T-SNE: T-distributed stochastic neighbourhood embedding

Reading material:

https://www.dailydoseofds.com/formulating-and-implementing-the-t-sne-algorithm-from-scratch/

Kevin Murphy’s book chapter: Section 20.4.10 in Probabilistic ML book.

Demos:

- https://projector.tensorflow.org/
- https://distill.pub/2016/misread-tsne/

T-SNE

- Another data projection method, specifically designed for visualizing high dimensional data in two dimensions.
- Preserves local similarities and clusters better than PCA

- Creates non-linear projection

First, SNE

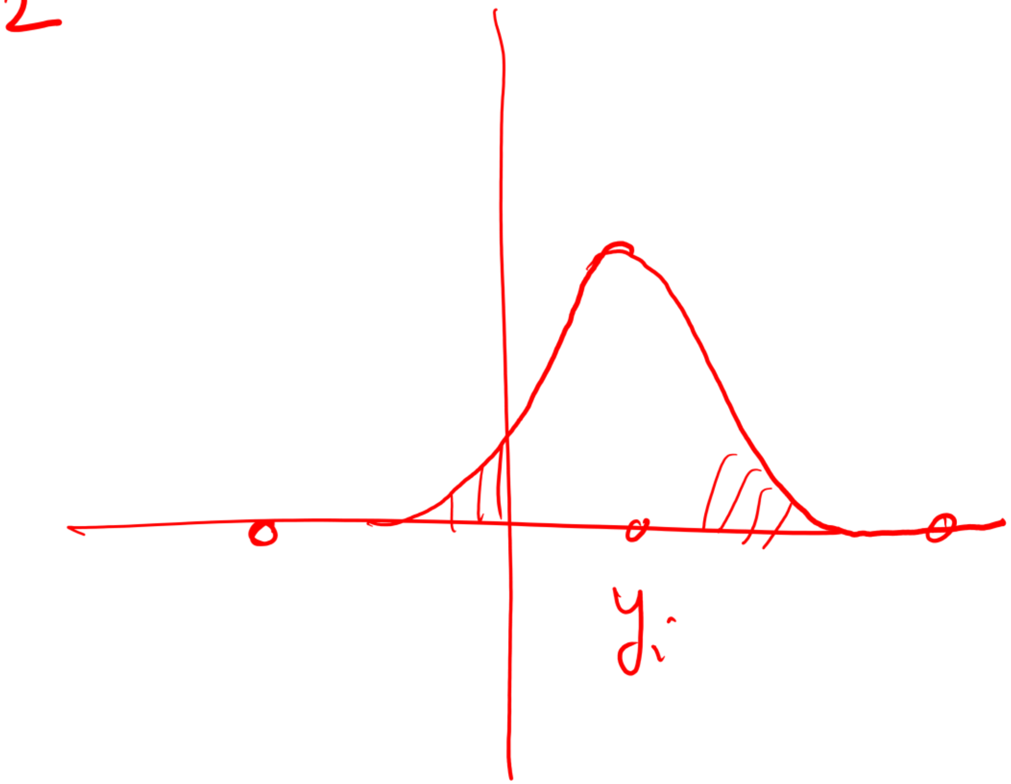
- Described here.
  - https://www.dailydoseofds.com/formulating-and-implementing-the-t-sne-algorithm-from-scratch

Limitations of SNE

- A fundamental problem with SNE and many other embedding techniques is that they tend to squeeze points that are relatively far away in the high dimensional space close together in the low dimensional (usually 2d) embedding space; this is called the crowding problem, and arises due to the use of squared errors (or Gaussian probabilities).

$$q(j|i) \propto e^{-\|y_i - y_j\|^2}$$

$$\underline{Z}$$



T-SNE

- Use a probability distribution in latent space that has heavier tails, which eliminates the unwanted attractive forces between points that are relatively far in the high dimensional space.

Student-t distribution - with one degree of freedom.

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

20.4.10.4 Choosing the length scale

An important parameter in t-SNE is the local bandwidth σ_i^2. This is usually chosen so that P_i has a perplexity chosen by the user.[7] This can be interpreted as a smooth measure of the effective number of neighbors.

Unfortunately, the results of t-SNE can be quite sensitive to the perplexity parameter, so it is wise to run the algorithm with many different values. This is illustrated in Figure 20.42. The input data is 2d, so there is no distortion generating by mapping to a 2d latent space. If the perplexity is too small, the method tends to find structure within each cluster which is not truly present. At perplexity 30 (the default for scikit-learn), the clusters seem equi-distant in embedding space, even though some are closer than others in the data space. Many other caveats in interpreting t-SNE plots can be found in [WVJ16].

[7] The perplexity is defined to be $2^{\mathbb{H}(P_i)}$, where $\mathbb{H}(P_i) = -\sum_j p_{j|i} \log_2 p_{j|i}$ is the entropy; see Section 6.1.5 for details. A big radius around each point (large value of σ_i) will result in a high entropy, and thus high perplexity.