Decision Trees

Dripta Mj
Department of Mathematics
Ramakrishna Mission Vivekananda Educational and Research Institute
Belur Math, India

Example

- Classification of public transport.

Auto

Class c_1

Taxi

Class c_2

Bus

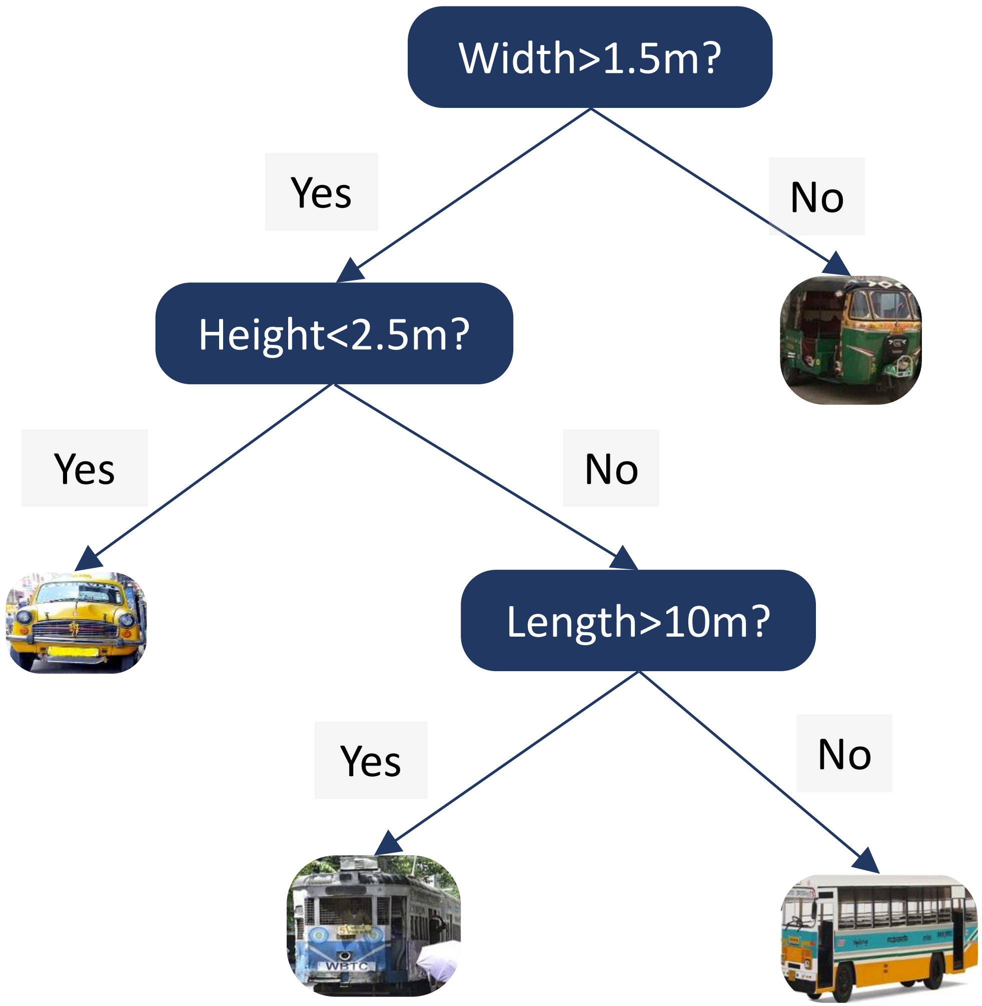
Class c_3

Tram

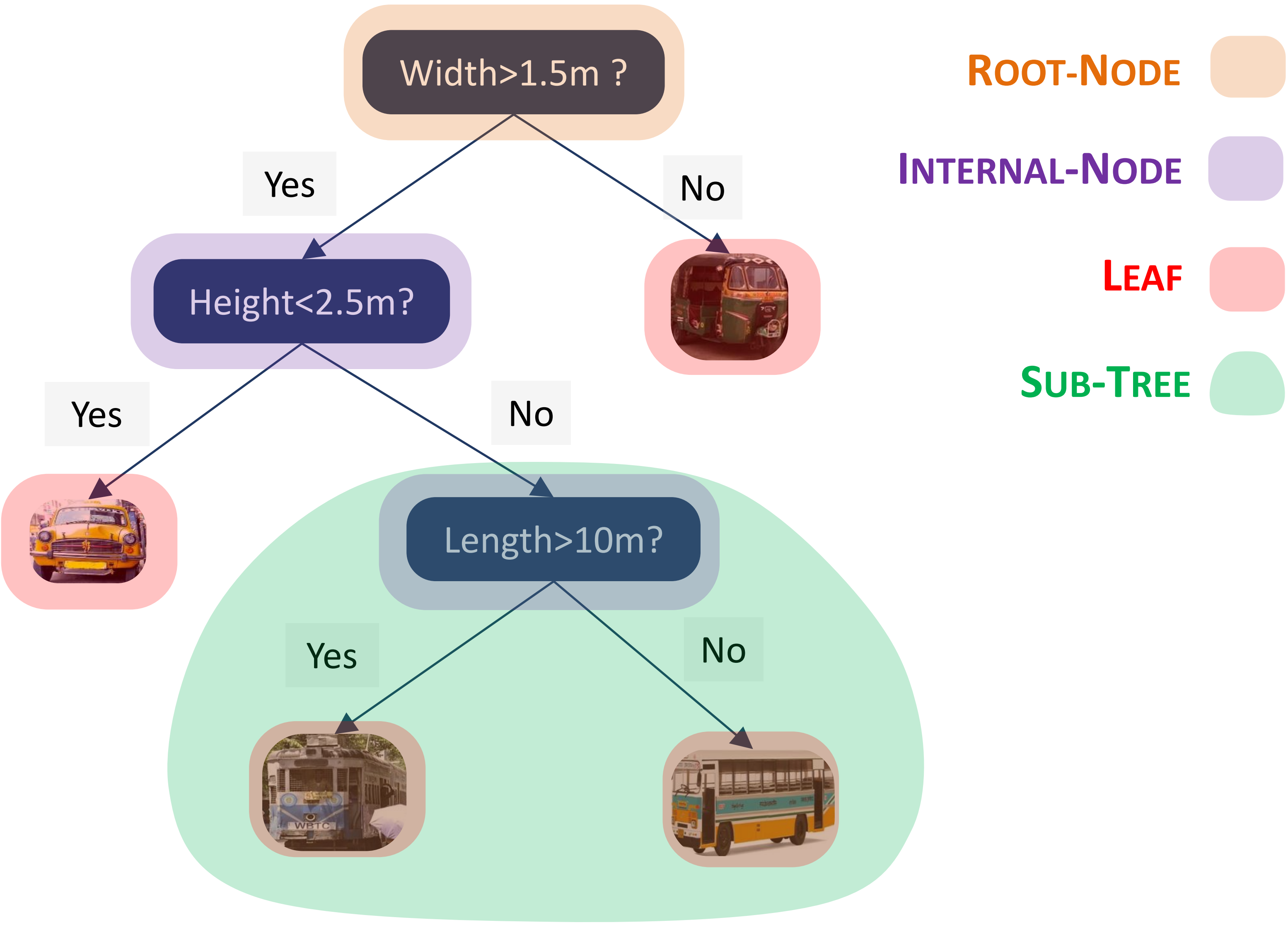
Class c_4

- Features:
  - Length (x_1)
  - Width (x_2)
  - Height (x_3)

Example

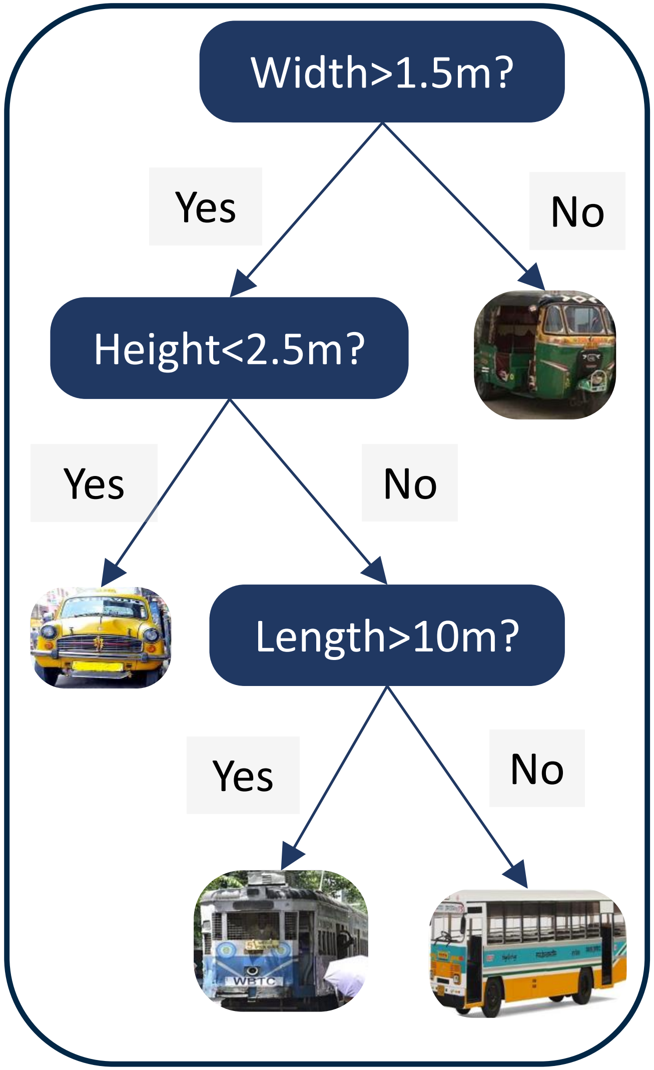


Terminology



Decision Tree

- Generate approximate solution through recursive top-down partitioning.
- The "informativeness" of the features are tested at each node.
- Criteria for quantifying feature informativeness:
  - Information Gain
  - Gain Ratio
  - Gini Index
- The most informative feature is selected for data-partitioning at a particular node.



Information: Intuition

- Are you going to class? Yes $\rightarrow 1$, No $\rightarrow 0$.
  - Specifying a choice $\rightarrow$ 1 bit of information
- Information conveyed by a sequence of 100 such independent events $\rightarrow$ 100 bits.
- Suppose you usually attend all the classes. If somebody tells me whether you are coming to the next class or not, then which of the following is more informative?
  - Yes
  - No ✓
- Another example: Temperature in Kolkata on December 20. Which of following is more informative?

22°C or 40°C ✓

- If the probability of an event is **high**, then the information conveyed by knowing that the event has occurred is **low**.

Information definition

- Suppose the probability of an event is p, then the information associated with it can be quantified as
$$I \equiv \log_2\left(\frac{1}{p}\right) = -\log_2(p)$$
- Note: $\log_2(1/p)$ is a decreasing function of p.
- Two questions:
  - Why log function?

    Ans.: log is a simple function with some nice properties, one of them being **additivity**. If x and y are two independent events, then the information conveyed through knowledge of the two events:
$$\begin{aligned} I_{x,y} &= \log_2\left(\frac{1}{P(x \text{ and } y)}\right) \\ &= \log_2\left(\frac{1}{p_x p_y}\right) = \log_2\left(\frac{1}{p_x}\right) + \log_2\left(\frac{1}{p_y}\right) \\ &= I_x + I_y \end{aligned}$$

Information definition

- Suppose the probability of an event is p, then the information associated with it can be quantified as

$$I \equiv \log_2\left(\frac{1}{p}\right) = -\log_2(p)$$

- Note: $\log_2(1/p)$ is a decreasing function of p.
- Two questions:
  - Why log function?
  - Why base 2?
    Ans.: This is from Shannon's convention. The unit of information is bits in this case
- Expected information in a set of K possible outcomes:

$$H(p_1, p_2, ..., p_K) = \sum_{k=1}^{K} p_k \log_2(1/p_k) = -\sum_{k=1}^{K} p_k \log_2(p_k)$$

- This is called **entropy**.

Entropy

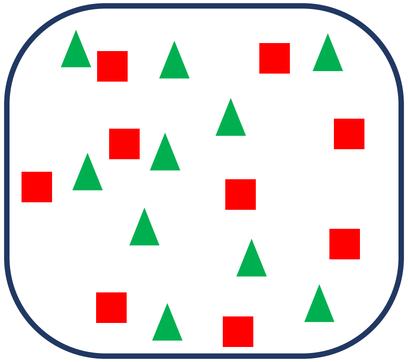
- Entropy is a measure of uncertainty/randomness in a dataset.
- Suppose we have a dataset $\mathcal{D}$ comprising of N points with K classes.
- The probability p_k of a data point to be in the kth class can be evaluated as

$$p_k = \frac{N_k}{N}$$

where N_k is the number of data points in class k.

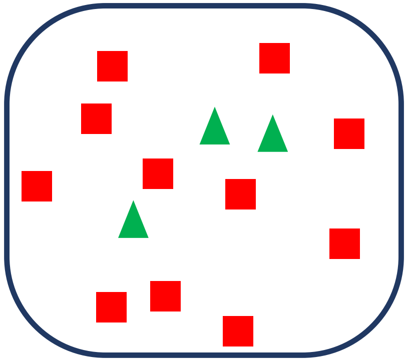
Entropy

High Uncertainty



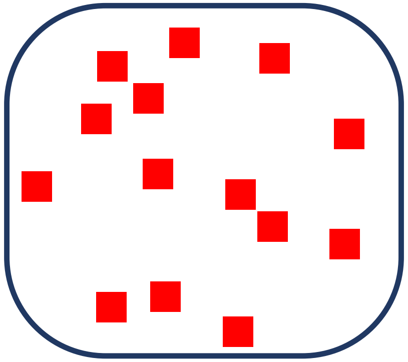
Entropy Level – High

Less Uncertainty



Entropy Level – Medium

No Uncertainty



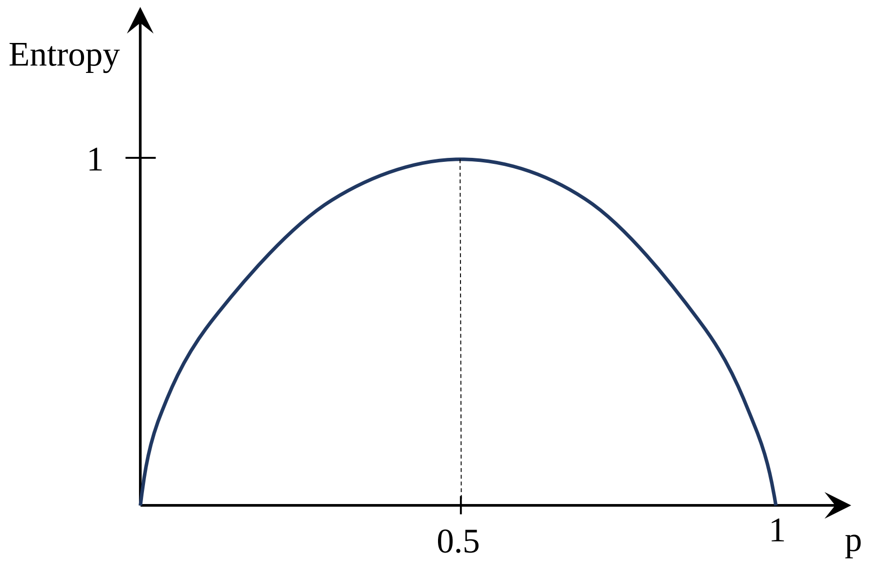
Entropy Level – Minimum

Entropy

- Entropy of the dataset:

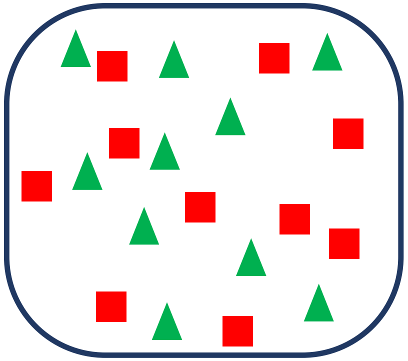
$$H(\mathcal{D}) = -\sum_{k=1}^{K} p_k \log_2 p_k$$

- Binary classification:



High entropy

- Entropy of the dataset is high if it comprise equally probable classes.



- Higher the entropy more the information content.
- For binary classification $\max H(\mathcal{D}) = 1$.
- For n-ary classification $\max H(\mathcal{D}) = \log_2 n$.

Information gain

- Information gain gives information on the importance of features.
- Suppose there are M features – $\{\mathcal{F}_1, \mathcal{F}_2, ..., \mathcal{F}_M\}$. Each of these features again takes different values.
- Suppose the mth feature can take any one of the β_m values from the set $\mathcal{V}_{\mathcal{F}_m} = \{v_{\mathcal{F}_m}^{(1)}, v_{\mathcal{F}_m}^{(2)}, ..., v_{\mathcal{F}_m}^{(\beta_m)}\}$.
- Let $\mathcal{D}_{\mathcal{F}_m,j}$ be the set comprising $n_{m,j}$ data points for which the mth feature $\mathcal{F}_m$ takes its jth value $v_{\mathcal{F}_m}^{(j)}$.
- The information gain after knowing the values of the mth feature $\mathcal{F}_m$ can then be given as:

$$IG(\mathcal{D}, \mathcal{F}_m) = H(\mathcal{D}) - \sum_{j=1}^{\beta_m} \left(\frac{n_{m,j}}{N}\right) H(\mathcal{D}_{\mathcal{F}_m,j})$$

Information gain

- $IG(\mathcal{D}, \mathcal{F}_m)$ is given as the entropy of $\mathcal{D}$ minus the weighted sum of entropy of its children.
- More information gain means less uncertainty on $\mathcal{D}$ after a particular feature is known.
- Information gain is used for identifying the best feature for discriminating between the output classes.
- Measures the amount of "information" a feature gives about the class.

Binary classification problem

Instance	Weather	Health	Teaching	Topic Importance	Going to class?
1	Hot	Good	Interesting	Medium	Yes
2	Cold	Average	Boring	High	Yes
3	Cold	Sick	Mediocre	Medium	No
4	Mild	Average	Interesting	High	Yes
5	Rainy	Sick	Mediocre	Low	No
6	Hot	Good	Boring	High	Yes
7	Rainy	Good	Mediocre	Medium	No
8	Mild	Good	Mediocre	Medium	Yes
9	Rainy	Good	Mediocre	High	Yes
10	Hot	Average	Interesting	Medium	Yes
11	Mild	Good	Boring	Low	No
12	Cold	Average	Interesting	Low	Yes
13	Mild	Sick	Interesting	High	Yes
14	Rainy	Average	Boring	Medium	No
15	Mild	Good	Interesting	Low	Yes

- 2 output classes:
  - Yes, going to class
  - No
- Dataset $\mathcal{D}$: 10 Yes and 5 No.
- Let output classes $\mathcal{C}_1$ and $\mathcal{C}_2$ correspond to Yes and No, respectively.

Binary classification problem

Instance	Weather	Health	Teaching	Topic Importance	Going to class?
1	Hot	Good	Interesting	Medium	Yes
2	Cold	Average	Boring	High	Yes
3	Cold	Sick	Mediocre	Medium	No
4	Mild	Average	Interesting	High	Yes
5	Rainy	Sick	Mediocre	Low	No
6	Hot	Good	Boring	High	Yes
7	Rainy	Good	Mediocre	Medium	No
8	Mild	Good	Mediocre	Medium	Yes
9	Rainy	Good	Mediocre	High	Yes
10	Hot	Average	Interesting	Medium	Yes
11	Mild	Good	Boring	Low	No
12	Cold	Average	Interesting	Low	Yes
13	Mild	Sick	Interesting	High	Yes
14	Rainy	Average	Boring	Medium	No
15	Mild	Good	Interesting	Low	Yes

- Four features:
  - Weather ($\mathcal{F}_1$) $\in$ { Hot, Cold, Rainy, Mild }
  - Health ($\mathcal{F}_2$) $\in$ { Good, Average, Sick }
  - Teaching ($\mathcal{F}_3$) $\in$ { Interesting, Mediocre, Boring }
  - Topic Importance ($\mathcal{F}_4$) $\in$ { High, Medium, Low }

Binary classification problem

Instance	Weather	Health	Teaching	Topic Importance	Going to class?
1	Hot	Good	Interesting	Medium	Yes
2	Cold	Average	Boring	High	Yes
3	Cold	Sick	Mediocre	Medium	No
4	Mild	Average	Interesting	High	Yes
5	Rainy	Sick	Mediocre	Low	No
6	Hot	Good	Boring	High	Yes
7	Rainy	Good	Mediocre	Medium	No
8	Mild	Good	Mediocre	Medium	Yes
9	Rainy	Good	Mediocre	High	Yes
10	Hot	Average	Interesting	Medium	Yes
11	Mild	Good	Boring	Low	No
12	Cold	Average	Interesting	Low	Yes
13	Mild	Sick	Interesting	High	Yes
14	Rainy	Average	Boring	Medium	No
15	Mild	Good	Interesting	Low	Yes

- Therefore $p(\mathcal{C}_1) = \frac{10}{15}$ and $p(\mathcal{C}_2) = \frac{5}{15}$
- Entropy of the dataset: $H(\mathcal{D}) = -p(\mathcal{C}_1)\log_2 p(\mathcal{C}_1) - p(\mathcal{C}_2)\log_2 p(\mathcal{C}_2)$

$$= -\frac{10}{15}\log_2\left(\frac{10}{15}\right) - \frac{5}{15}\log_2\left(\frac{5}{15}\right)$$

$$= 0.918$$

Binary classification problem

Instance	Weather	Health	Teaching	Topic Importance	Going to class?
1	Hot	Good	Interesting	Medium	Yes
2	Cold	Average	Boring	High	Yes
3	Cold	Sick	Mediocre	Medium	No
4	Mild	Average	Interesting	High	Yes
5	Rainy	Sick	Mediocre	Low	No
6	Hot	Good	Boring	High	Yes
7	Rainy	Good	Mediocre	Medium	No
8	Mild	Good	Mediocre	Medium	Yes
9	Rainy	Good	Mediocre	High	Yes
10	Hot	Average	Interesting	Medium	Yes
11	Mild	Good	Boring	Low	No
12	Cold	Average	Interesting	Low	Yes
13	Mild	Sick	Interesting	High	Yes
14	Rainy	Average	Boring	Medium	No
15	Mild	Good	Interesting	Low	Yes

- Now we will check the information gain for each of the (four) features to determine the feature yielding the largest information gain.
- Consider feature $\mathcal{F}_1$: Weather.
- It can take one of the 4 possible values: $\{\text{Hot}, \text{Cold}, \text{Rainy}, \text{Mild}\}$. The values are indexed from 1 to 4.

Binary classification problem

Instance	Weather	Health	Teaching	Topic Importance	Going to class?
1	Hot	Good	Interesting	Medium	Yes
2	Cold	Average	Boring	High	Yes
3	Cold	Sick	Mediocre	Medium	No
4	Mild	Average	Interesting	High	Yes
5	Rainy	Sick	Mediocre	Low	No
6	Hot	Good	Boring	High	Yes
7	Rainy	Good	Mediocre	Medium	No
8	Mild	Good	Mediocre	Medium	Yes
9	Rainy	Good	Mediocre	High	Yes
10	Hot	Average	Interesting	Medium	Yes
11	Mild	Good	Boring	Low	No
12	Cold	Average	Interesting	Low	Yes
13	Mild	Sick	Interesting	High	Yes
14	Rainy	Average	Boring	Medium	No
15	Mild	Good	Interesting	Low	Yes

Instance	Weather	Health	Teaching	Topic Importance	Going to class?
1	Hot	Good	Interesting	Medium	Yes
6	Hot	Good	Boring	High	Yes
10	Hot	Average	Interesting	Medium	Yes

- Therefore

$$
\begin{aligned}
H(\mathcal{D}_{\mathcal{F}_1,1}) &= -p(\mathcal{C}_1|\mathcal{D}_{\mathcal{F}_1,1}) \log_2 p(\mathcal{C}_1|\mathcal{D}_{\mathcal{F}_1,1}) - p(\mathcal{C}_2|\mathcal{D}_{\mathcal{F}_1,1}) \log_2 p(\mathcal{C}_2|\mathcal{D}_{\mathcal{F}_1,1}) \\
&= -\frac{3}{3} \log_2 \left(\frac{3}{3}\right) - \frac{0}{3} \log_2 \left(\frac{0}{3}\right) \\
&= 0
\end{aligned}
$$

Binary classification problem

Instance	Weather	Health	Teaching	Topic Importance	Going to class?
1	Hot	Good	Interesting	Medium	Yes
2	Cold	Average	Boring	High	Yes
3	Cold	Sick	Mediocre	Medium	No
4	Mild	Average	Interesting	High	Yes
5	Rainy	Sick	Mediocre	Low	No
6	Hot	Good	Boring	High	Yes
7	Rainy	Good	Mediocre	Medium	No
8	Mild	Good	Mediocre	Medium	Yes
9	Rainy	Good	Mediocre	High	Yes
10	Hot	Average	Interesting	Medium	Yes
11	Mild	Good	Boring	Low	No
12	Cold	Average	Interesting	Low	Yes
13	Mild	Sick	Interesting	High	Yes
14	Rainy	Average	Boring	Medium	No
15	Mild	Good	Interesting	Low	Yes

Instance	Weather	Health	Teaching	Topic Importance	Going to class?
1	Hot	Good	Interesting	Medium	Yes
6	Hot	Good	Boring	High	Yes
10	Hot	Average	Interesting	Medium	Yes

- Similarly $H(\mathcal{D}_{\mathcal{F}_1,2}) = 0.918$, $H(\mathcal{D}_{\mathcal{F}_1,3}) = 0.811$ and $H(\mathcal{D}_{\mathcal{F}_1,4}) = 0.722$.
- Therefore information gain after feature $\mathcal{F}_1$ is known:

$$IG(\mathcal{D}, \mathcal{F}_1) = H(\mathcal{D}) - \left(\left(\frac{n_{1,1}}{N}\right) H(\mathcal{D}_{\mathcal{F}_1,1}) + \left(\frac{n_{1,2}}{N}\right) H(\mathcal{D}_{\mathcal{F}_1,2}) + \left(\frac{n_{1,3}}{N}\right) H(\mathcal{D}_{\mathcal{F}_1,3}) + \left(\frac{n_{1,4}}{N}\right) H(\mathcal{D}_{\mathcal{F}_1,4})\right)$$

Binary classification problem

Instance	Weather	Health	Teaching	Topic Importance	Going to class?
1	Hot	Good	Interesting	Medium	Yes
2	Cold	Average	Boring	High	Yes
3	Cold	Sick	Mediocre	Medium	No
4	Mild	Average	Interesting	High	Yes
5	Rainy	Sick	Mediocre	Low	No
6	Hot	Good	Boring	High	Yes
7	Rainy	Good	Mediocre	Medium	No
8	Mild	Good	Mediocre	Medium	Yes
9	Rainy	Good	Mediocre	High	Yes
10	Hot	Average	Interesting	Medium	Yes
11	Mild	Good	Boring	Low	No
12	Cold	Average	Interesting	Low	Yes
13	Mild	Sick	Interesting	High	Yes
14	Rainy	Average	Boring	Medium	No
15	Mild	Good	Interesting	Low	Yes

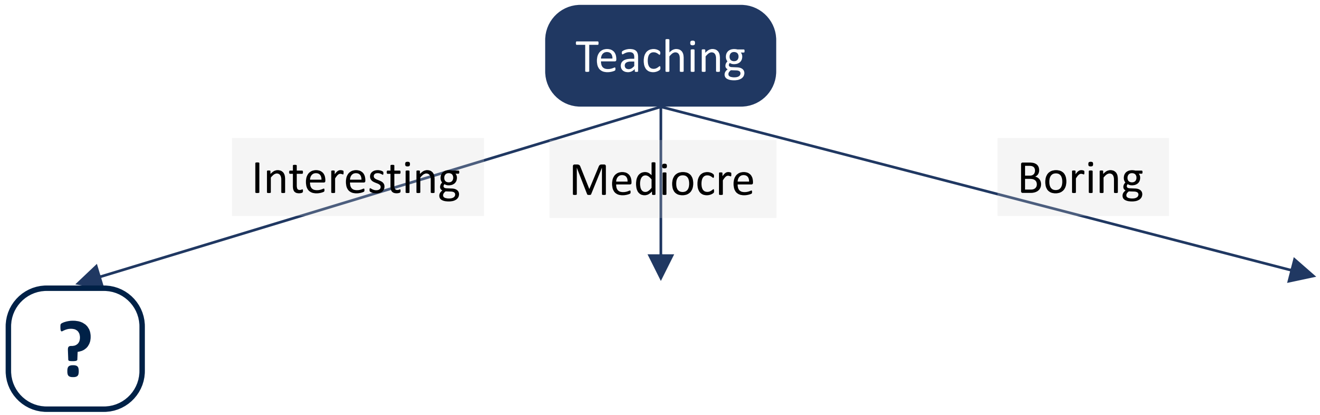
$$
\begin{aligned}
IG(\mathcal{D}, \mathcal{F}_1) &= 0.918 - \left(\frac{3}{15} \times 0 + \frac{3}{15} \times 0.918 + \frac{4}{15} \times 0.811 + \frac{5}{15} \times 0.722\right) \\
&= 0.277
\end{aligned}
$$

Binary classification problem

Instance	Weather	Health	Teaching	Topic Importance	Going to class?
1	Hot	Good	Interesting	Medium	Yes
2	Cold	Average	Boring	High	Yes
3	Cold	Sick	Mediocre	Medium	No
4	Mild	Average	Interesting	High	Yes
5	Rainy	Sick	Mediocre	Low	No
6	Hot	Good	Boring	High	Yes
7	Rainy	Good	Mediocre	Medium	No
8	Mild	Good	Mediocre	Medium	Yes
9	Rainy	Good	Mediocre	High	Yes
10	Hot	Average	Interesting	Medium	Yes
11	Mild	Good	Boring	Low	No
12	Cold	Average	Interesting	Low	Yes
13	Mild	Sick	Interesting	High	Yes
14	Rainy	Average	Boring	Medium	No
15	Mild	Good	Interesting	Low	Yes

- Similarly can compute the information gain for the other features:
  - $IG(\mathcal{D}, \mathcal{F}_2) = 0.091$
  - $IG(\mathcal{D}, \mathcal{F}_3) = 0.328$
  - $IG(\mathcal{D}, \mathcal{F}_4) = 0.251$
- Select the feature with the highest information gain as the root node as it is most informative.

Root node (Level-1)



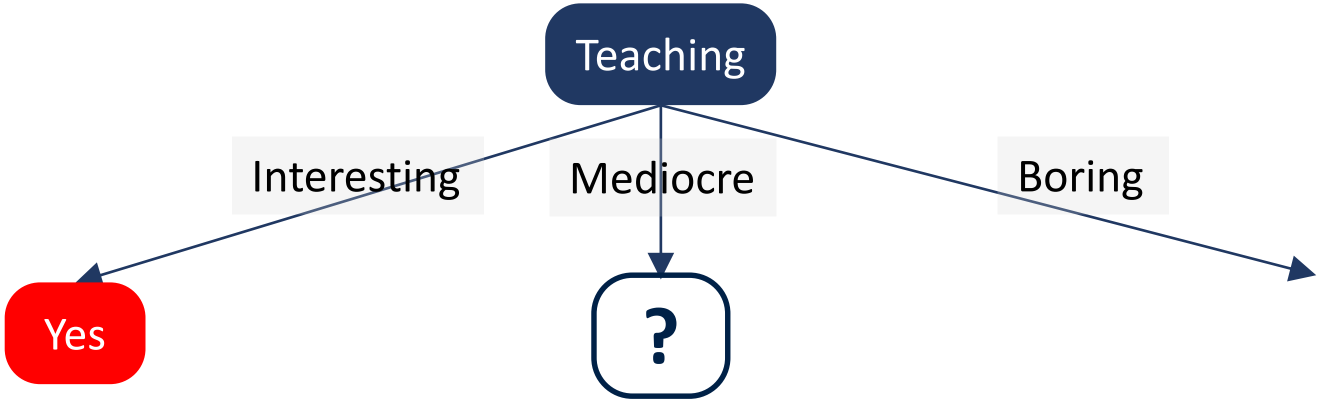
- In the present case feature $\mathcal{F}_3$ is taken as the root node.

Teaching – Interesting (Level-2, Node-1)

Instance	Weather	Health	Teaching	Topic Importance	Going to class?
1	Hot	Good	Interesting	Medium	Yes
4	Mild	Average	Interesting	High	Yes
10	Hot	Average	Interesting	Medium	Yes
12	Cold	Average	Interesting	Low	Yes
13	Mild	Sick	Interesting	High	Yes
15	Mild	Good	Interesting	Low	Yes

- Class labels of all the outputs are the same – Yes.
- Entropy is zero.
- No need for further subdivision.
- Expansion from a particular node is to be terminated when
  - all data points at that node belong to the same output class.
  - all features have been exhausted.

Level-2, Node-1



Teaching – Mediocre (Level-2, Node-2)

Instance	Weather	Health	Teaching	Topic Importance	Going to class?
3	Cold	Sick	Mediocre	Medium	No
5	Rainy	Sick	Mediocre	Low	No
7	Rainy	Good	Mediocre	Medium	No
8	Mild	Good	Mediocre	Medium	Yes
9	Rainy	Good	Mediocre	High	Yes

- Dataset $\mathcal{D}_{\mathcal{F}_3,2}$ – 2 Yes and 3 No.
- Entropy of this dataset:

$$H(\mathcal{D}_{\mathcal{F}_3,2}) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)$$
$$= 0.971$$

- Now compute the entropy of this dataset $\mathcal{D}_{\mathcal{F}_3,2}$ with respect to features $\mathcal{F}_1$, $\mathcal{F}_2$ and $\mathcal{F}_4$.
- For example

$$H(\mathcal{D}_{\mathcal{F}_1,2}|\mathcal{D}_{\mathcal{F}_3,2}) = p(\mathcal{C}_1)\log_2 p(\mathcal{C}_1|\mathcal{D}_{\mathcal{F}_1,2},\mathcal{D}_{\mathcal{F}_3,2}) + p(\mathcal{C}_2)\log_2 p(\mathcal{C}_2|\mathcal{D}_{\mathcal{F}_1,2},\mathcal{D}_{\mathcal{F}_3,2})$$
$$= -\frac{0}{1}\log_2\left(\frac{0}{1}\right) - \frac{1}{1}\log_2\left(\frac{1}{1}\right)$$
$$= 0$$

Level-2, Node-2

Instance	Weather	Health	Teaching	Topic Importance	Going to class?
3	Cold	Sick	Mediocre	Medium	No
5	Rainy	Sick	Mediocre	Low	No
7	Rainy	Good	Mediocre	Medium	No
8	Mild	Good	Mediocre	Medium	Yes
9	Rainy	Good	Mediocre	High	Yes

- Similarly can compute: $H(\mathcal{D}_{\mathcal{F}_1,3}|\mathcal{D}_{\mathcal{F}_3,2})$ =0.918 and $H(\mathcal{D}_{\mathcal{F}_1,4}|\mathcal{D}_{\mathcal{F}_3,2})$=0.

- Therefore information gain from $\mathcal{D}_{\mathcal{F}_3,2}$ after feature $\mathcal{F}_1$ is known:

$$IG(\mathcal{D}_{\mathcal{F}_3,2},\mathcal{F}_1) = 0.971 - \Big(\frac{1}{5}H(\mathcal{D}_{\mathcal{F}_1,2}|\mathcal{D}_{\mathcal{F}_3,2}) + \frac{3}{5}H(\mathcal{D}_{\mathcal{F}_1,3}|\mathcal{D}_{\mathcal{F}_3,2}) + \frac{1}{5}H(\mathcal{D}_{\mathcal{F}_1,4}|\mathcal{D}_{\mathcal{F}_3,2})\Big)$$

$$= 0.971 - (\frac{1}{5}\times 0 + \frac{3}{5}\times 0.918 + \frac{1}{5}\times 0)$$

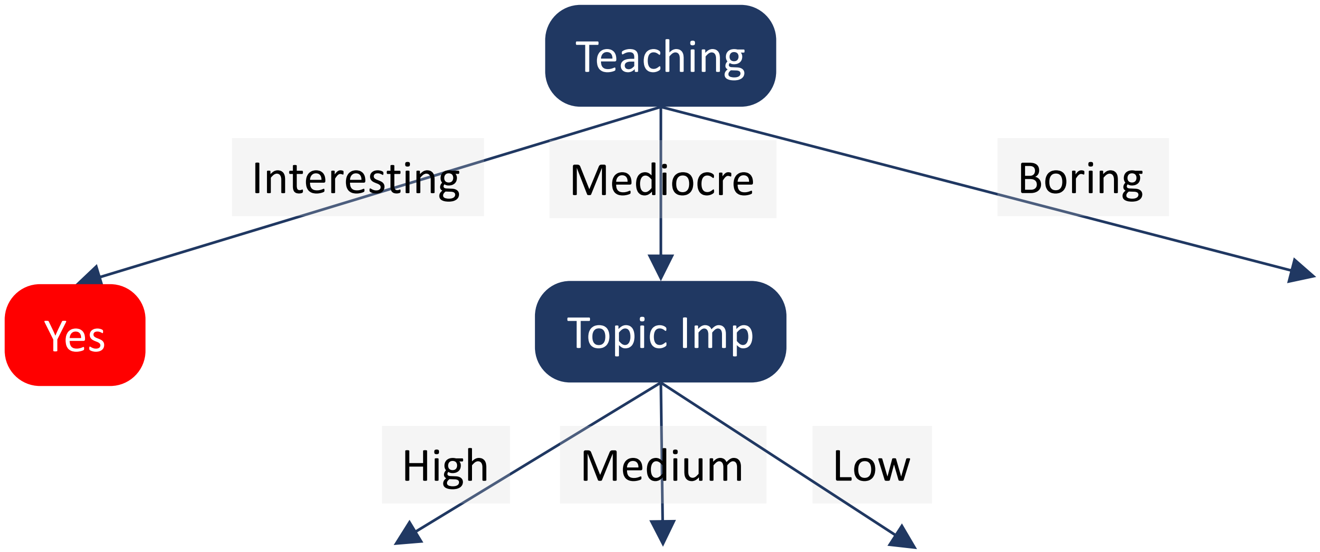
$$= 0.42$$

Level-2, Node-2

Instance	Weather	Health	Teaching	Topic Importance	Going to class?
3	Cold	Sick	Mediocre	Medium	No
5	Rainy	Sick	Mediocre	Low	No
7	Rainy	Good	Mediocre	Medium	No
8	Mild	Good	Mediocre	Medium	Yes
9	Rainy	Good	Mediocre	High	Yes

- Similarly can compute information gain from other features as:
  - $IG(\mathcal{D}_{\mathcal{F}_3,2}, \mathcal{F}_2) = 0.42$
  - $IG(\mathcal{D}_{\mathcal{F}_3,2}, \mathcal{F}_4) = 0.42$
- All the three features yield the same information gain, so can choose one of them.
- Take $\mathcal{F}_4$: Topic Importance.

Level-2, Node-2

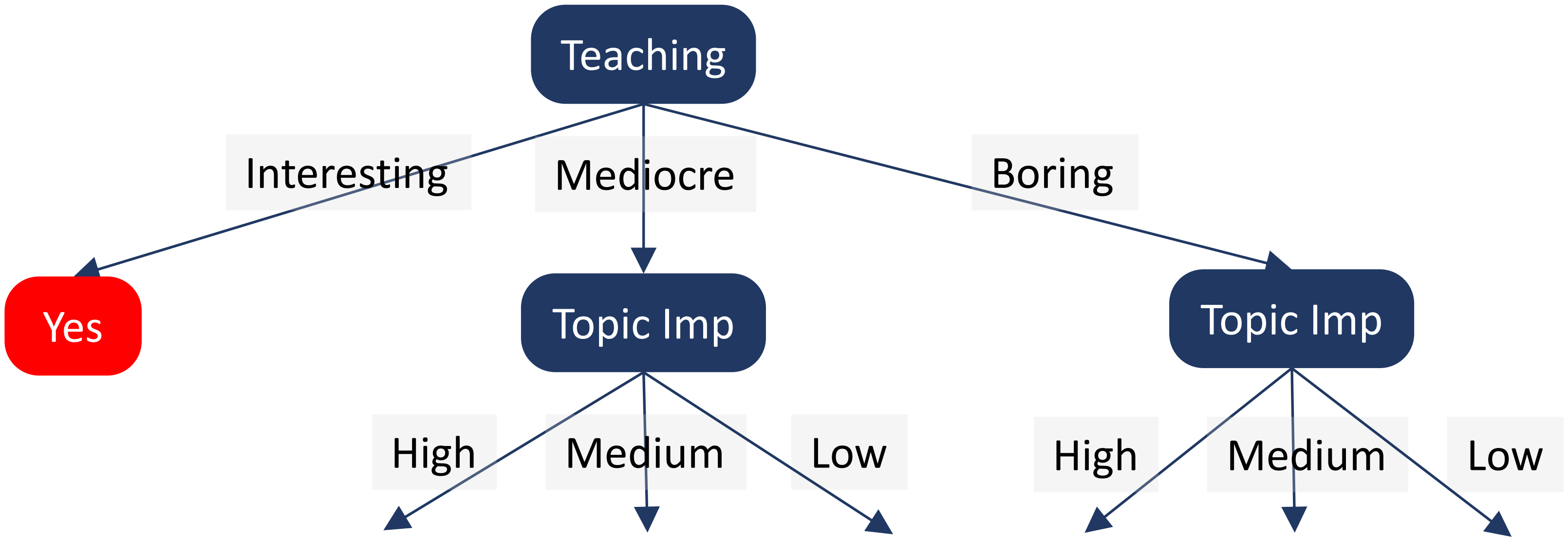


Level-2, Node-3

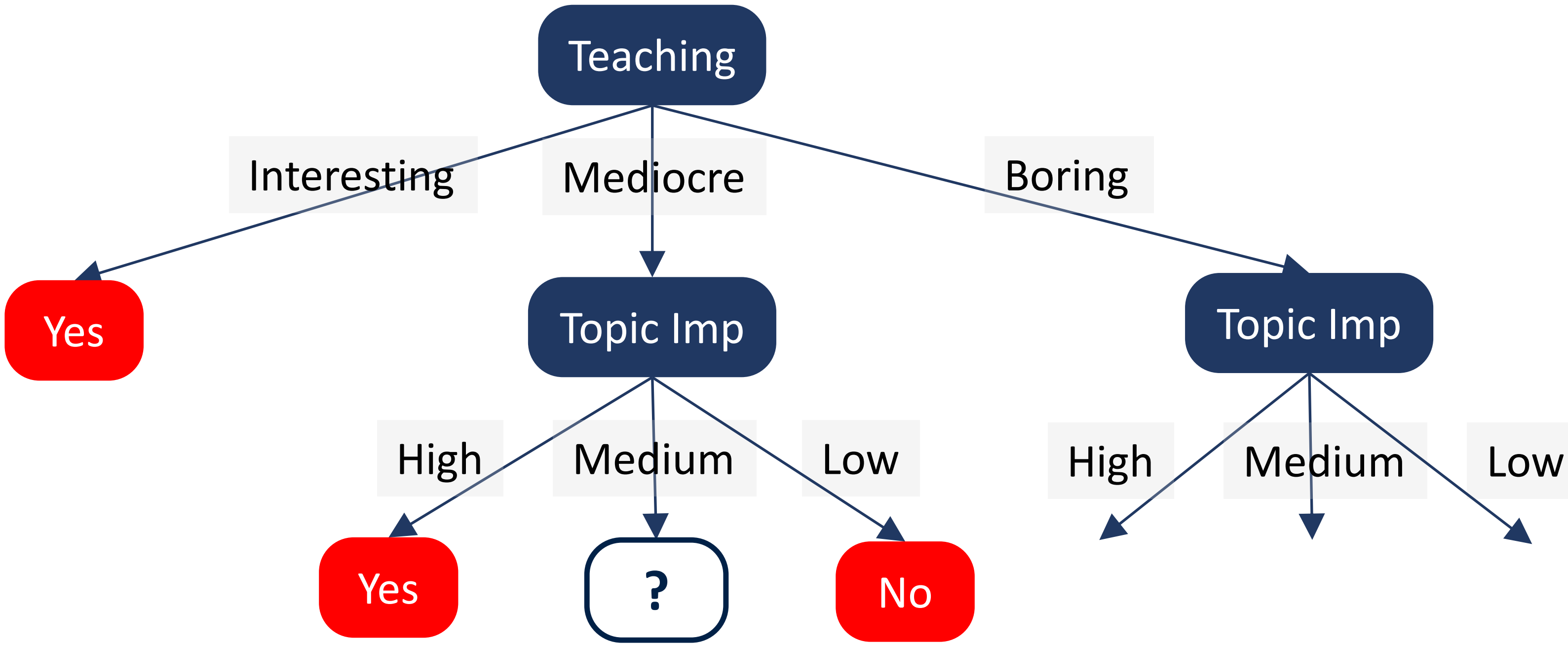
Instance	Weather	Health	Teaching	Topic Importance	Going to class?
2	Cold	Average	Boring	High	Yes
6	Hot	Good	Boring	High	Yes
11	Mild	Good	Boring	Low	No
14	Rainy	Average	Boring	Medium	No

- In a similar way we can compute information gain from $\mathcal{D}_{\mathcal{F}_3,3}$ with respect to the remaining features as:
  - $IG(\mathcal{D}_{\mathcal{F}_3,3}, \mathcal{F}_1) = 1$
  - $IG(\mathcal{D}_{\mathcal{F}_3,3}, \mathcal{F}_2) = 0$
  - $IG(\mathcal{D}_{\mathcal{F}_3,3}, \mathcal{F}_4) = 1$
- Choose $\mathcal{F}_4$.

Level-2, Node-3



Level-3



Instance	Weather	Health	Teaching	Topic Importance	Going to class?
3	Cold	Sick	Mediocre	Medium	No
5	Rainy	Sick	Mediocre	Low	No
7	Rainy	Good	Mediocre	Medium	No
8	Mild	Good	Mediocre	Medium	Yes
9	Rainy	Good	Mediocre	High	Yes

Level-3, Node-2

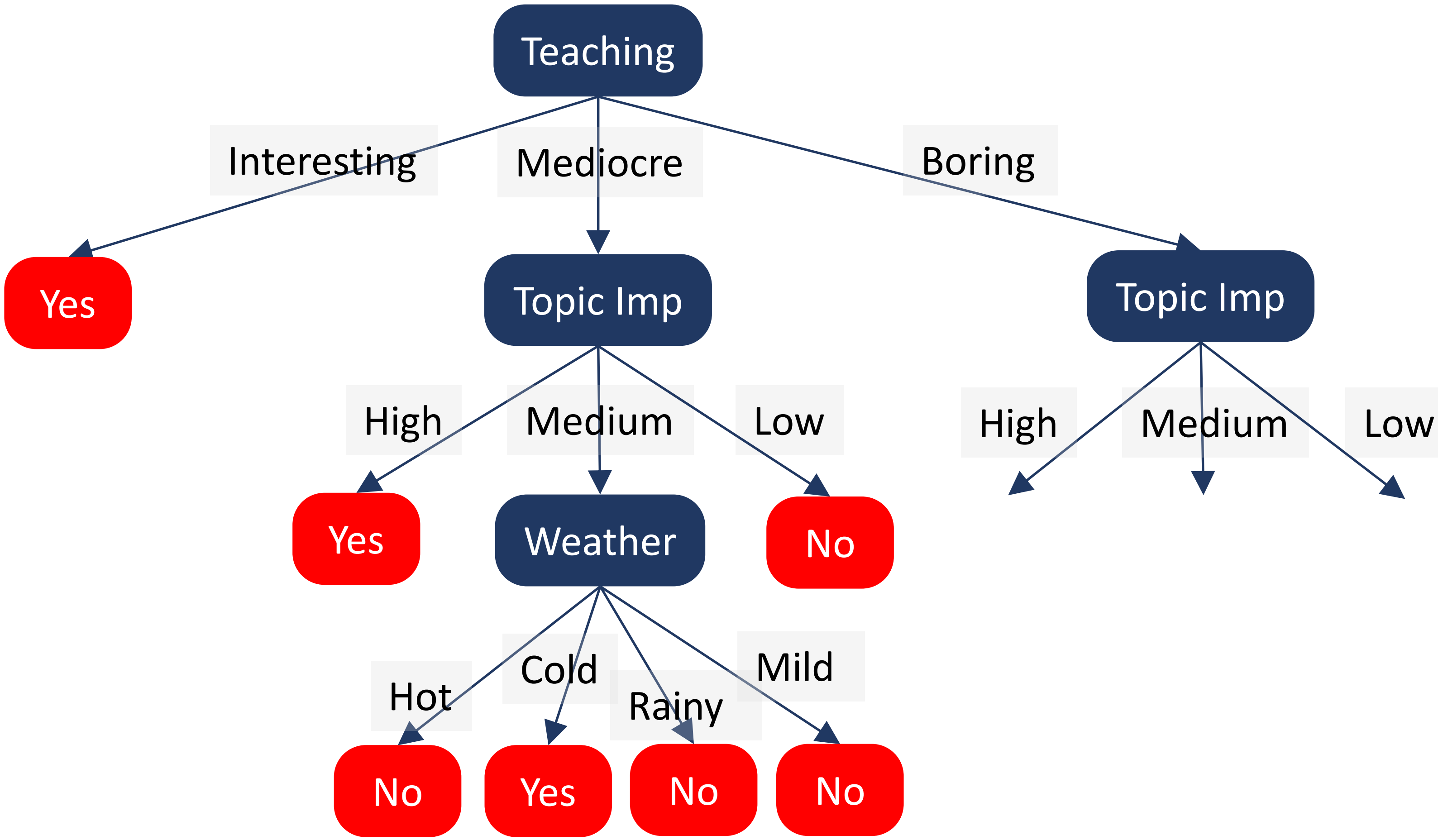
Instance	Weather	Health	Teaching	Topic Importance	Going to class?
3	Cold	Sick	Mediocre	Medium	No
7	Rainy	Good	Mediocre	Medium	No
8	Mild	Good	Mediocre	Medium	Yes

- Dataset $\mathcal{D}_{\mathcal{F}_3,2;\mathcal{F}_4,2}$: 1 Yes and 2 No.
- Entropy

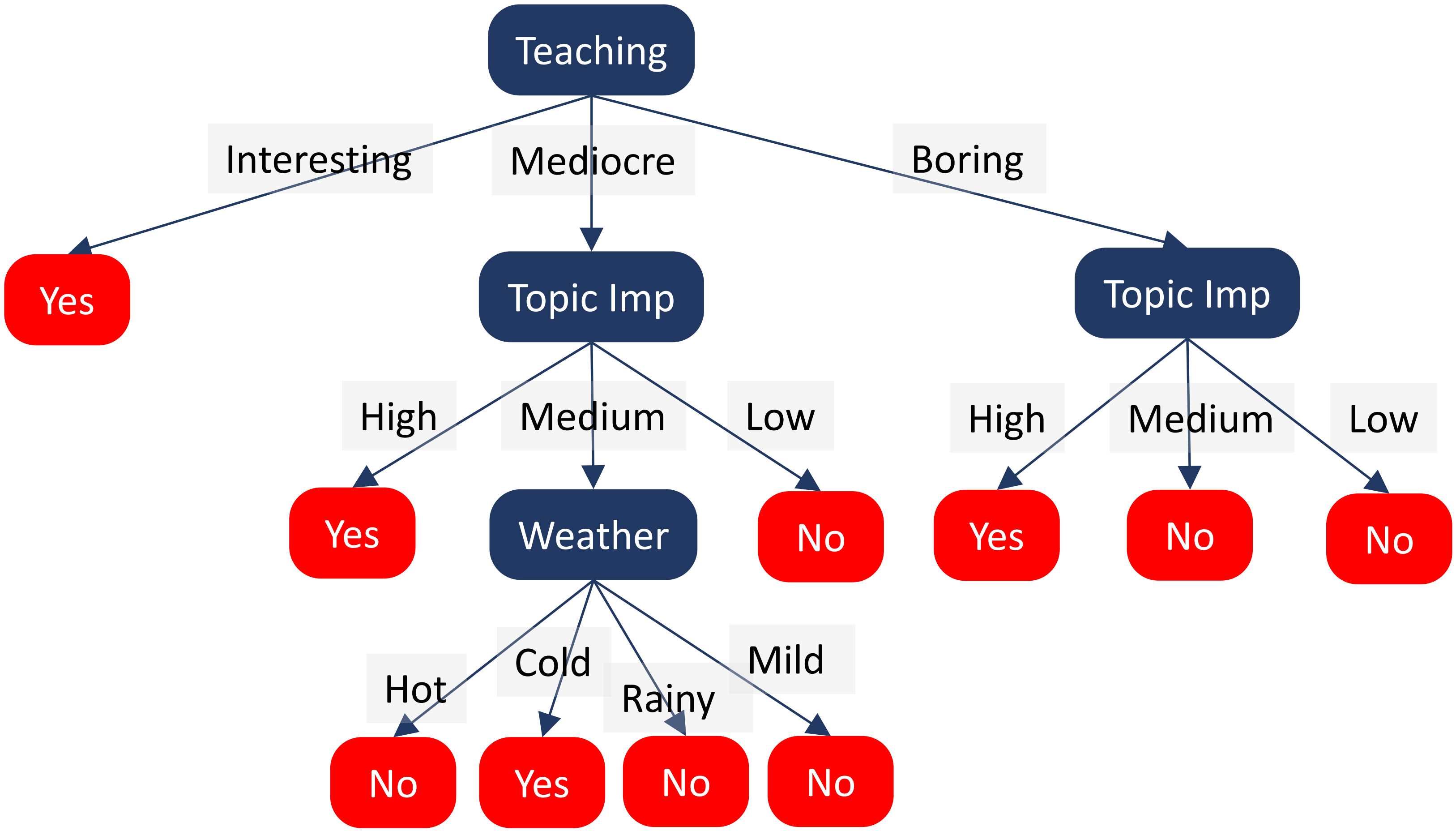
$$H(\mathcal{D}_{\mathcal{F}_3,2;\mathcal{F}_4,2}) = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)$$
$$= 0.918$$

- Information gain:
  - $IG(\mathcal{D}_{\mathcal{F}_3,2;\mathcal{F}_4,2}, \mathcal{F}_1) = 0.918$
  - $IG(\mathcal{D}_{\mathcal{F}_3,2;\mathcal{F}_4,2}, \mathcal{F}_2) = 0.251$
- $\mathcal{F}_1$ has the highest information gain.
- All nodes generated by $\mathcal{F}_1$ have the same class label. Therefore they are leaves of the tree.

Level-3, Node-2



Level-3, Node-4,5,6



Gain Ratio

- Gain Ratio reduces bias towards multi-valued attributes.
- It takes into account number and size of the branches when choosing a feature.
- Gain Ratio is defined as

$$\text{Gain Ratio}(\mathcal{D}) = \frac{\text{Information Gain}(\mathcal{D})}{\text{Intrinsic Info}(\mathcal{D})}$$

where Intrinsic Info($\mathcal{D}$) represents the potential information generated by splitting the datasets into J subsets:

$$\text{Intrinsic Info}(\mathcal{D}) = -\sum_{j=1}^{J} \frac{|\mathcal{D}_j|}{|\mathcal{D}|} \log_2\left(\frac{|\mathcal{D}_j|}{|\mathcal{D}|}\right)$$

  - Intrinsic Info is high when the subsets generated are of similar sizes.
  - Intrinsic Info is low when a few of the subsets contain most of the data.

Gini Index

- Gini Index is a measure of impurity and is defined as

$$\boxed{\mathrm{Gini}(\mathcal{D}) = 1 - \sum_{k=1}^{K} p(\mathcal{C}_k)^2}$$

  where $p(\mathcal{C}_k)$ is the probability that a tuple in $\mathcal{D}$ belongs to class $\mathcal{C}_k$.

- Maximum for a heterogeneous (impure) dataset when the records are equally distributed among all the classes. For such a case, if there are K classes in total, then the probability of the kth class is given as

$$p(\mathcal{C}_k) = \frac{1}{K}$$

- The Gini Index can then be computed as

$$\begin{aligned}\mathrm{Gini}(\mathcal{D}) &= 1 - \sum_{k=1}^{K} p(\mathcal{C}_k)^2 \\ &= 1 - K\left(\frac{1}{K}\right)^2 \\ &= 1 - \frac{1}{K}\end{aligned}$$

Gini Index

- Gini Index is a measure of impurity and is defined as

$$\boxed{\mathrm{Gini}(\mathcal{D}) = 1 - \sum_{k=1}^{K} p(\mathcal{C}_k)^2}$$

where $p(\mathcal{C}_k)$ is the probability that a tuple in $\mathcal{D}$ belongs to class $\mathcal{C}_k$.

- Minimum for a homogeneous (pure) dataset when all records belong to one class. For such a case

$$\mathrm{Gini}(\mathcal{D}) = 0$$

Average Gini Index

- Suppose the mth feature $\mathcal{F}_m$ is selected for splitting the set $\mathcal{D}$ into subsets $\mathcal{D}_1$ and $\mathcal{D}_2$.

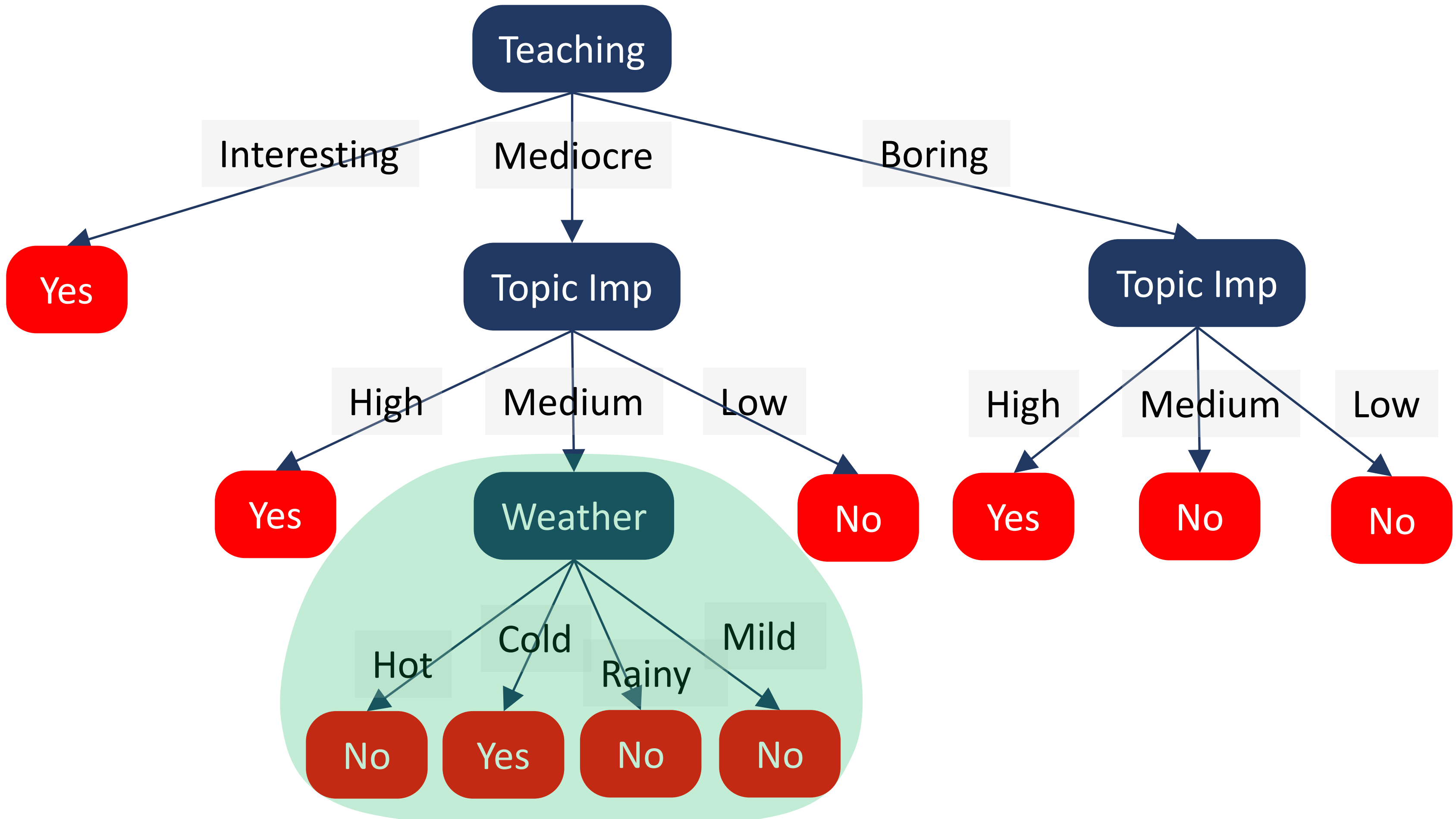
- Average Gini Index is defined as the weighted sum of the impurity measure of each subset produced after splitting:

$$\mathrm{Gini}_m(\mathcal{D}) = \frac{|\mathcal{D}_1|}{|\mathcal{D}|}\mathrm{Gini}(\mathcal{D}_1) + \frac{|\mathcal{D}_2|}{|\mathcal{D}|}\mathrm{Gini}(\mathcal{D}_2)$$

- The feature yielding the minimum value of the average Gini Index is selected for splitting the node.

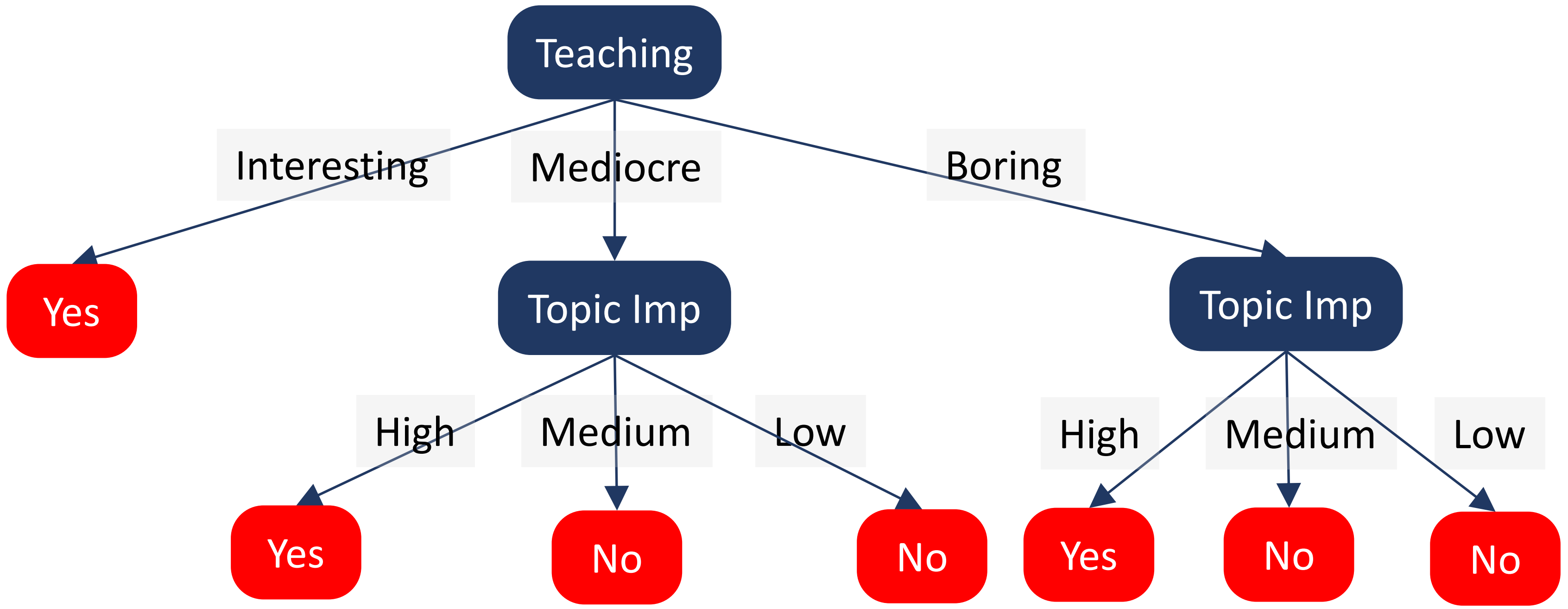
Pruning

- Bigger trees can lead to overfitting (capturing noise and outliers) of training data and poor generalization.
- Pruning refers to the class of techniques used to minimize the size of decision trees.



Pruning

- Bigger trees can lead to overfitting (capturing noise and outliers) of training data and poor generalization.
- Pruning refers to the class of techniques used to minimize the size of decision trees.



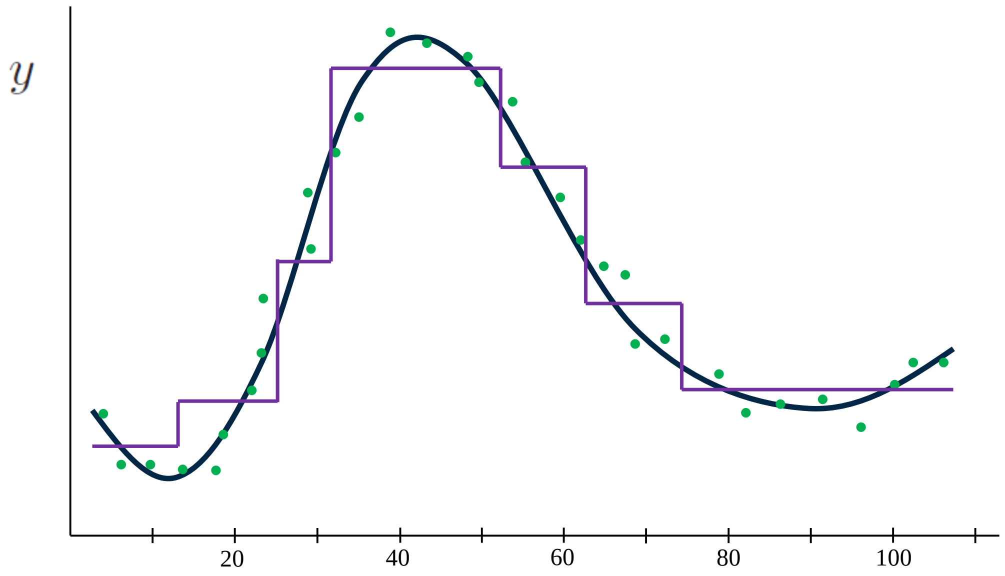
Pruning

- Prevent overfitting of training of data. Preference to smaller trees.
- Pruning approaches:
  - Post-pruning: Used after a full decision tree has been implemented. Example:
    - Reduced Error Pruning:
      - Classify all examples from a validation dataset (separate from training dataset).
      - Consider the nodes at the bottom of the tree.
      - Check the change in misclassifications if a node is replaced by the best possible leaf.
      - If the number of misclassifications is reduced or remains the same, then the node is replaced by the best leaf.
      - Repeat the same process with the new tree. Stop when the error (misclassifications) starts increasing.

Pruning

- Prevent overfitting of training of data. Preference to smaller trees.
- Pruning approaches:
  - Pre-pruning: Operates while the decision tree is being created. Example:
    - Minimum number of objects:
      - Pre-specify a value for minimum number of objects, say v.
      - If a node after splitting yields a child leaf with number of examples less than v, then that node is replaced by the best possible leaf.

Regression (Trees)



Procedure

- Data is split into subsets at each node.
- Prediction within each subset is (usually) taken to be the mean value of the output $\overline{y}$ in that subset.
- The mean-squared error (MSE) of each subset is computed.
- Partitioning is made at the location that yields the least weighted average of mean-squared error of the subsets.
  - Example: Want to partition set $\mathcal{S}$ into two subsets – $\mathcal{S}_1$ and $\mathcal{S}_2$.
  - Suppose subsets $\mathcal{S}_1$ and $\mathcal{S}_2$ have n_1 and n_2 number of examples, respectively. Total number of examples $n = n_1 + n_2$.
  - The weighted average of mean-squared error (WMSE) can be computed as:

$$\mathrm{WMSE} = \left(\frac{n_1}{n}\right)\mathrm{MSE}_1 + \left(\frac{n_2}{n}\right)\mathrm{MSE}_2$$

Splitting criteria