t-SNE (t distributed Stochastic Neighborhood Embedding)

Divy Kangeyan

May 26, 2017

Overview of t-SNE

- t-SNE is a dimension reduction/data visualization method

Overview of t-SNE

- t-SNE is a dimension reduction/data visualization method
- Proposed by Laurens van der Maaten & Geoffrey Hinton in 2008

Overview of t-SNE

- t-SNE is a dimension reduction/data visualization method
- Proposed by Laurens van der Maaten & Geoffrey Hinton in 2008
- t-SNE tends to preserve local structure at the same time preserving the global structure as much as possible

SNE (Stochastic Neighbor Embedding)

- Aim is to match distributions of distances between points in high and low dimensional space via conditional probabilities

SNE (Stochastic Neighbor Embedding)

- Aim is to match distributions of distances between points in high and low dimensional space via conditional probabilities
- Assume distances in both high and low dimensional space are Gaussian-distributed

SNE

- Let x_i be the i^{th} object in high dimensional space

SNE

- Let x_i be the i^{th} object in high dimensional space
- Let y_i be the i^{th} object in low dimensional space

SNE

- Let x_i be the i^{th} object in high dimensional space
- Let y_i be the i^{th} object in low dimensional space
- Construct:

$$p_{j|i} = \frac{exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k \neq i} exp(-||x_i - x_k||^2/2\sigma_i^2)}$$

SNE

- Let x_i be the i^{th} object in high dimensional space
- Let y_i be the i^{th} object in low dimensional space
- Construct:

$$p_{j|i} = \frac{exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k \neq i} exp(-||x_i - x_k||^2/2\sigma_i^2)}$$

$$q_{j|i} = \frac{exp(-||y_i - y_j||^2)}{\sum_{k \neq i} exp(-||y_i - y_k||^2)}$$

SNE

- Let x_i be the i^{th} object in high dimensional space
- Let y_i be the i^{th} object in low dimensional space
- Construct:

$$p_{j|i} = \frac{exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k \neq i} exp(-||x_i - x_k||^2/2\sigma_i^2)}$$

$$q_{j|i} = \frac{exp(-||y_i - y_j||^2)}{\sum_{k \neq i} exp(-||y_i - y_k||^2)}$$

- $p_{i|i} = q_{i|i} = 0$

SNE

- Let x_i be the i^{th} object in high dimensional space
- Let y_i be the i^{th} object in low dimensional space
- Construct:

$$p_{j|i} = \frac{exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k\neq i} exp(-||x_i - x_k||^2/2\sigma_i^2)}$$

$$q_{j|i} = \frac{exp(-||y_i - y_j||^2)}{\sum_{k\neq i} exp(-||y_i - y_k||^2)}$$

- $p_{i|i} = q_{i|i} = 0$
- Match these functions by minimizing sum of Kullback-Leibler divergences:

$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} log\left(\frac{p_{j|i}}{q_{j|i}}\right)$$

SNE

- Let x_i be the i^{th} object in high dimensional space
- Let y_i be the i^{th} object in low dimensional space
- Construct:

$$p_{j|i} = \frac{exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k \neq i} exp(-||x_i - x_k||^2/2\sigma_i^2)}$$

$$q_{j|i} = \frac{exp(-||y_i - y_j||^2)}{\sum_{k \neq i} exp(-||y_i - y_k||^2)}$$

- $p_{i|i} = q_{i|i} = 0$
- Match these functions by minimizing sum of Kullback-Leibler divergences:

$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} log \left(\frac{p_{j|i}}{q_{j|i}} \right)$$

SNE

- Since KL divergence is asymmetric,
  - large cost for representing nearby data points in high dimensional map by widely separated points in the low dimensional map
  - smaller coast for representing widely separated data points in high dimensional map by nearby points in the low dimension

SNE

- Since KL divergence is asymmetric,
  - large cost for representing nearby data points in high dimensional map by widely separated points in the low dimensional map
  - smaller coast for representing widely separated data points in high dimensional map by nearby points in the low dimension
- Hence local structure is highly preserved

SNE

- Since KL divergence is asymmetric,
  - large cost for representing nearby data points in high dimensional map by widely separated points in the low dimensional map
  - smaller coast for representing widely separated data points in high dimensional map by nearby points in the low dimension
- Hence local structure is highly preserved
- σ_i is associated with a parameter called perplexity which can be loosely interpreted as the number of close neighbors each point has

SNE

- Since KL divergence is asymmetric,
  - large cost for representing nearby data points in high dimensional map by widely separated points in the low dimensional map
  - smaller coast for representing widely separated data points in high dimensional map by nearby points in the low dimension
- Hence local structure is highly preserved
- σ_i is associated with a parameter called perplexity which can be loosely interpreted as the number of close neighbors each point has
- σ_i is found via binary search (More in the paper)

SNE

- Since KL divergence is asymmetric,
  - large cost for representing nearby data points in high dimensional map by widely separated points in the low dimensional map
  - smaller coast for representing widely separated data points in high dimensional map by nearby points in the low dimension
- Hence local structure is highly preserved
- σ_i is associated with a parameter called perplexity which can be loosely interpreted as the number of close neighbors each point has
- σ_i is found via binary search (More in the paper)
- Gradient of the cost function

$$\frac{\delta C}{\delta y_i} = 2\sum_j (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})(y_i - y_j)$$

SNE optimization

- Given the cost function SNE uses gradient descent for optimization

SNE optimization

- Given the cost function SNE uses gradient descent for optimization
- In addition to the gradient of the cost function it also has a momentum term to speed up the optimization and to avoid local optima

SNE optimization

- Given the cost function SNE uses gradient descent for optimization
- In addition to the gradient of the cost function it also has a momentum term to speed up the optimization and to avoid local optima
- $\mathcal{Y}^{(t)} = \mathcal{Y}^{(t-1)} + \eta \frac{\delta C}{\delta y} + \alpha(t)(\mathcal{Y}^{(t-1)} - \mathcal{Y}^{(t-2)})$
  where
  $\alpha(t)$: Momentum at iteration t
  $\mathcal{Y}^{(t)}$: Solution at iteration t
  η: Learning rate

Drawbacks of SNE and novel features of t-SNE

SNE has two main drawbacks:

- Cost function is difficult to optimize
- Crowding problem

Drawbacks of SNE and novel features of t-SNE

SNE has two main drawbacks:

- Cost function is difficult to optimize
- Crowding problem

Novel features in t-SNE

- t-SNE cost function has two distinct features:
  - Cost function is symmetrized version of that in SNE. i.e. ($p_{i|j} = p_{j|i}$ and $q_{i|j} = q_{j|i}$)
  - Student t-distribution is used to compute the similarities between data points in the low dimensional space.

Symmetric SNE

- The main feature in symmetric SNE is that $p_{ij} = p_{ji}$ and $p_{ii} = q_{ii} = 0$ for all i,j
- $$q_{ij} = \frac{exp(-||y_i - y_j||^2)}{\sum_{k \neq l} exp(-||y_k - y_l||^2)}$$

Symmetric SNE

- The main feature in symmetric SNE is that $p_{ij} = p_{ji}$ and $p_{ii} = q_{ii} = 0$ for all i,j
- $$q_{ij} = \frac{exp(-||y_i - y_j||^2)}{\sum_{k \neq l} exp(-||y_k - y_l||^2)}$$
- $$p_{ij} = \frac{exp(-||x_i - x_j||^2/2\sigma^2)}{\sum_{k \neq l} exp(-||x_k - x_l||^2/2\sigma^2)}$$
- Gradient of the cost function:

$$\frac{\delta C}{\delta y_i} = 4 \sum_j (p_{ji} - q_{ji})(y_i - y_j)$$

t-SNE mapping

- In t-SNE a student t distribution with one degree of freedom (Cauchy distribution) is used to represent the low dimensional map:

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

- t-distribution is robust to outliers and unlike a Gaussian distribution it doesn't have exponent in it so faster to evaluate
- Gradient of the cost function:

$$\frac{\delta C}{\delta y_i} = 4 \sum_j (p_{ji} - q_{ji})(y_i - y_j)(1 + ||y_i - y_j||^2)^{-1}$$

t-SNE algorithm

The general idea of t-SNE algorithm:

- Data: data set $\mathcal{X}$ with data points $= x_1, x_2, ..., x_n$ each of these points have very high dimension

t-SNE algorithm

The general idea of t-SNE algorithm:

- Data: data set $\mathcal{X}$ with data points $= x_1, x_2, ..., x_n$ each of these points have very high dimension
- cost function parameter: Perplexity *Perp*, perplexity is associated with variance σ in the cost function

t-SNE algorithm

The general idea of t-SNE algorithm:

- Data: data set $\mathcal{X}$ with data points $= x_1, x_2, ..., x_n$ each of these points have very high dimension
- cost function parameter: Perplexity *Perp*, perplexity is associated with variance σ in the cost function
- Optimization parameters: number of iterations T, learning rate η, momentum $\alpha(t)$

t-SNE algorithm

t-SNE algorithm:
begin

1. Compute pairwise affinities $p_{j|i}$ with perplexity $Perp$

t-SNE algorithm

t-SNE algorithm:
begin

1. Compute pairwise affinities $p_{j|i}$ with perplexity *Perp*
2. Set $p_{ij} = \frac{p_{i|j}+p_{j|i}}{2n}$ (n - Number of data points)

t-SNE algorithm

t-SNE algorithm:
begin

1. Compute pairwise affinities $p_{j|i}$ with perplexity *Perp*
2. Set $p_{ij} = \frac{p_{i|j}+p_{j|i}}{2n}$ (n - Number of data points)
3. Sample initial solution $Y^{(0)} = y_1, y_2, ..., y_n$ from $N(0, 10^{-4} I)$
   for t=1 **to** T **do**
   1. Compute low-dimensional affinities q_{ij}

t-SNE algorithm:
begin

1. Compute pairwise affinities $p_{j|i}$ with perplexity $Perp$
2. Set $p_{ij} = \frac{p_{i|j}+p_{j|i}}{2n}$ (n - Number of data points)
3. Sample initial solution $Y^{(0)} = y_1, y_2, ..., y_n$ from $N(0, 10^{-4}I)$
   for t=1 **to** T **do**
   1. Compute low-dimensional affinities q_{ij}
   2. Compute gradient $\frac{\delta C}{\delta y}$

t-SNE algorithm

t-SNE algorithm:
begin

1. Compute pairwise affinities $p_{j|i}$ with perplexity *Perp*
2. Set $p_{ij} = \frac{p_{i|j}+p_{j|i}}{2n}$ (n - Number of data points)
3. Sample initial solution $Y^{(0)} = y_1, y_2, ..., y_n$ from $N(0, 10^{-4}I)$
   for t=1 **to** T **do**
   1. Compute low-dimensional affinities q_{ij}
   2. Compute gradient $\frac{\delta C}{\delta y}$
   3. Set $\mathcal{Y}^{(t)} = \mathcal{Y}^{(t-1)} + \eta\frac{\delta C}{\delta y} + \alpha(t)(\mathcal{Y}^{(t-1)} - \mathcal{Y}^{(t-2)})$

   end

end

Software packages with t-SNE implementation

- *tsne* package in R:
  `tsne(dataset, k = number of lower level dimension, perplexity,max_iter)`

Software packages with t-SNE implementation

- *tsne* package in R:
  `tsne(dataset, k = number of lower level dimension, perplexity,max_iter)`
- *sklearn.manifold.TSNE* in sklearn module
  some of the arguments include `n_components`, `perplexity`, `learning_rate`, `n_iter`

Software packages with t-SNE implementation

- *tsne* package in R:
  `tsne(dataset, k = number of lower level dimension, perplexity,max_iter)`
- *sklearn.manifold.TSNE* in sklearn module
  some of the arguments include `n_components`, `perplexity`, `learning_rate`, `n_iter`
- Julia implementation in TSne

Software packages with t-SNE implementation

- *tsne* package in R:
  `tsne(dataset, k = number of lower level dimension, perplexity,max_iter)`
- *sklearn.manifold.TSNE* in sklearn module
  some of the arguments include `n_components`, `perplexity`, `learning_rate`, `n_iter`
- Julia implementation in TSne
- MATLAB implementation via `tsne` function
  input argument for the function include `X = dataset, labels if already known, no_dims = number of dimension expected in lower level manifold, init_dims = initial number of dimension in the data, perplexity`

Criticism of t-SNE

- There is a great paper by Wattenberg, Viegas and Johnson titled **How to Use t-SNE Effectively**, it explains some of the drawbacks of t-SNE with some interactive visualization tools

Criticism of t-SNE

- There is a great paper by Wattenberg, Viegas and Johnson titled **How to Use t-SNE Effectively**, it explains some of the drawbacks of t-SNE with some interactive visualization tools
- Different perplexity can lead to completely different clusters (too small - local variations dominate, too large - global change dominate)

Criticism of t-SNE

- There is a great paper by Wattenberg, Viegas and Johnson titled **How to Use t-SNE Effectively**, it explains some of the drawbacks of t-SNE with some interactive visualization tools
- Different perplexity can lead to completely different clusters (too small - local variations dominate, too large - global change dominate)
- Cluster size doesn't have any meaning to it. Naturally expands dense cluster and contracts sparse cluster

Criticism of t-SNE

- There is a great paper by Wattenberg, Viegas and Johnson titled **How to Use t-SNE Effectively**, it explains some of the drawbacks of t-SNE with some interactive visualization tools
- Different perplexity can lead to completely different clusters (too small - local variations dominate, too large - global change dominate)
- Cluster size doesn't have any meaning to it. Naturally expands dense cluster and contracts sparse cluster
- Distance between clusters might not have clear interpretation

Criticism of t-SNE

- There is a great paper by Wattenberg, Viegas and Johnson titled **How to Use t-SNE Effectively**, it explains some of the drawbacks of t-SNE with some interactive visualization tools
- Different perplexity can lead to completely different clusters (too small - local variations dominate, too large - global change dominate)
- Cluster size doesn't have any meaning to it. Naturally expands dense cluster and contracts sparse cluster
- Distance between clusters might not have clear interpretation
- Sometime random noise can lead to false positive structure in the t-SNE projection

Criticism of t-SNE

- There is a great paper by Wattenberg, Viegas and Johnson titled **How to Use t-SNE Effectively**, it explains some of the drawbacks of t-SNE with some interactive visualization tools
- Different perplexity can lead to completely different clusters (too small - local variations dominate, too large - global change dominate)
- Cluster size doesn't have any meaning to it. Naturally expands dense cluster and contracts sparse cluster
- Distance between clusters might not have clear interpretation
- Sometime random noise can lead to false positive structure in the t-SNE projection

Summary

- t-SNE is an incredibly successful tool for clustering and data visualization
- It provides better structure for very high dimensional data
- However higher flexibility leads to other drawback like lack of interpretability
- Not very intuitive to tune the parameters (perplexity, iterations, tolerance etc.)

  How to use t-SNE effectively