Problem 1

[1]:
```
import numpy as np
from scipy.optimize import fmin
import matplotlib.pyplot as plt

data = np.genfromtxt('data.csv', delimiter=',')
x, y = data[:, 0], data[:, 1]
```

(a)

[669]:
```
def mse(w1):
    w0 = 2
    h_x = w0 + w1 * x
    return np.mean((h_x - y) ** 2)
```

[670]:
```
mse(1)
```

[670]:
```
139.60271462
```

(b)

[675]:
```
def log_likelihood(w1):
    sigma = 4
    w0 = 2
    h_x = w0 + w1 * x
    return -np.sum(((h_x - y) ** 2) / (2 * sigma**2)) - len(x) * np.log(np.sqrt(2 * np.pi) * sigma)
```

[676]:
```
log_likelihood(1)
```

[676]:
```
-333.3908863099782
```

(c)

[686]:
```
w1_min = fmin(mse, 0)[0]
```

```
Optimization terminated successfully.
         Current function value: 14.069906
         Iterations: 28
         Function evaluations: 56
```

[687]:
```
print(f'The value of w1 that minimizes the mean squared error is w1 = {w1_min}')
```

```
The value of w1 that minimizes the mean squared error is w1 = 3.0071875000000032
```

[688]:
```
closed_form_sol_w1 = np.cov(x, y, bias=True)[0, 1] / np.var(x)
closed_form_sol_w1
```

[688]:
```
2.980895446904831
```

3.0072 is approximately 2.98 so I think I got it right

[ ]:
```

```

(d)

[689]:
```
def neg_log_likelihood(w1):
    return -log_likelihood(w1)
```

[692]:
```
w1_max = fmin(neg_log_likelihood, 0)[0]
```

```
Optimization terminated successfully.
         Current function value: 137.245873
         Iterations: 28
         Function evaluations: 56
```

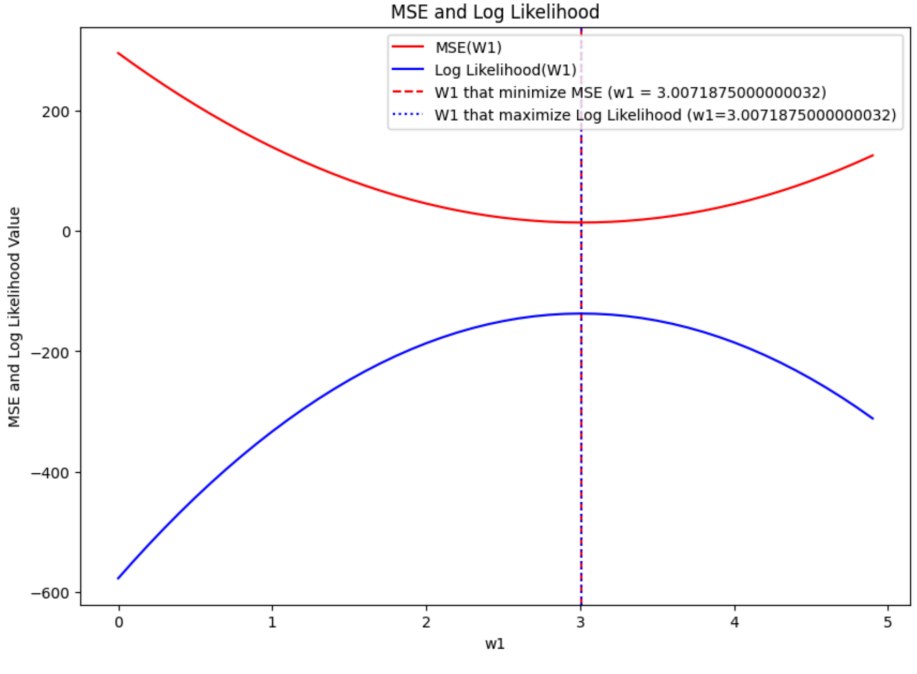
[695]:
```
print(f'The value of w1 that maximizes the log likelihood is w1 = {w1_max}')
```

```
The value of w1 that maximizes the log likelihood is w1 = 3.0071875000000032
```

(e)

710]:
```
w1s = np.arange(0,5,0.1)
mses = [mse(w1) for w1 in w1s]
log_likelihoods = [log_likelihood(w1) for w1 in w1s]
plt.figure(figsize=(10, 7))
plt.plot(w1s, mses, label="MSE(W1)", color="red")
plt.plot(w1s, log_likelihoods, label="Log Likelihood(W1)", color="blue")
plt.axvline(x=w1_min, color="red", linestyle="--", label=f"W1 that minimize MSE (w1 = {w1_min})")
plt.axvline(x=w1_max, color="blue", linestyle=":", label=f"W1 that maximize Log Likelihood (w1={w1_max})")
plt.xlabel("w1")
plt.ylabel("MSE and Log Likelihood Value")
plt.title("MSE and Log Likelihood")
plt.legend()
plt.show()
```

```
plt.show()
```



1+wq

problem 2,

Question is always: feature < threshold

Initial split (Left = Yes, Right = NO)

Temperature

55 | 62 | 65 | 71 | 72 | 73 | 79

58.5 63.5 68 71.5 72.5 76

Threshold	Left split	right split	$P_L u(S_L) + P_R u(S_R)$
58.5 Gini_idx	[Yes] $2 \cdot 1 \cdot 0 = 0$	[Yes, Yes, NO, NO, Yes, NO] $2 \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2} = 0.5$	$\frac{1}{7} \cdot 0 + \frac{6}{7} \cdot 0.5 = \frac{3}{7}$ ≈ 0.4286
63.5 Gini_idx	[Yes, NO] $2 \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2} = 0.5$	[Yes, Yes, NO, Yes, NO] $2 \cdot \frac{3}{5} \cdot \frac{2}{5} = \frac{12}{25} = 0.48$	$\frac{2}{7} \cdot \frac{1}{2} + \frac{5}{7} \cdot \frac{12}{25} = \frac{1}{7} + \frac{12}{35}$ $\approx 0.485\,7$
68 Gini_idx	[Yes, Yes, NO] $2 \cdot \frac{2}{3} \cdot \frac{1}{3} = \frac{4}{9} \approx 0.4444$	[Yes, NO, Yes, NO] $2 \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2} = 0.5$	$\frac{3}{7} \cdot 0.4444 + \frac{4}{7} \cdot \frac{1}{2}$ ≈ 0.4762
71.5 Gini_idx	[Yes, Yes, NO, Yes] $2 \cdot \frac{1}{4} \cdot \frac{3}{4} = \frac{6}{16} = 0.375$	[Yes, NO, NO] $2 \cdot \frac{1}{3} \cdot \frac{2}{3} = \frac{4}{9} \approx 0.4444$	$\frac{4}{7} \cdot 0.375 + \frac{3}{7} \cdot 0.4444$ ≈ 0.4048
72.5 Gini_idx	[Yes, Yes, Yes, NO, Yes] $2 \cdot \frac{1}{5} \cdot \frac{4}{5} = \frac{8}{25} = 0.32$	[NO, NO] $2 \cdot 1 \cdot 0 = 0$	$0.32 \cdot \frac{5}{7} + 0 \cdot \frac{2}{7}$ ≈ 0.2286
76 Gini_idx	[Yes, Yes, Yes, NO, Yes, NO] $2 \cdot \frac{2}{3} \cdot \frac{1}{3} = \frac{4}{9} \approx 0.4444$	[NO] $2 \cdot 1 \cdot 0 = 0$	$0.4444 \cdot \frac{6}{7} + 0 \cdot \frac{1}{7}$ ≈ 0.381

pressure:

1001 | 1003 | 1010 | 1011 | 1022 | 1025 | 1030

1002 1006.5 1010.5 1016.5 1023.5 1027.5

Threshold	Left Split	right split	$P_L u(S_L) + P_R u(S_R)$
1002 Gini_idx	[Yes] $2 \cdot 1 \cdot 0 = 0$	[Yes, NO, Yes, NO, Yes, NO] $2 \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2} = 0.5$	$\frac{1}{7} \cdot 0 + \frac{6}{7} \cdot \frac{1}{2}$ ≈ 0.4286
1006.5 Gini_idx	[Yes, Yes] $2 \cdot 1 \cdot 0 = 0$	[NO, Yes, NO, Yes, NO] $2 \cdot \frac{2}{5} \cdot \frac{3}{5} = \frac{12}{25} = 0.48$	$0 \cdot \frac{2}{7} + 0.48 \cdot \frac{5}{7}$ ≈ 0.3429

1010.5 Gini	[Yes, Yes, Yes] $2 \cdot 1 \cdot 0 = 0$	[NO, Yes, NO, NO] $2 \cdot \frac{1}{4} \cdot \frac{3}{4} = 0.375$	$\frac{3}{7} \cdot 0 + \frac{4}{7} \cdot 0.375$ $\approx \boxed{0.2143}$
1010.5 Gini	[Yes, Yes, Yes, NO] $2 \cdot \frac{3}{4} \cdot \frac{1}{4} = 0.375$	[Yes, NO, NO] $2 \cdot \frac{1}{3} \cdot \frac{2}{3} \approx 0.4444$	$\frac{4}{7} \cdot 0.375 + \frac{3}{7} \cdot 0.4444$ ≈ 0.4048
1023.5 Gini	[Yes, Yes, Yes, Yes, NO] $2 \cdot \frac{4}{5} \cdot \frac{1}{5} = \frac{8}{25} = 0.32$	[NO, NO] $2 \cdot 1 \cdot 0 = 0$	$\frac{5}{7} \cdot 0.32 + \frac{2}{7} \cdot 0$ ≈ 0.2286
1027.5 Gini	[Yes, Yes, Yes, NO, Yes, NO] $2 \cdot \frac{3}{3} \cdot \frac{1}{3} = \frac{4}{9} \approx 0.4444$	[NO] $2 \cdot 1 \cdot 0 = 0$	$\frac{6}{7} \cdot 0.4444 + \frac{1}{7} \cdot 0$ ≈ 0.381

The lowest possible Gini for the first split is 0.2143 where I squared it in the table, where the split is on pressure < 1010.5

then [4, 3]

pressure < 1010.5

Yes: [3, 0] [Yes, Yes, Yes] ✓

No: [1, 3] [Yes, NO, NO, NO]

Now work on this

I can see there only 1 Yes so I just need to isolate it out

We have

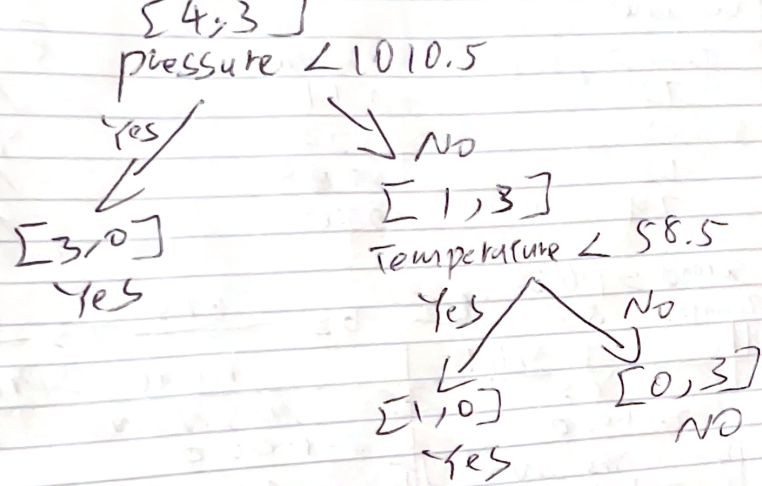
Temp	Pressure	Rain?
73	1011	No
55	1022	Yes
62	1025	No
79	1030	No

Second split

Temp: 55 — 58.5 — 62 — 67.5 — 73 — 76 — 79

Threshold	left split	right split	P(left) + P(right)
58.5	[Yes]	[No, No, No]	$0 \cdot \frac{1}{4} + 0 \cdot \frac{3}{4} = 0$
Gini	$2 - 1.0^2 - 0^2 = 0$	$2 - 1.0^2 - 0^2 = 0$	
67.5	[Yes, No]	[No, No]	$\frac{1}{2} \cdot 0.5 + \frac{1}{2} \cdot 0 = 0.25$
Gini	$2 - \frac{1}{2}^2 - \frac{1}{2}^2 = 0.5$	$2 - 1.0^2 - 0^2 = 0$	
76	[No, Yes, No]	[No]	$\frac{3}{4} \cdot 0.4444 + 0 \cdot \frac{1}{4} \approx 0.3333$
Gini	$2 - \frac{1}{3}^2 - \frac{2}{3}^2 = \frac{4}{9} \approx 0.4444$	$2 - 1.0^2 - 0^2 = 0$	
Pressure	1011 — 1016.5 — 1022	1023.5 — 1025 — 1027.5	1030
1016.5	[No]	[Yes, No, No]	$\frac{1}{4} \cdot 0 + \frac{3}{4} \cdot 0.4444 \approx 0.3333$
Gini	$2 - 1.0^2 - 0^2 = 0$	$2 - \frac{1}{3}^2 - \frac{2}{3}^2 = \frac{4}{9} \approx 0.4444$	
1023.5	[No, Yes]	[No, No]	$\frac{1}{2} \cdot 0.5 + \frac{1}{2} \cdot 0 = 0.25$
Gini	$2 - \frac{1}{2}^2 - \frac{1}{2}^2 = \frac{1}{2} = 0.5$	$2 \cdot 1 \cdot 0 = 0$	
1027.5	[No, Yes, No]	[No]	$\frac{3}{4} \cdot 0.4444 + \frac{1}{4} \cdot 0 \approx 0.3333$
Gini	$2 - \frac{1}{3}^2 - \frac{2}{3}^2 = \frac{4}{9} \approx 0.4444$	$2 - 1.0^2 - 0^2 = 0$	

The lowest possible Gini for the second split is 0 when I circled on the table, when the split is on Temperature < 58.5 then


[4,3]
pressure < 1010.5
Yes
No
[3,0]
Yes
[1,3]
Temperature < 58.5
Yes
No
[1,0]
Yes
[0,3]
No