

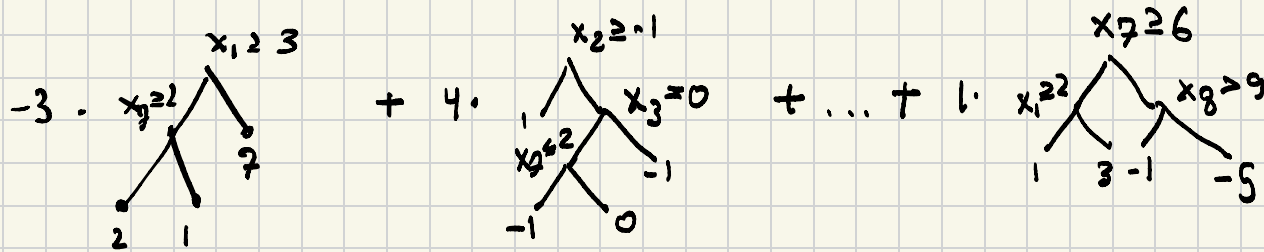
Tuesday, March 31

- Midterm 4/23
- Exam/project prep 4/21
- James Enown (USC)
 - ↳ Estimating Shapley values (my research interest)
 - ↳ Thursday 12pm in Roberts North 15

Today

- Autoencoders!!
😊

Gradient Boosting: $F_t = \alpha_1 f_1 + \alpha_2 f_2 + \dots + \alpha_t f_t$



Gradient boosting was SOTA on tabular data from 2014-2025

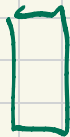
TabPFN ^{and friends} (Transformer for tabular data) is now SOTA

↳ in context learning

train features

train labels

test features



Learning

↳ supervised

- classification
- regression

↳ unsupervised

- autoencoders
- clustering

↳ semi-supervised

- reinforcement learning

Question: When our data doesn't have labels, what can we do with it?

- observations
- images
- text

Answer: Use the data as its own label

Autoencoder

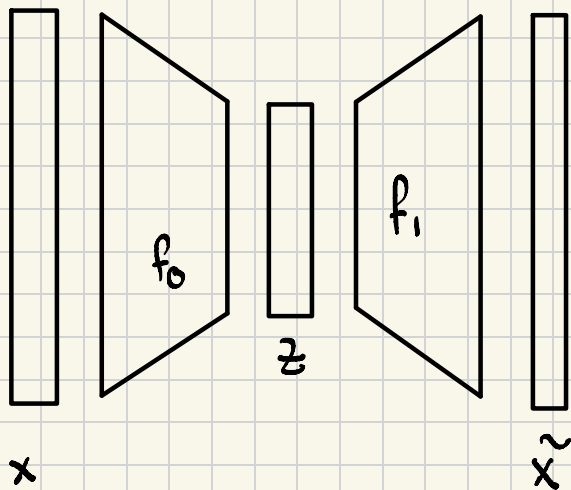
Encoder: $f_0: \mathbb{R}^d \rightarrow \mathbb{R}^k$

Decoder: $f_1: \mathbb{R}^k \rightarrow \mathbb{R}^d$

Input: $x \in \mathbb{R}^d$

Embedding: $z = f_0(x)$

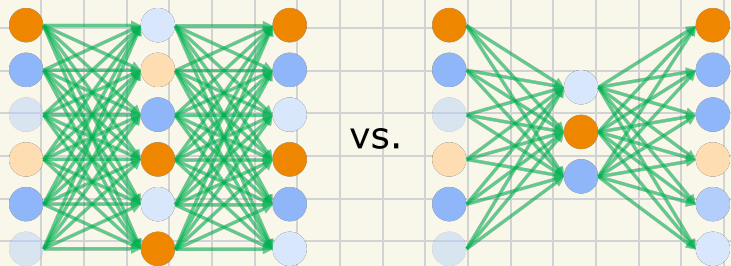
Output: $\tilde{x} = f_1(z) = f_1(f_0(x))$



Reconstruction Loss:

$$\mathcal{L}_{\text{recon}} = \frac{1}{n} \sum_{i=1}^n \|x^{(i)} - \tilde{x}^{(i)}\|_2^2$$

Q: How should we choose k ?



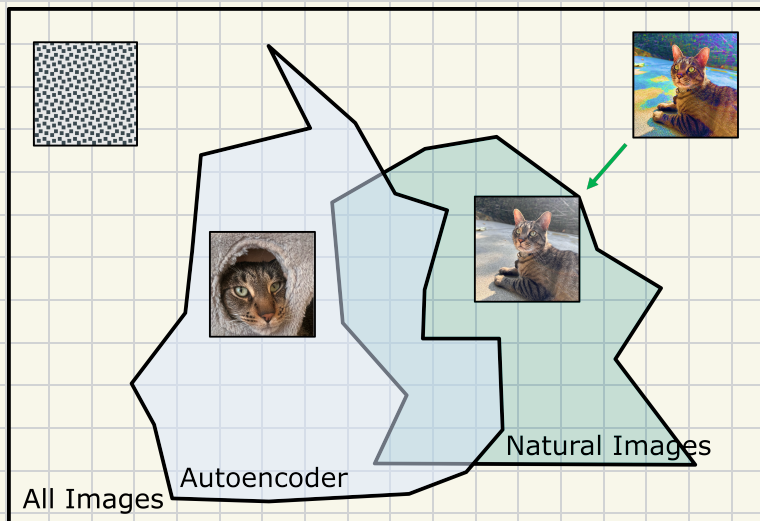
Autoencoder applications:

- data compression
- denoising
- inpainting
- representation learning

↑ depend on k rather than d

Manifold Perspective

Since $k \leq d$, we lose info.
But this is ok, because we're only interested in a manifold (subset) of \mathbb{R}^d



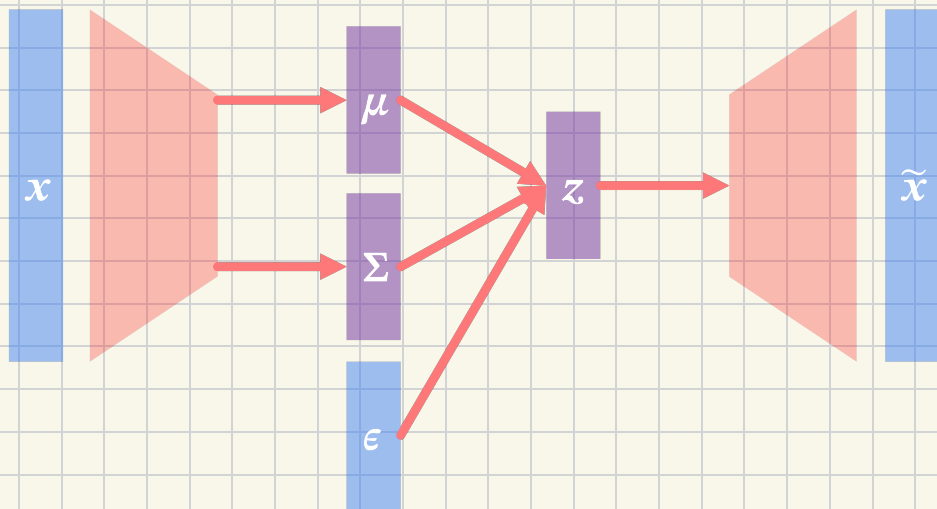
Variational Autoencoders

Goal: Evenly distribute embeddings $z = f_0(x) \in \mathbb{R}^k$, i.e.,

$z \sim \mathcal{N}(\mu_x, \Sigma_x)$ where

mean $\mu_x \in \mathbb{R}^k$ depends on x

variance $\Sigma_x \in \mathbb{R}^{k \times k}$ depends on x



$$f_0: \mathbb{R}^d \rightarrow \mathbb{R}^k \times \mathbb{R}^{k \times k}$$

$$\epsilon \sim \mathcal{N}(0, \mathbf{I}) \in \mathbb{R}^k$$

$$\mu_x, \Sigma_x = f_0(x)$$

$$z = \mu_x + \Sigma_x^{1/2} \epsilon$$

Variational Loss

tension!

reconstruction error + distribution distance

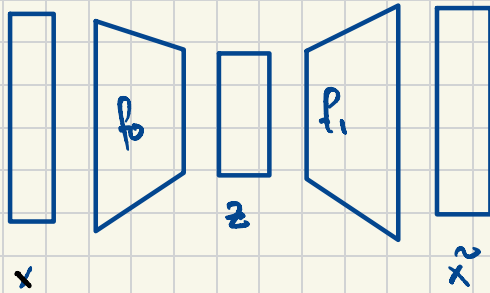
$$\frac{1}{n} \sum_{i=1}^n \|x^{(i)} - \tilde{x}^{(i)}\|_2^2 + \alpha \frac{1}{n} \sum_{i=1}^n D_{KL}(\mathcal{N}(\mu_{x^{(i)}}, \Sigma_{x^{(i)}}) \parallel \mathcal{N}(0, I))$$

↑
Kullback-Leibler divergence measures
"distance" between distributions

$$D_{KL}(P, Q) = \mathbb{E}_{z \sim P} \log \frac{P(z)}{Q(z)}$$

Thursday, April 2

Autoencoder



Principal Component Analysis

"linear regression" of autoencoders

$$f_0 = W_0 \in \mathbb{R}^{d \times k}$$

$$f_1 = W_1 \in \mathbb{R}^{k \times d}$$

A diagram illustrating the multiplication of a row vector x^T and a column vector w_0 to produce a scalar z . The row vector x^T is shown in a box on the left, followed by a vertical box labeled w_0 , and an equals sign followed by the scalar z .

A diagram illustrating the multiplication of a row vector x^T , a column vector w_0 , and a column vector w_1 to produce a row vector \tilde{x} . The row vector x^T is shown in a box on the left, followed by a vertical box labeled w_0 , a horizontal box labeled w_1 , an equals sign, and the row vector \tilde{x} .

$$\begin{aligned} \text{Goal: } & \min \sum_{i=1}^n \|x^{(i)} - \tilde{x}^{(i)}\|_2^2 \\ & = \min \sum_{i=1}^n \|x^{(i)} - x^{(i)T} W_0 W_1\|_2^2 \end{aligned}$$

A diagram illustrating the matrix multiplication $X W_0 W_1 = \tilde{X}$. A large vertical box labeled X is followed by a smaller vertical box labeled W_0 , then another smaller vertical box labeled W_1 , an equals sign, and a large vertical box labeled \tilde{X} . Below the X box is the dimension $n \times d$, and below the \tilde{X} box is the dimension $n \times d$.

$$= \min \|X - \tilde{X}\|_F^2$$

Sum of squared entries

Singular Value Decomposition

$$X \in \mathbb{R}^{n \times d} \quad n \geq d \quad \text{wlog}$$

$$X = U \Sigma V^T$$

$$= \begin{array}{c} U \\ \left[\begin{array}{c|c|c} | & | & | \\ u_1 & u_2 & \dots & u_d \\ | & | & & | \end{array} \right] \\ n \times d \end{array} \begin{array}{c} \Sigma \\ \left[\begin{array}{c|c} \sigma_1 & 0 \\ \sigma_2 & \\ \vdots & \\ 0 & \sigma_d \end{array} \right] \\ d \times d \end{array} \begin{array}{c} V^T \\ \left[\begin{array}{c} -v_1^T \\ -v_2^T \\ \vdots \\ -v_d^T \end{array} \right] \\ d \times d \end{array}$$

outer product
view

$$= \sum_{i=1}^d \sigma_i u_i v_i^T$$

left singular vectors $u_1, \dots, u_d \in \mathbb{R}^n$ orthonormal
right singular vectors $v_1, \dots, v_d \in \mathbb{R}^d$ orthonormal
singular values $\sigma_1 \geq \dots \geq \sigma_d \geq 0$

rank k matrix iff

$$k = |\{ \sigma_i : \sigma_i > 0 \}|$$

rank k version of X is

$$X_k = \sum_{i=1}^k \sigma_i u_i v_i^T$$

Eckart-Young-Mirsky

$$X_k = \underset{\text{rank-}k \tilde{X}}{\operatorname{argmin}} \|X - \tilde{X}\|_F^2$$

Tools

(A) Frobenius norm and trace: $\|X\|_F^2 = \operatorname{tr}(X^T X)$

sum of squared entries

sum of diagonal entries

$$\|X\|_F^2 = \sum_{i=1}^n \sum_{j=1}^d [X]_{ij}^2 = \sum_{i=1}^n \| [X]_{i,:} \|_2^2 = \operatorname{tr}(X^T X)$$

(B) Cyclic property: $\operatorname{tr}(AB) = \operatorname{tr}(BA)$ where $A \in \mathbb{R}^{n \times d}$, $B \in \mathbb{R}^{d \times n}$

$$\begin{aligned} \operatorname{tr}(AB) &= \sum_{i=1}^n [AB]_{i,i} = \sum_{i=1}^n \sum_{j=1}^d [A]_{i,j} [B]_{j,i} \\ &= \sum_{j=1}^d \sum_{i=1}^n [B]_{j,i} [A]_{i,j} = \sum_{j=1}^d [BA]_{j,j} = \operatorname{tr}(BA) \end{aligned}$$