Assignment 12

Yehonatan Keypur

January 12, 2025

Question 1 - Linear Regression (Rank of X) (Problem 3)

Part (a)

We aim to prove the relationships between the range space $\mathcal{R}(X)$, nullspace $\mathcal{N}(X)$, and their orthogonal complements:

1. $\mathcal{N}(X) = \mathcal{R}(X^T)^\perp$
2. $\mathcal{N}(X^T) = \mathcal{R}(X)^\perp$
3. $\mathcal{R}(X) = \mathcal{N}(X^T)^\perp$
4. $\mathcal{R}(X^T) = \mathcal{N}(X)^\perp$

Preliminaries

1. **Definition of Range Space $\mathcal{R}(X)$:**
$$\mathcal{R}(X) = \{Xa \mid a \in \mathbb{R}^{d+1}\}.$$
2. **Definition of Nullspace $\mathcal{N}(X)$:**
$$\mathcal{N}(X) = \{a \in \mathbb{R}^{d+1} \mid Xa = 0\}.$$
3. **Orthogonal Complement:** For a subspace $\mathcal{L} \subset \mathbb{R}^N$, its orthogonal complement is defined as:
$$\mathcal{L}^\perp = \{a \in \mathbb{R}^N \mid a^T b = 0, \text{ for all } b \in \mathcal{L}\}.$$

Proof of $\mathcal{N}(X) = \mathcal{R}(X^T)^\perp$:

1. By definition, $a \in \mathcal{N}(X)$ implies $Xa = 0$.
2. For any $b \in \mathcal{R}(X^T)$, there exists $c \in \mathbb{R}^n$ such that $b = X^T c$.
3. Consider the inner product:
$$a^T b = a^T (X^T c) = (Xa)^T c = 0^T c = 0.$$
This implies $a \perp b$, so $a \in \mathcal{R}(X^T)^\perp$.
4. Conversely, if $a \in \mathcal{R}(X^T)^\perp$, then $a^T(X^T c) = 0$ for all c. Thus, $(Xa)^T c = 0$ for all c, which implies $Xa = 0$, so $a \in \mathcal{N}(X)$.
5. Therefore, $\mathcal{N}(X) = \mathcal{R}(X^T)^\perp$.

Proof of $\mathcal{N}(X^T) = \mathcal{R}(X)^\perp$:

1. By definition, $b \in \mathcal{N}(X^T)$ implies $X^T b = 0$.
2. For any $a \in \mathcal{R}(X)$, there exists $c \in \mathbb{R}^{d+1}$ such that $a = Xc$.
3. Consider the inner product:

$$b^T a = b^T (Xc) = (X^T b)^T c = 0^T c = 0.$$

   This implies $b \perp a$, so $b \in \mathcal{R}(X)^\perp$.
4. Conversely, if $b \in \mathcal{R}(X)^\perp$, then $b^T(Xc) = 0$ for all c, which implies $(X^T b)^T c = 0$ for all c, so $X^T b = 0$, and $b \in \mathcal{N}(X^T)$.
5. Therefore, $\mathcal{N}(X^T) = \mathcal{R}(X)^\perp$.

Proof of $\mathcal{R}(X) = \mathcal{N}(X^T)^\perp$:

1. From the above, $\mathcal{N}(X^T) = \mathcal{R}(X)^\perp$.
2. Taking the orthogonal complement of both sides:

$$(\mathcal{N}(X^T))^\perp = (\mathcal{R}(X)^\perp)^\perp.$$

3. By the Fundamental Theorem of Linear Algebra, $(\mathcal{R}(X)^\perp)^\perp = \mathcal{R}(X)$.
4. Hence, $\mathcal{R}(X) = \mathcal{N}(X^T)^\perp$.

Proof of $\mathcal{R}(X^T) = \mathcal{N}(X)^\perp$:

1. From the above, $\mathcal{N}(X) = \mathcal{R}(X^T)^\perp$.
2. Taking the orthogonal complement of both sides:

$$(\mathcal{N}(X))^\perp = (\mathcal{R}(X^T)^\perp)^\perp.$$

3. By the Fundamental Theorem of Linear Algebra, $(\mathcal{R}(X^T)^\perp)^\perp = \mathcal{R}(X^T)$.
4. Hence, $\mathcal{R}(X^T) = \mathcal{N}(X)^\perp$.

Summary:

The claims are proved:

$$\mathcal{N}(X) = \mathcal{R}(X^T)^\perp, \quad \mathcal{N}(X^T) = \mathcal{R}(X)^\perp, \quad \mathcal{R}(X) = \mathcal{N}(X^T)^\perp, \quad \mathcal{R}(X^T) = \mathcal{N}(X)^\perp.$$

Part (b)

We Need To Prove:

$$\mathcal{R}(A^T A) = \mathcal{R}(A^T)$$

Proof. Using the relationships between the null space and the range, we have:

$$\mathcal{N}((A^T A)^T)^\perp = \mathcal{R}(A^T A) \quad \text{and} \quad \mathcal{R}(A^T) = \mathcal{N}(A)^\perp$$

Given that $\mathcal{N}(A^T A) = \mathcal{N}(A)$, it follows that:

$$\mathcal{N}((A^T A)^T)^\perp = \mathcal{N}(A)^\perp$$

Thus, we can conclude that:

$$\mathcal{R}(A^T A) = \mathcal{R}(A^T)$$

First Direction:

If $x \in \mathcal{N}(A)$, then $A^T Ax = 0$. This implies $Ax = 0$, meaning that $x^T A^T Ax = (Ax)^T (Ax) = 0$. Hence, $x \in \mathcal{N}(A^T A)$.

Second Direction:

If $x \in \mathcal{N}(A^T A)$, then $A^T Ax = 0$. This implies $x^T A^T Ax = (Ax)^T (Ax) = 0$, so $Ax = 0$. Thus, $x \in \mathcal{N}(A)$.

Conclusion: We have shown that $\mathcal{N}(A^T A) = \mathcal{N}(A)$. Therefore:

$$\mathcal{R}(A^T A) = \mathcal{R}(A^T)$$

Part (c)

When X is not full rank, the normal equations always have more than one solution, and any two solutions $\hat{w}_1$ and $\hat{w}_2$ differ by a vector in the nullspace of X, i.e.,

$$X(\hat{w}_1 - \hat{w}_2) = 0.$$

Proof. Assume the normal equations $(*)$:

$$X^T Xw = X^T y.$$

When X is not full rank, $\mathcal{N}(X^T X) \neq \{0\}$. Thus, there exist nonzero vectors $z \in \mathcal{N}(X^T X)$ such that:

$$X^T Xz = 0.$$

If $\hat{w}_1$ is a solution, then any $\hat{w}_2 = \hat{w}_1 + z$, where $z \in \mathcal{N}(X^T X)$, also satisfies the normal equations:

$$X^T X(\hat{w}_1 + z) = X^T X\hat{w}_1 + X^T Xz = X^T y.$$

For any two solutions $\hat{w}_1$ and $\hat{w}_2$, their difference is:

$$\hat{w}_2 - \hat{w}_1 = z,$$

where $z \in \mathcal{N}(X^T X)$. Since $\mathcal{N}(X^T X) = \mathcal{N}(X)$ (from part (b)), we have:

$$X(\hat{w}_2 - \hat{w}_1) = 0.$$

Therefore, the normal equations always have more than one solution, differing by a vector in $\mathcal{N}(X)$.

Part (d)

The projection of y onto $\mathcal{R}(X)$ is unique and is given by:

$$\hat{y} = X\hat{w},$$

where $\hat{w}$ is any solution to the normal equations.

The projection of y onto $\mathcal{R}(X)$ is unique and is defined by

$$\hat{y} = X\hat{w},$$

where $\hat{w}$ is any solution to the normal equations $(*)$.

Proof. Suppose the projection $\hat{y}$ of y onto $\mathcal{R}(X)$ is not unique. This implies there exist two distinct projections, $\hat{y}_1$ and $\hat{y}_2$, such that:

$$\hat{y}_1 \neq \hat{y}_2, \quad \hat{y}_1, \hat{y}_2 \in \mathcal{R}(X),$$

and each satisfies:

$$\hat{y}_1 = X\hat{w}_1, \quad \hat{y}_2 = X\hat{w}_2,$$

where $\hat{w}_1$ and $\hat{w}_2$ are solutions to the normal equations.

Since $\hat{y}_1, \hat{y}_2 \in \mathcal{R}(X)$, their difference $\hat{y}_1 - \hat{y}_2$ must also lie in $\mathcal{R}(X)$:

$$\hat{y}_1 - \hat{y}_2 = X(\hat{w}_1 - \hat{w}_2).$$

Moreover, if $\hat{w}_1$ and $\hat{w}_2$ are both solutions to the normal equations $(*)$, then:

$$X^T X\hat{w}_1 = X^T y, \quad X^T X\hat{w}_2 = X^T y.$$

Subtracting these equations gives:

$$X^T X(\hat{w}_1 - \hat{w}_2) = 0.$$

Thus, $\hat{w}_1 - \hat{w}_2 \in \mathcal{N}(X^T X)$.

From part (b), $\mathcal{R}(X^T X) = \mathcal{R}(X^T)$. Therefore, the nullspace of $X^T X$ is the same as the nullspace of X:

$$\mathcal{N}(X^T X) = \mathcal{N}(X).$$

This implies $\hat{w}_1 - \hat{w}_2 \in \mathcal{N}(X)$, so:

$$X(\hat{w}_1 - \hat{w}_2) = 0.$$

Hence:

$$\hat{y}_1 - \hat{y}_2 = X(\hat{w}_1 - \hat{w}_2) = 0.$$

Contradiction We assumed $\hat{y}_1 \neq \hat{y}_2$, but the above shows $\hat{y}_1 - \hat{y}_2 = 0$, i.e., $\hat{y}_1 = \hat{y}_2$. This contradiction implies the projection $\hat{y}$ is unique.

Conclusion The projection of y onto $\mathcal{R}(X)$ is unique and is defined by $\hat{y} = X\hat{w}$, where $\hat{w}$ is any solution to the normal equations.

Part (e)

When X has full column rank, the projection of y onto $\mathcal{R}(X)$ can be written as:

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

Proof:

1. When X has full column rank, $X^T X$ is invertible.
2. The normal equation $(*)$ has a unique solution:
$$\hat{w} = (X^T X)^{-1} X^T y.$$
3. Substituting $\hat{w}$ into $\hat{y} = X\hat{w}$, we have:
$$\hat{y} = X\left((X^T X)^{-1} X^T y\right).$$
4. Hence:
$$\hat{y} = (X^T X)^{-1} X^T y.$$

Question 2 - Linear Regression (properties of the hat matrix) (Problem 4)

(a) Prove that every eigenvalue of H is either 0 or 1

Key Facts:

1. The hat matrix $H = X(X^TX)^{-1}X^T$ is **symmetric**:

$$H^T = \left(X(X^TX)^{-1}X^T\right)^T = X\left((X^TX)^{-1}\right)^T X^T = X(X^TX)^{-1}X^T = H.$$

2. H is **idempotent**:

$$H^2 = H. \text{ This follows because:}$$

$$H^2 = \left[X(X^TX)^{-1}X^T\right]\left[X(X^TX)^{-1}X^T\right] = X(X^TX)^{-1}(X^TX)(X^TX)^{-1}X^T = X(X^TX)^{-1}X^T = H.$$

Eigenvalues of an idempotent matrix:

Let λ be an eigenvalue of H, and let v be the corresponding eigenvector, i.e., $Hv = \lambda v$. Since $H^2 = H$, substituting $Hv = \lambda v$ into H^2, we get:

$$H^2v = H(Hv) = H(\lambda v) = \lambda(Hv) = \lambda(\lambda v) = \lambda^2 v.$$

But since $H^2 = H$, we also have $H^2v = Hv = \lambda v$. Therefore:

$$\lambda^2 v = \lambda v.$$

Simplifying, we get:

$$(\lambda^2 - \lambda)v = 0.$$

Since $v \neq 0$ (eigenvectors are non-zero), we have $\lambda^2 - \lambda = 0$, or:

$$\lambda(\lambda - 1) = 0.$$

Thus, $\lambda = 0$ or $\lambda = 1$.

(b) Number of eigenvalues equal to 1 and the rank of H

Step 1: Trace of H

The trace of H equals the sum of its eigenvalues. Since H is idempotent and symmetric, its eigenvalues are either 0 or 1, as shown in part (a). To calculate the trace:

$$\text{tr}(H) = \text{tr}(X(X^TX)^{-1}X^T) = \text{tr}((X^TX)^{-1}X^TX) = \text{tr}(I_{d+1}) = d + 1.$$

Here, I_{d+1} is the identity matrix of size $d + 1$, and its trace is simply $d + 1$.

Step 2: Eigenvalues of H

Since $\text{tr}(H) = d + 1$, this implies there are $d + 1$ eigenvalues equal to 1 (as eigenvalues can only be 0 or 1).

Step 3: Rank of H

The rank of a matrix is the number of non-zero eigenvalues. Since $d+1$ eigenvalues are 1, the rank of H is $d+1$. This also matches the intuition that the rank of H equals the rank of X, which is $d+1$ when X has full column rank.

Conclusion

The rank of H is $d+1$, and there are $d+1$ eigenvalues of H equal to 1.

Final Answers:

- (a) Every eigenvalue of H is either 0 or 1.
- (b) There are $d+1$ eigenvalues of H equal to 1 and $\text{rank}(H) = d+1$.

Question 3 - Least Squares and Least Absolute Deviations (Problem 5)

(a) To show that the sample mean minimizes the sum of squared deviations:

To find the minimum of $E(h)$, we need to find where its derivative equals zero:

$$\frac{d}{dh}E(h) = \frac{d}{dh}\sum_{n=1}^{N}(h - y_n)^2 = 0$$

Using the chain rule:

$$\sum_{n=1}^{N} 2(h - y_n) = 0$$

Solving for h:

$$2h\sum_{n=1}^{N} 1 - 2\sum_{n=1}^{N} y_n = 0$$

$$2Nh - 2\sum_{n=1}^{N} y_n = 0$$

$$h = \frac{1}{N}\sum_{n=1}^{N} y_n = h_{\text{mean}}$$

To verify this is a minimum, check that the second derivative is positive:

$$\frac{d^2}{dh^2}E(h) = 2N > 0$$

Therefore, h_{mean} is indeed the minimizer.

(b) For the sum of absolute deviations:

The derivative of the absolute value function is the sign function:

$$\frac{d}{dh}|h - y_n| = \text{sign}(h - y_n)$$

Setting the derivative of $E(h)$ to zero:

$$\frac{d}{dh}E(h) = \sum_{n=1}^{N} \text{sign}(h - y_n) = 0$$

This equation means that the number of positive terms (where $h > y_n$) must equal the number of negative terms (where $h < y_n$).

By definition, this occurs at the median h_{med}, where half the data points are at most h_{med} and half are at least h_{med}.

(c) Impact of an outlier (where $y_N \to y_N + c$, and $c \to \infty$):

For h_{mean}:

- The mean is directly affected by extreme values.
- As $c \to \infty$:

$$h_{\text{mean}} = \frac{1}{N}\left(\sum_{n=1}^{N-1} y_n + (y_N + c)\right) \to \infty$$

For h_{med}:

- The median only depends on the relative ordering of values.
- If N is odd, h_{med} will be the middle value.
- If N is even, h_{med} will be between the two middle values.
- In either case, as long as the original y_N was not one of the middle values determining the median, h_{med} remains unchanged.
- Even if y_N was one of the middle values, h_{med} would only shift to the next ordered value, not to infinity.

Therefore:

- $h_{\text{mean}} \to \infty$ as $c \to \infty$.
- h_{med} remains bounded.

This demonstrates that the median is more robust to outliers than the mean.

Question 4 - Hard-Margin SVM (Problem 7)

Part (f): Lagrangian and KKT Conditions

Writing the Lagrangian

The primal problem is:

$$\min_{b,w} \frac{1}{2} w^T w$$
$$\text{subject to: } y_n(w^T x_n + b) \geq 1, \quad (n = 1, \ldots, N)$$

The Lagrangian function with multipliers $\alpha_n \geq 0$ is:

$$L(w, b, \alpha) = \frac{1}{2} w^T w + \sum_{n=1}^{N} \alpha_n (1 - y_n(w^T x_n + b))$$

KKT Conditions:

$$\frac{\partial L}{\partial w} = w - \sum_{n=1}^{N} \alpha_n y_n x_n = 0$$
$$\therefore w = \sum_{n=1}^{N} \alpha_n y_n x_n$$
$$\frac{\partial L}{\partial b} = -\sum_{n=1}^{N} \alpha_n y_n = 0$$

Complementary slackness:

$$\alpha_n (1 - y_n(w^T x_n + b)) = 0, \quad \forall n$$

Dual feasibility:

$$\alpha_n \geq 0, \quad \forall n$$

Primal feasibility:

$$y_n(w^T x_n + b) \geq 1, \quad \forall n$$

Deriving the Dual Problem

Substitute $w = \sum_{n=1}^{N} \alpha_n y_n x_n$ back into the Lagrangian:

$$L(\alpha) = \frac{1}{2} \left(\sum_{n=1}^{N} \alpha_n y_n x_n \right)^T \left(\sum_{m=1}^{N} \alpha_m y_m x_m \right) - \sum_{n=1}^{N} \alpha_n y_n \left(\left(\sum_{m=1}^{N} \alpha_m y_m x_m \right)^T x_n + b \right) + \sum_{n=1}^{N} \alpha_n$$

Expand the first term:

$$\frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m x_n^T x_m$$

From the stationarity condition $\sum_{n=1}^{N} \alpha_n y_n = 0$, the terms with b cancel out.

After simplification, we get the dual problem:

$$\max_{\alpha \in \mathbb{R}^N} -\frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} y_n y_m \alpha_n \alpha_m x_n^T x_m + \sum_{n=1}^{N} \alpha_n$$

$$\text{subject to: } \sum_{n=1}^{N} y_n \alpha_n = 0$$

$$\alpha_n \geq 0, \quad (n = 1, \ldots, N)$$

Part (g): Standard QP-Problem Form

Identifying Components

The standard QP-problem form is:

$$\min_{u \in \mathbb{R}^N} \frac{1}{2} u^T Q_D u + p_D^T u$$

$$\text{subject to: } A_D u \geq c_D$$

To convert our dual problem to this form:

Define Q_D matrix:

$$Q_D[n, m] = y_n y_m x_n^T x_m$$

The full matrix is:

$$Q_D = \begin{bmatrix} y_1 y_1 x_1^T x_1 & y_1 y_2 x_1^T x_2 & \cdots & y_1 y_N x_1^T x_N \\ y_2 y_1 x_2^T x_1 & y_2 y_2 x_2^T x_2 & \cdots & y_2 y_N x_2^T x_N \\ \vdots & \vdots & \ddots & \vdots \\ y_N y_1 x_N^T x_1 & y_N y_2 x_N^T x_2 & \cdots & y_N y_N x_N^T x_N \end{bmatrix}$$

Define p_D:

$$p_D = \begin{bmatrix} -1 \\ -1 \\ \vdots \\ -1 \end{bmatrix} \in \mathbb{R}^N$$

Define A_D and c_D:

$$A_D = \begin{bmatrix} y_1 & y_2 & \cdots & y_N \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}, \quad c_D = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Proving Q_D is Positive Semi-definite

For any vector $v \in \mathbb{R}^N$:

$$v^T Q_D v = \sum_{n=1}^{N} \sum_{m=1}^{N} v_n v_m y_n y_m x_n^T x_m$$

$$= \left(\sum_{n=1}^{N} v_n y_n x_n \right)^T \left(\sum_{m=1}^{N} v_m y_m x_m \right)$$

$$= \left\| \sum_{n=1}^{N} v_n y_n x_n \right\|^2 \geq 0$$

Therefore, Q_D is positive semi-definite, making the QP-problem convex.

Part (h): Optimal Value Analysis

Computing $L(\alpha^*)$

At the optimal point α^*:

$$L(\alpha^*) = -\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} y_n y_m \alpha_n^* \alpha_m^* x_n^T x_m + \sum_{n=1}^{N} \alpha_n^*$$

For support vectors (points where $\alpha_n^* > 0$):

$$y_n(w^{*T} x_n + b^*) = 1$$

Multiply by α_n^* and sum:

$$\sum_{n=1}^{N} \alpha_n^* y_n (w^{*T} x_n + b^*) = \sum_{n=1}^{N} \alpha_n^*$$

Using $w^* = \sum_{m=1}^{N} \alpha_m^* y_m x_m$:

$$\sum_{n=1}^{N}\sum_{m=1}^{N} y_n y_m \alpha_n^* \alpha_m^* x_n^T x_m + b^* \sum_{n=1}^{N} \alpha_n^* y_n = \sum_{n=1}^{N} \alpha_n^*$$

From KKT conditions, $\sum_{n=1}^{N} \alpha_n^* y_n = 0$, therefore:

$$L(\alpha^*) = \frac{1}{2}\sum_{n=1}^{N} \alpha_n^*$$

Computing Δ

The minimal distance Δ to the optimal separating hyperplane is:

$$\Delta = \frac{1}{\|w^*\|} = \frac{1}{\sqrt{(w^*)^T w^*}}$$

Substituting $w^* = \sum_{n=1}^{N} \alpha_n^* y_n x_n$:

$$\Delta = \frac{1}{\sqrt{\sum_{n=1}^{N} \alpha_n^*}} = \left(\sqrt{\sum_{n=1}^{N} \alpha_n^*}\right)^{-1}$$

Question 5 - SVM and KKT - (Problem 9)

(a) TRUE

Let's start with the premise that the sets are linearly separable and show this leads to a contradiction when $P \to \infty$ as $C \to \infty$.

If the sets are linearly separable:

- There exists a hyperplane (w', b') that perfectly separates the data
- This means $\forall n : y_n(w'^T x_n + b') \geq 1$ for some finite w' and b'
- Therefore, we can achieve $\xi_n = 0$ for all n

Consider the objective function:

$$P = \min_{b,w,\xi} \{\frac{1}{2} w^T w + C \sum_{n=1}^{N} \xi_n\}$$

under constraints:

$$y_n(w^T x_n + b) \geq 1 - \xi_n, \xi_n \geq 0$$

If the sets are linearly separable:

- We can use the separating hyperplane (w', b') as a feasible solution
- With this solution, all $\xi_n = 0$
- Therefore, $P \leq \frac{1}{2} w'^T w'$ (a finite value) regardless of C

Contradiction:

- We're given that $P \to \infty$ when $C \to \infty$
- But we just showed that P is bounded above by a finite value if the sets are linearly separable
- This is a contradiction

Therefore:

- Our initial assumption that the sets are linearly separable must be false
- The sets must not be linearly separable

The proof shows that when $P \to \infty$ as $C \to \infty$, the sets cannot be linearly separable, as a linearly separable solution would provide a finite upper bound for P regardless of C. Therefore, the statement is TRUE.

For linearly separable sets, P would remain bounded even as $C \to \infty$ because we could achieve $\xi_n = 0$ for all points. The fact that $P \to \infty$ implies that no such perfect separation exists, forcing some ξ_n to be positive and causing the penalty term $C \sum_{n=1}^{N} \xi_n$ to grow without bound as C increases.

(b) TRUE

Proof:
For a linearly separable case, consider the geometric interpretation:

- The margin width is $\frac{2}{\|w\|}$
- The objective function minimizes $\frac{1}{2}w^T w = \frac{1}{2}\|w\|^2$

For any separating hyperplane:

- Δ is the minimal distance to any data point
- Δ_0 is the maximum possible value of Δ

From the optimization problem:

- $\|w\| \cdot \Delta \leq 1$ (from constraint normalization)
- Δ_0 occurs at optimal w^*
- $\Delta_0 = \frac{1}{\|w^*\|}$

The optimal margin is:

- $\frac{2}{\|w^*\|} = 2\Delta_0$
- $P = \frac{1}{2}\|w^*\|^2$
- Therefore, $\Delta_0 = \frac{1}{\sqrt{2P}}$

This proves $\Delta_0 \leq \frac{1}{\sqrt{2P}}$

(c) TRUE

Proof:
From KKT conditions:

- Complementary slackness: $\alpha_n^*[y_n(w^{*T}x_n + b^*) - 1] = 0$
- If $\alpha_n^* > 0$, then $y_n(w^{*T}x_n + b^*) = 1$

The boundary of the optimal separating hyperplane is defined by:

$$w^{*T}x + b^* = \pm 1$$

Therefore, if $\alpha_n^* > 0$:

- The point (x_n, y_n) must satisfy $y_n(w^{*T}x_n + b^*) = 1$
- This places it exactly on the margin boundary

(d) TRUE

Proof:
Consider the dual formulation:

- Points can lie on the margin boundary even if they're not support vectors
- $\alpha_n^* = 0$ means the point doesn't contribute to defining w^*
- But it can still satisfy $y_n(w^{*T}x_n + b^*) = 1$ coincidentally

For Example: Consider a point that lies exactly on the margin boundary, if removing it doesn't affect the optimal solution then $\alpha_n^* = 0$ is possible while still being on the boundary.

(e) TRUE

Proof:
The optimal hyperplane is determined by:

- $w^* = \sum_{i=1}^{N} \alpha_i^* y_i x_i$
- b^* from points with $\alpha_i^* > 0$

If $\alpha_n^* = 0$:

- This point doesn't contribute to w^* calculation
- Removing it won't change the sum that defines w^*
- b^* is also unaffected as it's computed from support vectors

Therefore:

- The "fat" hyperplane (margin boundaries) remains unchanged
- Both w^* and b^* maintain their values
- The optimal solution remains optimal

This proves that removing a point with $\alpha_n^* = 0$ doesn't affect the optimal hyperplane.

Question 6 - Lasso - (Problem 12)

Part (a): Dual Function and KKT Conditions

Reformulating the Problem
The original lasso problem is:

$$L(\beta) = \frac{1}{2} \sum_i \left(y_i - \sum_j x_{ij} \beta_j \right)^2$$

and we minimize:

$$L(\beta) + \lambda \sum_j |\beta_j|,$$

where $\lambda > 0$.

We decompose β_j as:

$$\beta_j = \beta_j^+ - \beta_j^-, \quad \text{where } \beta_j^+, \beta_j^- \geq 0.$$

Then:

$$|\beta_j| = \beta_j^+ + \beta_j^-.$$

Substituting into the objective, the problem becomes minimizing:

$$L(\beta) + \lambda \sum_j (\beta_j^+ + \beta_j^-).$$

Lagrange Dual Function
We introduce Lagrange multipliers λ_j^+ and λ_j^- for the constraints $\beta_j^+ \geq 0$ and $\beta_j^- \geq 0$, respectively. The Lagrange function is:

$$L(\beta) + \lambda \sum_j (\beta_j^+ + \beta_j^-) - \sum_j \lambda_j^+ \beta_j^+ - \sum_j \lambda_j^- \beta_j^-,$$

where $\lambda_j^+, \lambda_j^- \geq 0$.

Karush-Kuhn-Tucker (KKT) Conditions
The KKT conditions ensure the optimality of the solution. The conditions are:

Stationarity: The gradient of the Lagrangian with respect to β_j^+ and β_j^- is zero:

$$\frac{\partial}{\partial \beta_j^+} \Big(L(\beta) + \lambda \sum_j (\beta_j^+ + \beta_j^-) - \sum_j \lambda_j^+ \beta_j^+ - \sum_j \lambda_j^- \beta_j^- \Big) = 0.$$

Expanding, we get:

$$\nabla L(\beta_j) + \lambda - \lambda_j^+ = 0,$$

and similarly:

$$-\nabla L(\beta_j) + \lambda - \lambda_j^- = 0.$$

Complementary Slackness:

$$\lambda_j^+ \beta_j^+ = 0, \quad \lambda_j^- \beta_j^- = 0.$$

Primal Feasibility: $\beta_j^+, \beta_j^- \geq 0$.

Dual Feasibility: $\lambda_j^+, \lambda_j^- \geq 0$.

Part (b): Implications of the KKT Conditions

Magnitude of $\nabla L(\beta_j)$
From the stationarity conditions:

$$\lambda_j^+ = \nabla L(\beta_j) + \lambda, \quad \lambda_j^- = -\nabla L(\beta_j) + \lambda.$$

Using the dual feasibility conditions ($\lambda_j^+, \lambda_j^- \geq 0$), we have:

$$\nabla L(\beta_j) + \lambda \geq 0 \quad \text{and} \quad -\nabla L(\beta_j) + \lambda \geq 0.$$

Adding these inequalities:

$$-\lambda \leq \nabla L(\beta_j) \leq \lambda.$$

Thus:

$$|\nabla L(\beta_j)| \leq \lambda.$$

Three Scenarios
The KKT conditions imply three possible scenarios:

Case 1: $\lambda = 0$:
From the stationarity conditions:

$$\nabla L(\beta_j) = 0 \quad \forall j.$$

Case 2: $\beta_j^+ > 0, \lambda > 0$:
From complementary slackness, $\lambda_j^+ = 0$. Then from the stationarity condition:

$$\nabla L(\beta_j) + \lambda = 0 \implies \nabla L(\beta_j) = -\lambda.$$

Since $\beta_j^- = 0$, $\beta_j = \beta_j^+ > 0$.

Case 3: $\beta_j^- > 0, \lambda > 0$:
From complementary slackness, $\lambda_j^- = 0$. Then from the stationarity condition:

$$-\nabla L(\beta_j) + \lambda = 0 \implies \nabla L(\beta_j) = \lambda.$$

Since $\beta_j^+ = 0$, $\beta_j = -\beta_j^- < 0$.

Active Predictors
For an active predictor ($\beta_j \neq 0$):

- If $\beta_j > 0$, $\nabla L(\beta_j) = -\lambda$. - If $\beta_j < 0$, $\nabla L(\beta_j) = \lambda$.

Relating λ to Correlation with Residuals
If predictors are standardized, then $\nabla L(\beta_j)$ corresponds to the correlation between the jth predictor and the residuals:

$$\nabla L(\beta_j) = x_j^\top (y - X\beta).$$

Thus, λ controls the maximum correlation between any predictor and the residuals.

Part (c): Linearity of the Lasso Solution Path

Piecewise Linearity

If the set of active predictors remains unchanged as λ varies between $\lambda_0 \geq \lambda \geq \lambda_1$, then the solution can be expressed as:

$$\hat{\beta}(\lambda) = \hat{\beta}(\lambda_0) - (\lambda - \lambda_0)\gamma_0.$$

Deriving γ_0

Let A denote the set of active predictors. For $j \in A$, the KKT conditions imply:

$$\nabla L(\beta_j) = \pm\lambda.$$

The residuals $r = y - X\beta$ must satisfy:

$$X_A^\top r = \pm\lambda.$$

Differentiating with respect to λ, we get:

$$\frac{d\hat{\beta}_A}{d\lambda} = -\gamma_0,$$

where γ_0 is constant for $\lambda_0 \geq \lambda \geq \lambda_1$.

Thus:

$$\hat{\beta}(\lambda) = \hat{\beta}(\lambda_0) - (\lambda - \lambda_0)\gamma_0.$$

—

Thank you for reviewing this assignment.