Q1

Q1a

A1a

The main difference between supervised learning and unsupervised learning lies in whether the optimization objective of the loss function is related to given true labels. Supervised learning requires each sample to correspond to a true label (optimization direction), such as SFT, support vector machines, logistic regression, etc., while unsupervised learning often determines optimization objectives through internal relationships between samples, such as contrastive learning, etc.

Q1b

A1b

1. False 2) False 3) False 4) True

Q1c

A1c

If X is a full column rank matrix $n \times d$, then for any non-zero vector x, $Xx = b$ has a unique solution, and b must be a non-zero vector.

Consider $(Xx)^T(Xx) = b^Tb > 0$,
Let $X^TX = Y$, then $x^TX^TXx = x^TYx > 0$ holds for any non-zero vector x, which means $X^TX = Y$ is a positive definite matrix.

Q2

Q2a

A2a

X can be decomposed by SVD as

$$X = V * [\Sigma, 0] * [U_1^T; U_2^T] = V\Sigma U_1^T$$

Let $V\Sigma = A$, where A is full rank.
Solving the least squares solution for X is equivalent to solving the least squares solution for A with respect to $U_1^T\theta = z$. This solution can be expressed as $(A^TA)^{-1}A^Ty = (\Sigma^TV^TV\Sigma)^{-1}\Sigma^TV^Ty$.
Since V is an orthogonal matrix, the above expression is equivalent to $z^* = (\Sigma^T\Sigma)^{-1}\Sigma^TV^Ty = \Sigma^{-1}V^Ty = U_1^T * \theta$.
Then for $U_1^T * \theta = z^*$, substituting $\theta_p = U_1z^*$, it satisfies $U_1^TU_1z^* = z^*$, meaning U_1z^* is a particular solution.
For the homogeneous equation $U_1^T\theta = 0$, U_2 can satisfy the homogeneous solution, so the general solution for θ is $U_1\Sigma^{-1}V^Ty + U_2w$, where w is an arbitrary vector.

Q2b

A2b

By directly taking the derivative of the optimization equation, we can obtain that when the derivative equals zero, it satisfies
$(X^TX + \lambda I)w = X^Ty$
Therefore, the optimal solution is $w^* = (X^TX + \lambda I)^{-1}X^Ty$

Q3

Q3a

A3a:

Maximum likelihood function: $L(\theta) = P(y|\theta) = P(\epsilon) = \prod(\epsilon_i) = \prod(e^{-|\epsilon_i|/b}/2b)$
Equivalent to log-likelihood maximization: $l(\theta) = \sum(log(1/2b) - |\epsilon_i|/b) = n * log(1/2b - \sum(|\epsilon_i|/b))$
$\epsilon_i = y_i - (X * \theta)_i$
Therefore: $l(\theta) = n * log(1/(2b)) - \sum(|y_i - (X * \theta)_i|)/b$
To maximize $L(\theta)$ is equivalent to minimizing $\sum(|y_i - (X * \theta) *_i |)$, i.e., $argmin_\theta||y - X * \theta||_1$

Q3b

A3b

$$h_\mu(z_j) = \begin{cases} z_j^2/2\mu & |z_j| < \mu \\ |z_j| - \mu/2 & |z_j| >= \mu \end{cases}$$
$$h'_j(z_j) = \begin{cases} z_j/\mu & |z_j| < \mu \\ 1 & z_j > \mu \\ -1 & z_j < -\mu \end{cases}$$
Therefore $L(\theta) = H_\mu(X\theta - y) = H_\mu(z)$, let $z = X\theta - y$
$\nabla L(\theta) = H'_\mu(z) = X^T \nabla H'(X\theta - y)$

Q3c

A3c

The code could been seen in the python file "../code_source/p3/p3.py"

The error plot are shown in "../code_source/p3/l1_estimator.png"

Q4

Q4a

A4a

For a perfect classifier θ^*: for $\forall i \in N$, we have $y_i(\theta^{*T})x_i > 0$ (same sign), so $\rho = \min y_i(\theta^* x_i) > 0$ always holds.

Q4b

A4b

Proof:
Given

$$\begin{cases} y_{k-1}(\theta_{k-1}^T x_{k-1}) < 0 \\ \theta_k = \theta_{k-1} + y_j x_j^T \end{cases}$$

Case analysis
Assume $\rho = \min_{1<=i<=n} y_i(\theta^{*T})x_i = y_j(\theta^{*T} x_i)$ represents the minimum index
①When k-1 = j, then $\theta = \theta_{k-1} + y_j x_j^T = y_j x_j^T$
i.e., $\theta_k^T \theta^* = \theta_{k-1}^T \theta^* + y_j x_j^T \theta^* = \theta_{k-1}^T \theta^* + y_j \theta^{*T} x_j = \theta_{k-1}^T \theta^* + \rho$
②When $k-1 \neq b$, then $\theta_k = \theta_{k-1} + y_j x_j^T$
i.e., $\theta_k^T \theta^* = \theta_{k-1}^T \theta^* + y_{k-1} x_{k-1}^T \theta^* = \theta_{k-1}^T \theta^* + y_{k-1} \theta^{*T} x_{k-1} \geq \theta_{k-1}^T \theta^* + y_j \theta^{*T} x_j = \theta_{k-1}^T \theta^* + \rho$
In conclusion, $\theta_k^T \theta^* \geq \theta_{k-1}^T \theta^* + \rho$, and by mathematical induction, $\theta_k^T \theta^* \geq k\rho$

Q4c

A4c

Given

$$\begin{cases} y_{k-1}(\theta_{k-1}^T x_{k-1}) < 0 \\ \theta_k = \theta_{k-1} + y_j x_j^T \end{cases}$$

$\theta_k\theta_{k-1} = \theta_{k-1}\theta_{k-1} + y_{k-1}x_{k-1}^T\theta_{k-1}$, since $y_{k-1}x_{k-1}^T\theta_{k-1} < 0$, $\theta_k\theta_{k-1} < \theta_{k-1}\theta_{k-1} = ||\theta_{k-1}||^2$
And $||\theta_k||^2 = ||(\theta_{k-1} + y_{k-1}x_{k-1}^T)||^2 = ||\theta_{k-1}||^2 + 2||\theta_{k-1}y_{k-1}x_{k-1}^T|| + ||y_{k-1}x_{k-1}^T||^2 + m$,
where $y \in \{-1, 1\}$, constant $m < 0$
Therefore $||\theta_k||^2 \leq ||\theta_{k-1}||^2 + ||x_{k-1}||^2$

Q4d

A4d

Since
$||\theta_k||^2 \leq ||\theta_{k-1}|| + ||x_{k-1}||^2 \leq ||\theta_{k-1}||^2 + \max_{1\leq i\leq n} ||x_i||^2$
Therefore
$||\theta_k||^2 \leq k * \max_{1\leq i\leq n} ||x_i||^2 = kR^2$

Q4e

A4e

Given

$$\begin{cases} \theta_k^T\theta^* \geq \theta_{k-1}^T\theta^* + \rho \\ ||\theta_k||^2 \leq k * R^2 \\ R = \max_{1\leq i\leq n} ||x_i|| \end{cases}$$

Then

$$\frac{\theta_k^T\theta^*}{||\theta_k||} \geq \frac{k\rho}{||\theta_k||} \geq \frac{k\rho}{\sqrt{k}R^2} = \sqrt{k}\frac{\rho}{R}$$

Also, since

$$\frac{\theta_k\theta^*}{||\theta_k||||\theta^*||} \leq 1$$

and for $\overline{k}$, $\theta_{\overline{k}} = \theta^*$
Therefore

$$\overline{k} \leq \frac{R^2||\theta^*||^2}{\rho^2}$$

Q5

A5

The answer is shown in "code_source/p5/p5/py"

The plot of 5b is shown in "code_source/p5/erro_comparision.png"

The plot of 5c is shown in "code_source/p5/feature_boundary.png"