$P(A|B) = \frac{P(A \cap B)}{P(B)}$

#1 Conditional Probability:

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

σ_i is the variance of the Gaussian that centered on x_i

Euclidean Distance

Coverting the shortest distance between the points into probability of similarity of points.

#2. For low-dimensional counterpars y_i and y_j of the high dimension datapoints x_i & x_j; Compute similar conditional probability:

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

Logically, the conditional probability $P_{j|i}$ & $q_{j|i}$ must be equal for a perfect representation of the similarity of data points in the different dimensional space.
So, SNE attempts to minimize this difference of conditional p.

#3. Minimize the difference of two p: minimize the distance function.
KL divergence between two

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

cost function.

#4 Variance σ_i of the student's t-distribution that is centered over each high-dimensional datapoints X_i.

t-SNE perform binary search for σ_i that produces P_i with fixed perplexity prep$(P_i) = 2^{H(P_i)}$, where $H(P_i)$ is Shannon entropy of P_i:

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

The perplexity can be interpreted as a smooth measure of effective number of neighbors. Typical values are 5~50.