# Machine Learning: Matrix factorization

Hao Tang

March 25, 2024

# K-means

# ~~K-means~~ Lloyd's algorithm

- Assign points to their nearest centroids

$$\gamma_i = \operatorname*{argmin}_{k=1,\dots,K} \|x_i - \mu_k\|_2^2 \qquad \text{for } i = 1, \dots, n \tag{1}$$

- Update centroids based on the assignment.

$$\mu_k = \frac{\sum_{i=1}^n \mathbb{1}_{\gamma_i = k} x_i}{\sum_{i=1}^n \mathbb{1}_{\gamma_i = k}} \qquad \text{for } k = 1, \dots, K \tag{2}$$

# K-means

# K-means

- The objective in the K-means lecture is

$$\sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{1}_{\gamma_i = k} \|x_i - \mu_k\|_2^2. \tag{3}$$

# K-means

- The objective in the K-means lecture is

$$\sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{1}_{\gamma_i=k}\|x_i-\mu_k\|_2^2. \tag{3}$$

- The goal is to find $\mu_1, \dots, \mu_K$ and $\gamma_1, \dots, \gamma_n$ so as to minimize the objective.

- Lloyd's algorithm only finds a local minimal.

# K-means

- If we pack everything into vectors and matrices,

$$z_i = \begin{bmatrix} \mathbb{1}_{\gamma_i=1} & \mathbb{1}_{\gamma_i=2} & \dots & \mathbb{1}_{\gamma_i=K} \end{bmatrix} \qquad W = \begin{bmatrix} - & \mu_1 & - \\ - & \mu_2 & - \\ & \vdots & \\ - & \mu_K & - \end{bmatrix} \tag{4}$$

  we can write

$$\mathbb{1}_{\gamma_i=k} \|x_i - \mu_k\|_2^2 = \|x_i - z_i W\|_2^2. \tag{5}$$

# K-means

- The final objective[1] is

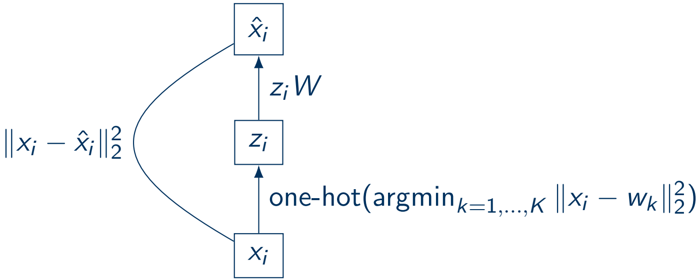
$$\min_{Z,W} \quad \|X - ZW\|_F^2 \tag{6}$$

$$\text{s.t.} \quad \sum_{k=1}^{K} z_{ik} = 1 \quad \text{for } i = 1, \ldots, n \tag{7}$$

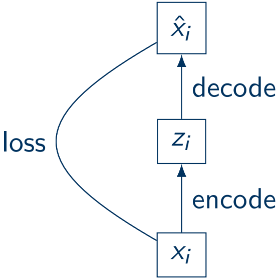
$$z_{ik} \in \{0, 1\} \quad \text{for } i = 1, \ldots, n \text{ and } k = 1, \ldots, K \tag{8}$$

[1] The Frobenius norm of $X$, written as $\|X\|_F$, is defined as the $L_2$ norm of the flattened matrix, or $\|\text{vec}(X)\|_2$.

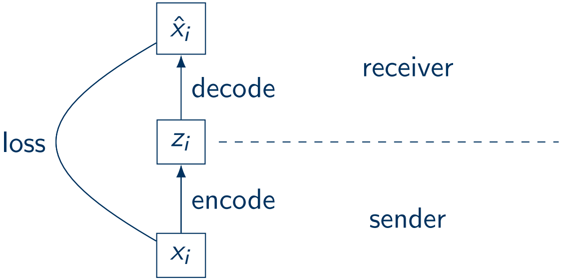
# K-means (a.k.a. vector quantization)



# Autoencoders



# Autoencoders



# Autoencoders

- A general autoencoder has the loss function

$$\|X - D(E(X))\|_F^2. \tag{9}$$

- The encoder $E$ and the decoder $D$ can be any function, including deep neural networks.

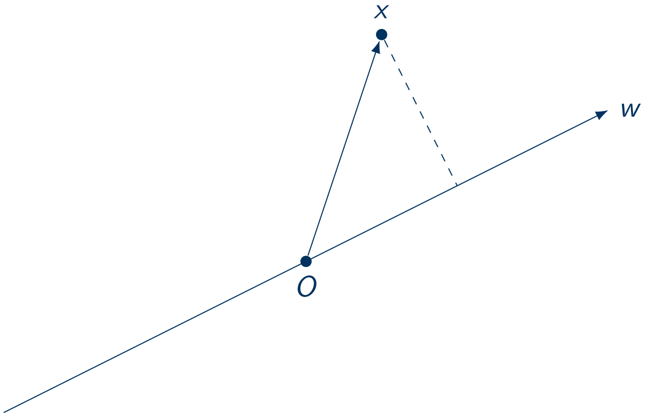
- When $E(x) = xW_1$ and $D(z) = zW_2$, we have

$$\|X - XW_1 W_2^\top\|_F^2. \tag{10}$$

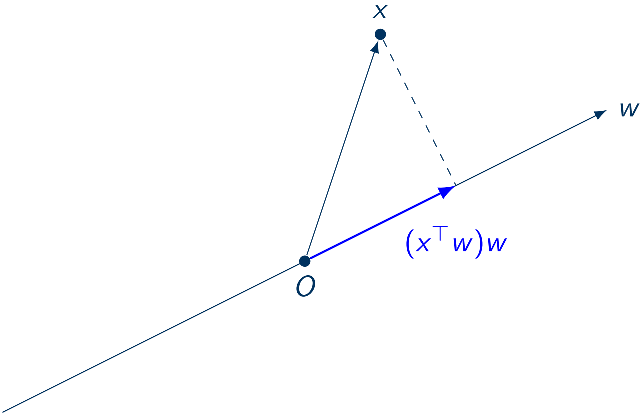
- When $E(x) = xW$ and $D(z) = zW^\top$, we have

$$\|X - XWW^\top\|_F^2. \tag{11}$$

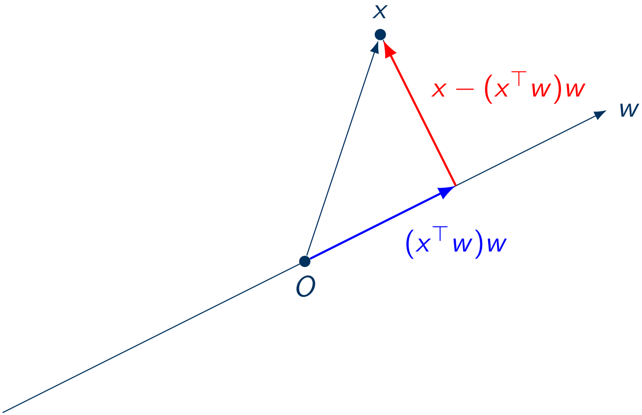
# PCA



# PCA



# PCA



# PCA

- Maximize spread (or variance)

$$\sum_{i=1}^{n} \|(x_i^\top w)w\|_2^2 = w^\top X^\top X w \tag{12}$$

- Minimize distance

$$\sum_{i=1}^{n} \|x_i - (x_i^\top w)w\|_2^2 = \|X - Xww^\top\|_F^2 \tag{13}$$

- Don't forget $\|w\|_2^2 = 1$.

# PCA

- The final objective is

$$\min_{W} \quad \|X - XWW^\top\|_F^2 \tag{14}$$

$$\text{s.t.} \quad W^\top W = I \tag{15}$$

# Singular value decomposition (SVD)

# Singular value decomposition (SVD)

- The singular value decomposition (SVD) of a matrix $X$ is $U\Sigma V^\top$, where $U^\top U = I$, $V^\top V = I$,

$$\Sigma = \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_d \end{bmatrix}, \tag{16}$$

and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$.

# Eckart-Young theorem

- Let $\Sigma_k = \text{diag}(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0)$ where $k \leq d$.
- The matrix $U\Sigma_k V^\top$ is the optimal solution to

$$\min_{\hat{X}} \quad \|X - \hat{X}\|_F^2 \tag{17}$$

$$\text{s.t.} \quad \text{rank}(\hat{X}) \leq k \tag{18}$$

# Eckart-Young theorem

- Let $\Sigma_k = \text{diag}(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0)$ where $k \leq d$.
- The matrix $U\Sigma_k V^\top$ is the optimal solution to

$$\min_{\hat{X}} \quad \|X - \hat{X}\|_F^2 \tag{17}$$

$$\text{s.t.} \quad \text{rank}(\hat{X}) \leq k \tag{18}$$

- The matrices $Z = U\Sigma_k$ and $W = V^\top$ are the optimal solution to

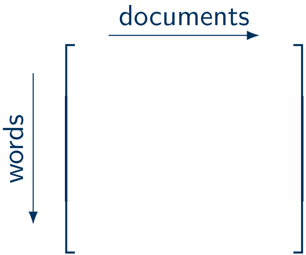
$$\min_{Z,W} \quad \|X - ZW\|_F^2 \tag{19}$$

$$\text{s.t.} \quad Z \in \mathbb{R}^{n \times k} \tag{20}$$

$$W \in \mathbb{R}^{k \times d} \tag{21}$$

# Latent semantic indexing

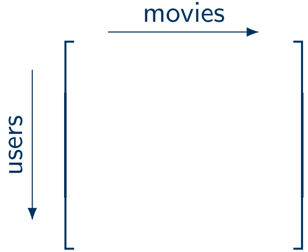
- Create a term-document matrix

$X =$ 


- Solve $\min_{Z,W} \|X - ZW\|_F^2$.
- The $Z$ matrix provides a vector for every word, and the $W$ matrix provides a vector for every document.

# Matrix completion

- Create a user-movie matrix

$X =$ 


- Solve $\min_{Z,W} \|X - ZW\|_F^2$.
- The reconstructed matrix $ZW$ provides a guess of the empty entries in $X$.

# Summary

- K-means = matrix factorization with assignment constraints
- Lloyd's algorithm = autoencoding with hard assignments
- PCA = linear autoencoder with encoder and decoder tied and orthogonality constraints
- SVD = low-rank matrix factorization

# Variants of autoencoders

- A regular autoencoder

$$\|X - D(E(X))\|_F^2. \tag{22}$$

- A denoising autoencoder

$$\|X - D(E(n(X)))\|_F^2, \tag{23}$$

where $n$ is a function that injects noise.

- A variational autoencoder

$$\mathbb{E}_{z\sim q(z|x)}[-\log p(x|z)] + \mathbb{E}_{z\sim q(z|x)}\left[\log \frac{q(z)}{p(z)}\right] \tag{24}$$