CS540 Introduction to Artificial Intelligence
Lecture 16

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

July 22, 2021

All Pay Auction

Admin

~~A1~~ not count

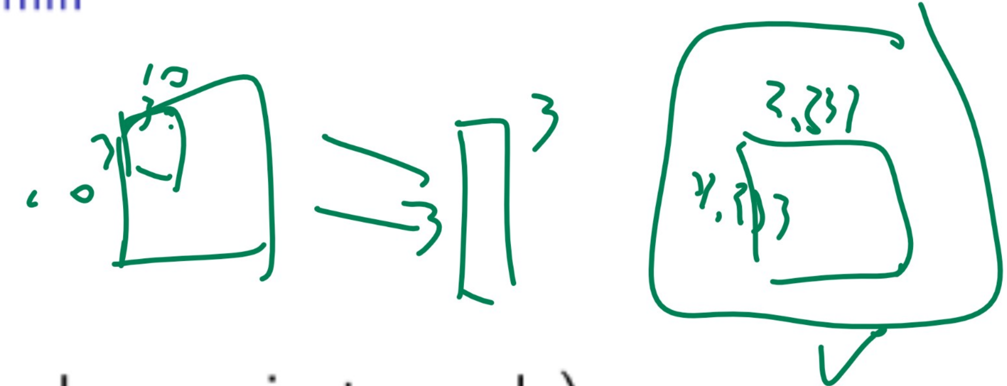
A2

- Write down a number x between 0 and 0.5 (two decimal places), you will lose x points from today's quiz grades.
- The people who wrote down the largest number will earn 0.5 bonus points.
- Any number that's not in range will be treated as 0.

Midterm Discussion

Admin

allow round up.



- Some bugs are fixed (clear cache or private mode).
- Grades are not updated on Canvas.
- No discussion session tomorrow.
- If you missed an exam, you can use P6 (in addition to P1-5) to replace the grade.

8%

one part 10%

Remind Me to Start Recording

Admin

- The messages you send in chat will be recorded: you can change your Zoom name now before I start recording.

K Means Clustering Demo

Description

min total distortion

$$\sum \|x_i - \text{center}\|$$

grad descent

Number of Clusters

Discussion

of possible y values

- There are a few ways to pick the number of clusters K.

1. K can be chosen using prior knowledge about X.
2. ~~K can be~~ the one that minimizes distortion? No, when $K = n$, distortion $= 0$.
3. K can be the one that minimizes distortion + regularizer.

features

$$K^{\star} = \arg\min_{k} \left(D_k + \lambda \cdot m \cdot k \cdot \log n \right)$$

cost

data points.

- λ is a fixed constant chosen arbitrarily.

Initial Clusters

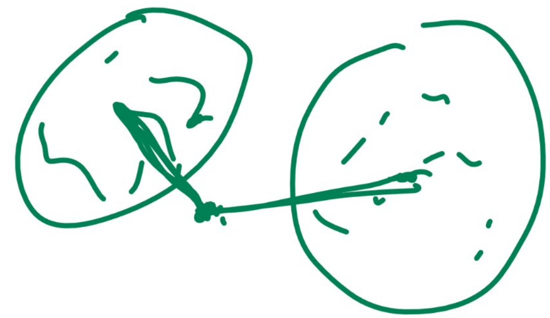
Discussion

- There are a few ways to initialize the clusters.

1. K uniform random points in $\{x_i\}_{i=1}^{n}$.
2. 1 uniform random point in $\{x_i\}_{i=1}^{n}$ as $c_1^{(0)}$, then find the farthest point in $\{x_i\}_{i=1}^{n}$ from $c_1^{(0)}$ as $c_2^{(0)}$, and find the farthest point in $\{x_i\}_{i=1}^{n}$ from the closer of $c_1^{(0)}$ and $c_2^{(0)}$ as $c_3^{(0)}$, and repeat this K times.

Gaussian Mixture Model

Discussion



- In K means, each instance belong to one cluster with certainty.
- One continuous version is called the Gaussian mixture model: each instance belongs to one of the clusters with a positive probability.
- The model can be trained using Expectation Maximization Algorithm (EM Algorithm).

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. 0, 1, 2 . . . k
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.

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

Low Dimension Representation

Motivation

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

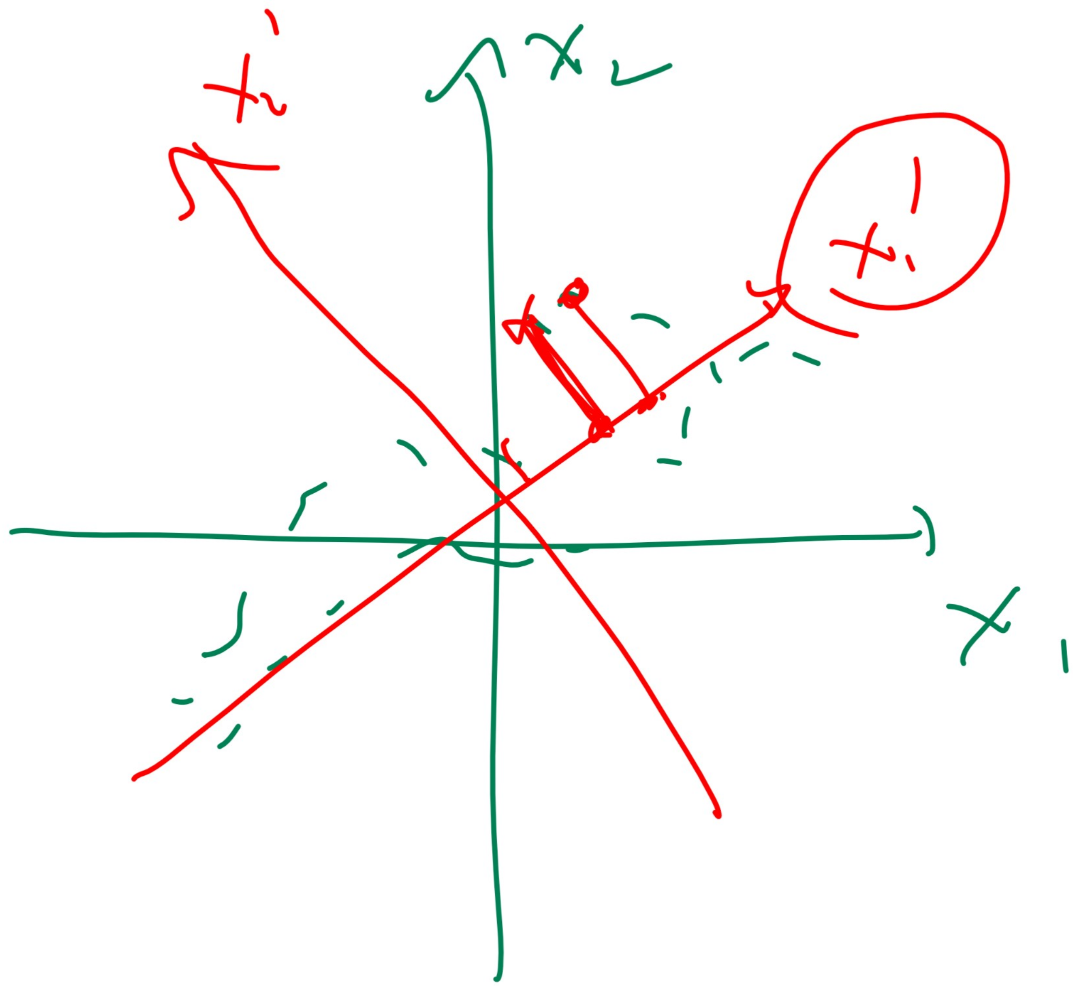
2D

1000D $\Longrightarrow$ 3D

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

Dimension Reduction Diagram

Motivation



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.

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

unit direction

length

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

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

Maximum Variance Directions

Definition

Variance matrix

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

$\frac{1}{n-1} \sum (x_i - \mu)(x_i - \mu)^T$

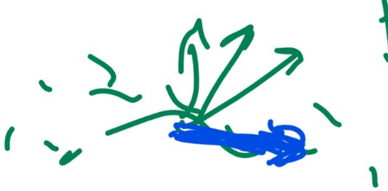
$\mu = \frac{1}{n} \sum x_i$

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

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

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

projected variance



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

u_k

variance

Eigenvalue

Definition

$u_k^T \sum u_k = u_k^T \lambda u_k$

- The λ represents the projected variance.

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

proj var $\Rightarrow \lambda$

- 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.

$$\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_m$$

$\sum u_k = \lambda u_k$

$u_1 = pc1$ $pc2$ k PC

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

SVD

Var(X)

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

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

λ_1, λ_2, λ_3

u_1 P v_2 u_3

$D \rightarrow$ eigenvalue

$P^{-1} = P^T$

$PC_1 = u_3 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ 0 \\ \frac{1}{\sqrt{2}} \end{pmatrix}$, proj var$_1 = 3$

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

Spectral Decomposition Example 2

Quiz

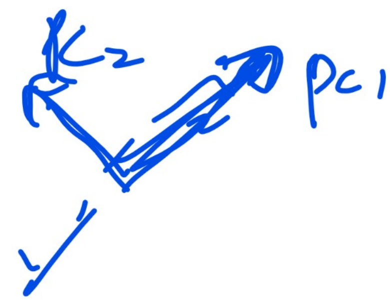
Q3

- 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: $\begin{bmatrix} 0 \\ 0 \\ 3 \end{bmatrix}$

PC1 PC2



Number of Dimensions

Discussion

PC_2 PC_1

$\lambda_2 = 0$

- 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.

$$x_i = (x_{i1}, x_{i2}, \ldots, x_{im})^T \quad \leftarrow \text{PCA}$$

1000

K

- The new feature space is K dimensional.

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

pc,

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

Eigenface

Discussion

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

PCA

$K \simeq 100$ general

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

m PC

$K+1, K+2 \ldots m$ not important

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

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

Reduced Space Example 1

Quiz

- 2017 Fall Final Q10
- If $u_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ 0 \\ \frac{1}{\sqrt{2}} \end{bmatrix}$ 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?

features

$$\left(u_1^T x, \; u_2^T x\right) = \left(\frac{4}{\sqrt{2}}, -\frac{2}{\sqrt{2}}\right)$$

Reduced Space Example 1 Diagram

Quiz

$$x \approx \frac{4}{\sqrt{2}}\begin{pmatrix}1/\sqrt{2}\\ 0\\ 1/\sqrt{2}\end{pmatrix} + \left(-\frac{2}{\sqrt{2}}\right)\begin{pmatrix}1/\sqrt{2}\\ 0\\ -1/\sqrt{2}\end{pmatrix}$$

$$= \qquad \approx \begin{pmatrix}1\\ 2\\ 3\end{pmatrix}$$

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 new representation using only the first two principal components?
- 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: $\begin{bmatrix} 2 \\ 3 \end{bmatrix}$

Reduced Space Example 3

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: $\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$

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.