CS540 Introduction to Artificial Intelligence

Lecture 16

Young Wu

Based on lecture slides by Jerry Zhu, Yingyu Liang, and Charles Dyer

July 18, 2022

Random Choice 1

Quiz

P1, P2
P3 due tonight + 1 week
P4 can start, not submit.

P6

one week after final

Q7

- Pick a random choice.
- A :
- B :
- C :
- D :
- E :

Random Choice 2

Quiz

Q8

- Pick the choice you think is the least popular.
- A :
- B :
- C :
- D :
- E :

Q9

Random Choice 3

Quiz

- Pick the choice based on the last digit of your ID.
- $A : 0 - 1$
- $B : 2 - 3$
- $C : 4 - 5$
- $D : 6 - 7$
- $E : 8 - 9$

Q10

Unsupervised Learning

Motivation

- Supervised learning: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.
- Unsupervised learning: $x_1, x_2, \ldots, x_n$.
- There are a few common tasks without labels.

1. Clustering: separate instances into groups.
2. Novelty (outlier) detection: find instances that are different.
3. Dimensionality reduction: represent each instance with a lower dimensional feature vector while maintaining key characteristics.

High Dimensional Data

Motivation

- High dimensional data are training set with a lot of features.

1. Document classification.
2. MEG brain imaging.
3. Handwritten digits (or images in general).

Low Dimension Representation

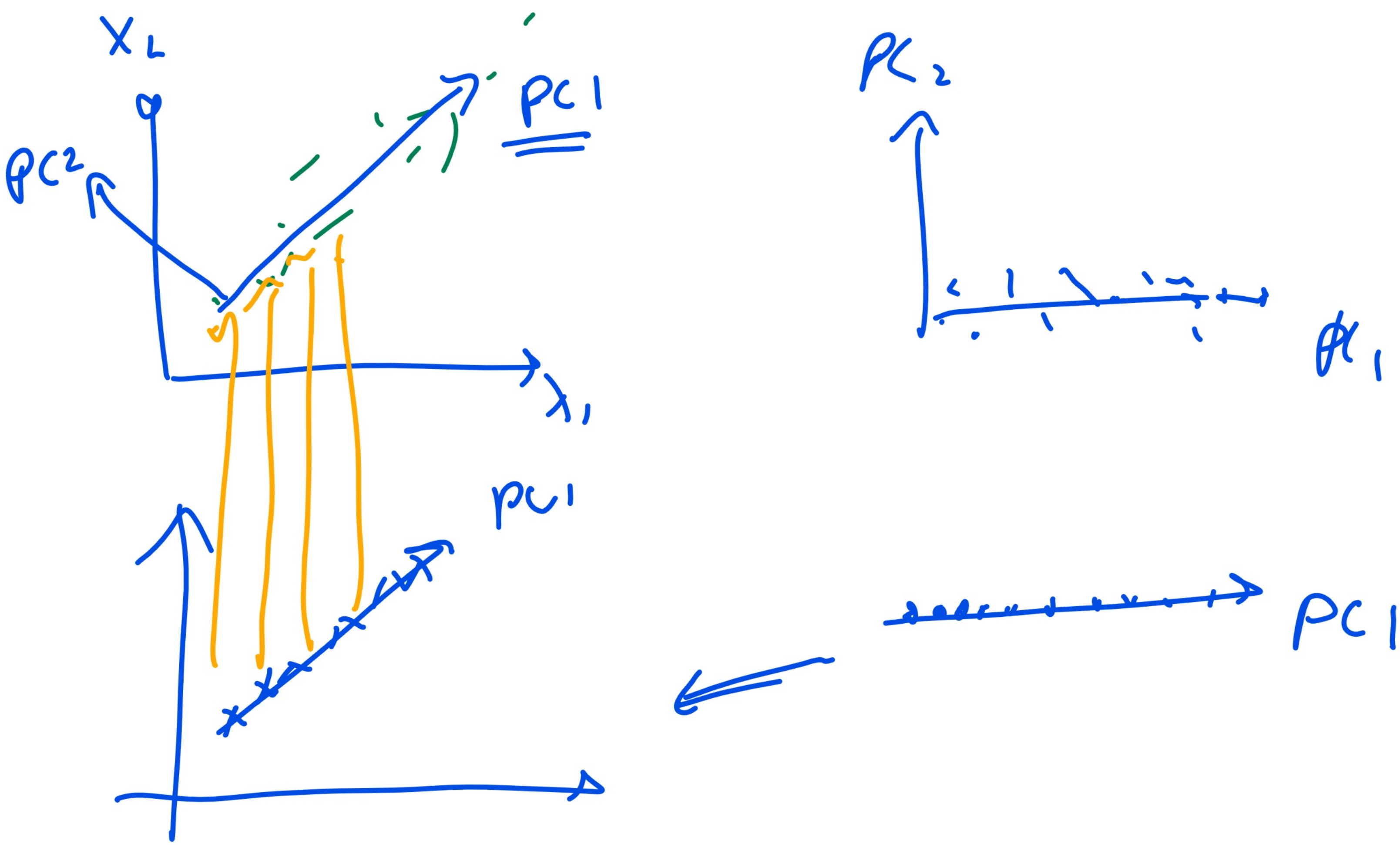
Motivation

- Unsupervised learning techniques are used to find low dimensional representation.

1. Visualization.
2. Efficient storage.
3. Better generalization.
4. Noise removal.

Dimension Reduction Demo

Motivation



Projection

Definition

- The projection of x_i onto a unit vector u_k is the vector in the direction of u_k that is the closest to x_i.

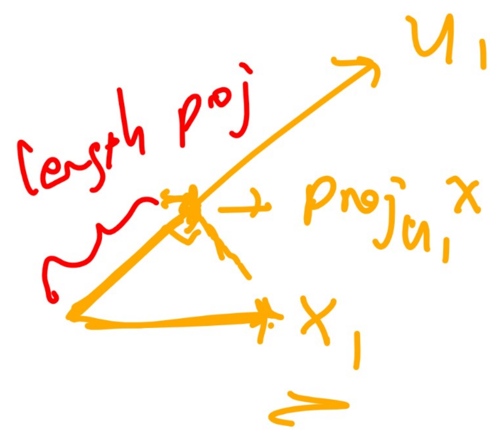
$$\text{proj}_{u_k} x_i = \boxed{\left(\frac{u_k^T x_i}{u_k^T u_k}\right)} u_k = u_k^T x_i u_k$$

$(1 = u_{k1}^2 + u_{k2}^2 + \dots + u_{kn}^2)$

if u_k is not unit vector

- The length of the projection of x_i onto a unit vector u_k is $u_k^T x_i$.

$$\left\| \text{proj}_{u_k} x_i \right\|_2 = u_k^T x_i$$



Variance

Definition

- The sample variance of a data set $\{x_1, x_2, ..., x_n\}$ is the sum of the squared distance from the mean.

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{bmatrix}$$

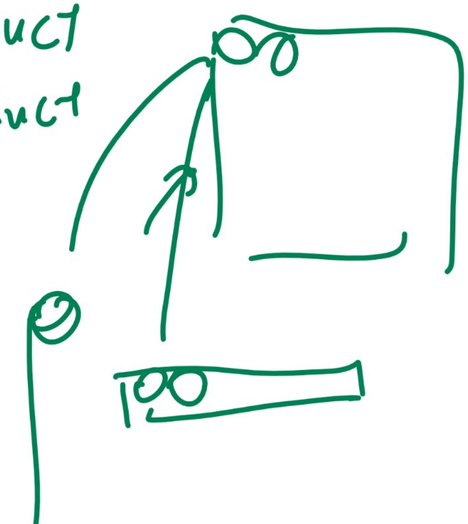
$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n} \underline{(x_i - \hat{\mu})(x_i - \hat{\mu})^T}$$

$(x_i - \hat{\mu})^T (x_i - \hat{\mu})$

inner product

dot product



outer product.

Projection Example 1

Quiz

$$\frac{1}{2}\left[\left(\sqrt{2}-\bar{\mu}\right)^2+\left(\frac{\sqrt{2}}{2}-\bar{\mu}\right)^2\right] \rightarrow \text{proj variance}$$

where $\bar{\mu}=\underline{\frac{1}{2}\left(\sqrt{2}+\frac{\sqrt{2}}{2}\right)}$

$$\frac{u^T x}{u^T u}\,u$$

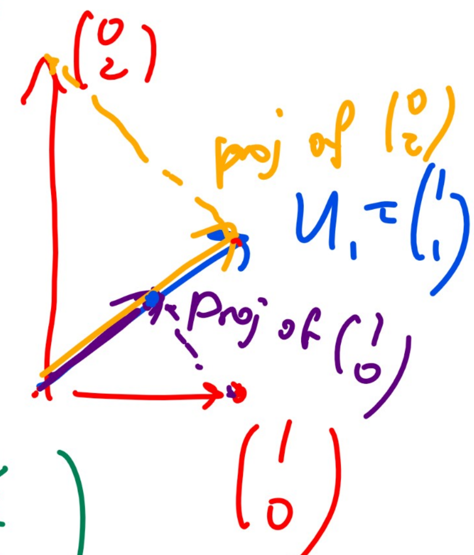
$$\text{proj}_{\begin{pmatrix}1\\1\end{pmatrix}}\begin{pmatrix}0\\2\end{pmatrix}=\frac{(1\ \ 1)\begin{pmatrix}0\\2\end{pmatrix}}{(1\ \ 1)\begin{pmatrix}1\\1\end{pmatrix}}\begin{pmatrix}1\\1\end{pmatrix}=\frac{2}{2}\begin{pmatrix}1\\1\end{pmatrix}=\underline{\begin{pmatrix}1\\1\end{pmatrix}}$$

not unit

- What is the projection of $\underbrace{\begin{bmatrix}1\\0\end{bmatrix} \text{ and } \begin{bmatrix}0\\2\end{bmatrix}}$ onto $\underline{\begin{bmatrix}1\\1\end{bmatrix}}$ and what is the projected variance?

unit vector $\begin{pmatrix}\frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}}\end{pmatrix}$

$\sqrt{2}\,,\ \frac{\sqrt{2}}{2}$



$$\underline{\text{proj}_{\begin{pmatrix}\frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}}\end{pmatrix}}\begin{pmatrix}1\\0\end{pmatrix}}=\underbrace{\left(\frac{1}{\sqrt{2}}\ \ \frac{1}{\sqrt{2}}\right)}_{u^T}\underbrace{\begin{pmatrix}1\\0\end{pmatrix}}_{x}\underbrace{\begin{pmatrix}\frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}}\end{pmatrix}}_{u}=\left(\frac{1}{\sqrt{2}}\cdot 1+\cancel{\frac{1}{\sqrt{2}}\cdot 0}\right)\begin{pmatrix}\frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}}\end{pmatrix}=\begin{pmatrix}\frac{1}{2}\\ \frac{1}{2}\end{pmatrix}$$

Projection Example 3

Quiz

Q11

- What is the projection of $\begin{bmatrix}1\\2\\3\end{bmatrix}$ onto $\begin{bmatrix}1\\1\\1\end{bmatrix}$? unit?
- $A : [2 \quad 2 \quad 2]^T$
- $B : [3 \quad 3 \quad 3]^T$
- $C : [4 \quad 4 \quad 4]^T$
- $D : [6 \quad 6 \quad 6]^T$
- E : *I* don't understand.

$\sqrt{1^2+1^2+1^2} = \sqrt{3}$

$\frac{u^T x}{u^T u} u = proj_u x$

$u \begin{pmatrix}\frac{1}{\sqrt{3}}\\\frac{1}{\sqrt{3}}\\\frac{1}{\sqrt{3}}\end{pmatrix} \rightarrow \sqrt{\left(\frac{1}{\sqrt{3}}\right)^2+\left(\frac{1}{\sqrt{3}}\right)^2+\left(\frac{1}{\sqrt{3}}\right)^2}$

$= \sqrt{\frac{1}{3}+\frac{1}{3}+\frac{1}{3}}$

$= \sqrt{1} = 1$

$\frac{6}{3} \cdot \begin{pmatrix}1\\1\\1\end{pmatrix} = 2\begin{pmatrix}1\\1\\1\end{pmatrix}$

Projection Example 4

Quiz

$\bar{\mu} = \sqrt{12}$

$$\frac{1}{2}\left((\sqrt{12}-\sqrt{12})^2 + (\sqrt{12}-\sqrt{12})^2 \right) = 0$$

$\cancel{\frac{1}{n-1}}$ Q12

- What is the ~~projection~~ projected variance of $\begin{bmatrix}1\\2\\3\end{bmatrix}$ and $\begin{bmatrix}3\\2\\1\end{bmatrix}$ onto $\begin{bmatrix}1\\1\\1\end{bmatrix}$?

$\begin{pmatrix}2\\2\\2\end{pmatrix} \to \sqrt{12}$ $\begin{pmatrix}2\\2\\2\end{pmatrix} \to \sqrt{12}$ $\frac{6}{3}\begin{pmatrix}1\\1\\1\end{pmatrix} = \begin{pmatrix}2\\2\\2\end{pmatrix}$

- *A* : 0
- *B* : 12
- *C* : 24
- *D* : 48
- *E* : *I* don't understand.

Maximum Variance Directions

Definition

- The goal is to find the direction that maximizes the projected variance.

$$\boxed{\max_{u_k} u_k^T \hat{\Sigma} u_k \text{ such that } u_k^T u_k = 1}$$

$$\Rightarrow \max_{u_k} u_k^T \hat{\Sigma} u_k - \lambda u_k^T u_k$$

$$\Rightarrow \boxed{\hat{\Sigma} u_k = \lambda u_k}$$

derivative

eigenvalue.

eigenvector

Eigenvalue

Definition

- The λ represents the projected variance.

$$u_k^T \hat{\Sigma} u_k = u_k^T \lambda u_k = \lambda$$

- The larger the variance, the larger the variability in direction u_k. There are m eigenvalues for a symmetric positive semidefinite matrix (for example, $X^T X$ is always symmetric PSD). Order the eigenvectors u_k by the size of their corresponding eigenvalues λ_k.

$$\lambda_1 \geqslant \lambda_2 \geqslant ... \geqslant \lambda_m$$

Eigenvalue Algorithm

Definition

- Solving eigenvalue using the definition (characteristic polynomial) is computationally inefficient.
$$\left(\hat{\Sigma} - \lambda_k I\right) u_k = 0 \Rightarrow \det \left(\hat{\Sigma} - \lambda_k I\right) = 0$$

- There are many fast eigenvalue algorithms that computes the spectral (eigen) decomposition for real symmetric matrices. Columns of Q are unit eigenvectors and diagonal elements of D are eigenvalues.
$$\begin{aligned} \hat{\Sigma} &= PDP^{-1}, D \text{ is diagonal} \\ &= QDQ^T, \text{ if } Q \text{ is orthogonal, i.e. } Q^T Q = I \end{aligned}$$

Spectral Decomposition Example 1

700 → k=3

Quiz

$PC_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ 0 \\ \frac{1}{\sqrt{2}} \end{pmatrix}$ $PC_2 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ 0 \\ -\frac{1}{\sqrt{2}} \end{pmatrix}$ $PC_3 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$

- Given the following spectral decomposition of $\hat{\Sigma}$, what are the first two principal components?

$$\hat{\Sigma} = \underbrace{\begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix}}_{P} \underbrace{\begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix}}_{\text{diag}} \underbrace{\begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix}}_{P^{-1}}$$

eigenvectors

eigenvalues

PC3, PC2, PC1

$PC_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ 0 \\ \frac{1}{\sqrt{2}} \end{pmatrix}$

$PC_2 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ 0 \\ -\frac{1}{\sqrt{2}} \end{pmatrix}$

$PC_3 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$

Spectral Decomposition Example 2

Quiz

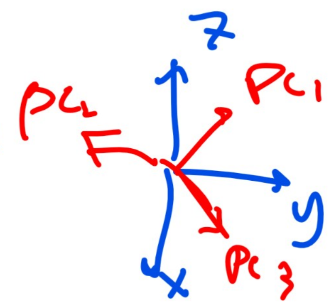
unit vector

Q1

- Given the following $\hat{\Sigma}$, what are the first two principal components?

e.v   e.v = variance

$$\Sigma = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \hat{\Sigma} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}^{-1}$$



- $A: \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$, B: $\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$, C: $\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$, ~~D: $\begin{bmatrix} 0 \\ 5 \\ 0 \end{bmatrix}$, E: $\begin{bmatrix} 0 \\ 0 \\ 3 \end{bmatrix}$~~

PC1   PC2

unit vector

Number of Dimensions

Discussion

- There are a few ways to choose the number of principal components K. $= 2, 3$
- K can be selected given prior knowledge or requirement.
- K can be the number of non-~~zero~~ eigenvalues.
- K can be the number of eigenvalues that are large (larger than some threshold).

>1

validation set

compare K.

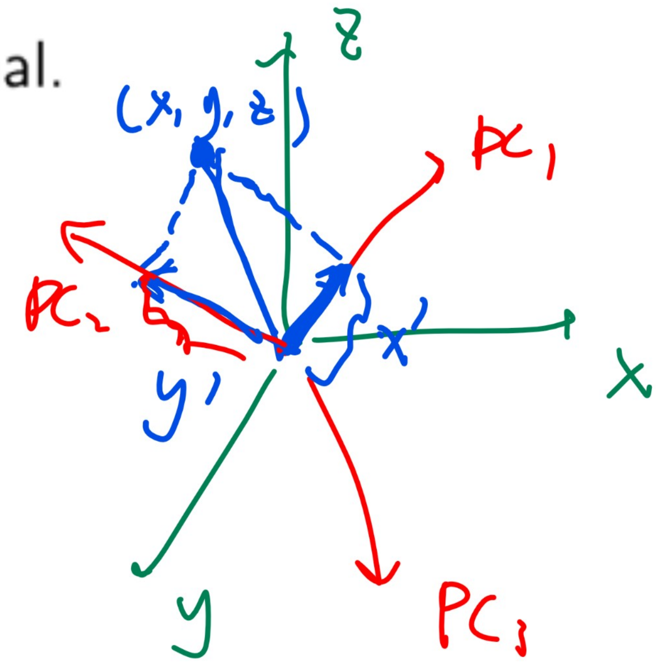
Reduced Feature Space

Discussion

- The original feature space is m dimensional.

(x, y, z)

$$(x_{i1}, x_{i2}, \ldots, x_{im})^T$$



- The new feature space is K dimensional.

length of projection

(x', y')

$$(u_1^T x_i, u_2^T x_i, \ldots, u_K^T x_i)^T$$

PC1 PC2 PCK

- Other supervised learning algorithms can be applied on the new features.

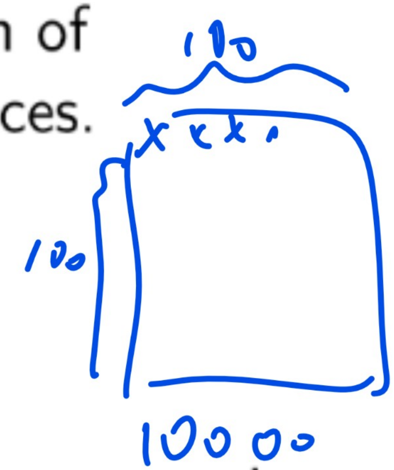
Eigenface

Discussion

- Eigenfaces are eigenvectors of face images (pixel intensities or HOG features).
- Every face can be written as a linear combination of eigenfaces. The coefficients determine specific faces.

$K = m$ $\quad\quad\quad\quad\quad\quad$ $< m$

$$x_i = \sum_{k=1}^{m} \left(u_k^T x_i\right) u_k \approx \sum_{k=1}^{K} \left(u_k^T x_i\right) u_k$$



- Eigenfaces and SVM can be combined to detect or recognize faces.

$K = 112$

$(x, y, z) \Rightarrow (\underline{x', y'}, \cancel{z})$

PC1 PC2

Reduced Space Example 1

Quiz

- If $u_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ 0 \\ \frac{1}{\sqrt{2}} \end{bmatrix}$ and $u_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ 0 \\ -\frac{1}{\sqrt{2}} \end{bmatrix}$. If one original item is $x = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$. What is its new representation and the reconstructed vector using only the two principal components?

$\Rightarrow u_3 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$

reconstruction

approx x

$\boxed{(2\sqrt{2}, -\sqrt{2})}$

$= 2\sqrt{2} \cdot u_1 + -\sqrt{2} \cdot u_2$

$= \begin{pmatrix} 2 \\ 0 \\ 2 \end{pmatrix} + \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}$

$= \begin{pmatrix} 1 \\ 0 \\ 3 \end{pmatrix}$

(x'_1, x'_2)

length of proj

$u_1^T x$, $u_2^T x$

$1 \cdot \frac{1}{\sqrt{2}} + 2 \cdot 0 + 3 \cdot \frac{1}{\sqrt{2}} = 2\sqrt{2}$

$= 1 \cdot \frac{1}{\sqrt{2}} + 2 \cdot 0 + 3 \cdot (-\frac{1}{\sqrt{2}}) = -\sqrt{2}$

Reduced Space Example 1 Diagram

Quiz

Reduced Space Example 2

Quiz

Q2

- $\hat{\Sigma} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 3 \end{bmatrix}$. If one original data is $x = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$. What is the reconstructed vector using only the first two principal components?
- $A: \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}$, B: $\begin{bmatrix} 2 \\ 3 \\ 0 \end{bmatrix}$, C: $\begin{bmatrix} 0 \\ 2 \\ 3 \end{bmatrix}$, D: $\begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix}$, E: I don't understand.

$PC_1 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$ $PC_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$

$x' = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$ $\quad 2 \to PC_1^T x$, $\quad 3 \to PC_2^T x$

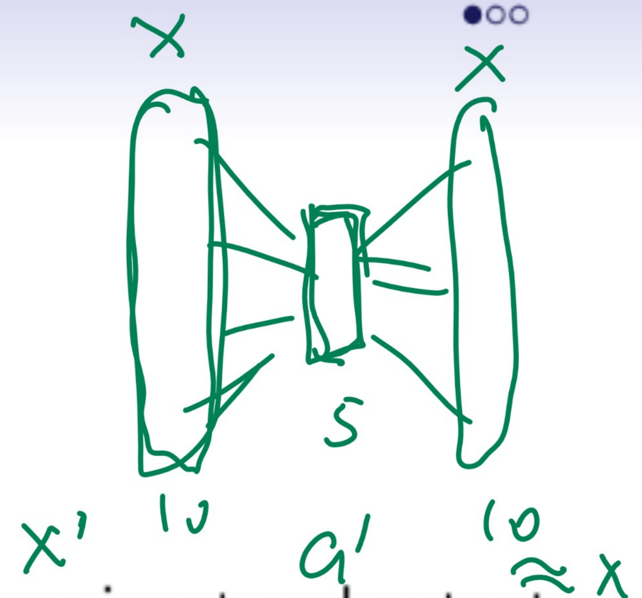
$$2 \cdot PC_1 + 3\, PC_2 = \begin{pmatrix} 0 \\ 2 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 3 \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \\ 3 \end{pmatrix}$$

Autoencoder

Discussion

non-linear "PCA" $\Leftarrow a = g(W^T x + b)$

PCA $\Leftarrow a = W^T x + b$



- A multi-layer neural network with the same input and output $y_i = x_i$ is called an autoencoder.
- The hidden layers have fewer units than the dimension of the input m.
- The hidden units form an encoding of the input with reduced dimensionality.

Autoencoder Diagram

Discussion

Kernel PCA

Discussion

SVM

- A kernel can be applied before finding the principal components.

back 6:30

$$\hat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} \varphi(x_i) \varphi(x_i)^T$$

feature map

svm

$\varphi = \infty$ dim

- The principal components can be found without explicitly computing $\varphi(x_i)$, similar to the kernel trick for support vector machines.

K

- Kernel PCA is a non-linear dimensionality reduction method.