Thứ
Ngày . . Nơi

1) Biến đổi lại công thức toán t-SNE, SNE, có tính đạo hàm bởi các parameter.

Bài toán t-SNE:

The similarity of data point x_j to data point x_i is the conditional probability $p(j|i)$. For nearby data points, $p(j|i)$ is high whereas for widely separated data points, $p(j|i)$ will be almost infinitesimal.

The conditional probability is given by:

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

where σ_i is the variance of the Gaussian that is centered on data point x_i. For the low-dimensional counterpart y_i & y_j of the high-dimensional data points x_i & x_j. It is possible to compute a similar conditional probability, which we denote by $q(j|i)$:

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

We set $q(i|i) = 0$. If the map points y_i and y_j correctly model the similarity between the high-dimensional data point x_i and x_j the conditional probabilities $p_{j|i}$ and $q_{j|i}$ will be equal. SNE minimise the sum of Kullback-Leibler ~~convergen~~ divergence over all data points using a gradient descent method. The cost function C is given by:

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

KOKUYO

in which P_i represents the conditional probability distribution over all other data points given data point x_i. Q_i represents the conditional probability distribution over all other map points given map point y_i. Because the Kullback-Leiber divergence is not symmetric different types of error in the pairwise distances in the low-dimensional map are not weighted equal.

SNE performs a binary search for the value of σ_i that produces a P_i with a fixed perplexity that is specified by the user. The perplexity is defined as:

$$Per(P_i) = 2^{H(P_i)}$$

where $H(P_i)$ is the Shannon entropy of P_i measured in:

$$H(P_i) = -\sum_j p_{j|i} \log_2 p_{j|i}$$

The perplexity can be interpreted as a measure of the effective number of neighbors. The performance of SNE is fairly robust to changes in the perplexity & typical values are between 5 & 50. The minimization of the cost function is performed using a gradient descent method. The gradient has a surprisingly simple form:

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

Mathematically, the gradient update with a momen term is given by:

$$y^{(t)} = y^{(t-1)} + \eta \frac{\delta C}{\delta y} + \alpha(t)(y^{(t-1)} - y^{(t-2)})$$

where $y^{(t)}$ indicates the solution at iteration t, η indicates the learning rate & $\alpha(t)$ represents the momen at iteration t.

As an alternative to ~~minimizing~~ minimize the sum of the Kullback-Leibler divergence between a joint prob distribution P in the high-dimensional space & a joint distribution Q in the two-dimensional space

$$C = \sum_i KL(P||Q) = \sum_i \sum_j p(j|i) \log \frac{p_{ij}}{q_{ij}}$$

Set p_{ii} & q_{ii} to 0. We refer to this type of SNE as symmetric SNE because it has the property that $p_{ij} = p_{ji}$ & $q_{ij} = q_{ji}$ $\forall i, j$. In symmetric SNE, the pairwise similarities in the two-dimensional map q_{ij} are given by:

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

The obvious way to define the pairwise similarities in the high dimensional space $p(ij)$ is:

$$p_{ij} = \frac{e^{-\|x_i - x_j\|^2 / 2\sigma^2}}{\sum_{k \neq l} e^{-\|x_k - x_l\|^2 / 2\sigma^2}}$$

The gradient of symmetric SNE is fairly similar to that of asymmetric SNE & is given:

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