CS540 Introduction to Artificial Intelligence
Lecture 12

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

July 7, 2020

Midterm Discussion

Admin

Final
23, 24

Q1 pick A.

due July 28 ← $\boxed{P1 - P6}$ M8 – M12 $\boxed{\text{exclude M11}}$

- Bug fix in auto-grading, grades updated.
- Did not fix individual grades.
- Version A Part 1 average: 4.5, Part 2 average: 3.1
- Version B Part 1 average: 3.3, Part 2 average: 3.2
- None of the questions has PROB < 0.25, RPBI < 0.
- No curve for all versions.

A = B/A

1) if no submission to P1S, P2S ... ⇒ 0

2) same code/output as someone else ⇒ 0

3) P6

Low Dimension Representation

Motivation

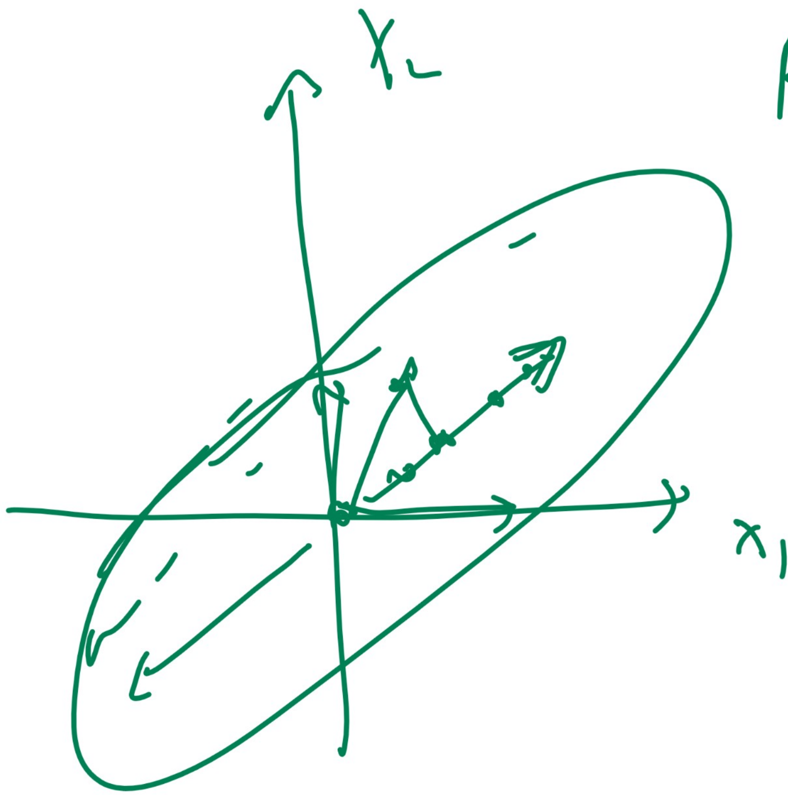
- Unsupervised learning techniques are used to find low dimensional representation.

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

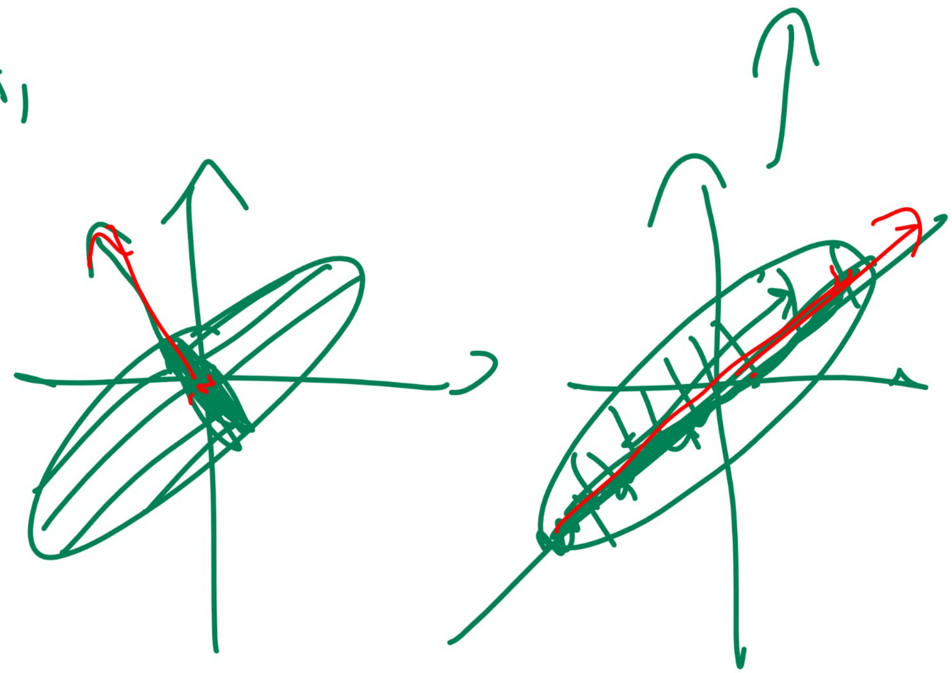
regularization

Dimension Reduction Diagram

Motivation



PCA find direction captures largest variation.

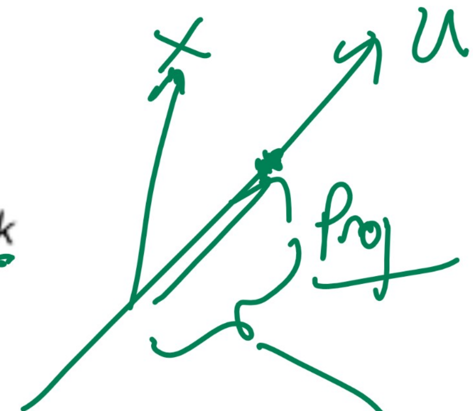


Projection

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.

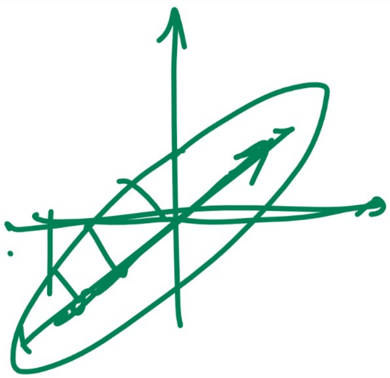
$$\text{proj}_{\,u_k} x_i = \boxed{\left(\frac{u_k^T x_i}{u_k^T u_k}\right)} u_k = u_k^T x_i u_k$$



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

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

Maximum Variance Directions



Definition

projected variance
length

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

$u^T_k \lambda u_k = \lambda$

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

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

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

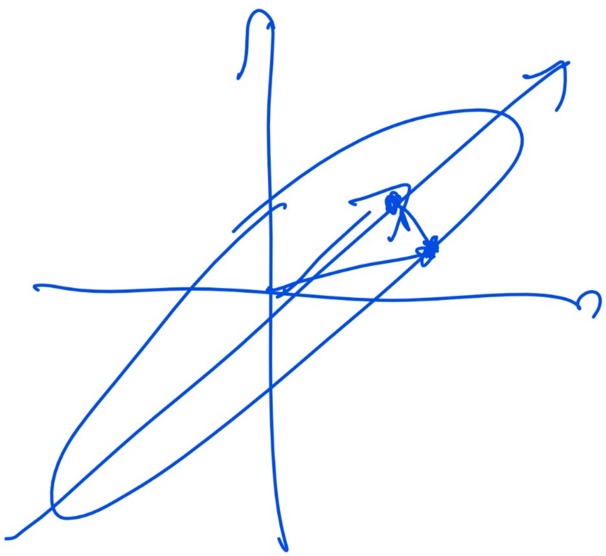
λ

λ = eigenvalue.

max $\underline{\lambda}$ by $\Sigma u_k = \lambda u_k$

Projection Example 1

Quiz

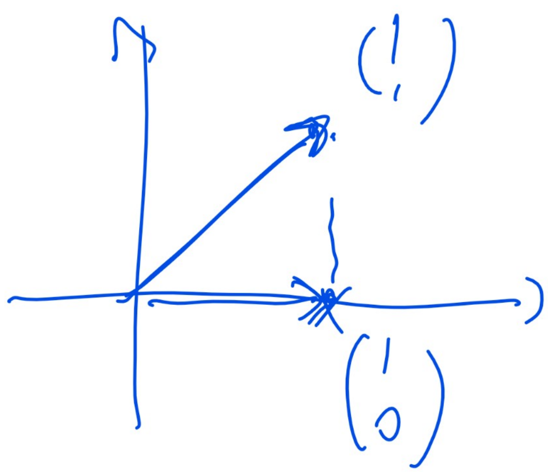


$u_1 \quad u_2 \quad u_3$

$PC1 \quad PC2$

- What is the projection of $\underset{x}{\begin{bmatrix} 1 \\ 1 \end{bmatrix}}$ onto $\underset{u}{\begin{bmatrix} 1 \\ 0 \end{bmatrix}}$?

$(x^T u)\, u$



$\boxed{(u^T x)}\, u$

$= \underbrace{(1, 0)\begin{pmatrix} 1 \\ 1 \end{pmatrix}}_{1} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$

$w^T x = x^T w = w \cdot x = x \cdot w$

Projection Example 2

Quiz

$$\frac{(1,0)\begin{pmatrix}1\\1\end{pmatrix}}{(1,1)\begin{pmatrix}1\\1\end{pmatrix}}\begin{pmatrix}1\\1\end{pmatrix} = \frac{1}{2}\cdot\begin{pmatrix}1\\1\end{pmatrix} = \begin{pmatrix}\frac{1}{2}\\ \frac{1}{2}\end{pmatrix}$$

x   u → not unit

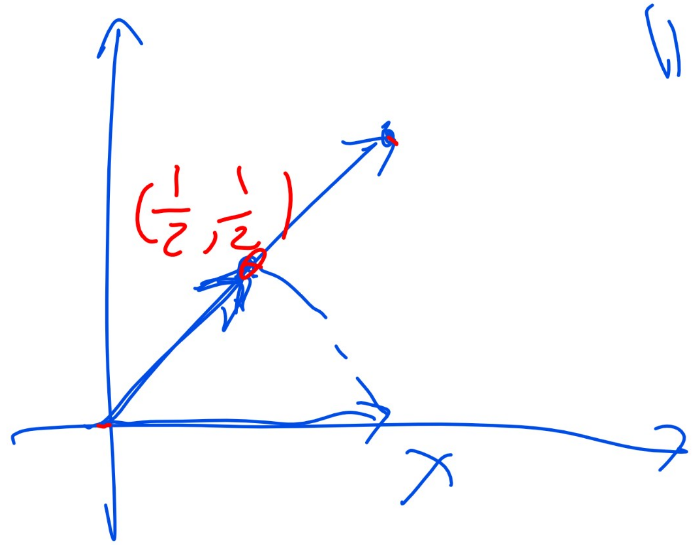
- What is the projection of $\begin{bmatrix}1\\0\end{bmatrix}$ onto $\begin{bmatrix}1\\1\end{bmatrix}$?

$\begin{bmatrix}1\\1\end{bmatrix} \rightarrow \begin{pmatrix}\frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}}\end{pmatrix}$ u

$\|u\| = 1$

$$\boxed{u^T x \; u} = \cancel{\begin{pmatrix}1\\1\end{pmatrix}}$$

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



Projection Example 3

Quiz

Q2

- What is the projection of $\overset{x}{\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}}$ onto $\overset{u}{\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}}$?
- A: $[1 \quad 1 \quad 1]^T$
- B: $[2 \quad 2 \quad 2]^T$
- C: $[3 \quad 3 \quad 3]^T$
- D: $[4 \quad 4 \quad 4]^T$
- E: $[6 \quad 6 \quad 6]^T$

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

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

Projection Example 4

Quiz

Q3

- What is the projection of $\begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$ onto $\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$?
- A: $[1 \quad 1 \quad 1]^T$
- B: $[2 \quad 2 \quad 2]^T$
- C: $[3 \quad 3 \quad 3]^T$
- D: $[4 \quad 4 \quad 4]^T$
- E: $[6 \quad 6 \quad 6]^T$

$\frac{6}{3}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$

Spectral Decomposition Example 1

Quiz

- Given the following spectral decomposition of $\hat{\Sigma}$, what is the first principal component?

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

$\downarrow$ eigenvalues. max λ

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

direction

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

Spectral Decomposition Example 2

Quiz

@4

eig(_)

- Given the following spectral decomposition of $\hat{\Sigma}$, what is the first principal component?

$$\hat{\Sigma} = \overset{I}{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}} \begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}^{-1}$$

- A: $\underset{PC2}{\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}}$, B: $\underset{PC1}{\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}}$, C: $\underset{PC3}{\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}}$, D: $\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$, E: $\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$

Spectral Decomposition Example 3

Quiz

Q5 (last)

- $\hat{\Sigma} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, what is the second principal component?
- 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: $\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$, E: $\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$

Principal Component Analysis

Algorithm

- Input: instances: $\{x_i\}_{i=1}^{n}$, the number of dimensions after reduction $K < m$.
- Output: K principal components.
- Find the largest K eigenvalues $\lambda_1 \geqslant \lambda_2 \geqslant ... \geqslant \lambda_K$.
- Return the corresponding unit orthogonal eigenvectors $u_1, u_2 ... u_K$.

Number of Dimensions

Discussion

- There are a few ways to choose the number of principal components K.
- 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).

Reduced Feature Space

Discussion

- The original feature space is m dimensional.

786 → 3

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

- The new feature space is K dimensional.

PC2

PC1

$$(u_1^T x_i, u_2^T x_i, ..., u_K^T x_i)^T$$

3 786

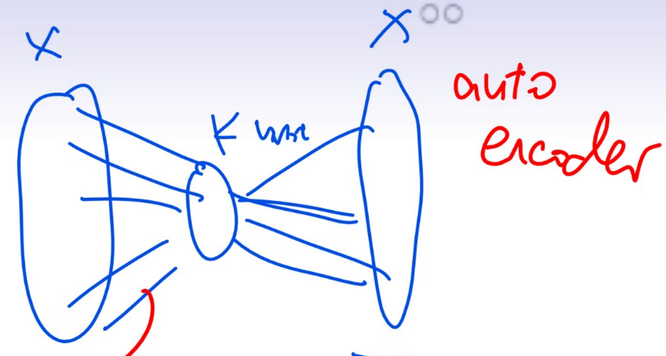
$K << m$

length of proj

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

Eigenface

Discussion



approx PCs e.vector activation

- Eigenfaces are eigenvectors of face images (pixel intensities or HOG features).

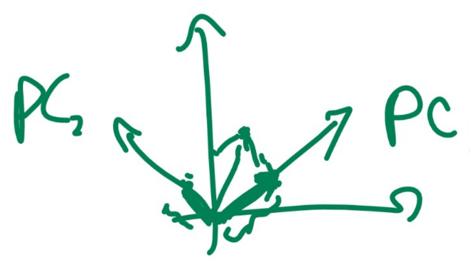
PC

- Every face can be written as a linear combination of eigenfaces. The coefficients determine specific faces.

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

$u_1 \sim u_K$

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



Reduced Space Example 1

Quiz

- 2017 Fall Final Q10
- If $u_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{bmatrix}$ and $u_2 = \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{bmatrix}$. If one original data is $x = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$. What is the new representation?

PC1 PC2

u_1, u_2, u_3 orthogonal

$u_1^T u_3 = 0$

$u_2^T u_3 = 0$

$$x' = \begin{pmatrix} u_1^T x \\ u_2^T x \end{pmatrix} = \begin{pmatrix} 3/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$$

$$u_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = u_1 \times u_2$$ cross prod

Reduced Space Example 1 Diagram

Quiz

$$x' = \begin{pmatrix} 3/\sqrt{2} \\ 1/\sqrt{2} \\ 3 \end{pmatrix} \rightarrow \boxed{\frac{3}{\sqrt{2}} \cdot u_1 + \frac{1}{\sqrt{2}} u_2} + 3 u_3$$

using 3 PCs.

$$= \underbrace{\begin{pmatrix} \frac{3}{2} \\ \frac{3}{2} \\ 0 \end{pmatrix} + \begin{pmatrix} -\frac{1}{2} \\ \frac{1}{2} \\ 0 \end{pmatrix}}_{\begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}} + \begin{pmatrix} 0 \\ 0 \\ 3 \end{pmatrix}$$

$$= \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = x$$

Reduced Space Example 2

Quiz

Q1

$\hat{\Sigma} = PDP^{-1}$

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

$m = 3$

$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ P^{-1}

- $\hat{\Sigma} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{bmatrix}$. If one original data is $x = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$. What is the new representation using only the first two principal components?

K

- A: $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$, B: $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$, C: $\begin{bmatrix} 1 \\ 3 \end{bmatrix}$, D: $\begin{bmatrix} 3 \\ 1 \end{bmatrix}$, E: $\begin{bmatrix} 2 \\ 3 \end{bmatrix}$

$PC_1 \cdot x = 2$, $PC_2 \cdot x = 1$

Reduced Space Example 3

Quiz

Q2

$u_1 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$, $u_2 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$

$x \approx \underbrace{u_1^T x}_{2}\, u_1 + \underbrace{u_2^T x}_{1}\, u_2$

- $\hat{\Sigma} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{bmatrix}$. 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: $\begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}$, C: $\begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix}$, D: $\begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix}$, E: $\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$

$\begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}$

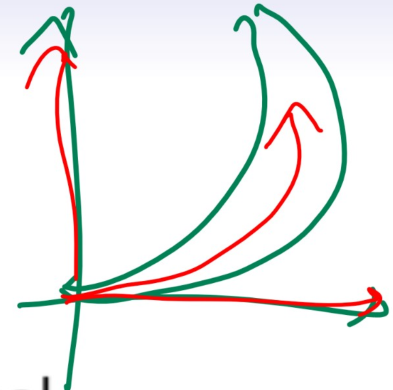
Autoencoder

Discussion

- 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.

Kernel PCA

Discussion



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

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

- The principal components can be found without explicitly computing $\varphi(x_i)$, similar to the kernel trick for support vector machines.
- Kernel PCA is a non-linear dimensionality reduction method.