<T-SNE Visualization>

고차원 데이터를 저차원으로 시각화 (2D or 3D)
↳ 이때 원래 data의 local structure 및 Manifold를 유지할것.

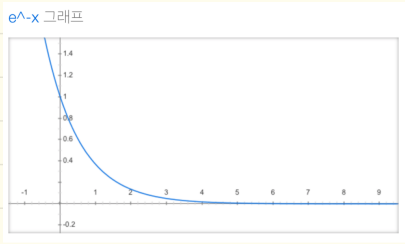
Manifold란? 국소적으로 유클리드 공간과 닮은 위상공간

- SNE (Stochastic Neighbor Embedding)

가까운 점에 대한 확률은 높고 먼 점에 대한 확률은 낮게

점 x_i 에대한 점 x_j 의 조건부 확률은 다음과 같이 표현됨.

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



↓
σ 를 어떻게 적용할 것인가?

Perplexity : # of nearest neighbors 라고 생각하면 된다고 함

옆의 그래프에서 볼수있듯이 거리가 가까울수록 커지고 멀수록 낮은수

x_i, x_j 가 고차원 상이면 y_i, y_j 는 저차원상의 점.

저차원상의 조건부 확률로 식으로 표현하면

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

$\Rightarrow$ 두 확률분포의 KL Divergence를 Minimize 하면 되는건가?

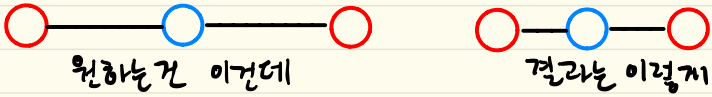
↓

하지만 KL Div. 는 Symmetric 하지않은데?

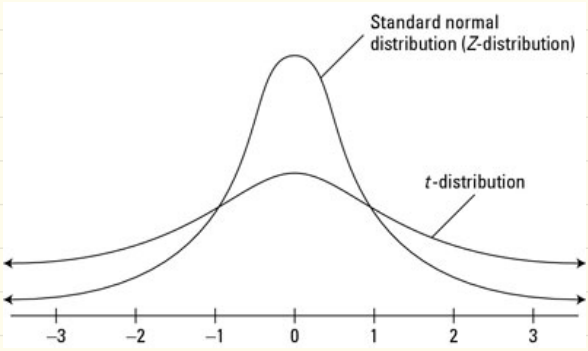
$p_{i|j} \neq p_{j|i}$ 니까 p_{ij}를 아예 $\frac{p_{j|i} + p_{i|j}}{2N}$ 로 정의하자

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

근데 해봐도 잘되거나 않았음. 클러스터끼리 너무 붙는 Crowding Problem발생



q_{ij}에 Gaussian Distribution 대신 Student-t distribution을 쓰자



Student's **t-distribution** has the probability density function given by

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

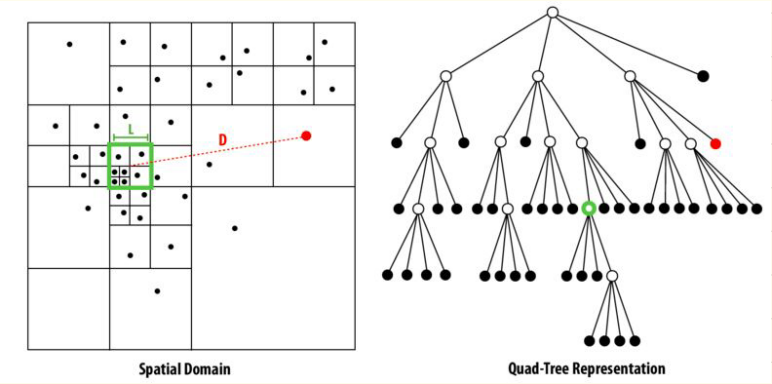
where ν is the number of *degrees of freedom* and Γ is the gamma function. This may also be written as

$$f(t) = \frac{1}{\sqrt{\nu}\,\mathrm{B}(\frac{1}{2},\frac{\nu}{2})}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

이렇게 되면 거리가 먼(확률이 낮은) 부분의 가중치가 커져서 Crowding problem을 해결할 수 있음.

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

데이터가 너무 많으면 연산면에서 너무 오래걸리지 않을까?



그래서 Barnes Hut Algorithm을 적용하여 일정크기 이하의 거리를 갖는 클러스터를 묶고 그 지점 이하 트리는 prune 해서 연산속도를 올림.

↳ Barnes Hut t-SNE $O(N^2) \rightarrow O(N \log N)$