Assignment No. 2 - Solution

Machine Learning Assignment

1 Problem 1: Gini Index Calculation

1.1 Part (a): Gini Index Before Splitting

Given dataset: 300 samples with 220 positive and 80 negative samples.

The Gini index is calculated as:

$$\text{Gini}(D) = 1 - \sum_{i=1}^{c} p_i^2 \tag{1}$$

where p_i is the proportion of samples belonging to class i.

For our dataset:

$$p_{\text{positive}} = \frac{220}{300} = \frac{11}{15} \approx 0.733 \tag{2}$$

$$p_{\text{negative}} = \frac{80}{300} = \frac{4}{15} \approx 0.267 \tag{3}$$

Therefore:

$$\text{Gini}(D) = 1 - \left(\frac{11}{15}\right)^2 - \left(\frac{4}{15}\right)^2 \tag{4}$$

$$= 1 - \frac{121}{225} - \frac{16}{225} \tag{5}$$

$$= 1 - \frac{137}{225} \tag{6}$$

$$= \frac{88}{225} \approx 0.391 \tag{7}$$

1.2 Part (b): Weighted Gini Index After Splitting

After splitting:

- Left subset: 90 positive, 10 negative (total = 100)
- Right subset: 100 positive, 100 negative (total = 200)

For the left subset:

$$p_{\text{pos,left}} = \frac{90}{100} = 0.9 \tag{8}$$

$$p_{\text{neg,left}} = \frac{10}{100} = 0.1 \tag{9}$$

$$\text{Gini}(D_{\text{left}}) = 1 - (0.9)^2 - (0.1)^2 = 1 - 0.81 - 0.01 = 0.18 \tag{10}$$

For the right subset:

$$p_{\text{pos,right}} = \frac{100}{200} = 0.5 \tag{11}$$

$$p_{\text{neg,right}} = \frac{100}{200} = 0.5 \tag{12}$$

$$\text{Gini}(D_{\text{right}}) = 1 - (0.5)^2 - (0.5)^2 = 1 - 0.25 - 0.25 = 0.5 \tag{13}$$

The weighted Gini index after splitting:

$$\begin{aligned}
\text{Gini}_{\text{weighted}} &= \frac{|D_{\text{left}}|}{|D|} \cdot \text{Gini}(D_{\text{left}}) + \frac{|D_{\text{right}}|}{|D|} \cdot \text{Gini}(D_{\text{right}}) && (14)\\
&= \frac{100}{300} \cdot 0.18 + \frac{200}{300} \cdot 0.5 && (15)\\
&= \frac{1}{3} \cdot 0.18 + \frac{2}{3} \cdot 0.5 && (16)\\
&= 0.06 + 0.333 = 0.393 && (17)
\end{aligned}$$

Conclusion: Since the weighted Gini index after splitting (0.393) is slightly higher than the original Gini index (0.391), this split does **not** improve purity. The split makes the dataset slightly less pure.

2 Problem 2: Regression Tree Construction

Given dataset:

X_1	X_2	Y
1	5	10
2	6	12
3	8	15
4	10	18
5	12	21
6	15	25
7	18	28
8	20	30

2.1 Part (a): Finding Best Splitting Point for X_1

First, calculate the overall mean of Y:

$$\bar{Y} = \frac{10 + 12 + 15 + 18 + 21 + 25 + 28 + 30}{8} = \frac{159}{8} = 19.875 \tag{18}$$

The Sum of Squared Errors (SSE) before splitting:

$$\begin{aligned}
\text{SSE}_{\text{total}} &= \sum_{i=1}^{8} (y_i - \bar{Y})^2 && (19)\\
&= (10 - 19.875)^2 + (12 - 19.875)^2 + \ldots + (30 - 19.875)^2 && (20)\\
&= 97.516 + 62.016 + 23.766 + 3.516 + 1.266 + 26.266 + 66.016 + 102.516 && (21)\\
&= 382.875 && (22)
\end{aligned}$$

For regression trees, we consider splitting points between consecutive values of X_1. The possible splitting points are: 1.5, 2.5, 3.5, 4.5, 5.5, 6.5, 7.5.

Let's calculate SSE for each potential split:

Split at $X_1 = 3.5$:

- Left: $X_1 \leq 3.5 \to (1, 10), (2, 12), (3, 15) \to \bar{Y}_L = 12.33$
- Right: $X_1 > 3.5 \to (4, 18), (5, 21), (6, 25), (7, 28), (8, 30) \to \bar{Y}_R = 24.4$

$$\begin{aligned}
\text{SSE}_L &= (10 - 12.33)^2 + (12 - 12.33)^2 + (15 - 12.33)^2 = 5.44 + 0.11 + 7.11 = 12.66 && (23)\\
\text{SSE}_R &= (18 - 24.4)^2 + (21 - 24.4)^2 + (25 - 24.4)^2 + (28 - 24.4)^2 + (30 - 24.4)^2 && (24)\\
&= 40.96 + 11.56 + 0.36 + 12.96 + 31.36 = 97.2 && (25)\\
\text{SSE}_{\text{total}} &= 12.66 + 97.2 = 109.86 && (26)
\end{aligned}$$

Split at $X_1 = 4.5$:

- Left: $X_1 \leq 4.5 \rightarrow (1, 10), (2, 12), (3, 15), (4, 18) \rightarrow \bar{Y}_L = 13.75$
- Right: $X_1 > 4.5 \rightarrow (5, 21), (6, 25), (7, 28), (8, 30) \rightarrow \bar{Y}_R = 26$

$$\begin{aligned}
\text{SSE}_L &= (10 - 13.75)^2 + (12 - 13.75)^2 + (15 - 13.75)^2 + (18 - 13.75)^2 && (27) \\
&= 14.06 + 3.06 + 1.56 + 18.06 = 36.75 && (28) \\
\text{SSE}_R &= (21 - 26)^2 + (25 - 26)^2 + (28 - 26)^2 + (30 - 26)^2 && (29) \\
&= 25 + 1 + 4 + 16 = 46 && (30) \\
\text{SSE}_{\text{total}} &= 36.75 + 46 = 82.75 && (31)
\end{aligned}$$

After checking all possible splits, the split at $X_1 = 4.5$ gives the minimum SSE of 82.75.

2.2 Part (b): First Split of Regression Tree

The first split of the regression tree using SSE as the impurity measure is:

Root Node: Split on $X_1 \leq 4.5$

- **Left Branch** ($X_1 \leq 4.5$): Contains samples $(1, 10), (2, 12), (3, 15), (4, 18)$
  - Predicted value: $\bar{Y}_L = 13.75$
  - SSE: 36.75
- **Right Branch** ($X_1 > 4.5$): Contains samples $(5, 21), (6, 25), (7, 28), (8, 30)$
  - Predicted value: $\bar{Y}_R = 26$
  - SSE: 46

The reduction in SSE achieved by this split is:

$$\Delta\text{SSE} = 382.875 - 82.75 = 300.125 \tag{32}$$

This represents a significant improvement in the model's ability to predict the target variable.