) \tag{40}$$

For derivation of gradients, see [2].

Lecture Outline

1. Probabilistic Approach of Embedding
2. Stochastic Neighbor Embedding
3. Symmetric Stochastic Neighbor Embedding
4. The Crowding Problem
5. t-Stochastic Neighbor Embedding
6. Early Exaggeration
7. General Degrees of Freedom in t-SNE
8. Examples

Examples

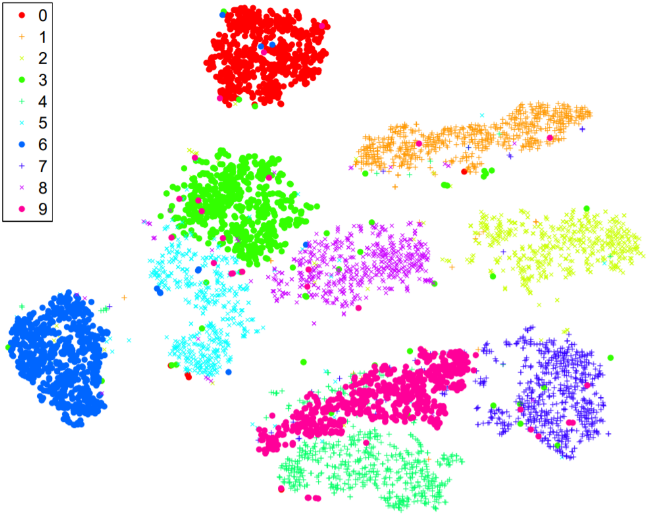


(a) (b) (c) (d)

(e) (f) (g) (h)

The embeddings of training data are shown in (a) SNE, (b) symmetric SNE, (c) t-SNE (Cauchy-SNE), and (d) t-SNE with general degrees of freedom. The out-of-sample embeddings are shown in (e) SNE, (f) symmetric SNE, (g) t-SNE (Cauchy-SNE), and (h) t-SNE with general degrees of freedom.

Examples on MNIST dataset



The credit of this image is for [5].

Useful Resources To Read

- Tutorial paper: "Stochastic Neighbor Embedding with Gaussian and Student-t Distributions: Tutorial and Survey" [2]
- Tutorial YouTube videos by Prof. Ali Ghodsi at University of Waterloo: [Click here]

References

[1] G. E. Hinton and S. T. Roweis, “Stochastic neighbor embedding,” in *Advances in neural information processing systems*, pp. 857–864, 2003.

[2] B. Ghojogh, A. Ghodsi, F. Karray, and M. Crowley, “Stochastic neighbor embedding with Gaussian and Student-t distributions: Tutorial and survey,” *arXiv preprint arXiv:2009.10301*, 2020.

[3] L. van der Maaten, “Learning a parametric embedding by preserving local structure,” in *Artificial Intelligence and Statistics*, pp. 384–391, 2009.

[4] P. Jain and P. Kar, “Non-convex optimization for machine learning,” *arXiv preprint arXiv:1712.07897*, 2017.

[5] L. van der Maaten and G. Hinton, “Visualizing data using t-SNE,” *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.