t-SNE → Dimensionality Red^n

01 November 2024 17:35

t-distributed | Stochastic | Neighbor | Embedding

NON LINEAR dimensionality reduction

① Model similarity in higher dimensional space
② Maintain that similarity in lower dimension.

Effective in visualizing LOCAL STRUCTURES

→ who is usually high dimensional w/ complex relationships? Text + image data.

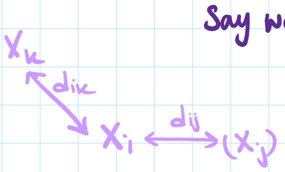
- effective in preserving LOCAL structure
- may distort global relationship

→ DO NOT USE if global structure is important.

Analysing the t-SNE algorithm:

We've N points in high dimension → $x_1, x_2, x_3 \ldots x_N$

Say we choose a point x_i.
What's the probability that we choose x_j as its neighbor?



$$P_{j|i} = f\left(d_{ij}^2, \; d_{ik}^2\right) \quad k \neq i,j$$

d_{ij}: distance b/w i + j

d_{ik}: distance b/w i and every point but itself + j.

now we look for some function f that can satisfy these conditions —

- If $d_{ij} \uparrow\uparrow$ $P_{j|i} \downarrow\downarrow$ — cuz probability of choosing j as a neighbor goes down if you're far from i.
- If $d_{ik} \uparrow\uparrow$ $P_{j|i} \uparrow\uparrow$ — you're isolating i from everyone such that j is her only close friend which makes the probability of choosing j as a neighbor ↑. — think of stockholm syndrome

$$P_{j|i} = \frac{\exp\left[-|x_i - x_j|^2 / 2\sigma_i^2\right]}{\exp\left[-|x_i - x_k|^2 / 2\sigma_i^2\right]}$$

gaussian distribution

numerator → here our mean is x_j, we look around her

denominator → here our mean is x_k, so we look around him

We're assigning probability proportional to the probability density of a normal distribution of distance.

So we covered probability of picking x_j as a neighbor of x_i. $(P_{j|i})$

Let's move further.

What's the probability that i + j are grouped together, ASSUMING that the datapoints are picked at random? P_{ij}

$P_{ij} = \frac{1}{\ldots} P_{j|i} + \frac{1}{\ldots} P_{i|j}$

p_{ij}

$$p_{ij} = \frac{1}{2N} p_{j|i} + \frac{1}{2N} p_{i|j}$$

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

→ prob that j is chosen as i's neighbor

→ prob that i is chosen as j's neighbor.

2N → cuz I have N data points and I'm counting them once for $p_{j|i}$ + then again for $p_{i|j}$.

→ divide by 2N so that p_{ij} sums to 1 and remove redundancies.

Now, your chances of picking i + j together is based off of the spread (σ_i) of the normal distribution. Remember? From earlier?

So this is suitable if x_i is a SPARSE PORTION of the sample space.

Assume that x_i, x_j is from a high dimensional space.

↓

This is transferred to a lower dimensional space (y_i, y_j)

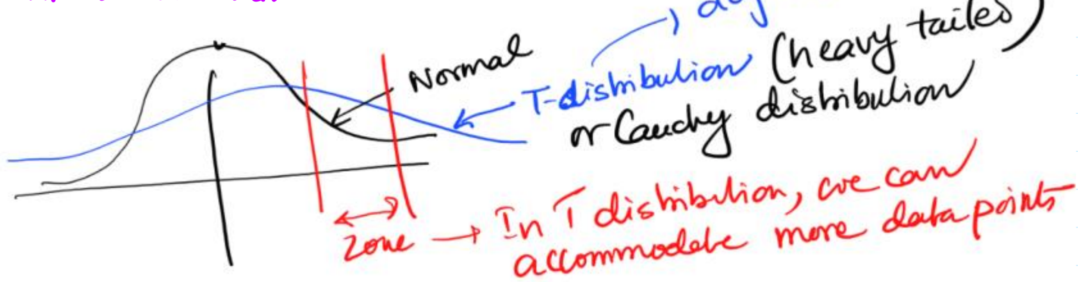
See, the problem here is that when we move from a higher dimensional to lower dimensional space, we struggle to find the same data in a similar region of the graph

↓

It's like I used a shrink-inator on my house and expected to find my toaster in the kitchen but couldn't.

Can we use a similar gaussian distribution to estimate probability q_{ij}?

WE CAN.



As we go up in dimensionality, a MUCH larger number of datapoints are in the region AWAY from the zero point.

Take a look. Assume $r_2 = 2$ $r_1 = 1$

(3D) volume $= \frac{4}{3}\pi r^3$ ⇒ change in vol $= \frac{\frac{4}{3}\pi(r_2^3 - r_1^3)}{\frac{4}{3}\pi(r_1^3)} = \frac{(2^3 - 1^3)}{1^3} = 7$

area $= 4\pi r^2$ ⇒ change in area $= \frac{4\pi(r_2^2 - r_1^2)}{4\pi(r_1^2)} = \frac{(2^2 - 1^2)}{(1^2)} = 3$

And that's why we adopt the t-dist. It's more compact.

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

$$q_{ij} = \frac{(1 + |y_i - y_j|^2)^{-1}}{\sum(1 + |y_k - y_i|^2)^{-1}}$$

P_{ij} describes similarity b/w X_i + X_j in the higher dimensions.
q_{ij} describes the same but in lower dimensions.

Now,
as per the t-sne objective, P_{ij} + q_{ij} are supposed to be close to each other.
↓
Capture closeness using KL Divergence!

$$\text{Loss} = KL(p\|q) = \sum_{i \neq j} P_{ij} \log\left(\frac{P_{ij}}{q_{ij}}\right)$$

We find our y_i + y_j by minimizing this loss (typically using GD)

Perplexity Hyperparameter.

To choose the gaussian in the high dimensional space, we need a σ.

This is defined by:
$$\text{perp}(p) = 2^{-\sum p(x) \log_2 p(x)}$$

→ this essentially reflects our opinion of the data
→ value is b/w 5 + 50.

When the perplexity is set, the σ corresponding to it is chosen by the algo.

Note: CROWDING PROBLEM!
↓
cuz we decrease the no. of dimensions,
the things that can be represented in this space also ↓
↓
so it may not feel entirely right in the lower dimension.

Namita Achyuthan