Lecture 19: PCA

Nisha Chandramoorthy

October 31, 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

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

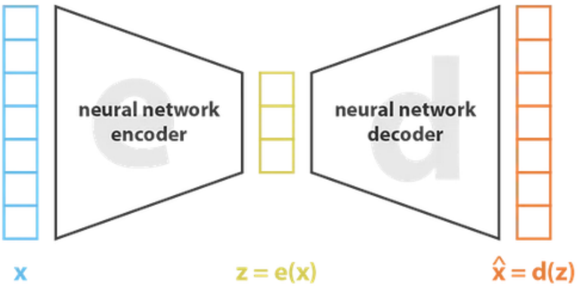
- Posed as ERM problem.
- E is encoder, D is decoder.
- E maps x to z (latent space), D maps z to $\hat{x}$ (reconstruction).
- Both parameterized as Neural Networks.

Variational autoencoders

- Probabilistic encoder and decoder.

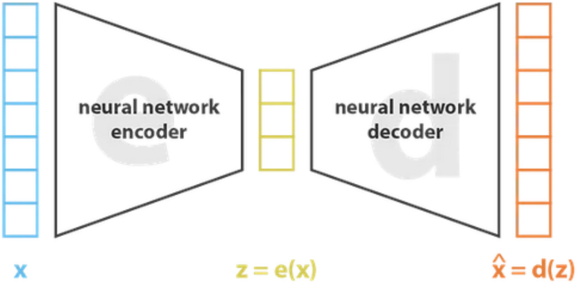
Variational autoencoders

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



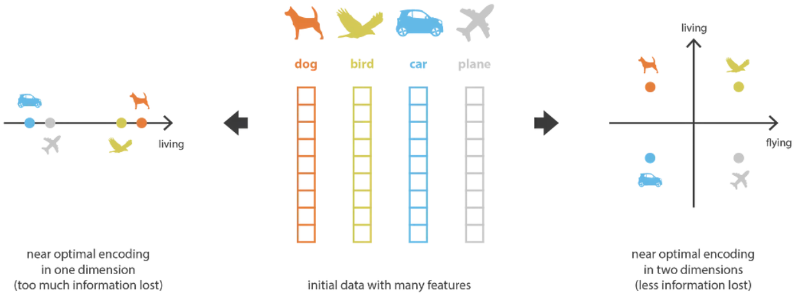
$$\text{loss} = \| x - \hat{x} \|^2 = \| x - d(z) \|^2 = \| x - d(e(x)) \|^2$$

- tends to overfit as a Generative model

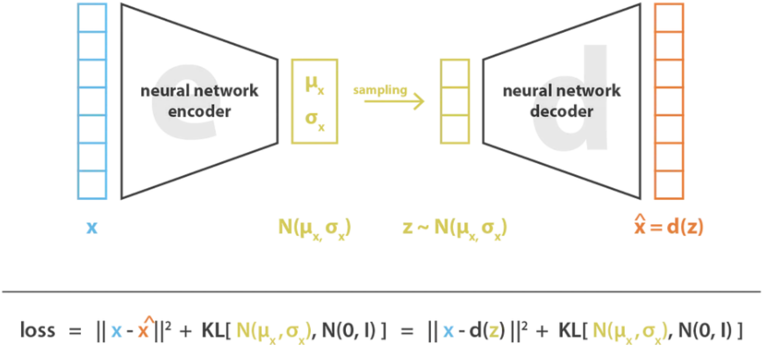


$$\text{loss} = \| x - \hat{x} \|^2 = \| x - d(z) \|^2 = \| x - d(e(x)) \|^2$$

- 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.
- $E(x) = Wx$, $D(z) = W^\top z$.

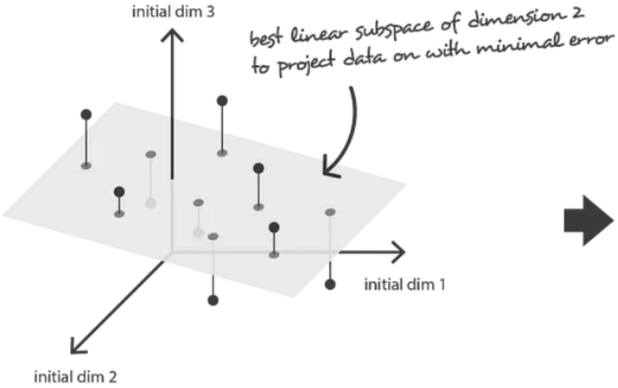
PCA

- 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

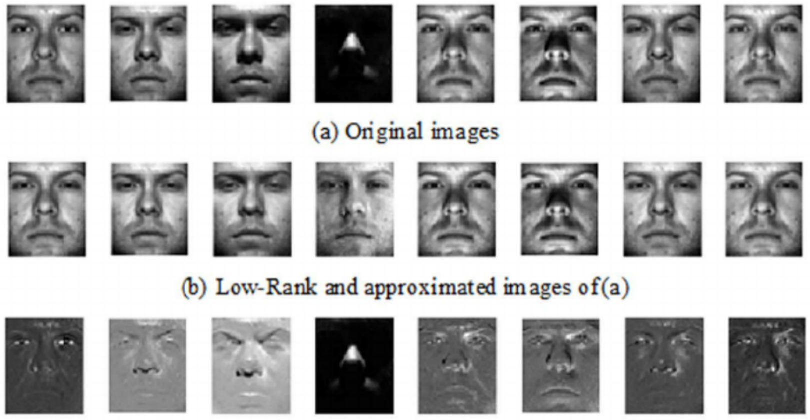
- when E and D are linear $\rightarrow$ PCA.
- $E(x) = Wx$, $D(z) = W^\top z$.
- Let $C = \sum_{i=1}^{m} x_i x_i^\top = X^\top X$ be the data correlation matrix, neglecting the $1/m$ factor.
- C is symmetric and positive semi-definite, $C = V \Lambda V^\top$.
- Theorem PCA: among linear hypothesis classes,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

Best linear subspace



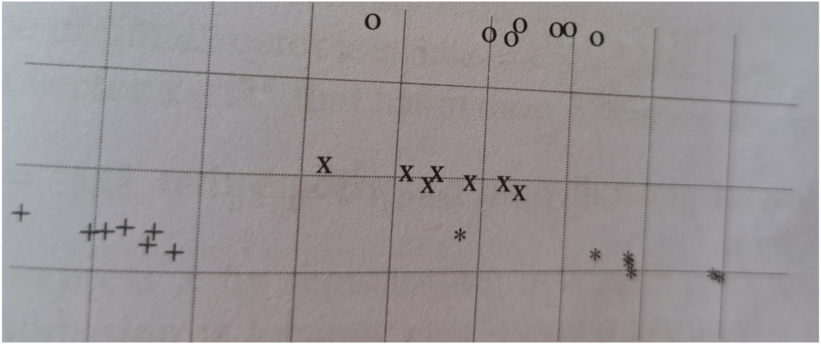
Courtesy: https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73

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,

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

Rayleigh Quotient

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

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

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: $O(\min(m^2 d, md^2))$.

Convolutional Neural Networks (source: cs231n.stanford.edu)

- Suitable for image recognition. Won the 2012 ImageNet competition and subsequent ones.
- Three types of layers: convolutional, FC, pooling
- Convolutional layer: accepts a volume of size $W_1 \times H_1 \times D_1$ and outputs a volume of size $W_2 \times H_2 \times D_2$ where $W_2 = (W_1 - F + 2P)/S + 1$ and $H_2 = (H_1 - F + 2P)/S + 1$ and $D_2 = K$.
- K is number of filters, F is filter size, S is stride, P is padding.
- Pooling layer: downsamples along width and height, and optionally along depth.
- FC layer: computes class scores, resulting in volume of size $1 \times 1 \times K$.