Iterative Dichotomiser 3

The **ID3** algorithm is a decision tree algorithm used for classification tasks. It constructs a decision tree by recursively selecting the attribute that best splits the dataset based on **Information Gain** (IG). Here are the steps involved in ID3:

Here are the steps involved in ID3:

1. **Calculate the entropy of the total dataset (Entropy(S)):** Entropy is a measure of the disorder or uncertainty in a dataset. It is calculated as:

$$H(S) = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

Where p_i is the proportion of the class in the dataset. For example, for a dataset with 4 female (F) and 5 male (M) classes:

$$Entropy(4F, 5M) = -\left(\frac{4}{9}\log_2\frac{4}{9}\right) - \left(\frac{5}{9}\log_2\frac{5}{9}\right) = 0.9911$$

2. **Choose an attribute and split the dataset:** ID3 splits the dataset based on different attributes (features) to form branches.

3. **Calculate the entropy of each branch:** After splitting, the entropy for each subset (branch) is calculated. For example, if one branch has 1F and 3M:

$$Entropy(1F, 3M) = -\left(\frac{1}{4}\log_2\frac{1}{4}\right) - \left(\frac{3}{4}\log_2\frac{3}{4}\right) = 0.8113$$

Similarly, if another branch has 3F and 2M:

$$Entropy(3F, 2M) = -\left(\frac{3}{5}\log_2\frac{3}{5}\right) - \left(\frac{2}{5}\log_2\frac{2}{5}\right) = 0.9710$$

4. **Calculate the Information Gain (IG):** Information Gain is the reduction in uncertainty (entropy) after splitting the dataset on an attribute. It is calculated as:

$$IG(A) = H(S) - \sum_{i=1}^{n} p(S_i)H(S_i)$$

For example, if the dataset is split on hair length ($\text{Hair Length} \leq 5$):

$$Gain(\text{Hair Length} \leq 5) = 0.9911 - \left(\frac{4}{9} \times 0.8113 + \frac{5}{9} \times 0.9710\right) = 0.0911$$

5. **Repeat steps 2-4:** The attribute with the highest Information Gain is chosen as the decision node, and this process is repeated for sub-datasets until each sub-dataset contains a single class.

Giải thích:

- $p(S_i)$ là xác suất của nhánh S_i, tức là tỷ lệ số mẫu trong nhánh S_i trên tổng số mẫu trong tập dữ liệu S.
- Giả sử S có tổng cộng n mẫu, và nhánh S_i có n_i mẫu, thì xác suất của nhánh S_i được tính là:

$$p(S_i) = \frac{n_i}{n}$$

Áp dụng vào ví dụ

Giả sử ban đầu tập dữ liệu có 9 mẫu với 4 **Female** và 5 **Male**:

1. **Nhánh S_1 với 1 Female và 3 Male:**
   - Tổng số mẫu trong nhánh S_1 là $1 + 3 = 4$.
   - Tổng số mẫu ban đầu là 9.
   - Xác suất $p(S_1)$ sẽ là:

$$p(S_1) = \frac{4}{9}$$

2. **Nhánh S_2 với 3 Female và 2 Male:**
   - Tổng số mẫu trong nhánh S_2 là $3 + 2 = 5$.
   - Tổng số mẫu ban đầu là 9.
   - Xác suất $p(S_2)$ sẽ là:

$$p(S_2) = \frac{5}{9}$$

To make the example clearer and more consistent, let's build a simple table based on the **Humidity** attribute, showing both **High** and **Low** humidity and their corresponding outcomes (Yes/No). This table will help explain the entropy and information gain calculations

outlook	temperature	humidity	windy	play
sunny	hot	high	FALSE	yes
sunny	hot	low	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	TRUE	yes
rainy	cool	low	FALSE	yes
rainy	cool	high	TRUE	no
overcast	cool	low	TRUE	yes
sunny	mild	low	FALSE	no
sunny	cool	low	FALSE	no

Splitting on Attribute: **Humidity**

Now, we split the dataset on the attribute **Humidity** (High, Low):

- For **High Humidity**: 4 records → 3 Yes, 1 No

$$Entropy(High) = -\left(\frac{3}{4}\log_2\frac{3}{4}\right) - \left(\frac{1}{4}\log_2\frac{1}{4}\right) = 0.8113$$

- For **Low Humidity**: 5 records → 2 Yes, 3 No

$$Entropy(Low) = -\left(\frac{2}{5}\log_2\frac{2}{5}\right) - \left(\frac{3}{5}\log_2\frac{3}{5}\right) = 0.9709$$

Now, calculate the weighted average entropy after the split:

$$H(S_{Humidity}) = \left(\frac{4}{9}\times 0.8113\right) + \left(\frac{5}{9}\times 0.9709\right) = 0.9005$$

Finally, calculate **Information Gain (IG)** for the **Humidity** split:

$$IG(Humidity) = 0.9911 - 0.9005 = 0.0906$$

Splitting on Attribute: Windy

Next, we split the dataset on the attribute **Windy** (True, False):

- For **Windy = True**: 4 records → 2 Yes, 2 No

$$Entropy(Windy = True) = -\left(\frac{2}{4}\log_2\frac{2}{4}\right) - \left(\frac{2}{4}\log_2\frac{2}{4}\right) = 1.0$$

- For **Windy = False**: 5 records → 3 Yes, 2 No

$$Entropy(Windy = False) = -\left(\frac{3}{5}\log_2\frac{3}{5}\right) - \left(\frac{2}{5}\log_2\frac{2}{5}\right) = 0.9709$$

The weighted average entropy for this split is:

$$H(S_{Windy}) = \left(\frac{4}{9} \times 1.0\right) + \left(\frac{5}{9} \times 0.9709\right) = 0.9848$$

Now, calculate **Information Gain (IG)** for the **Windy** split:

$$IG(Windy) = 0.9911 - 0.9848 = 0.0063$$

Conclusion: Choose the Attribute with the Highest IG

Between **Humidity** (IG = 0.0906) and **Windy** (IG = 0.0063), **Humidity** has the higher Information Gain. Therefore, **Humidity** will be chosen for the next decision node.