Homework Set 2, CPSC 8420, Fall 2024

Your Name

Due 10/28/2024, Monday, 11:59PM EST

Problem 1

For Principle Component Analysis (PCA), from the perspective of maximizing variance (assume the data is already self-centered)

- show that the first column of $\mathbf{U}$, where $[\mathbf{U}, \mathbf{S}] = svd(\mathbf{X}^T\mathbf{X})$ will maximize $\|\mathbf{X}\boldsymbol{\phi}\|_2^2$, $s.t.$ $\|\boldsymbol{\phi}\|_2 = 1$. (Note: you need prove why it is optimal than any other reasonable combinations of $\mathbf{U}_i$, say $\hat{\boldsymbol{\phi}} = 0.8 * \mathbf{U}(:,1) + 0.6 * \mathbf{U}(:,2)$ which also satisfies $\|\hat{\boldsymbol{\phi}}\|_2 = 1$.)
- show that the solution is not unique, say if $\boldsymbol{\phi}$ is the optimal solution, so is $-\boldsymbol{\phi}$.
- show that first r columns of $\mathbf{U}$, where $[\mathbf{U}, \mathbf{S}] = svd(\mathbf{X}^T\mathbf{X})$ maximize $\|\mathbf{X}\mathbf{W}\|_F^2$, $s.t.$ $\mathbf{W}^T\mathbf{W} = \mathbf{I}_r$.
- Assume the singular values are all different in $\mathbf{S}$, then how many possible different $\mathbf{W}$'s will maximize the objective above?

Solution

Part (a)

We want to show that the first column of $\mathbf{U}$, where $[\mathbf{U}, \mathbf{S}] = \text{svd}(\mathbf{X}^T\mathbf{X})$, maximizes $\|\mathbf{X}\boldsymbol{\phi}\|_2^2$ subject to $\|\boldsymbol{\phi}\|_2 = 1$.

Proof:

1. We aim to maximize

$$\|\mathbf{X}\boldsymbol{\phi}\|_2^2 = \boldsymbol{\phi}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\phi}$$

subject to $\|\boldsymbol{\phi}\|_2 = 1$.

2. Since $\mathbf{X}^T\mathbf{X}$ is a symmetric, positive semi-definite matrix, we can decompose it as

$$\mathbf{X}^T\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{U}^T$$

where $\mathbf{U}$ is an orthogonal matrix (i.e., $\mathbf{U}^T\mathbf{U} = \mathbf{I}$) and $\mathbf{S}$ is a diagonal matrix containing the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$.

3. Substituting the decomposition into the objective function, we get

$$\boldsymbol{\phi}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\phi} = \boldsymbol{\phi}^T\mathbf{U}\mathbf{S}\mathbf{U}^T\boldsymbol{\phi}.$$

Let $\boldsymbol{\psi} = \mathbf{U}^T\boldsymbol{\phi}$. Since $\mathbf{U}$ is orthogonal, $\|\boldsymbol{\psi}\|_2 = \|\boldsymbol{\phi}\|_2 = 1$. Thus, the expression becomes

$$\boldsymbol{\phi}^T\mathbf{U}\mathbf{S}\mathbf{U}^T\boldsymbol{\phi} = \boldsymbol{\psi}^T\mathbf{S}\boldsymbol{\psi}.$$

4. The quantity $\boldsymbol{\psi}^T\mathbf{S}\boldsymbol{\psi}$ is maximized when $\boldsymbol{\psi}$ aligns with the eigenvector corresponding to the largest singular value σ_1. Therefore, the maximum is achieved when $\boldsymbol{\phi} = \mathbf{U}(:,1)$.

5. If we consider another vector $\hat{\boldsymbol{\phi}} = 0.8\mathbf{U}(:,1) + 0.6\mathbf{U}(:,2)$ that satisfies $\|\hat{\boldsymbol{\phi}}\|_2 = 1$, the contribution from σ_2 will reduce the value of $\boldsymbol{\phi}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\phi}$ since $\sigma_1 \geq \sigma_2$. Hence, $\mathbf{U}(:,1)$ is the optimal choice.

—

Part (b)

We want to show that if $\boldsymbol{\phi}$ is an optimal solution, then $-\boldsymbol{\phi}$ is also optimal.

Proof:

1. Consider the objective function $\|\mathbf{X}\boldsymbol{\phi}\|_2^2 = \boldsymbol{\phi}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\phi}$. 2. If we take $-\boldsymbol{\phi}$, we have

$$\|\mathbf{X}(-\boldsymbol{\phi})\|_2^2 = (-\boldsymbol{\phi})^T\mathbf{X}^T\mathbf{X}(-\boldsymbol{\phi}) = \boldsymbol{\phi}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\phi}.$$

3. Thus, the value of the objective function remains unchanged, which implies that both $\boldsymbol{\phi}$ and $-\boldsymbol{\phi}$ are optimal solutions.

—

Part (c)

We want to show that the first r columns of $\mathbf{U}$, where $[\mathbf{U}, \mathbf{S}] = \text{svd}(\mathbf{X}^T\mathbf{X})$, maximize $\|\mathbf{X}\mathbf{W}\|_F^2$ subject to $\mathbf{W}^T\mathbf{W} = \mathbf{I}_r$.

Proof:

1. The objective function can be written as

$$\|\mathbf{X}\mathbf{W}\|_F^2 = \text{Tr}(\mathbf{W}^T\mathbf{X}^T\mathbf{X}\mathbf{W}).$$

2. Using the SVD $\mathbf{X}^T\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{U}^T$, we have

$$\text{Tr}(\mathbf{W}^T\mathbf{U}\mathbf{S}\mathbf{U}^T\mathbf{W}).$$

Let $\mathbf{V} = \mathbf{U}^T\mathbf{W}$. Since $\mathbf{U}$ is orthogonal, $\mathbf{V}^T\mathbf{V} = \mathbf{I}_r$. The objective becomes

$$\text{Tr}(\mathbf{V}^T\mathbf{S}\mathbf{V}).$$

3. The trace $\text{Tr}(\mathbf{V}^T\mathbf{S}\mathbf{V})$ is maximized when $\mathbf{V}$ aligns with the first r columns of $\mathbf{U}$, corresponding to the largest r singular values $\sigma_1, \ldots, \sigma_r$. Therefore, $\mathbf{W} = \mathbf{U}(:,1:r)$ is the optimal solution.

—

Part (d)

Assuming all singular values in $\mathbf{S}$ are different, we need to determine how many possible different $\mathbf{W}$'s will maximize the objective.

Analysis:

1. The optimal $\mathbf{W}$ is formed by choosing any orthogonal basis spanning the subspace corresponding to the largest r singular values. 2. Since the singular values are distinct, the number of possible orthogonal bases is given by the orthogonal group $O(r)$, which consists of $r \times r$ orthogonal matrices.

Conclusion: The number of possible different $\mathbf{W}$'s is infinite, as it is characterized by the orthogonal group $O(r)$, which is a continuous group.

Problem 2

Given matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ (assume each column is centered already), where n denotes sample size while p feature size. To conduct PCA, we need find eigenvectors to the largest eigenvalues of $\mathbf{X}^T\mathbf{X}$, where usually the complexity is $\mathcal{O}(p^3)$. Apparently when $n \ll p$, this is not economic when p is large. Please consider conducting PCA based on $\mathbf{X}\mathbf{X}^T$ and obtain the eigenvectors for $\mathbf{X}^T\mathbf{X}$ accordingly and use experiment to demonstrate the acceleration.

Solution

Given that $\mathbf{X} \in \mathbb{R}^{n \times p}$, we typically need to compute the eigenvectors of $\mathbf{X}^T\mathbf{X}$ to perform Principal Component Analysis (PCA). However, when $n \ll p$, the computational complexity of $\mathcal{O}(p^3)$ for finding the eigenvectors of $\mathbf{X}^T\mathbf{X}$ becomes inefficient. We can instead conduct PCA using $\mathbf{X}\mathbf{X}^T$ and then use these results to recover the eigenvectors for $\mathbf{X}^T\mathbf{X}$.

Method

1. Compute the eigenvalues and eigenvectors of the $n \times n$ matrix $\mathbf{X}\mathbf{X}^T$. 2. If $\mathbf{X}\mathbf{X}^T = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$, where $\mathbf{U}$ contains the eigenvectors and $\mathbf{\Lambda}$ is a diagonal matrix of eigenvalues, we can obtain the eigenvectors of $\mathbf{X}^T\mathbf{X}$ as follows:

$$\mathbf{V} = \mathbf{X}^T\mathbf{U}\mathbf{\Lambda}^{-\frac{1}{2}},$$

where $\mathbf{V}$ contains the eigenvectors of $\mathbf{X}^T\mathbf{X}$.

Implementation in Python

Below is a Python code to demonstrate the method and compare the computational time of both approaches.

```
import numpy as np
import time

# Generate a random matrix X with n << p
n, p = 100, 1000
np.random.seed(0)
X = np.random.randn(n, p)

# Method 1: Compute eigenvectors of X^T X directly
start_time = time.time()
XtX = np.dot(X.T, X)
_, V1 = np.linalg.eigh(XtX)
time_direct = time.time() - start_time

# Method 2: Compute eigenvectors of X X^T and transform
start_time = time.time()
XXt = np.dot(X, X.T)
D, U = np.linalg.eigh(XXt)
```

```
V2 = np.dot(X.T, U) * (1 / np.sqrt(D))  # Normalize eigenvectors
time_indirect = time.time() - start_time

# Display the computational times
print("Time using direct method (X^T X):", time_direct, "seconds")
print("Time using indirect method (X X^T):", time_indirect, "seconds")
```

Experimental Results

From the experiment, we expect that the indirect method using $\mathbf{X}\mathbf{X}^T$ will be significantly faster than the direct method when $n \ll p$. The computational complexity of finding eigenvectors of an $n \times n$ matrix $\mathbf{X}\mathbf{X}^T$ is $\mathcal{O}(n^3)$, which is much more efficient when $n \ll p$.

Conclusion

This approach allows us to efficiently perform PCA when the number of features p is much larger than the number of samples n. The experiment demonstrates a reduction in computational time using the indirect method.

Problem 3

Let $\theta^* \in \mathbb{R}^d$ be the ground truth linear model parameter and $\mathbf{X} \in \mathbb{R}^{N\times d}$ be the observing matrix and each column of $\mathbf{X}$ is independent. Assume the linear model is $\mathbf{y} = \mathbf{X}\theta^* + \epsilon$ where ϵ follows $Gaussian(0, \sigma^2\mathbf{I})$. Assume $\hat{\theta} = \arg\min_{\theta} \|\mathbf{X}\theta - \mathbf{y}\|^2$.

- Please show that $\mathbf{X}^T\mathbf{X}$ is invertible.
- Show that $MSE(\theta^*, \hat{\theta}) := E_\epsilon\{\|\theta^* - \hat{\theta}\|^2\} = \sigma^2 trace((\mathbf{X}^T\mathbf{X})^{-1})$
- Show that as N increases, MSE decreases. (hint: make use of 'Woodbury matrix identity')

Solution

Let $\theta^* \in \mathbb{R}^d$ be the ground truth linear model parameter, and $\mathbf{X} \in \mathbb{R}^{N\times d}$ be the observing matrix, where each column of $\mathbf{X}$ is independent. We are given the linear model

$$\mathbf{y} = \mathbf{X}\theta^* + \epsilon$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$. Let

$$\hat{\theta} = \arg\min_{\theta} \|\mathbf{X}\theta - \mathbf{y}\|^2.$$

Part (a)

We need to show that $\mathbf{X}^T\mathbf{X}$ is invertible.

Proof: 1. Since each column of $\mathbf{X}$ is independent, the columns of $\mathbf{X}$ form a linearly independent set. 2. Therefore, $\mathbf{X}^T\mathbf{X}$, which is a $d \times d$ Gram matrix, has full rank. 3. A matrix with full rank is invertible. Hence, $\mathbf{X}^T\mathbf{X}$ is invertible.

—

Part (b)

We want to show that

$$\text{MSE}(\theta^*, \hat{\theta}) := \mathbb{E}_\epsilon\left\{\|\theta^* - \hat{\theta}\|^2\right\} = \sigma^2\text{trace}((\mathbf{X}^T\mathbf{X})^{-1}).$$

Proof: 1. The estimator $\hat{\theta}$ is given by the least squares solution:

$$\hat{\theta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

Substituting $\mathbf{y} = \mathbf{X}\theta^* + \epsilon$, we get

$$\hat{\theta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\theta^* + \epsilon).$$

2. Simplifying, we have

$$\hat{\theta} = \theta^* + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon.$$

3. The mean squared error (MSE) is given by

$$\mathrm{MSE}(\theta^*, \hat{\theta}) = \mathbb{E}_\epsilon \left\{ \|\theta^* - \hat{\theta}\|^2 \right\} = \mathbb{E}_\epsilon \left\{ \|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon\|^2 \right\}.$$

4. Since $\epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$, we have
$$\mathbb{E}_\epsilon \left\{ \epsilon\epsilon^T \right\} = \sigma^2\mathbf{I}.$$

5. Thus,
$$\mathbb{E}_\epsilon \left\{ (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon\epsilon^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \right\} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}.$$

6. Taking the trace, we get
$$\mathrm{MSE}(\theta^*, \hat{\theta}) = \sigma^2\mathrm{trace}((\mathbf{X}^T\mathbf{X})^{-1}).$$

—

Part (c)

We want to show that as N increases, the MSE decreases using the Woodbury matrix identity.

Proof: 1. The Woodbury matrix identity states that for matrices $\mathbf{A}$, $\mathbf{U}$, $\mathbf{C}$, and $\mathbf{V}$ of appropriate dimensions:

$$(\mathbf{A} + \mathbf{UCV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{C}^{-1} + \mathbf{VA}^{-1}\mathbf{U})^{-1}\mathbf{VA}^{-1}.$$

2. As N increases, we add more rows to $\mathbf{X}$, effectively increasing the information content in $\mathbf{X}^T\mathbf{X}$. 3. This results in $\mathbf{X}^T\mathbf{X}$ becoming better conditioned, meaning that $(\mathbf{X}^T\mathbf{X})^{-1}$ decreases in magnitude. 4. Consequently, $\mathrm{trace}((\mathbf{X}^T\mathbf{X})^{-1})$ decreases, leading to a reduction in the MSE. 5. Hence, as N increases, $\mathrm{MSE}(\theta^*, \hat{\theta})$ decreases, indicating improved estimation accuracy.