1.

The dataset on the right has higher MSE.

Left: $E_{in} = \frac{1}{N}\sum_{n=1}^{N}(h(x_n) - y_n)^2$

$= \frac{1}{10}[(2.5-2.5)^2+(4-3)^2+(3.5-3.5)^2+(3-4)^2+(4.5-4.5)^2+(4-5)^2+(5.5-5.5)^2+(7-6)^2$

$+(6.5-6.5)^2+(7-7)^2]$

$= \frac{1}{10}(0+1+0+1+0+1+0+1+0+0)$

$= \frac{1}{10} \times 4$

$= \frac{2}{5}$

Right: $E_{in} = \frac{1}{N}\sum_{n=1}^{N}(h(x_n) - y_n)^2$

$= \frac{1}{10}[(2.5-2.5)^2+(3-3)^2+(3.5-3.5)^2+(6-4)^2+(4.5-4.5)^2+(5-5)^2+(5.5-5.5)^2+(4-6)^2+$

$(6.5-6.5)^2+(7-7)^2]$

$= \frac{1}{10}(0+0+0+4+0+0+0+4+0+0)$

$= \frac{1}{10} \times 8$

$= \frac{4}{5}$

As E_{in} of the right $= \frac{4}{5} > E_{in}$ of the left $= \frac{2}{5}$, the dataset on the right has higher MSE.

2.

(1). $\tanh(s)=\frac{e^s-e^{-s}}{e^s+e^{-s}}=\frac{e^s-\frac{1}{e^s}}{e^s+\frac{1}{e^s}}=\frac{\frac{e^s\cdot e^s-1}{e^s}}{\frac{e^s\cdot e^s+1}{e^s}}=\frac{e^{2s}-1}{e^{2s}+1}$

$\theta(s)=\frac{e^s}{1+e^s}$

so $\theta(2s)=\frac{e^{2s}}{1+e^{2s}}$, $2\theta(2s)-1=\frac{2e^{2s}}{1+e^{2s}}-1=\frac{2e^{2s}-1-e^{2s}}{1+e^{2s}}=\frac{e^{2s}-1}{e^{2s}+1}=\tanh(s)$

So $\tanh(s)=2\theta(2s)-1$

(2). From (1). $\tanh(s)=2\theta(2s)-1$

① When $|s|$ is extremely large ($|s|\to+\infty$):

If s is positive, $s\to+\infty$. then $\theta(2s)\to1$, $\tanh(s)\to2\cdot1-1=1$,

and as $\tanh(s)=\frac{e^s-e^{-s}}{e^s+e^{-s}}$, $e^{-s}\to0$ when s is large, so $\tanh(s)\to\frac{e^s}{e^s}=1$

If s is negative, $s\to-\infty$, then $\theta(2s)\to0$, $\tanh(s)\to2\cdot1-1=-1$,

and as $\tanh(s)=\frac{e^s-e^{-s}}{e^s+e^{-s}}$, $e^s\to0$ when $-s$ is large, so $\tanh(s)\to\frac{e^{-s}}{-e^{-s}}=-1$

② When $|s|$ is extremely small ($s\to0$):

$\tanh(s)=\frac{e^s-e^{-s}}{e^s+e^{-s}}$ and both e^s and e^{-s} are close to 1, so cannot dominate.

Thus, $\tanh(s)$ converges to a hard threshold of 1 or -1 for large $|s|$,

and no threshold for small $|s|$

3.

(1). $E_{in}(w) = \sum_{n=1}^{N} (y=+1) \ln \frac{1}{h(x_n)} + (y=-1) \ln \frac{1}{1-h(x_n)}$

$P(y|x) = \begin{cases} \ln \frac{1}{h(x_n)} & \text{for } y=+1 \\ \ln \frac{1}{1-h(x_n)} & \text{for } y=-1 \end{cases}$

Substitute $h(x) = \theta(w^T x)$, $P(y|x) = \theta(y w^T x)$

Maximum likelihood: $L(h) = \prod_{n=1}^{N} P(y_n|x_n) = \prod_{n=1}^{N} (h(x_n))^{(y_n=+1)} (1-h(x_n))^{(y_n=-1)}$

$\prod_{n=1}^{N} P(y_n|x_n) = \prod_{n=1}^{N} \theta(y_n w^T x_n)$

$$\begin{aligned} \max \prod_{n=1}^{N} P(y_n|x_n) &\Leftrightarrow \max(\ln \prod_{n=1}^{N} P(y_n|x_n)) \\ &\equiv \max \sum_{n=1}^{N} \ln P(y_n|x_n) \\ &\Leftrightarrow \min -\frac{1}{N} \sum_{n=1}^{N} \ln P(y_n|x_n) \\ &\equiv \min \frac{1}{N} \sum_{n=1}^{N} \ln \frac{1}{P(y_n|x_n)} \\ &\equiv \min \frac{1}{N} \sum_{n=1}^{N} \ln \frac{1}{\theta(y_n w^T x_n)} \\ &\equiv \min \frac{1}{N} \sum_{n=1}^{N} \ln(1+e^{-y_n w^T x_n}) \end{aligned}$$

$\ln L(h) = \sum_{n=1}^{N} (y_n=+1) \ln(h(x_n)) + (y_n=-1) \ln(1-h(x))$

$-\ln L(h) = \sum_{n=1}^{n} (y_n=+1) \ln \frac{1}{h(x_n)} + (y_n=-1) \ln \frac{1}{1-h(x)} = E_{in}(w)$

So cross-entropy error measure is equal to minimum method of likelihood

4.

(1). Derivative of $f(x)$: $f'(x) = \frac{df(x)}{dx} = \frac{d\left(\frac{a}{1+e^{kx+b}}\right)}{dx}$

$$= a\frac{d(1+e^{kx+b})^{-1}}{dx}$$

$$= -a(k \cdot e^{kx+b} \cdot (1+e^{kx+b})^{-2}$$

$$= \frac{-ake^{kx+b}}{(1+e^{kx+b})^2}$$

① When $a=0$ or $k=0$, $b \in R$:

$f'(x) = 0$, so $f(x)$ is a constant

② When $a \cdot k > 0$ ($a,k>0$ or $a,k<0$), $b \in R$:

$e^{kx+b} > 0$ regardless of $(kx+b)$, $(1+e^{kx+b})^2 > 0$, $a \cdot k > 0$, $-ak < 0$, so $f'(x) < 0$ for all x.

As x getting larger, $f(x)$ keeps decreasing.

③ When $a \cdot k < 0$ ($a>0, k<0$ or $a<0, k>0$), $b \in R$:

$e^{kx+b} > 0$ regardless of $(kx+b)$, $(1+e^{kx+b})^2 > 0$, $a \cdot k < 0$, $-ak > 0$, so $f'(x) < o$ for all x.

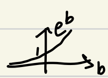
As x getting larger, $f(x)$ keeps increasing.

In conclusion, b will not influence the monotonicity of $f(x)$, and when $a=0$ or $k=0$, $f(x)$ becomes a constant; when $a,k>0$ or $a,k<0$, $f(x)$ decreases when x increases; when ($a>0, k<0$) or ($a<0, k>0$), $f(x)$ increases when x increases;

(2). Based on (1),

① When $a=0$, $k, b \in R$: $f(x) = 0$ for all x.

② When $a \neq 0$, $k=0$, $b \in R$: $f(x) = \frac{a}{1+e^b}$ for all x.



When $b \to -\infty$, $f(x) \to a$ $(a \neq 0)$

When $b = 0$, $f(x) = \frac{a}{2}$ $(a \neq 0)$

When $b \to +\infty$, $f(x) \to 0$

③ When $a \cdot k > 0$ ($a,k>0$ or $a,k<0$) and $b \in R$: $f'(x) < 0$ for all x, $x \uparrow$, $f(x) \downarrow$.

when $x \to +\infty$, $\begin{cases} \text{if } a,k>0,\ e^{kx+b} \to +\infty,\ 1+e^{kx+b} \to +\infty,\ f(x)_{min} \to 0 \\ \text{if } a,k<0,\ e^{kx+b} \to 0,\ 1+e^{kx+b} \to 1,\ f(x)_{min} \to a\ (a<0) \end{cases}$

when $x \to -\infty$ $\begin{cases} \text{if } a,k>0,\ e^{kx+b} \to 0,\ 1+e^{kx+b} \to 1,\ f(x)_{max} = \frac{a}{1+e^{kx+b}} \to a\ (a>0) \\ \text{if } a,k<0,\ e^{kx+b} \to +\infty,\ 1+e^{kx+b} \to +\infty,\ f(x)_{max} \to 0 \end{cases}$

④ When $a \cdot k < 0$ ($a>0, k<0$ or $a<0, k>0$) and $b \in R$: $f'(x) > 0$ for all x, $x \uparrow$, $f(x) \uparrow$

So when $x \to -\infty$, $\begin{cases} \text{if } a>0, k<0, e^{kx+b} \to +\infty, 1+e^{kx+b} \to +\infty, f(x)\min \to 0 \\ \text{if } a<0, k>0, e^{kx+b} \to 0, 1+e^{kx+b} \to 1, f(x)\min \to a \ (a<0) \end{cases}$

when $x \to +\infty$, $\begin{cases} \text{if } a>0, k<0, e^{kx+b} \to 0, 1+e^{kx+b} \to 1, f(x)\max \to a \ (a>0) \\ \text{if } a<0, k>0, e^{kx+b} \to +\infty, 1+e^{kx+b} \to +\infty, f(x)\max \to 0 \end{cases}$

In conclusion, when $a=0$, $k, b \in R$: $f(x)=0$ for all x;

when $a \neq 0$, $k=0$, $b \in R$: $f(x) = \frac{a}{1+e^b}$ ($a \neq 0, b \in R$) $\neq 0$ for all x;

when $a>0$, $k \neq 0$, $b \in R$: $f(x) \in (0, a)$ ($a>0$);

when $a<0$, $k \neq 0$, $b \in R$: $f(x) \in (a, 0)$ ($a<0$).

(3). From (1), I got $f'(x) = \frac{df(x)}{dx} = \frac{-ake^{kx+b}}{(1+e^{kx+b})^2}$

$$\begin{aligned} \frac{k}{a} f(x)(f(x)-a) &= \frac{k}{a} \cdot \frac{a}{1+e^{kx+b}} \cdot \left(\frac{a}{1+e^{kx+b}} - a\right) \\ &= \frac{k}{1+e^{kx+b}} \cdot \frac{a-(a+a \cdot e^{kx+b})}{1+e^{kx+b}} \\ &= \frac{k}{1+e^{kx+b}} \cdot \frac{-a \cdot e^{kx+b}}{1+e^{kx+b}} \\ &= \frac{-ake^{kx+b}}{(1+e^{kx+b})^2} = \frac{d}{dx} f(x) \end{aligned}$$

5.

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

$= ((X(X^TX)^{-1})X^T)^T$

$= (X^T)^T((X^TX)^{-1})^T X^T$

$= X(X^T(X^T)^T)^{-1}X^T$

$= X(X^TX)^{-1}X^T$

$= H$

As $H^T = H$, H is symmetric.

(2). ① Base case: when $k=1$, $H^k = H^1 = H$, true.

② Assume that $H^k = H$ for positive integer $k>1$, then we can prove $H^k = H$ by proving $H^{k+1} = H$.

$H^{k+1} = H^k \cdot H$

$= H \cdot H$

$= (X(X^TX)^{-1}X^T)(X(X^TX)^{-1}X^T)$

$= X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T$

$= X(X^TX)^{-1}(X^TX)(X^TX)^{-1}X^T$

$= X((X^TX)^{-1}(X^TX))(X^TX)^{-1}X^T$

$= X \cdot 1 \cdot (X^TX)^{-1}X^T$

$= X(X^TX)^{-1}X^T$

$= H$

As $H^{k+1} = H$, assumption holds, so $H^k = H$ for any positive integer k.

(3). ① Base case: when $k=1$, $(I-H)^k = (I-H)^1 = I-H$, true.

② Assume that $(I-H)^k = I-H$ for integer $k>1$, we can prove $(I-H)^k = I-H$ by proving $(I-H)^{k+1} = I-H$

$(I-H)^{k+1} = (I-H)^k(I-H)$

$= (I-H)(I-H)$

$= I^2 - IH - HI + H^2$

$= I - H - H + H^2$

$= I - 2H + H \cdot H$

$= I - 2H + H$ As $H \cdot H = H$ from (2)

$= I - H$

As $(I-H)^{k+1} = I - H$, assumption holds, so $(I-H)^k = I - H$ for any positive integer k.

(4). $\text{trace}(H) = \text{trace}(X(X^T X)^{-1} X^T)$

$= \text{trace}((X(X^T X)^{-1}) X^T)$

$= \text{trace}(X^T (X(X^T X)^{-1}))$

$= \text{trace}((X^T X)(X^T X)^{-1})$

$= \text{trace}(I)$

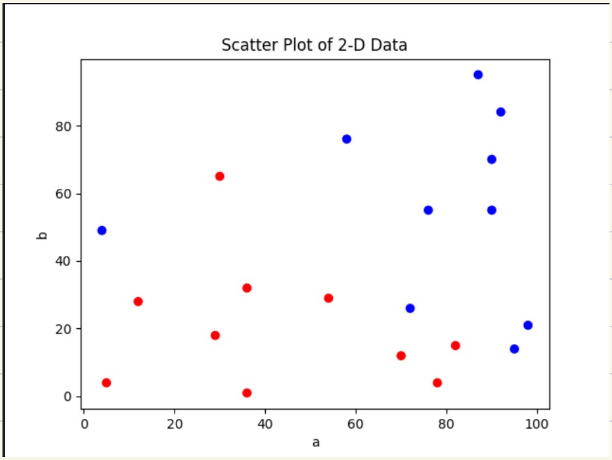
As X is an N by $d+1$ matrix, $X^T X$ is a $d+1$ by $d+1$ matrix,

then $I = (X^T X)(X^T X)^{-1}$ is a $d+1$ by $d+1$ identity matrix.

So $\text{trace}(I) = 1 \times (d+1) = d+1$

6.

(1).



(2)

```
import numpy as np
X = np.array([[4, 49],
              [5, 4],
              [12, 28],
              [29, 18],
              [30, 65],
              [36, 32],
              [36, 1],
              [54, 29],
              [58, 76],
              [70, 12],
              [72, 26],
              [76, 55],
              [78, 4],
              [82, 15],
              [87, 95],
              [90, 70],
              [90, 55],
              [92, 84],
              [95, 14],
              [98, 21]])
y = np.array([0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0,
1, 1, 0, 0, 0, 0, 0, 0])
np.random.seed(42)
W = np.random.randn(X.shape[1], 1)
b = np.zeros((1, 1))

def sigmoid(z):
   return 1 / (1 + np.exp(-z))

def forward_propagation(X, W, b):
   Z = np.dot(X, W) + b
   A = sigmoid(Z)
   return A

def backward_propagation(X, A, y):
   dZ = A - y.reshape(-1, 1)
   dW = np.dot(X.T, dZ)
   db = np.sum(dZ, axis=0, keepdims=True)
   return dW, db

def gradient_descent(X, y, W, b, learning_rate, num_iterations):
   for i in range(num_iterations):
      A = forward_propagation(X, W, b)
      dW, db = backward_propagation(X, A, y)
      W -= learning_rate * dW
      b -= learning_rate * db
   return W, b

learning_rate = 0.01
num_iterations = 1000
W, b = gradient_descent(X, y, W, b, learning_rate,
num_iterations)
# Calculate accuracy:
y_pred = forward_propagation(X, W, b)
y_pred_class = np.round(y_pred)
accuracy = np.mean(y_pred_class == y.reshape(-1, 1)) * 100
print(f"Accuracy on the training dataset: {accuracy}%")
```

(3). Learning rate: 0.001 Error: 30.0%
Learning rate: 0.01Error: 10.0%
Learning rate: 0.1 Error: 25.0%
Learning rate: 1 Error: 50.0%
Learning rate: 10 Error: 25.5%

(4).