PCA Wrap-Up

Projection perspective

$$\tilde{x}_i = Bz_i \qquad B \text{ again orthonormal}$$

Want to minimize reconstruction error

$$J_M \overset{\text{def}}{=} \frac{1}{N}\sum_{i=1}^{N} \|x_i - \tilde{x}_i\|^2$$

What are the optimal coordinates z_i for x_i w.r.t. B?

$$\frac{\partial J_M}{\partial z_{ji}} = \frac{\partial J_M}{\partial \tilde{x}_i}\frac{\partial \tilde{x}_i}{\partial z_{ji}}$$

$$\frac{\partial J_M}{\partial \tilde{x}_i} = -\frac{2}{N}(x - \tilde{x}_i)^T \in \mathbb{R}^{1\times D}$$

$$\frac{\partial \tilde{x}_i}{\partial z_{ji}} = \frac{\partial}{\partial z_{ji}}\left(\sum_{m=1}^{M} z_{mi} b_m\right) = b_j$$

So

$$\frac{\partial J_M}{\partial z_{ji}} = -\frac{2}{N}\left(x_i - \tilde{x}_i\right)^T b_j$$

$$= -\frac{2}{N}\left(x_i - \sum_{m=1}^{M} z_{mi}\, b_m\right)^T b_j \qquad \left(b_m^T \cdot b_j = \emptyset \text{ if } m \neq j\right)$$

$$= -\frac{2}{N}\left(x_i^T b_j - z_{ji}\, b_j^T b_j\right) = -\frac{2}{N}\left(x_i^T b_j - z_{ji}\right)$$

Set to $\emptyset \rightarrow z_{ji} = b_j^T x_i$

A similar argument for choice of B (basis) can be made, yielding again the M largest Eigenvectors (see reading!)

Stochastic Neighbor Embedding (SNE)

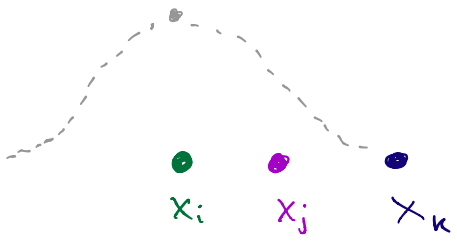
Aim: $X \longrightarrow y$

(X: (very) high dim; y: (very) low dim)

Define a conditional probability that encodes similarity

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

This is in high dim space



Similarly in the map y

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

Ideally: $p_{j|i} \cong q_{j|i} \quad \forall i, j$

Formalize this with KL-Divergence (KL)

$$KL(q||p) \doteq \sum_x q(x) \log \frac{q(x)}{p(x)}$$

"How different is the distribution q from p?"

Properties:

(1) $KL(q||p) \geq 0$

(2) if $KL(q|p) = 0 \rightarrow q = p$

(3) $KL(q||p) \neq KL(p||q)$

Cost function:

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

Q_i: Same, but in M.

P_i: Conditional distribution over all other points j given i in D

to place points, find y to minimize C

using gradient descent $\nabla_y C$.

Symmetric SNE

Instead of Conditionals $p_{i|j}$, $q_{i|j}$ define

Joint distributions p_{ij}, q_{ij}

$$q_{ij} = \frac{\exp\{-\|y_i - y_j\|^2\}}{\sum_{k \neq l} \exp\{-\|y_k - y_l\|^2\}}$$

Normalizes dist by all pairs

$$\rightarrow q_{ij} = q_{ji}$$

For high-dim space, outliers pose a challenge because denominator will be large $\rightarrow p_{ij}$ small $\rightarrow y_{ij}$ unimportant.

Instead:

$$p_{ij} \stackrel{\text{def}}{=} \frac{p_{j|i} + p_{i|j}}{2n}$$

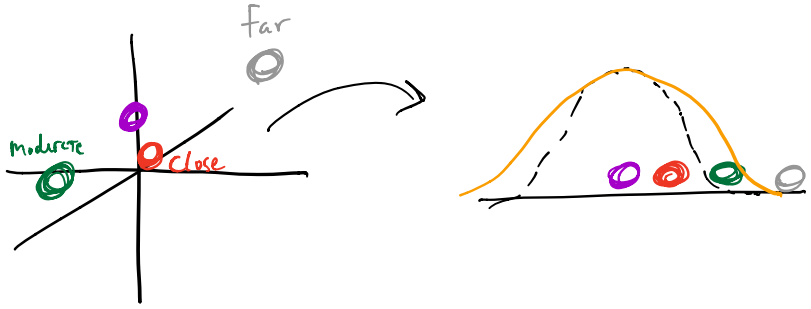
$$\rightarrow \sum_j p_{ij} \geq \frac{1}{2n} \quad \forall x_i$$

Yields a nicer ∇_y

$$\nabla_{y_i} C_{\text{symmetric}} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)$$

The Crowding problem

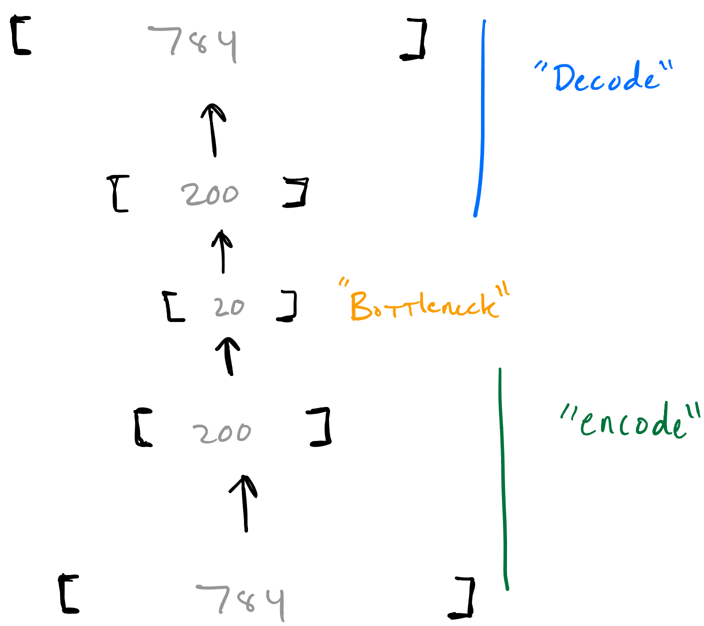
Not enough space in lower dims



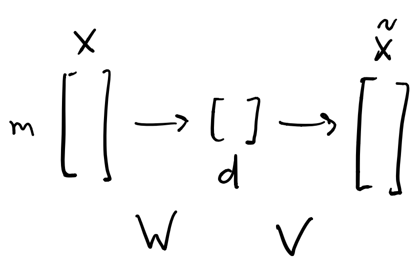
In t-SNE we model the joint probability q_{ij} using a Student t-distribution which has heavier tails → moderate distance in X → big distance in Y is ok -- does not force "moderate" distances in X to yield small distances in Y

Auto-encoders

Design a network that consumes x, and then re-constructs it.



Simplest version



$$\mathcal{L}(x, \tilde{x}) = \|x - \tilde{x}\|^2$$

Note: This is just linear dim reduction, and should look familiar

$$Wx \to z \; ; \quad \tilde{x} = Vz$$

$d \times m \quad m \times 1 \quad d \times 1 \qquad m \times d$

More next time!