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

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

Random Choice 2

Quiz

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

Random Choice 3

Quiz

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

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.

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.

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

- 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.

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

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

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

Projection Example 1

Quiz

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

Projection Example 3

Quiz

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

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.

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

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

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

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

Quiz

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

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

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

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.

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

- The new feature space is K dimensional.

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

- 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.

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

- 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.