Lecture 19: PCA, SVD, Rayleigh quotient review

Nisha Chandramoorthy

November 2, 2023

Autoencoder decoder

$$(E^*, D^*) = \arg\min_{E,D} \sum_{i=1}^{m} \|x_i - D(E(x_i))\|^2 \tag{1}$$

- Posed as ERM problem.

Autoencoder decoder

$$(E^*, D^*) = \arg\min_{E,D} \sum_{i=1}^{m} \|x_i - D(E(x_i))\|^2 \quad (1)$$

- Posed as ERM problem.
- E is encoder, D is decoder.

Autoencoder decoder

$$(E^*, D^*) = \arg\min_{E,D} \sum_{i=1}^{m} \|x_i - D(E(x_i))\|^2 \tag{1}$$

- Posed as ERM problem.
- E is encoder, D is decoder.
- E maps x to z (latent space), D maps z to $\hat{x}$ (reconstruction).

Autoencoder decoder

$$(E^*, D^*) = \arg\min_{E,D} \sum_{i=1}^{m} \|x_i - D(E(x_i))\|^2 \tag{1}$$

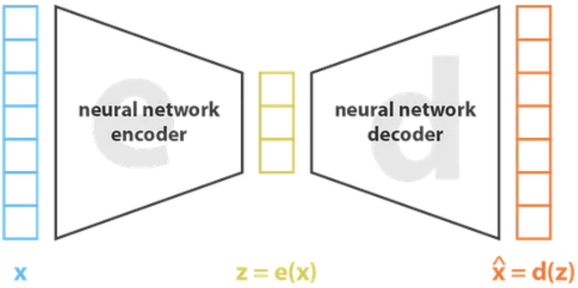
- Posed as ERM problem.
- E is encoder, D is decoder.
- E maps x to z (latent space), D maps z to $\hat{x}$ (reconstruction).
- Both parameterized as Neural Networks.

Variational autoencoders

- Probabilistic encoder and decoder.

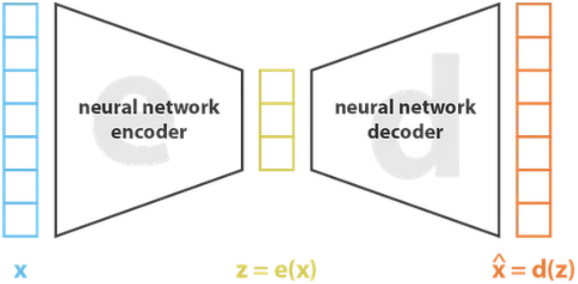
Variational autoencoders

- Probabilistic encoder and decoder.
- Encoder: $q(z|x)$, Decoder: $p(x|z)$



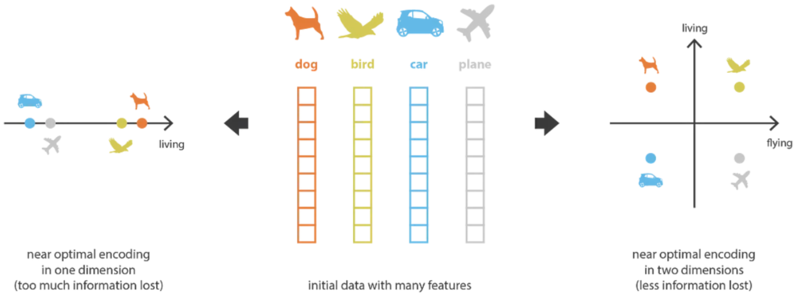
$$\text{loss} = \| x - \hat{x} \|^2 = \| x - d(z) \|^2 = \| x - d(e(x)) \|^2$$

- tends to overfit as a Generative model

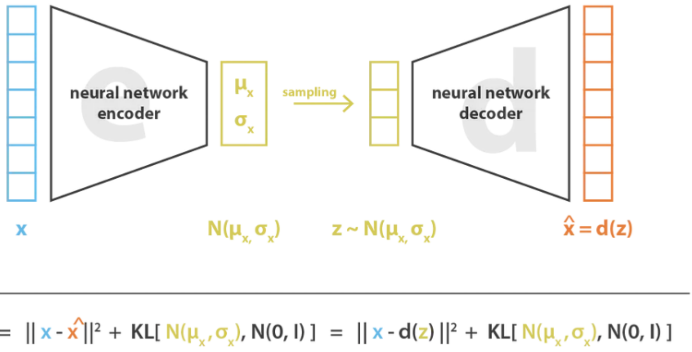


$$\text{loss} = \|x - \hat{x}\|^2 = \|x - d(z)\|^2 = \|x - d(e(x))\|^2$$

- tends to overfit as a Generative model
- VAE: uses VI to regularize the latent space.



Courtesy: https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73



Courtesy: https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73

PCA

- when E and D are linear $\rightarrow$ PCA.

PCA

- when E and D are linear $\to$ PCA.
- $E(x) = Wx$, $D(z) = W^\top z$.

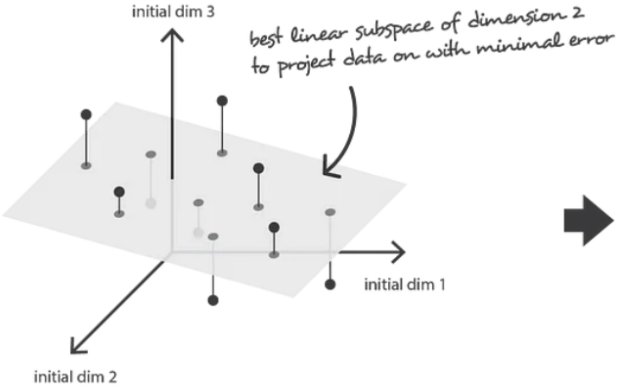
PCA

- when E and D are linear $\rightarrow$ PCA.
- $E(x) = Wx$, $D(z) = W^\top z$.
- Let $C = \sum_{i=1}^{m} x_i x_i^\top = X^\top X$ be the data correlation matrix, neglecting the $1/m$ factor.

PCA

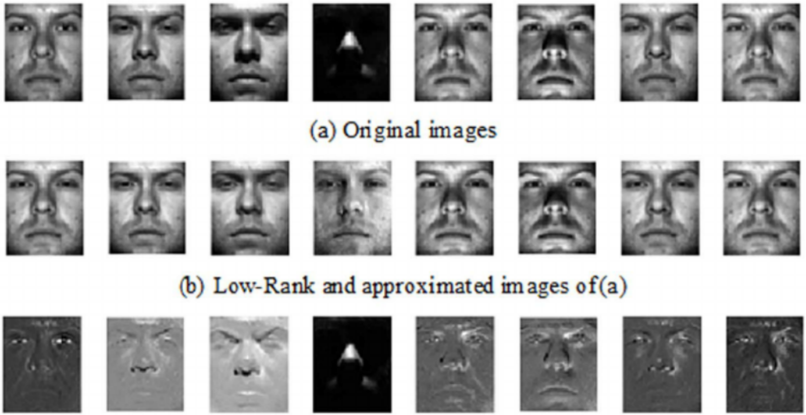
- when E and D are linear $\rightarrow$ PCA.
- $E(x) = Wx$, $D(z) = W^\top z$.
- Let $C = \sum_{i=1}^{m} x_i x_i^\top = X^\top X$ be the data correlation matrix, neglecting the $1/m$ factor.
- C is symmetric and positive semi-definite, $C = V \Lambda V^\top$.
- Theorem PCA: among linear hypothesis classes, $E^* = V^\top$, $D^* = V$, where V is the matrix of eigenvectors of $C = X^\top X$.

Best linear subspace



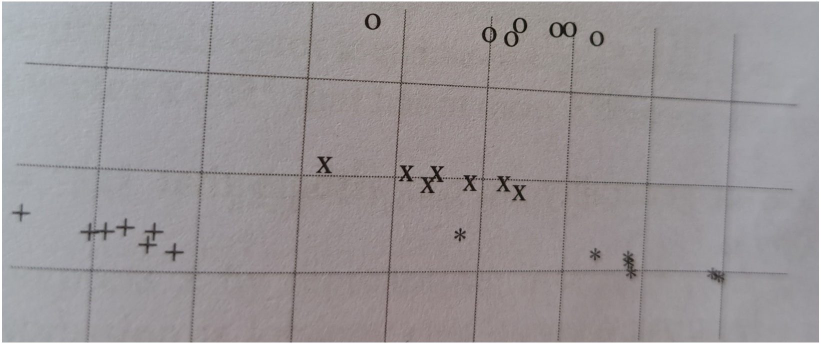
Courtesy: https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73

PCA applied to Yale dataset



Courtesy: Hou, Sun, Chong, Zheng 2014

PCA applied to Yale dataset



Courtesy: Shalev-Schwartz and Ben-David 2014

Linear algebra review: SVD

- for any matrix $X \in \mathbb{R}^{m \times d}$, $X = U \Sigma V^\top$, where $U \in \mathbb{R}^{m \times m}$, $V \in \mathbb{R}^{d \times d}$ are orthogonal matrices, and $\Sigma \in \mathbb{R}^{m \times d}$ is a diagonal matrix.

Linear algebra review: SVD

- for any matrix $X \in \mathbb{R}^{m \times d}$, $X = U \Sigma V^\top$, where $U \in \mathbb{R}^{m \times m}$, $V \in \mathbb{R}^{d \times d}$ are orthogonal matrices, and $\Sigma \in \mathbb{R}^{m \times d}$ is a diagonal matrix.
- U and V are the left and right singular vectors of X, and Σ is the matrix of singular values of X.

Linear algebra review: SVD

- for any matrix $X \in \mathbb{R}^{m \times d}$, $X = U \Sigma V^\top$, where $U \in \mathbb{R}^{m \times m}$, $V \in \mathbb{R}^{d \times d}$ are orthogonal matrices, and $\Sigma \in \mathbb{R}^{m \times d}$ is a diagonal matrix.
- U and V are the left and right singular vectors of X, and Σ is the matrix of singular values of X.
- U and V are the eigenvectors of $XX^\top$ and $X^\top X$ respectively.

Linear algebra review: SVD

- for any matrix $X \in \mathbb{R}^{m \times d}$, $X = U \Sigma V^\top$, where $U \in \mathbb{R}^{m \times m}$, $V \in \mathbb{R}^{d \times d}$ are orthogonal matrices, and $\Sigma \in \mathbb{R}^{m \times d}$ is a diagonal matrix.
- U and V are the left and right singular vectors of X, and Σ is the matrix of singular values of X.
- U and V are the eigenvectors of $XX^\top$ and $X^\top X$ respectively.
- Σ is the square root of the eigenvalues of the SPSD matrices $X^\top X$ and $XX^\top$.

Eigenvalue decomposition, SPSD matrices, SVD

- for a square non-defective or diagonalizable matrix $A \in \mathbb{R}^{d \times d}$, $A = Q \Lambda Q^{-1}$, where Q is the matrix of eigenvectors of A, and Λ is the diagonal matrix of eigenvalues of A.

Eigenvalue decomposition, SPSD matrices, SVD

- for a square non-defective or diagonalizable matrix $A \in \mathbb{R}^{d \times d}$, $A = Q \Lambda Q^{-1}$, where Q is the matrix of eigenvectors of A, and Λ is the diagonal matrix of eigenvalues of A.
- for an SPSD matrix, like $XX^\top$ or $X^\top X$, the eigenvalue decomposition is the same as SVD. Left and right singular vectors are the same and equal to the eigenvectors.

Eigenvalue decomposition, SPSD matrices, SVD

- for a square non-defective or diagonalizable matrix $A \in \mathbb{R}^{d \times d}$, $A = Q\Lambda Q^{-1}$, where Q is the matrix of eigenvectors of A, and Λ is the diagonal matrix of eigenvalues of A.
- for an SPSD matrix, like $XX^\top$ or $X^\top X$, the eigenvalue decomposition is the same as SVD. Left and right singular vectors are the same and equal to the eigenvectors.
- Reduced SVD: $X = U\Sigma V^\top$, where $U \in \mathbb{R}^{m \times r}$, $V \in \mathbb{R}^{d \times r}$ are orthogonal matrices, and $\Sigma \in \mathbb{R}^{r \times r}$ is a diagonal matrix (having non-zero values), when X has rank r.

SVD optimality

- Geometric interpretation: if S is the unit sphere in $\mathbb{R}^d$, XS is the ellipsoid in $\mathbb{R}^m$. The vectors $\sigma_i u_i$ are the semi-axes of the ellipsoid; v_i are the pre-images, i.e., $Xv_i = \sigma_i u_i$.

SVD optimality

- Geometric interpretation: if S is the unit sphere in $\mathbb{R}^d$, XS is the ellipsoid in $\mathbb{R}^m$. The vectors $\sigma_i u_i$ are the semi-axes of the ellipsoid; v_i are the pre-images, i.e., $Xv_i = \sigma_i u_i$.
- Theorem 5.8 (Trefethen and Bau): For any k-dimensional subspace W, the best rank-k approximation to X is given by $X_k = \sum_{i=1}^{k} \sigma_i u_i v_i^\top$. That is,

$$\mathrm{argmin}_{\hat{X}:\mathrm{rank}(\hat{X})\leqslant k}\|X-\hat{X}\|_F = \mathrm{argmin}_{\hat{X}:\mathrm{rank}(\hat{X})\leqslant k}\|X-\hat{X}\| = X_k.$$

Rayleigh Quotient

- For a square matrix $A \in \mathbb{R}^{d \times d}$, the Rayleigh quotient is a scalar function,

$$r(x) = \frac{x^\top Ax}{x^\top x}.$$

Rayleigh Quotient

- For a square matrix $A \in \mathbb{R}^{d \times d}$, the Rayleigh quotient is a scalar function,

$$r(x) = \frac{x^\top A x}{x^\top x}.$$

- Eigenvalues of A are the stationary points of $r(x)$.

Rayleigh Quotient

- For a square matrix $A \in \mathbb{R}^{d \times d}$, the Rayleigh quotient is a scalar function,

$$r(x) = \frac{x^\top A x}{x^\top x}.$$

- Eigenvalues of A are the stationary points of $r(x)$.
- $\nabla r(x) = \frac{2}{x^\top x}(Ax - r(x)x)$.

PCA by SVD

- When $m > d$, do eigenvalue decomposition of $X^\top X$ or SVD of X.

PCA by SVD

- When $m > d$, do eigenvalue decomposition of $X^\top X$ or SVD of X.
- When $m < d$, do eigenvalue decomposition of $XX^\top$. If $v_1, v_2, \cdots, v_n$ are the n largest eigenvectors, principal vectors are $\frac{1}{\|X^\top v_i\|} X^\top v_i$.

PCA by SVD

- When $m > d$, do eigenvalue decomposition of $X^\top X$ or SVD of X.
- When $m < d$, do eigenvalue decomposition of $XX^\top$. If $v_1, v_2, \cdots, v_n$ are the n largest eigenvectors, principal vectors are $\frac{1}{\|X^\top v_i\|} X^\top v_i$.
- Computational complexity: $O(\min(m^2 d, md^2))$.

→ Midterm 1 – collect at 4 pm
(10/31) CODA Costa coffee

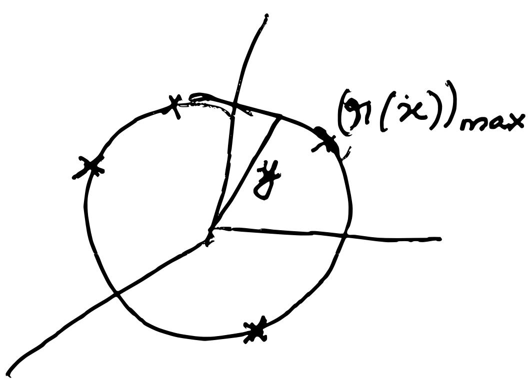
→ 1 page Nov 2nd

Principal Component Analysis

→ Matrix induced lp norm

$$\|A\| = \sup_{\|x\|_p \leq 1} \|Ax\|_p$$

→ Rayleigh Quotient



$$r(x) = (y(x))^T A \, y(x)$$

$$y(x) = \frac{x}{\|x\|}$$

$$r_A(x) = x^T A x$$

$$\max_{\{x : \|x\|=1\}} r(x)$$

$$\rightarrow \mathcal{L}(x, \lambda) = -x^T A x + \lambda (x^T x - 1)$$

$$\nabla \mathcal{L}(x, \lambda) = 2Ax = 2\lambda x$$

$$\lambda = 0 \text{ or } x^T x = 1$$

A is SPSD

u_1 : largest eigenvector

$$\rightarrow A_1 = (I - u_1 u_1^T) A$$

$$(\langle A_1, u_1 \rangle = 0)$$

$$\max_{\substack{\{x : \|x\|=1 \\ \langle x, u_1 \rangle = 0\}}} r_{A_1}(x) = \lambda_2$$

$$\rightarrow u_1, u_2, \ldots, u_n, \quad \langle u_i, u_j \rangle = 0$$

$$\max_{\|x\|=1} \left(r_A(x) + r_A(y_1(x)) + \cdots + r_A(y_n(x)) \right)$$

$>$

s.t.

$$\langle y_i(x), u_1 \rangle = 0, \ldots, \langle y_i(x), u_i \rangle = 0$$

$$\|u_i\| = 1$$

$$= \sum_{i=1}^{n} \lambda_i$$

$\lambda_1 > \cdots \lambda_d$ are the eigenvalues of A (SPSD)

Complexity

$\underline{m > d}$

X^T $\quad$ X $\qquad \longrightarrow \underbrace{X^T X}$

$d \times m$ $\quad$ $m \times d$ $\qquad$ Cost: $d^2 m$

$\longrightarrow$ $eig(X^T X)$

Cost: d^3

$$O(d^2 m + d^3) = \underline{O(d^2 m)}$$

$\underline{d > m}$

$\rightarrow$ $X X^T$

Cost: $m^2 d$

$\rightarrow$ $eig(X X^T)$

Cost: m^3

$$\underline{O(m^2 d)}$$

$$\rightarrow \underset{\substack{\rightarrow E, D \\ \mathbb{R}^{n\times d} \quad \mathbb{R}^{d\times n}}}{\operatorname{argmin}} \| \underset{\mathbb{R}^{m\times d}}{X} - DEX \|^2$$

$$=$$

$$\underset{\substack{D \in \mathbb{R}^{d\times n} \\ D^T D = I_n}}{\operatorname{argmin}} \| X - DD^T X \|^2$$

$$= \underset{D \in \mathbb{R}^{d\times n}}{\operatorname{argmin}} \sum_{i=1}^{m} \| x_i - DD^T x_i \|^2$$

$$\min l(D) = \overset{\min_D}{\sum_{i=1}^{m}} \| x_i - DD^T x_i \|^2 = \overset{\min_D}{\sum_{i=1}^{m}} \| x_i \|^2 - x_i^T D D^T x_i$$

$$= \max_D \sum_{i=1}^{m} x_i^T DD^T x_i \quad \left(\sum_{i=1}^{m} \| D x_i \|^2 \right)$$

$$= \max_D Tr(D^T \underline{X^T X} D)$$

$$\left(\| y \|^2 = Tr(y y^T) \right)$$

$$Tr(D^T \underline{\sum_{i=1}^{m} x_i x_i^T} D)$$

We have shown

$$\min_D l(D) = \max_D Tr(D^T X^T X D)$$

$$= \max_D Tr(D^T C D)$$

$$= \sum_{i=1}^{n} \lambda_i$$

Have to show

$$\underset{\substack{D, E \\ \in \\ \mathbb{R}^{d\times n}}}{\operatorname{argmin}} \| X - \underbrace{DEX}\|^2 = \underset{D}{\operatorname{argmin}} \| X - DD^T X \|^2$$

$$DEx \in \operatorname{Ran}(D) \subset (\text{linear subspace of dim } n \text{ in } \mathbb{R}^d)$$

Say $V \in \mathbb{R}^{d\times n}$ is orthogonal basis for $\mathbb{R}^n$

Vy for any $y \in \mathbb{R}^n$. For any x, D, E, V,

$$\| x - DEx \|^2 \geq \min_y \| x - \underbrace{Vy}_{VV^T x} \|^2$$

$$\min_y \| x - Vy \|^2 \qquad V^T V = I_n$$

$$\rightarrow \quad y = V^T x.$$

$$\min_{D, E} \| x - DEx \|^2 = \min_{\substack{V \\ V^T V = I}} \| x - VV^T x \|^2$$