Asignment5

1. Question1

(a) Calculate the Gini index

Split #1:

left child: $p_+ = 300/400 = 0.75,\ p_- = 100/400 = 0.25$

right child: $p_+ = 100/400 = 0.25,\ p_- = 300/400 = 0.75$

$$Gini_{left} = 1 - (0.75)^2 - (0.25)^2 = 0.375$$

$$Gini_{right} = 1 - (0.25)^2 - (0.75)^2 = 0.375$$

$$Gini_{all} = (400/800) * 0.375 + (400/800) * 0.375 = 0.375$$

Split #2:

left child: $p_+ = 200/600 = 1/3,\ p_- = 400/600 = 2/3$

right child: $p_+ = 200/200 = 1,\ p_- = 0$

$$Gini_{left} = 1 - (1/3)^2 - (2/3)^2 \approx 0.444$$

$$Gini_{right} = 1 - 1^2 - 0^2 = 0$$

$$Gini_{all} = (600/800) * 0.444 + (200/800) * 0 = 0.333$$

(b) Calculate the Misclassification index

Split #1:

left child: $p_+ = 300/400 = 0.75,\ p_- = 100/400 = 0.25$

right child: $p_+ = 100/400 = 0.25,\ p_- = 300/400 = 0.75$

$$\text{Misclassification}_{left} = 1 - \max(0.75, 0.25) = 0.25$$

$$\text{Misclassification}_{right} = 1 - \max(0.25, 0.75) = 0.25$$

$$\text{Misclassification}_{all} = (400/800) * 0.25 + (400/800) * 0.25 = 0.25$$

Split #2:

left child: $p_+ = 200/600 = 1/3,\ p_- = 400/600 = 2/3$

right child: $p_+ = 200/200 = 1,\ p_- = 0$

$$\text{Misclassification}_{left} = 1 - \max(0.667, 0.333) = 0.333$$

$$\text{Misclassification}_{right} = 1 - \max(1, 0) = 0$$

$$\text{Misclassification}_{all} = (600/800) * 0.333 + (200/800) * 0 = 0.25$$

(c) Which one of the two splitting methods is better and why?

For split #1, the Gini index is 0.375 and the Misclassification index is 0.25. For split #2, the Gini index is 0.333 and the Misclassification index is 0.25.

Although the Misclassification indexes of the two splits are the same (0.25), the Gini index is lower in Split #2 (0.334 vs. 0.375). The Gini index is more sensitive to changes in the distribution of categories, and therefore is more often used in decision trees to measure the purity of nodes.

Additionally, the right subset of split #2 is pure (all +ve), which is an ideal outcome for the decision tree. This means that we can clearly classify this part of the data.

Based on these two reasons, it can be concluded that split #2 is better.

2. Question 2

(1) Calculate the Entropy of the Whole Dataset

In 14 samples, there are 9 "Yes" and 5 "No". So we have:

$p_+ = 9/14,\ p_- = 5/14$

$$\begin{aligned}\text{Entropy} &= -p_+\log_2 p_+ - p_-\log_2 p_- \\ &\approx -0.643 \cdot \log_2 0.643 - 0.357 \cdot \log_2 0.357 \\ &\approx 0.940\end{aligned}$$

(2) Calculate the information gain for each of the four attributes.

Attribute1: Price (values: 250k, 300k, 350k)

Price=250k: House 3, 7, 12, 13, all "Yes":

$$\text{Entropy(Price = 250k)} = 0$$

Price=300k: House 1, 2, 8, 9, 11. 3 "No" and 2 "Yes":

$$\text{Entropy(Price = 300k)} = -0.4 \cdot \log_2 0.4 - 0.6 \cdot \log_2 0.6 \approx 0.971$$

Price=350k: House 4, 5, 6, 10, 14. 2 "No" and 3 "Yes":

$$\text{Entropy}(\text{Price} = 350\text{k}) \approx 0.971$$

So we have: $\text{Entropy}_{\text{Price}} = 4/14 * 0 + 5/14 * 0.971 + 5/14 * 0.971 \approx 0.694$

Information gain is:

$$\text{IG}_{\text{Price}} = 0.940 - 0.694 = 0.246$$

Attribute2: Number of Bedrooms (values: 1, 2, 3)

Number of Bedrooms=1: House 1, 2, 3, 13. 2 "No" and 2 "Yes":

$$\text{Entropy}(\text{NoB} = 1) = -0.5 \cdot \log_2 0.5 - 0.5 \cdot \log_2 0.5 = 1$$

Number of Bedrooms=2: House 4, 8, 10, 11, 12, 14. 2 "No" and 4 "Yes":

$$\text{Entropy}(\text{NoB} = 2) = -1/3 \cdot \log_2(1/3) - 2/3 \cdot \log_2(2/3) \approx 0.918$$

Number of Bedrooms=3: House 5, 6, 7, 9. 1 "No" and 3 "Yes":

$$\text{Entropy}(\text{NoB} = 3) = -0.25 \cdot \log_2 0.25 - 0.75 \cdot \log_2 0.75 = 0.811$$

So we have: $\text{Entropy}_{\text{NoB}} = 4/14 * 1 + 6/14 * 0.918 + 4/14 * 0.811 \approx 0.911$

Information gain is:

$$\text{IG}_{\text{NoB}} = 0.940 - 0.911 = 0.029$$

Attribute3: Size (sqft) (values: 3500, 5000)

Size=3500: House 1, 2, 3, 4, 8, 12, 14. 4 "No" and 3 "Yes":

$$\text{Entropy}(\text{Size} = 3500) = -4/7 \cdot \log_2(4/7) - 3/7 \cdot \log_2(3/7) \approx 0.985$$

Size=5000: House 5, 6, 7, 9, 10, 11, 13. 1 "No" and 6 "Yes":

$$\text{Entropy}(\text{Size} = 5000) = -1/7 \cdot \log_2(1/7) - 6/7 \cdot \log_2(6/7) \approx 0.592$$

So we have: $\text{Entropy}_{\text{Size}} = 7/14 * 0.985 + 7/14 * 0.592 = 0.7885$

Information gain is:

$$\text{IG}_{\text{Size}} = 0.940 - 0.7885 = 0.1515$$

Attribute4: Distance to Bus-Stop (values: far, near)

Distance to Bus-Stop=far: House 1, 3, 4, 5, 8, 9, 10, 13. 2 "No" and 6 "Yes":

$$\text{Entropy}(\text{DtoB} = \text{far}) = -2/8 \cdot \log_2(2/8) - 6/8 \cdot \log_2(6/8) \approx 0.811$$

Distance to Bus-Stop=near: House 2, 6, 7, 11, 12, 14. 3 "No" and 3 "Yes":

$$\text{Entropy}(\text{DtoB} = \text{near}) = -3/6 \cdot \log_2(3/6) - 3/6 \cdot \log_2(3/6) = 1$$

So we have: $\text{Entropy}_{\text{DtoB}} = 8/14 * 0.811 + 6/14 * 1 \approx 0.892$

Information gain is:

$$\mathrm{IG}_{\mathrm{DtoB}} = 0.940 - 0.892 = 0.048$$

Therefore, the first attribute to split on is "Price", which has the maximal informatin gain (0.246).

(3) For the first attribute, we use "Price" to split and have three subsets of houses:

Subset1: Price=250k: House 3, 7, 12, 13, all "Yes"

Subset2: Price=300k: House 1, 2, 8, 9, 11. 3 "No" and 2 "Yes"

Subset3: Price=350k: House 4, 5, 6, 10, 14. 2 "No" and 3 "Yes"

Subset1 is pure, so we only need to deal with subset2 and subset3.

Subset2: Price=300k: House 1, 2, 8, 9, 11. 3 "No" and 2 "Yes"

$$\mathrm{Entropy(Subset2)} = -0.4 \cdot \log_2 0.4 - 0.6 \cdot \log_2 0.6 \approx 0.971$$

Attribute: Number of Badrooms (values: 1, 2, 3)

Number of Bedrooms=1: House 1, 2. 2 "No":

$$\mathrm{Entropy(NoB = 1)} = 0$$

Number of Bedrooms=2: House 8, 11. 1 "No" and 1 "Yes":

$$\mathrm{Entropy(NoB = 2)} = 1$$

Number of Bedrooms=3: House 9. 1 "Yes":

$$\mathrm{Entropy(NoB = 3)} = 0$$

So we have: $\mathrm{Entropy}_{\mathrm{NoB}} = 2/5 * 0 + 2/5 * 1 + 1/5 * 0 = 0.4$

Information gain is:

$$\mathrm{IG}_{\mathrm{NoB}} = 0.971 - 0.4 = 0.571$$

Attribute: Size (sqft) (values: 3500, 5000)

Size=3500: House 1, 2, 8. 3 "No":

$$\mathrm{Entropy(Size = 3500)} = 0$$

Size=5000: House 9, 11. 2 "Yes":

$$\mathrm{Entropy(Size = 5000)} = 0$$

So we have: $\mathrm{Entropy}_{\mathrm{Size}} = 0$

Information gain is:

$$\mathrm{IG}_{\mathrm{Size}} = 0.971$$

Attribute: Distance to Bus-Stop (values: far, near)

Distance to Bus-Stop=far: House 1, 8, 9. 2 "No" and 1 "Yes":

$$\mathrm{Entropy}(\mathrm{DtoB} = \mathrm{far}) = -2/3 \cdot \log_2(2/3) - 1/3 \cdot \log_2(1/3) \approx 0.918$$

Distance to Bus-Stop=near: House 2, 11. 1 "No" and 1 "Yes":

$$\mathrm{Entropy}(\mathrm{DtoB} = \mathrm{near}) = 1$$

So we have: $\mathrm{Entropy}_{\mathrm{DtoB}} = 3/5 * 0.918 + 2/5 * 1 \approx 0.9508$

Information gain is:

$$\mathrm{IG}_{\mathrm{DtoB}} = 0.971 - 0.9508 = 0.0202$$

Therefore, **the attribute to split subset2 is "Size"**, which has the maximal informatin gain (0.971). The two resulting subsets are pure, so no further division is necessary.

Subset3: Price=350k: House 4, 5, 6, 10, 14. 2 "No" and 3 "Yes"

$$\mathrm{Entropy}(\mathrm{Subset3}) = -0.4 \cdot \log_2 0.4 - 0.6 \cdot \log_2 0.6 \approx 0.971$$

Attribute: Number of Badrooms (values: 2, 3)

Number of Bedrooms=2: House 4, 10, 14. 1 "No" and 2 "Yes":

$$\mathrm{Entropy}(\mathrm{NoB} = 2) = -1/3 \cdot \log_2(1/3) - 2/3 \cdot \log_2(2/3) \approx 0.918$$

Number of Bedrooms=3: House 5, 6. 1 "No" and 1 "Yes":

$$\mathrm{Entropy}(\mathrm{NoB} = 3) = 1$$

So we have: $\mathrm{Entropy}_{\mathrm{NoB}} = 3/5 * 0.918 + 2/5 * 1 = 0.9508$

Information gain is:

$$\mathrm{IG}_{\mathrm{NoB}} = 0.971 - 0.9508 = 0.0202$$

Attribute: Size (sqft) (values: 3500, 5000)

Size=3500: House 4, 14. 1 "No" and 1 "Yes":

$$\mathrm{Entropy}(\mathrm{Size} = 3500) = 1$$

Size=5000: House 5, 6, 10. 1 "No" and 2 "Yes":

$$\mathrm{Entropy}(\mathrm{Size} = 5000) = -1/3 \cdot \log_2(1/3) - 2/3 \cdot \log_2(2/3) \approx 0.918$$

So we have: $\mathrm{Entropy}_{\mathrm{Size}} = 2/5 * 1 + 3/5 * 0.918 = 0.9508$

Information gain is:

$$\mathrm{IG}_{\mathrm{Size}} = 0.971 - 0.9508 = 0.0202$$

Attribute: Distance to Bus-Stop (values: far, near)

Distance to Bus-Stop=far: House 4, 5, 10. 3 “Yes”:

$$\mathrm{Entropy}(\mathrm{DtoB} = \mathrm{far}) = 0$$

Distance to Bus-Stop=near: House 6, 14. 2 “No”:

$$\mathrm{Entropy}(\mathrm{DtoB} = \mathrm{near}) = 0$$

So we have: $\mathrm{Entropy}_{\mathrm{DtoB}} = 0$

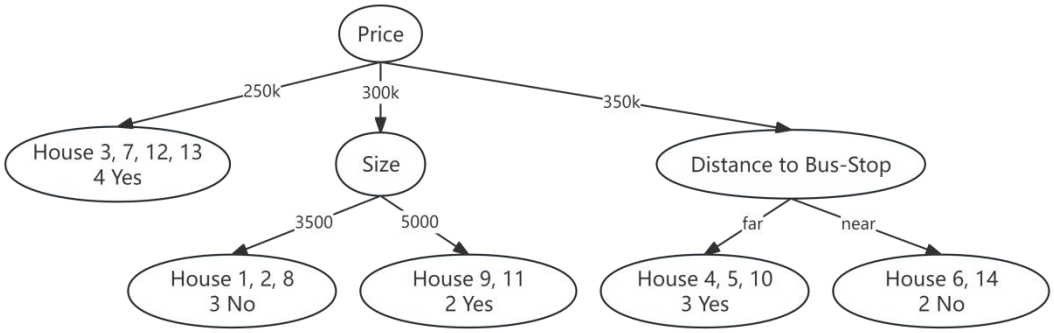
Information gain is:

$$\mathrm{IG}_{\mathrm{DtoB}} = 0.971$$

Therefore, **the attribute to split subset3 is “Distance to Bus-Stop”**, which has the maximal informatin gain (0.971). The two resulting subsets are pure, so no further division is necessary.

(4) Draw the final tree.

The final tree is as follows:



3. Question3

From the problem description, we know that:

input: $[\mathrm{i}_1 = 1,\ \mathrm{i}_2 = 1]$

weights of the hidden layer:

weights of h_1: $[\mathrm{w}_{11} = 0.8,\ \mathrm{w}_{21} = 0.2]$

weights of h_2: $[w_{12} = 0.4,\ w_{22} = 0.9]$

weights of h_3: $[w_{13} = 0.3,\ w_{23} = 0.5]$

weights of the output layer: $[w_{h1o} = 0.3,\ w_{h2o} = 0.5,\ w_{h3o} = 0.9]$

The activation function is sigmoid $\sigma(x) = 1/(1 + e^{-x})$ and the target output is 0. We use learning rate $\eta = 0.5$. For error function, we use the MSE error.

The first round:

Forward Propagation:

Hidden layer outputs:

$$h_1 = \sigma(1 * 0.8 + 1 * 0.2) = \sigma(1) \approx 0.731$$

$$h_2 = \sigma(1 * 0.4 + 1 * 0.9) = \sigma(1.3) \approx 0.786$$

$$h_3 = \sigma(1 * 0.3 + 1 * 0.5) = \sigma(0.8) \approx 0.690$$

Network output: $o = \sigma(0.731 * 0.3 + 0.786 * 0.5 + 0.690 * 0.9) = \sigma(1.2333) \approx 0.774$

Prediction Error: $E = \frac{1}{2}(o - target)^2 = \frac{1}{2}(0.774 - 0)^2 \approx 0.2995$

Back Propagation:

To calculate the gradient, we first write the gradient expression of the activation function. For the sigmoid function $\sigma(x)$, its gradient can be expressed as:

$$\partial\sigma(x)/\partial x = \sigma(x) \cdot (1 - \sigma(x))$$

Output Layer Gradients:

$$\partial E/\partial o = o - target = 0.774$$

$$\partial E/\partial w_{h1o} = (o - target) \cdot o \cdot (1 - o) \cdot h_1 = 0.599076 * 0.226 * 0.731 \approx 0.099$$

$$\partial E/\partial w_{h2o} = (o - target) \cdot o \cdot (1 - o) \cdot h_2 = 0.599076 * 0.226 * 0.786 \approx 0.106$$

$$\partial E/\partial w_{h3o} = (o - target) \cdot o \cdot (1 - o) \cdot h_3 = 0.599076 * 0.226 * 0.690 \approx 0.093$$

Hidden layer Gradients:

$$\partial E/\partial w_{11} = \frac{\partial E}{\partial o} \cdot \frac{\partial o}{\partial h_1} \cdot \frac{\partial h_1}{\partial w_{11}} = (o - target) \cdot [o \cdot (1 - o) \cdot w_{h1o}] \cdot [h_1 \cdot (1 - h_1) \cdot i_1]$$

$$= 0.774 * (0.774 * 0.226 * 0.3) * (0.731 * 0.269 * 1) \approx 0.0080$$

$$\partial \mathrm{E}/\partial \mathrm{w}_{21} = \frac{\partial \mathrm{E}}{\partial \mathrm{o}} \cdot \frac{\partial \mathrm{o}}{\partial h_1} \cdot \frac{\partial h_1}{\partial \mathrm{w}_{21}} = (\mathrm{o} - \mathrm{target}) \cdot [\mathrm{o} \cdot (1 - \mathrm{o}) \cdot \mathrm{w}_{\mathrm{h1o}}] \cdot [h_1 \cdot (1 - h_1) \cdot \mathrm{i}_2]$$

$$= 0.774 * (0.774 * 0.226 * 0.3) * (0.731 * 0.269 * 1) \approx 0.0080$$

$$\partial \mathrm{E}/\partial \mathrm{w}_{12} = \frac{\partial \mathrm{E}}{\partial \mathrm{o}} \cdot \frac{\partial \mathrm{o}}{\partial h_2} \cdot \frac{\partial h_2}{\partial \mathrm{w}_{12}} = (\mathrm{o} - \mathrm{target}) \cdot [\mathrm{o} \cdot (1 - \mathrm{o}) \cdot \mathrm{w}_{\mathrm{h2o}}] \cdot [h_2 \cdot (1 - h_2) \cdot \mathrm{i}_1]$$

$$= 0.774 * (0.774 * 0.226 * 0.5) * (0.786 * 0.214 * 1) \approx 0.0114$$

$$\partial \mathrm{E}/\partial \mathrm{w}_{22} = \frac{\partial \mathrm{E}}{\partial \mathrm{o}} \cdot \frac{\partial \mathrm{o}}{\partial h_2} \cdot \frac{\partial h_2}{\partial \mathrm{w}_{22}} = (\mathrm{o} - \mathrm{target}) \cdot [\mathrm{o} \cdot (1 - \mathrm{o}) \cdot \mathrm{w}_{\mathrm{h2o}}] \cdot [h_2 \cdot (1 - h_2) \cdot \mathrm{i}_2]$$

$$= 0.774 * (0.774 * 0.226 * 0.5) * (0.786 * 0.214 * 1) \approx 0.0114$$

$$\partial \mathrm{E}/\partial \mathrm{w}_{13} = \frac{\partial \mathrm{E}}{\partial \mathrm{o}} \cdot \frac{\partial \mathrm{o}}{\partial h_3} \cdot \frac{\partial h_3}{\partial \mathrm{w}_{13}} = (\mathrm{o} - \mathrm{target}) \cdot [\mathrm{o} \cdot (1 - \mathrm{o}) \cdot \mathrm{w}_{\mathrm{h3o}}] \cdot [h_3 \cdot (1 - h_3) \cdot \mathrm{i}_1]$$

$$= 0.774 * (0.774 * 0.226 * 0.9) * (0.690 * 0.310 * 1) \approx 0.0260$$

$$\partial \mathrm{E}/\partial \mathrm{w}_{23} = \frac{\partial \mathrm{E}}{\partial \mathrm{o}} \cdot \frac{\partial \mathrm{o}}{\partial h_3} \cdot \frac{\partial h_3}{\partial \mathrm{w}_{23}} = (\mathrm{o} - \mathrm{target}) \cdot [\mathrm{o} \cdot (1 - \mathrm{o}) \cdot \mathrm{w}_{\mathrm{h3o}}] \cdot [h_3 \cdot (1 - h_3) \cdot \mathrm{i}_2]$$

$$= 0.774 * (0.774 * 0.226 * 0.9) * (0.690 * 0.310 * 1) \approx 0.0260$$

Then we can do weight updates:

The new weights of inputs to hidden layer neurons:

$$\mathrm{w}_{11}^{\mathrm{new}} = \mathrm{w}_{11} - \eta \cdot \frac{\partial \mathrm{E}}{\partial \mathrm{w}_{11}} = 0.8 - 0.5 * 0.0080 = 0.796$$

$$\mathrm{w}_{21}^{\mathrm{new}} = \mathrm{w}_{21} - \eta \cdot \frac{\partial \mathrm{E}}{\partial \mathrm{w}_{21}} = 0.2 - 0.5 * 0.0080 = 0.196$$

$$\mathrm{w}_{12}^{\mathrm{new}} = \mathrm{w}_{12} - \eta \cdot \frac{\partial \mathrm{E}}{\partial \mathrm{w}_{12}} = 0.4 - 0.5 * 0.0114 = 0.3943$$

$$\mathrm{w}_{22}^{\mathrm{new}} = \mathrm{w}_{22} - \eta \cdot \frac{\partial \mathrm{E}}{\partial \mathrm{w}_{22}} = 0.9 - 0.5 * 0.0114 = 0.8943$$

$$\mathrm{w}_{13}^{\mathrm{new}} = \mathrm{w}_{13} - \eta \cdot \frac{\partial \mathrm{E}}{\partial \mathrm{w}_{13}} = 0.3 - 0.5 * 0.0260 = 0.287$$

$$w_{23}^{new} = w_{23} - \eta \cdot \frac{\partial E}{\partial w_{23}} = 0.5 - 0.5 * 0.0260 = 0.487$$

The new weights of hidden layer neurons to the output neuron:

$$w_{h1o}^{new} = w_{h1o} - \eta \cdot \frac{\partial E}{\partial w_{h1o}} = 0.3 - 0.5 * 0.099 = 0.2505$$

$$w_{h2o}^{new} = w_{h2o} - \eta \cdot \frac{\partial E}{\partial w_{h2o}} = 0.5 - 0.5 * 0.106 = 0.447$$

$$w_{h3o}^{new} = w_{h3o} - \eta \cdot \frac{\partial E}{\partial w_{h3o}} = 0.9 - 0.5 * 0.093 = 0.8535$$

The second round:

Forward Propagation:

Hidden layer outputs:

$$h_1 = \sigma(1 * 0.796 + 1 * 0.196) = \sigma(0.992) \approx 0.729$$

$$h_2 = \sigma(1 * 0.3943 + 1 * 0.8943) = \sigma(1.2886) \approx 0.784$$

$$h_3 = \sigma(1 * 0.287 + 1 * 0.487) = \sigma(0.774) \approx 0.684$$

Network output: $o = \sigma(0.729 * 0.2505 + 0.784 * 0.447 + 0.684 * 0.8535) = \sigma(1.1168565) \approx 0.7534$

Prediction Error: $E = \frac{1}{2}(o - target)^2 = \frac{1}{2}(0.7534 - 0)^2 \approx 0.2838$

By comparing the results of the two forward propagations and the predicted losses, it can be observed that the loss corresponding to the updated parameters has decreased, which proves the correctness of the calculation.