CS540 Introduction to Artificial Intelligence

Lecture 16

Young Wu

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

July 29, 2022

Dark Knight Boat Game

Quiz

Q 2

$0.5 - 0.5 = 0$

$0 - 0.5 = -0.5$

-0.5

- Two groups: in person group and Zoom group.
- A : I am in person, -0.5 quiz grade to everyone on Zoom.
- B : I am in person, do nothing.
- C : I am on Zoom, -0.5 quiz grade to everyone in person.
- D : I am on Zoom, do nothing.
- If both groups vote to do nothing, both groups will get -0.5 quiz grade.

Sharing Solutions

Admin

Q2, Q3

- *M*8 is not announced. *P*4 too, but you can start.
- For sharing solutions: the important thing is writing clear solutions that help other students with homework and exams.
- Posts after the deadline and exam are not helpful.
- Posts before I cover the topic during the lecture may or may not be helpful: should use the convention in the lectures.
- If you copy another student's solution (without consent), it's considered cheating. First time: warning, second time: talk to the department.

Unsupervised Learning

Motivation

- Supervised learning: $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$.
- Unsupervised learning: $x_1, x_2, ..., x_n$.
- 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.

bag of words feature $\begin{pmatrix}1\\0\\0\end{pmatrix}\begin{pmatrix}0\\1\\0\end{pmatrix}\begin{pmatrix}0\\0\\1\end{pmatrix}$

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

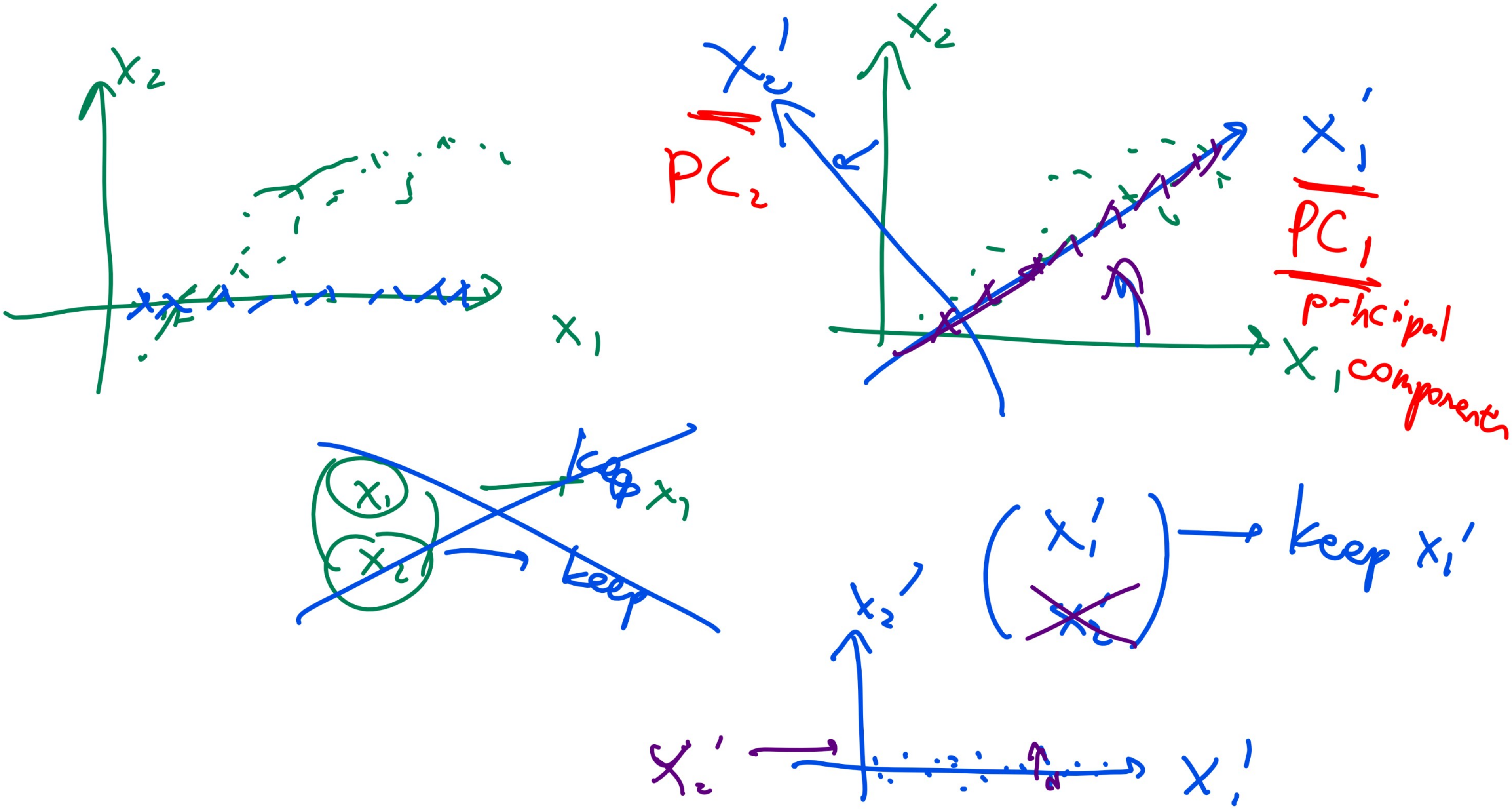
Motivation

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

1. Visualization. ← $K = 2, 3$
2. Efficient storage. ← reconstruct
3. Better generalization. ← regularization
4. Noise removal. ←

Dimension Reduction Demo

Motivation

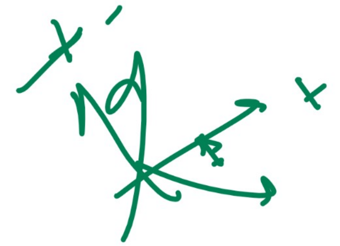


Projection

Definition

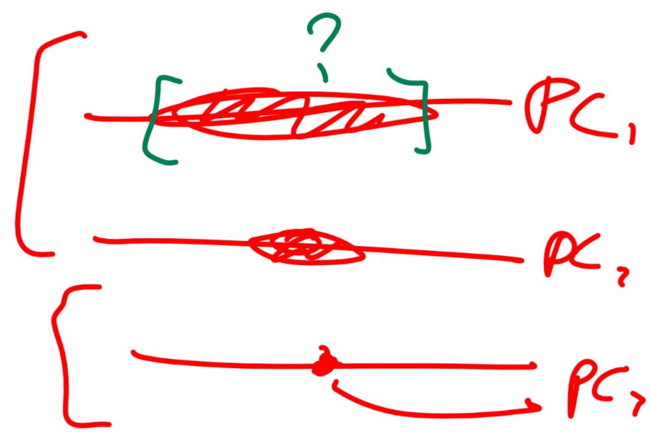
- The projection of x_i onto a unit vector u_k is the vector in the direction of u_k that is the closest to x_i.

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



- 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.

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

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

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

$$\frac{1}{n}$$

Projection Example 1

Quiz

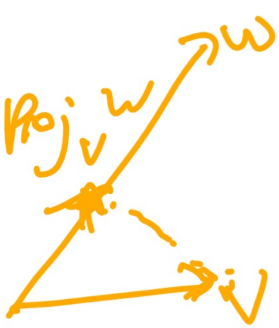
$$\frac{1}{2}\left((l_1-\mu)^2+(l_2-\mu)^2\right)$$

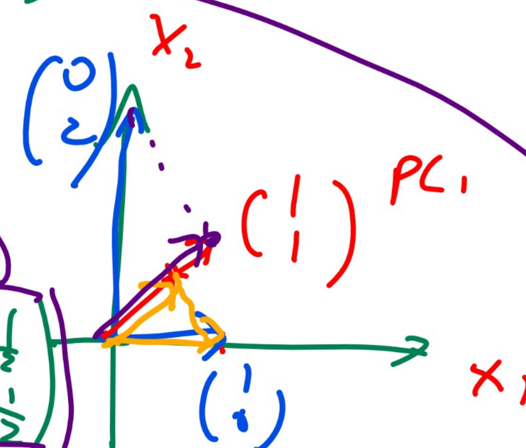
unit vector $\begin{pmatrix}\frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}}\end{pmatrix}$

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

$$\text{proj} = \frac{u^T x}{u^T u}\, u$$

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





$$\text{proj} = u^T x\, u = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)\begin{pmatrix}0\\2\end{pmatrix}\begin{pmatrix}\frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}}\end{pmatrix} = \frac{2}{\sqrt{2}}\begin{pmatrix}\frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}}\end{pmatrix} = \begin{pmatrix}1\\1\end{pmatrix}$$

Projection Example 3

Quiz

$$\frac{(1,1,1)\begin{pmatrix}1\\2\\3\end{pmatrix}}{(1,1,1)\begin{pmatrix}1\\1\\1\end{pmatrix}} = \frac{6}{3}\begin{pmatrix}1\\1\\1\end{pmatrix} = 2\begin{pmatrix}1\\1\\1\end{pmatrix} = \begin{pmatrix}2\\2\\2\end{pmatrix}$$

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

Q3

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

- $A : [2 \;\; 2 \;\; 2]^T$
- $B : [3 \;\; 3 \;\; 3]^T$
- $C : [4 \;\; 4 \;\; 4]^T$
- $D : [6 \;\; 6 \;\; 6]^T$
- E : I don't understand.

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

unit

$$u^T x \, u$$

$$u = \begin{pmatrix}1\\1\\1\end{pmatrix} \qquad \mathrm{proj}_u x$$

Projection Example 4

Quiz

- What is the projection variance of $\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$ and $\begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$ onto $\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$?
- $A : 0$
- $B : 12$
- $C : 24$
- $D : 48$
- E : *I* don't understand.

Maximum Variance Directions

Definition

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

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

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

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

u_k

Lagrange multiplier

eigenvector

eigenvalue.

$$\boxed{\max \lambda \quad \text{s.t} \quad \hat{\Sigma} u_k = \lambda u_k}$$

Eigenvalue

Definition

- The λ represents the projected variance.

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

- 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.

$$\boxed{\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_m}$$

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

Quiz

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

$$\hat{\Sigma} = \overset{P}{\begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix}} \overset{D}{\begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix}} \overset{P^{-1}}{\begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix}}$$

e.$\vec{v}$ e.v

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

Spectral Decomposition Example 2

Quiz

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

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

- $A: \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$, 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: $\begin{bmatrix} 0 \\ 0 \\ 3 \end{bmatrix}$

C_1	C_2	C_3
1	5	100
2	4	99
3	3	1
4	2	0
/4	/5	/100

$$\frac{X - \min}{\max - \min}$$

C_1	C_2	C_3
0.25	1	1
0.5	0.8	0.99
0.75	0.6	0.01
1	0.4	0

Number of Dimensions

Discussion

$$10\% \cdot \max(\underset{100\%}{M}, \underset{90\% + 95\%}{X}) + 10\% \max(\underset{0\%}{Q}, \underset{90\%}{X}) + 30\% X$$
$$+ 10\% \max(D, X) + 40\% P$$

$100\% \cdot \quad 100\%$

- 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.

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

- The new feature space is K dimensional.

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

- 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.

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

PC

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

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?

Reduced Space Example 1 Diagram

Quiz

Reduced Space Example 2

Quiz

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

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.

Autoencoder Diagram

Discussion

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.

Summary

Description

- Unsupervised learning:

1. Clustering: Hierachical.
2. Clustering: K-Means.
3. Dimensionality Reduction: Principal Component Analysis $\rightarrow$ Find varinaces $\rightarrow$ Find directions (principal components) with the largest projected variances (eigenvalues) $\rightarrow$ Find projection onto the principal direction (original points can be reconstructed).