EECE 568 Assignment 1

Jason Li (37727179)

October 11, 2023

1 Math Review - 36 points

1.1 Eigenvalue - 6 points

Proof 1: Eigenvalues of A and A^T

Let A be an $m \times n$ matrix where $m = n$. We want to show that A and A^T have the same set of eigenvalues.

Let λ be an eigenvalue of A with an associated eigenvector v. This means:

$$A \cdot v = \lambda \cdot v$$

The eigenvalue equation for A^T:

$$A^T \cdot u = \mu \cdot u$$

where μ is an eigenvalue of A^T, and u is the associated eigenvector.

Taking the transpose of the equation $A \cdot v = \lambda \cdot v$, we get:

$$v^T \cdot A^T = \lambda \cdot v^T$$

Then multiply both sides of the equation by u:

$$v^T \cdot A^T \cdot u = \lambda \cdot v^T \cdot u$$

We also know that $A^T \cdot u = \mu \cdot u$, so we can substitute that in:

$$v^T \cdot (\mu \cdot u) = \lambda \cdot v^T \cdot u$$

Now, since μ is just a scalar because it is an eigenvalue, we can pull it out of the left side of the equation:

$$\mu \cdot v^T \cdot u = \lambda \cdot v^T \cdot u$$

Now, cancel out $v^T \cdot u$ on both sides of the equation (assuming $v^T \cdot u$ is nonzero, which is true for nonzero eigenvalues):

$$\mu = \lambda$$

This shows that if λ is an eigenvalue of A, it is also an eigenvalue of A^T, and vice versa. Therefore, A and A^T have the same set of eigenvalues when $m = n$.

Proof 2: Eigenvalues of AB and BA

Let's prove this statement by contradiction. Suppose AB and BA do not have the same set of eigenvalues. This means there exists an eigenvalue λ of AB such that λ is not an eigenvalue of BA.

Let v be the associated eigenvector of λ for AB:

$$AB \cdot v = \lambda \cdot v$$

Now, let's consider the equation for BA:

$$BA \cdot w = \mu \cdot w$$

where μ is an eigenvalue of BA, and w is the associated eigenvector.

Now, multiply both sides of the equation $BA \cdot w = \mu \cdot w$ by v from the left:

$$AB \cdot (vw) = \mu \cdot (vw)$$

Since λ is an eigenvalue of AB, we know that $AB \cdot v = \lambda \cdot v$. So, we can rewrite the equation as:

$$\lambda \cdot (vw) = \mu \cdot (vw)$$

Divide both sides by (vw):

$$\lambda = \mu$$

This implies that the eigenvalue λ of AB is also an eigenvalue of BA, which contradicts our initial assumption that λ is not an eigenvalue of BA.

Therefore, our assumption that AB and BA do not have the same set of eigenvalues must be false. Hence, we conclude that AB and BA have the same set of eigenvalues.

1.2 Rank - 6 points

Let A be a full-rank matrix in $\mathbb{R}^{n \times n}$. We want to prove that matrices B and $A^{-1}BA$ have the same set of eigenvalues.

First, let λ be an eigenvalue of B, and let $\mathbf{v}$ be the corresponding eigenvector. This means that $B\mathbf{v} = \lambda\mathbf{v}$.

Now, consider $A^{-1}BA$. We can multiply both sides of the equation by A^{-1} from the left:

$$A^{-1}(A^{-1}BA)\mathbf{v} = A^{-1}(\lambda\mathbf{v})$$
$$(A^{-1}BA)(A^{-1}\mathbf{v}) = \lambda(A^{-1}\mathbf{v})$$

Let $\mathbf{w} = A^{-1}\mathbf{v}$. Then, we have

$$(A^{-1}BA)\mathbf{w} = \lambda\mathbf{w}$$

This shows that λ is also an eigenvalue of $A^{-1}BA$ with the corresponding eigenvector $\mathbf{w}$.

Since we've shown that any eigenvalue of B is also an eigenvalue of $A^{-1}BA$, we can repeat the same argument in the reverse direction to show that any eigenvalue of $A^{-1}BA$ is also an eigenvalue of B.

Therefore, matrices B and $A^{-1}BA$ have the same set of eigenvalues.

1.3 Definite & Semidefinite - 6 points

Statement 1: If every eigenvalue of matrix A is positive, then A is a positive definite matrix Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix with all positive eigenvalues. Then, A is a positive definite matrix.

Suppose A has all positive eigenvalues. Let $x \in \mathbb{R}^n$ be a non-zero vector. Since A is symmetric, it can be diagonalized as $A = PDP^T$, where P is orthogonal and D is a diagonal matrix with positive eigenvalues.

Now, let $y = P^T x$. Since P is orthogonal, it preserves inner products, so $y^T y = x^T x$. We can express $x^T Ax$ as follows:

$$x^T Ax = x^T (PDP^T)x = (Px)^T D(Px) = y^T Dy \tag{1}$$

Since D is a diagonal matrix with all positive entries, and y is a non-zero vector, $y^T Dy$ is the sum of products of positive numbers, and therefore, it is positive.

Thus, we have shown that for any non-zero vector x, $x^T Ax > 0$, which means that matrix A is positive definite.

Statement 2: If matrix A is positive definite, it is full rank and invertible If matrix A is positive definite, it is full rank and invertible.

Let's prove this statement by contrapositive. Suppose A is not full rank. That means there exists a non-zero vector x such that $Ax = 0$ because the nullity of A is non-zero.

Now, consider $x^T Ax$. We have:

$$x^T Ax = x^T 0 = 0 \tag{2}$$

Since $x^T Ax = 0$ for a non-zero vector x, this contradicts the definition of a positive definite matrix, which requires that $x^T Ax$ must be greater than zero for all non-zero vectors x.

Therefore, if matrix A is not full rank, it cannot be positive definite.

Conversely, if A is full rank, it means that the null space of A is trivial (i.e., it contains only the zero vector), and this implies that $Ax = 0$ only has the trivial solution ($x = 0$). In other words, A is invertible because it has a unique solution for every b in $Ax = b$. Additionally, since all eigenvalues of A are positive (as given in Statement 1), it follows that A is positive definite.

Thus, we have proven that if A is positive definite, it is also full rank and invertible.

1.4 Gram Matrix - 6 points

Statement 1: G is always positive semidefinite

Let $A \in \mathbb{R}^{m \times n}$ be a matrix. We want to show that $G = A^T A$ is positive semidefinite.

For any vector $x \in \mathbb{R}^n$, we can consider the quadratic form:

$$x^T Gx = x^T (A^T A)x = (Ax)^T (Ax) = \|Ax\|^2 \geq 0$$

Since $x^T Gx$ is non-negative for all $x \in \mathbb{R}^n$, this implies that $G = A^T A$ is positive semidefinite.

Statement 2: If $m \geq n$ and A is full rank, then G is positive definite

Now, let's prove that if $m \geq n$ and A is full rank, then $G = A^T A$ is positive definite.

First, we know that A being full rank implies that the columns of A are linearly independent.

Consider any nonzero vector $x \in \mathbb{R}^n$. We want to show that $x^T G x > 0$.

Since A is full rank, we can consider the linear transformation $y = Ax$. Since A is $m \times n$, y is in $\mathbb{R}^m$, and because A is full rank, y can take any nonzero vector in $\mathbb{R}^m$.

Now, we have:

$$x^T G x = x^T (A^T A) x = (Ax)^T (Ax) = \|Ax\|^2 = \|y\|^2$$

Since y can take any nonzero vector in $\mathbb{R}^m$, and the norm $\|y\|^2$ is always positive for nonzero vectors, we can conclude that $x^T G x > 0$ for all nonzero $x \in \mathbb{R}^n$.

Therefore, $G = A^T A$ is positive definite.

1.5 Matrix Derivative - 6 points

Jacobian Matrix of f:

The function $f(x, y)$ is defined as follows:

$$f(x, y) = \begin{bmatrix} f_1(x, y) \\ f_2(x, y) \\ f_3(x, y) \end{bmatrix} = \begin{bmatrix} x^2 y^3 \\ 4x^2 + \cos(y) \\ 4y^2 - 2x^2 \end{bmatrix}$$

The Jacobian matrix of f is given by:

$$J_f = \begin{bmatrix} \frac{\partial f_1}{\partial x} & \frac{\partial f_1}{\partial y} \\ \frac{\partial f_2}{\partial x} & \frac{\partial f_2}{\partial y} \\ \frac{\partial f_3}{\partial x} & \frac{\partial f_3}{\partial y} \end{bmatrix}$$

The dimensions of the Jacobian matrix J_f in (row, column) format are (3, 2) because there are 3 rows corresponding to the 3 components of f and 2 columns corresponding to the partial derivatives with respect to x and y.

Partial Derivatives:

Partial Derivative of f_2 with respect to y:

$$\frac{\partial f_2}{\partial y} = -\sin(y)$$

Second Mixed Partial Derivative of f_1 with respect to x and y:

$$\frac{\partial^2 f_1}{\partial x \partial y} = \frac{\partial}{\partial x}\left(\frac{\partial f_1}{\partial y}\right) = \frac{\partial}{\partial x}\left(3x^2 y^2\right) = 6xy^2$$

1.6 Push-through Identity - 6 points

Given that A is an $m \times n$ matrix, B is an $n \times m$ matrix, and the $m \times m$ matrix $(\lambda I + AB)$ is invertible.

Part 1: Invertibility of $(\lambda I + BA)$

We want to prove that the $n \times n$ matrix $(\lambda I + BA)$ is invertible.

Consider the equation $(\lambda I + BA)x = 0$, where x is an $n \times 1$ vector. We will show that this implies $(\lambda I + AB)y = 0$, where $y = Ax$.

Multiplying both sides of $(\lambda I + BA)x = 0$ by A, we get:

$$A(\lambda I + BA)x = A \cdot 0 = 0$$

Now, let $y = Ax$. Therefore, we have:

$$(\lambda I + AB)y = (\lambda I + AB)(Ax)$$

Since $\lambda I + AB$ is invertible, we can multiply both sides by its inverse:

$$(\lambda I + AB)^{-1}(\lambda I + AB)(Ax) = (\lambda I + AB)^{-1} \cdot 0 = 0$$

So, $(\lambda I + AB)y = 0$.

Now, suppose $(\lambda I + BA)$ is not invertible. Then, there exists a nonzero vector x such that $(\lambda I + BA)x = 0$. But we have shown that this implies $(\lambda I + AB)y = 0$, where $y = Ax$. This contradicts the invertibility of $(\lambda I + AB)$.

Therefore, if $(\lambda I + AB)$ is invertible, then $(\lambda I + BA)$ must also be invertible.

Part 2: The Push-Through Identity

Now, let's prove the push-through identity $B(\lambda I + AB)^{-1} = (\lambda I + BA)^{-1}B$.

We will start by proving that $B(\lambda I + AB) = (\lambda I + BA)B$.

Multiplying both sides of $(\lambda I + AB)y = 0$ by B, we get:

$$B(\lambda I + AB)y = B \cdot 0 = 0$$

Now, let $z = By$. Therefore, we have:

$$(\lambda I + BA)z = (\lambda I + BA)(By)$$

Since $\lambda I + BA$ is invertible, we can multiply both sides by its inverse:

$$(\lambda I + BA)^{-1}(\lambda I + BA)(By) = (\lambda I + BA)^{-1} \cdot 0 = 0$$

So, $(\lambda I + BA)z = 0$.

This shows that $B(\lambda I + AB) = (\lambda I + BA)B$.

Now, let's multiply both sides of the equation by $(\lambda I + AB)^{-1}$ from the right to obtain the push-through identity:

$$B(\lambda I + AB)(\lambda I + AB)^{-1} = (\lambda I + BA)B(\lambda I + AB)^{-1}$$

Since we know that $B(\lambda I + AB) = (\lambda I + BA)B$, we can simplify:

$$(\lambda I + BA)(\lambda I + AB)^{-1} = B(\lambda I + AB)(\lambda I + AB)^{-1}$$

Now, we can multiply both sides by $(\lambda I + AB)^{-1}$ from the left:

$$(\lambda I + BA) = B(\lambda I + AB)(\lambda I + AB)^{-1}$$

Therefore, we have shown that $B(\lambda I + AB)^{-1} = (\lambda I + BA)^{-1}B$, which is the push-through identity.

2 Theoretical Questions - 25 points

2.1 Regression Coefficients - 5 points

Answer:

The correct answer is (a) This feature has a strong effect on the model (should be retained).

Explanation:

In linear regression, the coefficient of a feature represents the change in the target variable (dependent variable) for a one-unit change in that feature, while holding all other features constant. A high negative coefficient indicates that as this feature decreases, the target variable tends to increase.

Therefore, a feature with a relatively high negative coefficient has a strong effect on the model. It means that this feature is negatively correlated with the target variable and is providing important information for predicting the target variable. It should be retained in the model because removing it may lead to a loss of valuable predictive power.

Option (b) is incorrect because a high negative coefficient indicates the opposite - that the feature is indeed impactful.

Option (c) is also incorrect because, based on the information provided (a relatively high negative coefficient), we can make reasonable conclusions about the importance of the feature without additional information.

2.2 Independent Features - 10 points

Part (a): Training on One Feature

Given the target value vector $Y \in \mathbb{R}^n$ and the data samples matrix $X \in \mathbb{R}^{n \times d}$, we want to show that when training a regressor on just one of the features (say, the j-th feature), we have $w_j = x_j^T Y$, where w_j is the j-th element of the parameter vector.

Let's consider the regressor that is trained on only the j-th feature:

$$Y = X_j w_j + \text{error}$$

Where: - Y is the target vector. - X_j is the j-th feature column vector from X. - w_j is the weight (parameter) associated with the j-th feature. - "error" represents the error term.

We want to find the value of w_j.

Solving for w_j:

$$w_j = \frac{X_j^T Y}{X_j^T X_j}$$

Now, $w_j = x_j^T Y$ where x_j is the j-th column of matrix X.

So, we have shown that when training on just one feature, $w_j = x_j^T Y$.

Part (b): Orthogonal Columns in Matrix X

Suppose the columns of matrix X are orthogonal. We want to prove that the optimal parameters obtained by training the regressor on all features are the same as the optimal parameters obtained by training on each feature independently.

Let W be the parameter vector when training on all features, and W_j be the parameter vector when training on just the j-th feature. Also, let X_j be the j-th column of matrix X.

The objective of linear regression is to minimize the sum of squared errors:

$$\text{minimize } \sum_{i=1}^{n}(y_i - X_i^T W)^2$$

Where y_i is the i-th element of vector Y, and X_i is the i-th row of matrix X.

When the columns of X are orthogonal, the cross-terms in the expansion of the squared error vanish because orthogonal vectors have zero dot products:

$$X_i^T X_j = 0 \text{ for } i \neq j$$

Now, consider training the regressor on just the j-th feature:

$$\text{minimize } \sum_{i=1}^{n}(y_i - X_j^T W_j)^2$$

Since the columns are orthogonal, the sum can be simplified as follows:

$$\sum_{i=1}^{n}(y_i - X_j^T W_j)^2 = \sum_{i=1}^{n}(y_i - x_j^T W_j)^2$$

This is equivalent to training a regressor on just the j-th feature independently.

Therefore, when the columns of matrix X are orthogonal, the optimal parameters obtained by training the regressor on all features (W) are the same as the optimal parameters obtained by training on each feature independently (W_j).

2.3 Regularization - 10 points

Part (a): Invertibility of $(X^T X + \lambda I_d)$

We want to determine if the matrix $(X^T X + \lambda I_d)$ is always invertible. If yes, we will prove it; otherwise, we will identify the conditions under which it is invertible.

The matrix $(X^T X + \lambda I_d)$ is invertible if and only if it is positive definite. To ensure its invertibility, the matrix must satisfy the following conditions:

1. λ (the regularization parameter) must be greater than zero. 2. The columns of matrix X must be linearly independent.

Explanation:

1. If $\lambda > 0$, then λI_d is a positive definite matrix, and adding a positive definite matrix to another positive definite matrix results in a positive definite matrix. Therefore, $(X^T X + \lambda I_d)$ is positive definite.

2. If the columns of matrix X are linearly independent, then $X^T X$ is positive definite. Adding a positive definite matrix to another positive definite matrix results in a positive definite matrix. Therefore, $(X^T X + \lambda I_d)$ is positive definite.

In conclusion, $(X^T X + \lambda I_d)$ is invertible if $\lambda > 0$ and the columns of matrix X are linearly independent.

Part (b): Efficient Expression for w when $n \ll d$

When $n \ll d$, it is computationally more efficient to calculate an equivalent expression for w by working with an $n \times n$ matrix instead of a $d \times d$ matrix. We can use the push-through identity to derive the efficient expression for w.

Given the minimization problem:

$$\arg\min_{w \in \mathbb{R}^d} \frac{1}{2}\|Xw - y\|_2^2 + \frac{\lambda}{2}\|w\|_2^2$$

Taking the gradient with respect to w and setting it to zero, we have:

$$X^T(Xw - y) + \lambda I_d w = 0$$

Now, we want to express w efficiently in terms of X and y. Using the push-through identity, we can rewrite the equation as follows:

$$X^T X w - X^T y + \lambda I_d w = 0$$

Grouping terms with w on one side:

$$(X^T X + \lambda I_d)w = X^T y$$

Now, we can solve for w:

$$w = (X^T X + \lambda I_d)^{-1} X^T y$$

This expression allows us to compute w using an $n \times n$ matrix $(X^T X + \lambda I_n)$ instead of the larger $d \times d$ matrix, which is more computationally efficient when $n \ll d$.