WEEK-3 REPORT P-1

DASARI SRINITH (CS21BTECH11015)

t-SNE:

Theory:

- t-Distributed Stochastic Neighbor Embedding (t-SNE) is an unsupervised, non-linear technique primarily used for data exploration and visualizing high-dimensional data.
- t-SNE gives you a feel or intuition of how the data is arranged in a high-dimensional space.
- The t-SNE algorithm calculates a similarity measure between pairs of instances in the high dimensional space and in the low dimensional space.
- Then it optimizes the similarity values using the cost function.
- The similarity of high-dimentional datapoint x_j to datapoint x_i is the conditional probability, $p_{j|i}$, that x_i would pick x_j as its neighbor if neighbors were picked in proportion to their probability density under a Gaussian centered at x_i.
- Mathematically ,$p_{j|i}$ is given by

$$p_{j|i} = \frac{\exp\left(-||x_i - x_j||^2/2\sigma^2\right)}{\sum_{k\neq i}\exp\left(-||x_i - x_k||^2/2\sigma^2\right)} \tag{1}$$

  where σ_i is the variance of the Gaussian that is centered on datapoint

- For the low-dimensional counterparts y_i and y_j of the high-dimensional datapoints x_i and x_j , it is possible to compute a similar conditional probability, which we denote by $q_{j|i}$
- Hence ,

$$q_{j|i} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k\neq i}(1 + ||y_i - y_k||^2)^{-1}} \tag{2}$$

- If the map points y_i and y_j correctly model the similarity between the high-dimensional datapoints x_i and x_j , the conditional probabilities $p_{j|i}$ and $q_{j|i}$ will be equal.

- SNE minimizes the sum of Kullback-Leibler divergences over all datapoints using a gradient descent method. The cost function C is given by ,

$$C = \sum_i \sum_j p_{j|i} \log \left(\frac{p_{j|i}}{q_{j|i}} \right) \tag{3}$$

- The minimization of the cost function is performed using a gradient descent method. The gradient has a surprisingly simple form

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

- The gradient update is given by ,

$$\gamma^t = \gamma^{t-1} + \eta \frac{\delta C}{\delta y_i} + \alpha(t) \left(\gamma^{t-1} - \gamma^{t-2} \right) \tag{5}$$

  where γ^t indicates the solution at iteration t, η indicates the learning rate, and $\alpha(t)$ represents the momentum at iteration t

1. What is perplexity?

   Solution:
   It describes the expected density around each point or, in other words, relates to the target number of nearest neighbors from the point of interest.

2. Why does t-SNE takes so long to calculate?

   Solution:
   t-SNE is a resource-intensive algorithm because it inspects every single data point and measures the distances between every pair of points.

3. What is the value of $p_{i|i}$ taken?

   Solution:
   0

4. When is t-SNE misleading?

   Solution:
   If you get a T-Sne graph with lots of overlapping data, there is a high chance that the classifier will perform badly.

5. If we take two points and try to calculate the conditional probability between them then values of $p_{j|i}$ and $p_{i|j}$ will be different , then which value should be taken ?

   Solution:

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2N} \tag{6}$$