Lecture 4 21/10/2024

AUTOENCODERS

PCA
MDS } Classical MDS $\Delta_{il}^2 = \|x^i - x^l\|^2$

↳ ISOMAP $\Delta_{il} \to$ geodesic distance

Gram matrix $\xrightarrow[\text{trick}]{\text{kernel}}$ kernel PCA

Not classical MDS → Diffusion Maps
→ Sketch Map

MDS $L = f(\Delta, \Theta)$

Probabilistic Embeddings

$$\begin{array}{ccc} \Delta & \rightarrow & P(N) \\ \downarrow{\scriptstyle \text{MDS}} & & \downarrow \\ \Theta & \leftarrow & Q(N) \end{array}$$

$\Delta \to P(N)$ Original space

Θ (y) $\leftarrow Q(N)$ Projected Space

Autoencoders

latent space

$$x \to \boxed{\text{ENCODER}} \to y \to \boxed{\text{DECODER}} \to \tilde{x}$$

$q_\theta(x) = y$ $\qquad$ $p_\phi(y) = \tilde{x}$

$D \;\cdots\cdots\; d \;\cdots\cdots\; D$

<u>q_θ & p_ϕ are linear functions</u>

$$\left.\begin{array}{ll} q_\theta : W & y^i = W x^i \\ p_\phi : U & \tilde{x}^i = U y^i \end{array}\right\}$$

$$\tilde{x}^i = \underbrace{U W}_{\text{mapping function}} x^i$$

$$L = \sum_i \| x^i - \tilde{x}^i \|^2 \quad \text{Reconstruction error}$$

$$\min_{U \in \mathbb{R}^{D,d},\; W \in \mathbb{R}^{d,D}} L = \sum_i \| x^i - U W x^i \|^2 \quad \text{with respect}$$

(A) $\operatorname*{argmin}_{U,W} \sum_i \| x^i - U W x^i \|^2$ implies

$$W = U^T \;;\; U^T U = \mathbb{I}$$

(B) minimizing L is equivalent to maximize the Tr of the covariance matrix on the latent space

(A) $\tilde{x}^i = UWx^i$

$\therefore V \quad \boxed{UWx^i = V\alpha^i}$

$$\ell = \|x - V\alpha\|^2 = \|x\|^2 + \alpha^T \underbrace{V^T V}_{\mathbb{I}} \alpha - 2\alpha^T V^T x$$

$$= \|x\|^2 + \|\alpha\|^2 - 2\alpha^T (V^T x)$$

$\alpha^T\alpha = \|\alpha\|^2$

$$\frac{\partial \ell}{\partial \alpha} = 2\alpha - 2(V^T x) \xrightarrow{(\min)} \underline{V^T x = \alpha}$$

$$\operatorname{argmin} \ell \quad \|x - V\alpha\|^2 = \|x - V(V^T x)\|^2 \quad \forall x \in x^i$$

$$\boxed{V = U \,;\; V^T = W \,;\; U^T U = \mathbb{I}}$$

$$L = \sum_i \|x^i - UWx^i\|^2 = \sum_i \|x^i - UU^T x^i\|^2$$

(B) $$\sum_i \|x^i\|^2 - 2x^{iT} UU^T x^i + x^{iT} U \underbrace{U^T U}_{\mathbb{I}} U^T x^i$$

$$= \sum_i \|x^i\|^2 - 2x^{iT} UU^T x^i + x^{iT} UU^T x^i =$$

$$\sum_i \|x^i\|^2 - \underbrace{x^{iT} U}_{y^{iT}} \underbrace{U^T x^i}_{y^i} = \sum_i \|x^i\|^2 - \sum_i y^{iT} y^i$$

$$= \underbrace{\sum_i \|x^i\|^2}_{\text{constant}} - \underbrace{\operatorname{Tr}(y^T y)}_{\hookrightarrow \text{PCA}}$$

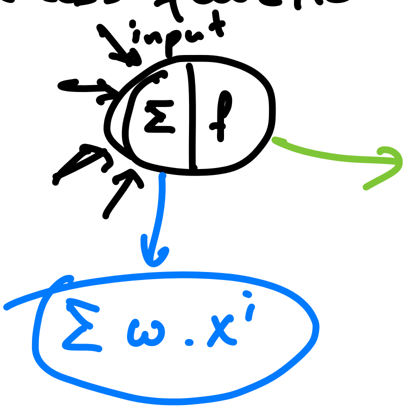
Made with Goodnotes

Three ways of deriving PCA:

1. Diagonalize Covariance matrix
2. Reproducing the euclidean distances in the original space by minimizing the Frobenious Norm of the difference $\|\Delta - \Theta\|^2$
3. Linear Autoencoders

$p_\phi : q_\theta$ are not linear $\rightarrow$ Neural Networks

Find functions (not-linear) that minimize a loss function



Non-linear function (Activation)

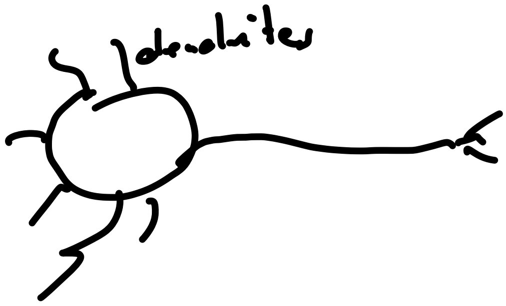
(x_1^i, x_2^i) input

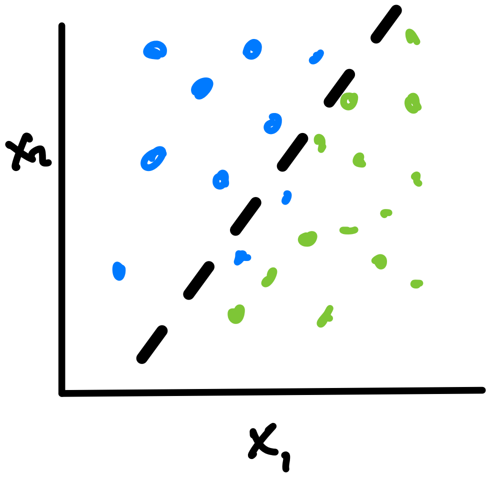
$\downarrow$

$\tanh(\underbrace{\omega_1 x_1} + \underbrace{\omega_2 x_2} + \underbrace{\omega_3}_{\text{bias}})$

$\downarrow$

output $\in [-1, 1]$

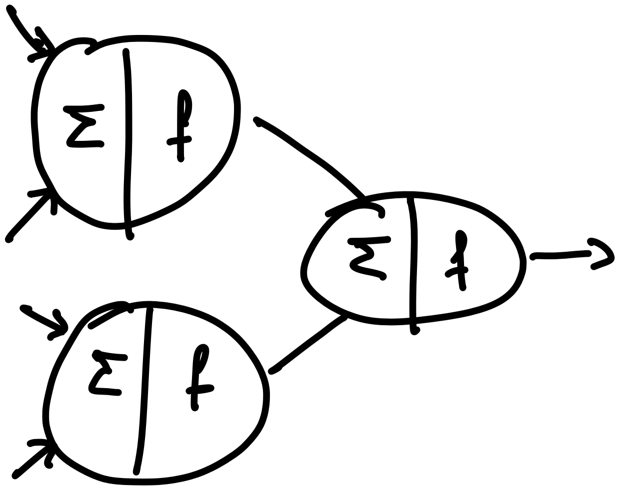


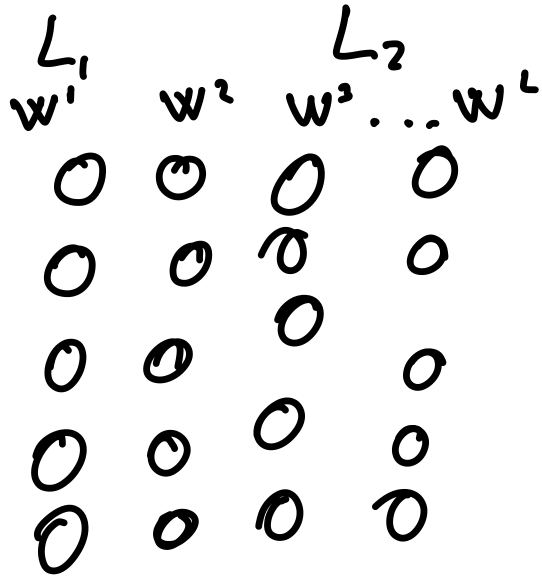


blue → -1

green → 1

$$L = \sum_i \| y^i - \tanh(\underbrace{w_1 x_1 + w_2 x_2 + w_3}) \|^2$$





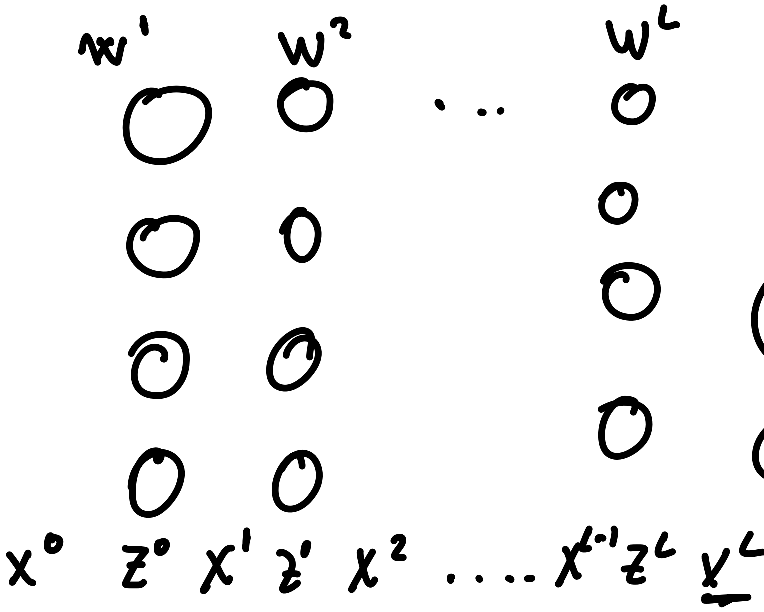
$X^0 Z^0 \to X^1$

$$[w]_{t+1} = [w]_t - \alpha \underbrace{\left[\frac{\partial L}{\partial w}\right]_t}$$

$$\nabla_{w_l} L = (X^{l+1})^T \cdot \Delta^{l+1}$$

$$\Delta^{l+1} = \underbrace{\nabla_{X^{l+1}} L} \odot f'^{\,l+1}(Z^{l+1})$$

$$\nabla_{X^l} L = \Delta^{l+1} \cdot (W^{l+1})^T$$



① Compute $\nabla_{x_L} L$

② $\nabla_{W^L} L$ & Δ^L

③ $\nabla_{x^{L-1}} L$

④ $\nabla_{W^{L-1}} L$ & Δ^{L-1}

...

Stochastic Gradient Descent

$$\nabla = \frac{1}{N} \sum_i^N \delta_i \xrightarrow{n} \nabla \sim \tilde{\nabla} = \frac{1}{n} \sum^n \delta_i$$

GD

$t = \emptyset$

While not converged
- $t{+}{+}$
- $\nabla_{W^L} L = \emptyset$
- for i in N
  - accumulate ∇
- update weights

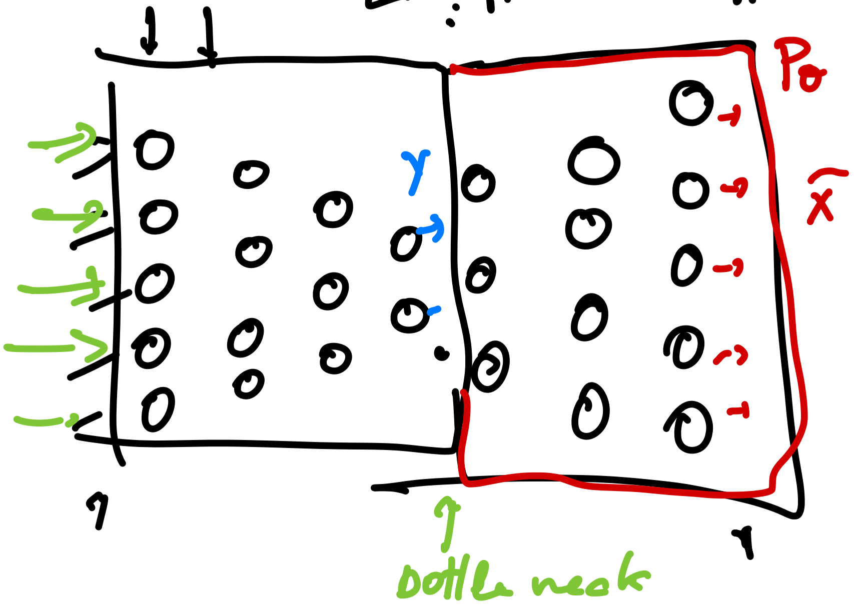
SGD

$t = \emptyset$

While not converged
- $t{+}{+}$
- shuffle my data
- for j in mini batches
  - $\nabla_{W^L} L = 0$
  - for i in N/m
    - accumulate ∇
  - update weights

$$X \to \underset{q_\theta}{\boxed{\text{ENCO.}}} \to y \to \underset{P_\phi}{\boxed{\text{DECO.}}} \to \tilde{X}$$

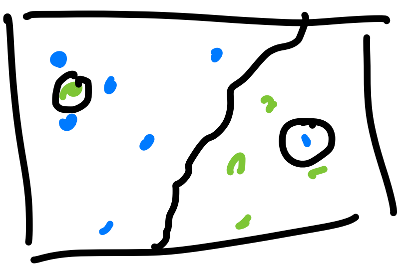
$$L = \sum \| x^i - \tilde{x}^i \|^2$$



$$L = \sum \| x^i - P_\phi (q_\theta (x^i)) \|^2$$

$$y^i = q_\theta (x^i)$$

$\theta, \phi \Rightarrow$ parameters that minimize



$$L = \| x^i - \tilde{x}^i \|^2 + \lambda \sum_k \| \omega_k \|^p$$

$p = 1 \to$ LASSO reg. $\qquad p_1$ & p_2

$p = 2 \to$ Ridge reg. $\qquad + \lambda \left(\alpha \| \omega_k \| + (1-\alpha) \| \omega_k \|^2 \right)$

Elastic nets

VARIATIONAL AUTOENCODERS

p_ϕ & q_θ $\longrightarrow$ VARIATIONAL STOCHASTIC MAPPING

TRADITIONAL

$q_\theta(x) = y$

$p_\phi(y) = \tilde{x}$

$q_\theta(y|x)$

$p_\phi(\tilde{x}|y)$

Max. Likelihood for the loss

prior to $p(y) \sim N(\phi, 1)$

$KL(p_\phi(x|y) \mid \underbrace{N(\phi, 1)})$