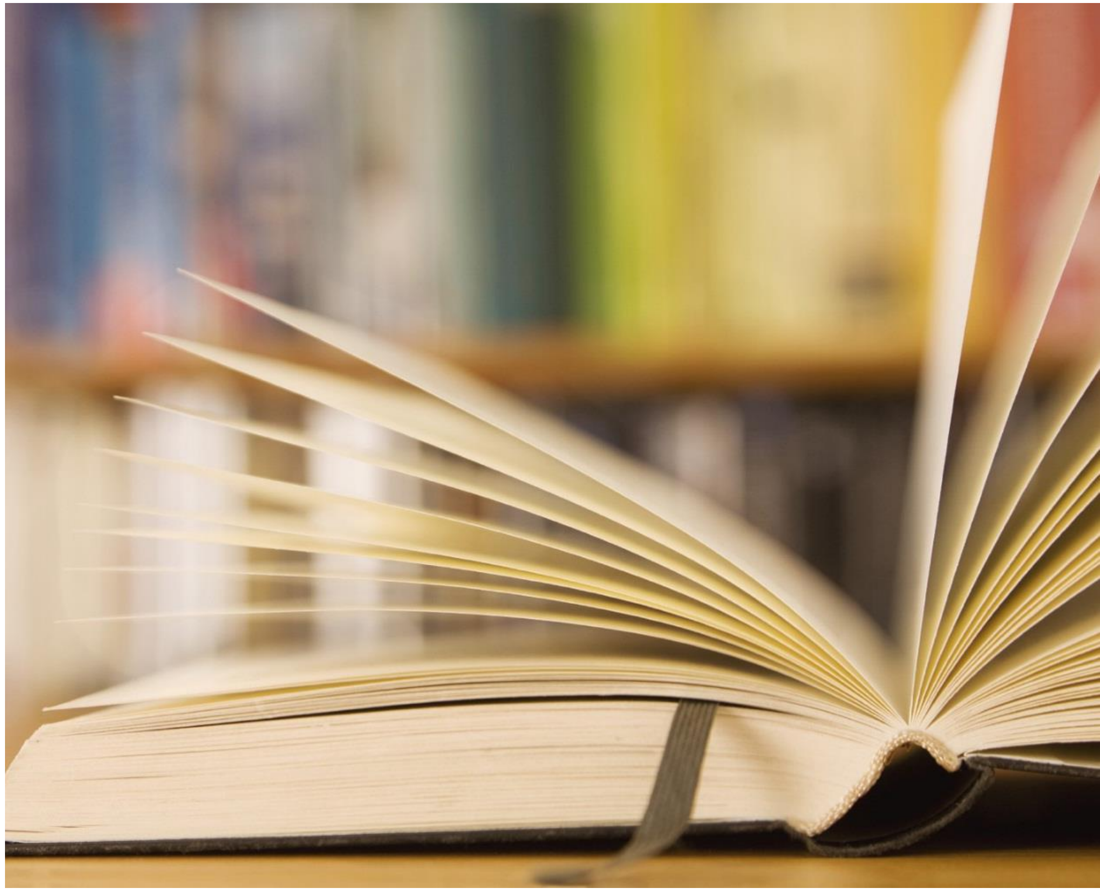
Classification Tree

Discussion



Classification and Regression Tree

- Classification and Regression Trees or CART for short is an acronym introduced by Leo Breiman to refer to Decision Tree algorithms that can be used for classification or regression predictive modelling problems.

Classification and Regression Tree

- A node represents a single input variable (X) and a split point on that variable, assuming the variable is numeric. The leaf nodes (also called terminal nodes) of the tree contain an output variable (y) which is used to make a prediction.
- Once created, a tree can be navigated with a new row of data following each branch with the splits until a final prediction is made.

Classification and Regression Tree

- Creating a binary decision tree is actually a process of dividing up the input space. A greedy approach is used to divide the space called recursive binary splitting. This is a numerical procedure where all the values are lined up and different split points are tried and tested using a cost function.
- The split with the best cost (lowest cost because we minimize cost) is selected. All input variables and all possible split points are evaluated and chosen in a greedy manner based on the cost function.

Classification and Regression Tree

- The split with the best cost (lowest cost because we minimize cost) is selected. All input variables and all possible split points are evaluated and chosen in a greedy manner based on the cost function.
    - **Regression**: The cost function that is minimized to choose split points is the sum squared error across all training samples that fall within the rectangle.
    - **Classification**: The Gini cost function is used which provides an indication of how pure the nodes are, where node purity refers to how mixed the training data assigned to each node is.
- Splitting continues until nodes contain a minimum number of training examples or a maximum tree depth is reached

Classification and Regression Tree

- Decision on splitting of nodes can be done using:
    - Information Gain
    - Gain Ratio
    - Gini Index
    - Variance
    - Chi-Square

Information Gain

- Information gain is the measure of ‘informativeness’ of a features in a dataset that provides the most information to make decision. For example, the information about the ‘gender’ of a person will allow one to make certain decision.

Information Gain

- Information gain computes the difference between entropy (uncertainty) before split (decision making) and average entropy after split of a dataset based on given attribute values.

$$IG(S,A) = E_{(s)} - \sum_{v \in values(A)}^{c} \frac{|S_v|}{|S|} E_{(S_v)}$$

An attribute in set S.

The entropy of the superset S before split.

Resulted subsets from S that share the same value in A.

The size or the number of instances in S.

The entropy of a subset Sv. This should be computed for each value of A (assuming it is a discrete attribute)

Entropy

- Entropy is a measurement of impurity or uncertainty of a dataset. The formula of entropy is:

$$Entropy = -\sum_{i=1}^{n} p_i \times \log_2 p_i$$

Where p_i is the probability of getting the i^{th} value when randomly selecting one from a dataset.

- For example, in a dataset if there were only ‘female’ values exist, then the entropy of getting the value ‘female’ randomly from the dataset is 0.

Entropy

- If there exists a mixture of ‘male’ and ‘female’ in a dataset, then the entropy (uncertainty) of getting the value ‘female’ randomly from the dataset will be higher because the impurity of the dataset has increase.
- In other higher the impurity (uncertainty) in a dataset, the entropy of the dataset increase; the lower the impurity (uncertainty) in a dataset, the entropy of the dataset decrease.

Calculation of Entropy

- Example: In a class consisting of 10 students, there are 6 males students and 4 female students.

$$dataset\text{: } \{male, male, female, male, female, female, female, male, male, male\}$$

$$Entropy(male) = -\frac{6}{10} \times \log_2 \frac{6}{10} = -0.6 \times (-0.73697) = 0.44218$$

$$Entropy(female) = -\frac{4}{10} \times \log_2 \frac{4}{10} = -0.4 \times (-1.32193) = 0.52877$$

$$Entropy(dataset) = 0.44218 + 0.52877 = 0.970952$$

Information gain

- We switch to use a slightly different dataset as follows:

Gender	SelfEmployed	YearlyIncome	CreditRating
Male	False	100000	2
Male	True	80000	2
Female	False	84000	1
Male	False	42000	2
Female	True	48000	2
Female	False	24000	3
Female	False	12000	3
Male	True	60000	1
Male	True	57600	1
Male	True	74400	1

Information gain

- In our example, the target class is the feature creditRating with class values {1, 2, 3}. Hence, the entropy of the dataset regarding the target feature will be: (Note: I use E to represent Entropy from here on).

$$\begin{aligned} E(ds, creditRating) = {} & E(ds, creditRating)_{cr=1} + \\ & E(ds, creditRating)_{cr=2} + \\ & E(ds, creditRating)_{cr=3} \end{aligned}$$

Information gain

$There\ are\ 4\ instances\ of\ creditRating\ 1,$
$4\ instances\ of\ creditRating\ 2, and$
$2\ instances\ of\ creditRating\ 3.$

$$E(ds, creditRating)_{cr=1} = -\frac{4}{10}\left(\log_2 \frac{4}{10}\right) = 0.52877$$

$$E(ds, creditRating)_{cr=2} = -\frac{4}{10}\left(\log_2 \frac{4}{10}\right) = 0.52877$$

$$E(ds, creditRating)_{cr=3} = -\frac{2}{10}\left(\log_2 \frac{2}{10}\right) = 0.46439$$

Information gain

$$E(ds, creditRating) = 0.52877 + 0.52877 + 0.46439$$

$$E(ds, creditRating) = 1.52193$$

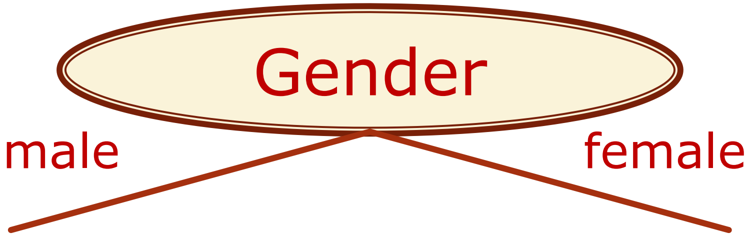
The total entropy of the dataset is 1.52193.

Starting from here, we compute information gain for all features (attributes) of the dataset to find the best features in terms of information gain to determine which features are the splitting node.

Highest reduction in entropy means highest gain in certainty.

Information gain

- Feature: Gender



Gender

male

female

1 x 3, 2 x 3, 3 x 0

1 x 1, 2 x 1, 3 x 2

$$E(male, cr = 1) = -\frac{3}{6}\left(\log_2 \frac{3}{6}\right) = 0.5$$

$$E(male, cr = 2) = -\frac{3}{6}\left(\log_2 \frac{3}{6}\right) = 0.5$$

$$E(male, cr = 3) = -\frac{0}{6}\left(\log_2 \frac{0}{10}\right) = 0.0$$

$$E(Gender, male) = 0.5 + 0.5 + 0.0 = 1.0$$

$$E(female, cr = 1) = -\frac{1}{4}\left(\log_2 \frac{1}{4}\right) = 0.5$$

$$E(female, cr = 2) = -\frac{1}{4}\left(\log_2 \frac{1}{4}\right) = 0.5$$

$$E(female, cr = 3) = -\frac{2}{4}\left(\log_2 \frac{2}{4}\right) = 0.5$$

$$E(Gender, female) = 0.5 + 0.5 + 0.5 = 1.5$$

Information gain

$$\mathrm{IG}(ds, Gender) = E(ds, creditRating) - \left[\frac{6}{10} \times E(Gender, male) + \frac{4}{10} \times E(Gender, female)\right]$$

$$= 1.52193 - \left[\left(\frac{6}{10} \times 1\right) + \left(\frac{4}{10} \times 1.5\right)\right]$$

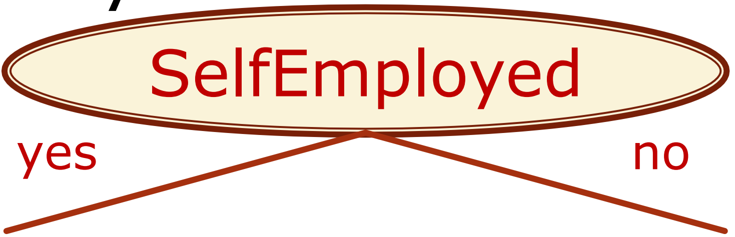
$$= 1.52193 - [(0.6) + (0.6)]$$

$$= 1.52193 - [1.2]$$

$$\mathrm{IG}(ds, Gender) = 0.32193$$

Information gain

- Feature: SelfEmployed



SelfEmployed

yes

no

1 x 3, 2 x 2, 3 x 0

1 x 1, 2 x 2, 3 x 2

$$E(SelfEmployed = yes, cr = 1) = -\frac{3}{5}\left(\log_2 \frac{3}{5}\right) = 0.44218$$

$$E(SelfEmployed = no, cr = 1) = -\frac{1}{5}\left(\log_2 \frac{1}{5}\right) = 0.46439$$

$$E(SelfEmployed = yes, cr = 2) = -\frac{2}{5}\left(\log_2 \frac{2}{5}\right) = 0.52877$$

$$E(SelfEmployed = no, cr = 2) = -\frac{2}{5}\left(\log_2 \frac{2}{5}\right) = 0.52877$$

$$E(SelfEmployed = yes, cr = 3) = -\frac{0}{5}\left(\log_2 \frac{0}{5}\right) = 0.0$$

$$E(SelfEmployed = no, cr = 3) = -\frac{2}{5}\left(\log_2 \frac{2}{5}\right) = 0.52877$$

$$E(SelfEmployed, yes) = 0.44218 + 0.52877 + 0.0 = 0.0$$

$$\begin{aligned} E(SelfEmployed, no) = 0.46439 + 0.52877 + \\ 0.52877 \\ = 1.52193 \end{aligned}$$

Information gain

$$\text{IG}(ds, SelfEmployed) = E(ds, creditRating) - \left[\frac{5}{10} \times E(SelfEmployed, yes) + \frac{5}{10} \times E(SelfEmployed, no)\right]$$

$$= 1.52193 - \left[\left(\frac{5}{10} \times 0.97095\right) + \left(\frac{5}{10} \times 1.52193\right)\right]$$

$$= 1.52193 - [(0.485475) + 760965]$$

$$= 1.52193 - [1.24644]$$

$$IG(ds, SelfEmployed) = 0.27549$$

Information gain

- Feature: YearlySalary
- The feature YearlySalary is a continuous attributes. The computation of the entropy for the feature is slightly complicated.
- The following are the steps:
  1. Arrange the values of the features in ascending order
  2. For each pair of the adjacent values, determine the average
  3. Compute the entropy for each of the average values and determine the information gains.

Information gain

2	100000
2	80000
1	84000
2	42000
2	48000
3	24000
3	12000
1	60000
1	57600
1	74400

3	12000
3	24000
2	42000
2	48000
1	57600
1	60000
1	74400
2	80000
1	84000
2	100000

18000
33000
45000
52800
58500
67200
77200
82000
92000

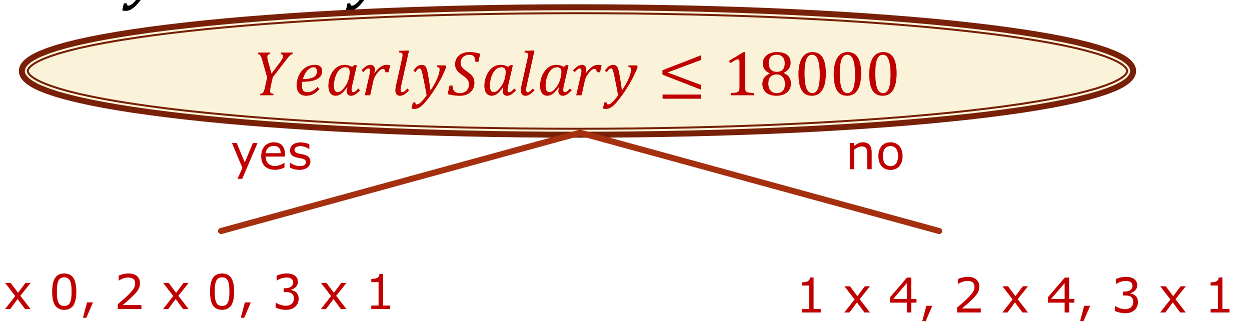
Arrange the yearlySalary in ascending order, and for each two adjacent yearlySalary, compute the average.

IG_{18000}
IG_{33000}
IG_{45000}
IG_{52800}
IG_{58500}
IG_{67200}
IG_{77200}
IG_{82000}
IG_{92000}

Compute the IG for each of the average yearlySalary. The one give the highest information gain will be the split.

Information gain

- Feature: $YearlySalary \leq 18000$



$YearlySalary \leq 18000$

yes

no

1 x 0, 2 x 0, 3 x 1

1 x 4, 2 x 4, 3 x 1

$$E(YearSal \leq 18000, cr = 1) = -\frac{0}{1}\left(\log_2 \frac{0}{1}\right) = 0.0$$

$$E(YearSal > 18000, cr = 1) = -\frac{4}{9}\left(\log_2 \frac{4}{9}\right) = 0.51997$$

$$E(YearSal \leq 18000, cr = 2) = -\frac{0}{1}\left(\log_2 \frac{0}{1}\right) = 0.0$$

$$E(YearSal > 18000, cr = 2) = -\frac{4}{9}\left(\log_2 \frac{4}{9}\right) = 0.51997$$

$$E(YearSal \leq 18000, cr = 3) = -\frac{1}{1}\left(\log_2 \frac{1}{1}\right) = 0.0$$

$$E(YearSal > 18000, cr = 3) = -\frac{1}{9}\left(\log_2 \frac{1}{9}\right) = 0.35221$$

$$E(YearSal \leq 18000, yes) = 0.0 + 0.0 + 0.0 = 0.0$$

$$E(YearSal \leq 18000, no) = 0.51997 + 0.51997 + 0.35221 = 1.39215$$

Information gain

$$\mathrm{IG}(ds, YearSal \leq 18000) = E(ds, YearSal \leq 18000) - \left[\frac{1}{10} \times E(YearSal \leq 18000, yes) + \frac{9}{10} \times E(YearSal \leq 18000, no)\right]$$

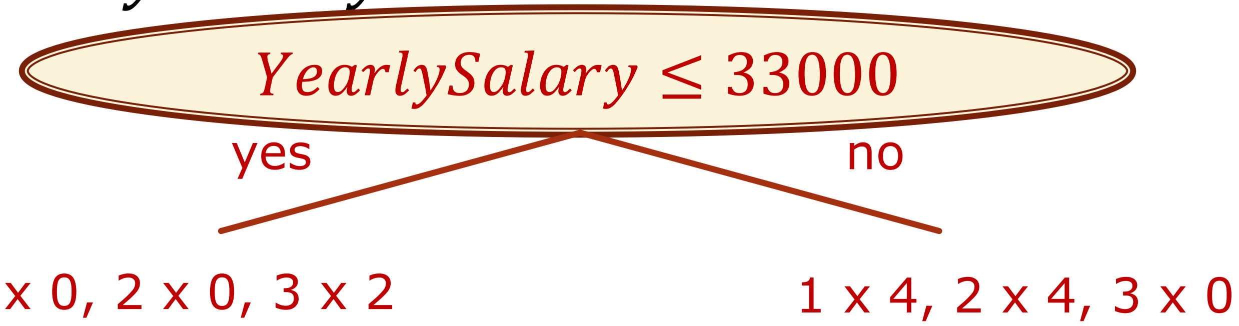
$$= 1.52193 - \left[\left(\frac{1}{10} \times 0.0\right) + \left(\frac{9}{10} \times 1.39215\right)\right]$$

$$= 1.52193 - [1.252935]$$

$$IG(ds, YearSal \leq 18000) = 0.268995$$

Information gain

- Feature: $YearlySalary \leq 33000$



$$E(YearSal \leq 33000, cr = 1) = -\frac{0}{1}\left(\log_2 \frac{0}{1}\right) = 0.0$$

$$E(YearSal > 33000, cr = 1) = -\frac{4}{8}\left(\log_2 \frac{4}{8}\right) = 0.5$$

$$E(YearSal \leq 33000, cr = 2) = -\frac{0}{1}\left(\log_2 \frac{0}{1}\right) = 0.0$$

$$E(YearSal > 33000, cr = 2) = -\frac{4}{8}\left(\log_2 \frac{4}{8}\right) = 0.5$$

$$E(YearSal \leq 33000, cr = 3) = -\frac{2}{2}\left(\log_2 \frac{2}{2}\right) = 0.0$$

$$E(YearSal > 33000, cr = 3) = -\frac{0}{8}\left(\log_2 \frac{0}{8}\right) = 0.0$$

$$E(YearSal \leq 33000, yes) = 0.0 + 0.0 + 0.0 = 0.0$$

$$\begin{aligned} E(YearSal \leq 33000, no) &= 0.5 + 0.5 + 0.0 \\ &= 1.0 \end{aligned}$$

Information gain

$$\mathrm{IG}(ds, YearSal \leq 33000) = E(ds, YearSal \leq 33000) - \left[\frac{2}{10} \times E(YearSal \leq 33000, yes) + \frac{8}{10} \times E(YearSal \leq 33000, no)\right]$$

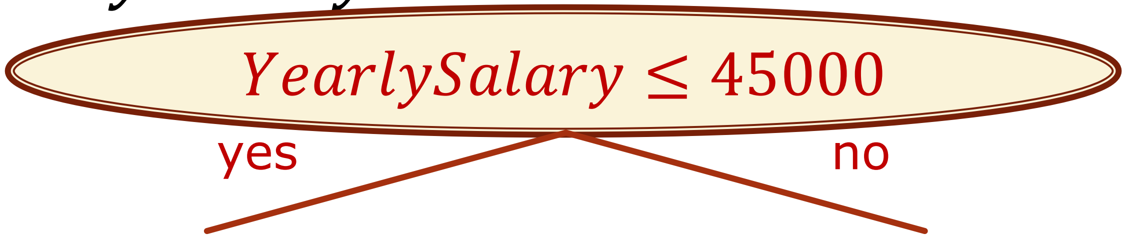
$$= 1.52193 - \left[\left(\frac{2}{10} \times 0.0\right) + \left(\frac{8}{10} \times 1.0\right)\right]$$

$$= 1.52193 - [0.8]$$

$$IG(ds, YearSal \leq 33000) = 0.72193$$

Information gain

- Feature: $YearlySalary \leq 45000$



$YearlySalary \leq 45000$

yes

no

1 x 0, 2 x 1, 3 x 2

1 x 4, 2 x 3, 3 x 0

$$E(YearSal \leq 45000, cr = 1) = -\frac{0}{3}\left(\log_2 \frac{0}{3}\right) = 0.0$$

$$E(YearSal \leq 45000, cr = 2) == -\frac{1}{3}\left(\log_2 \frac{1}{3}\right) = 0.52832$$

$$E(YearSal \leq 45000, cr = 3) == -\frac{2}{3}\left(\log_2 \frac{2}{3}\right) = 0.38998$$

$$E(YearSal > 45000, cr = 1) = -\frac{4}{7}\left(\log_2 \frac{4}{7}\right) = 0.46135$$

$$E(YearSal > 45000, cr = 2) = -\frac{3}{7}\left(\log_2 \frac{3}{7}\right) = 0.52388$$

$$E(YearSal > 45000, cr = 3) = -\frac{0}{7}\left(\log_2 \frac{0}{7}\right) = 0.0$$

$$E(YearSal \leq 45000, yes) = 0.0 + 0.52832 + 0.38998 = 0.9183$$

$$E(YearSal \leq 45000, no) = 0.46135 + 0.52388 + 0.0 = 0.98523$$

Information gain

$$\mathrm{IG}(ds, YearSal \leq 45000) = E(ds, YearSal \leq 45000) - \left[\frac{3}{10} \times E(YearSal \leq 45000, yes) + \frac{7}{10} \times E(YearSal \leq 45000, no)\right]$$

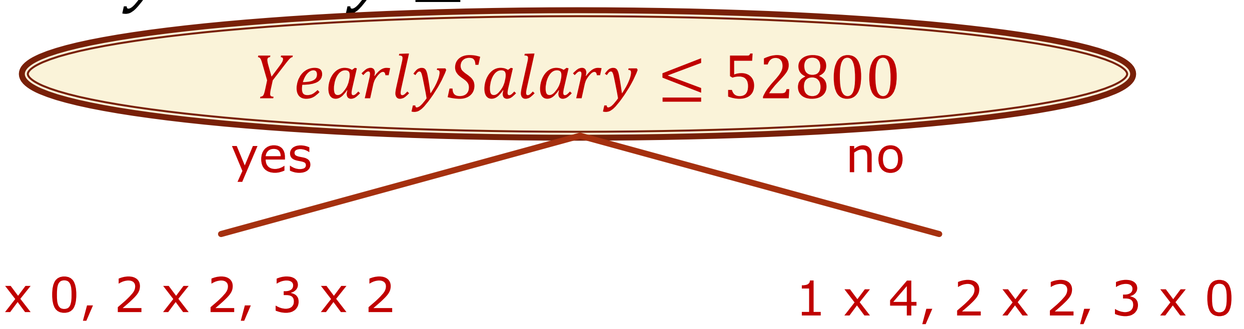
$$= 1.52193 - \left[\left(\frac{3}{10} \times 0.9183\right) + \left(\frac{7}{10} \times 0.98523\right)\right]$$

$$= 1.52193 - [0.27549 + 0.68966]$$

$$IG(ds, YearSal \leq 45000) = 0.55678$$

Information gain

- Feature: $YearlySalary \leq 52800$



$YearlySalary \leq 52800$

yes

no

1 x 0, 2 x 2, 3 x 2

1 x 4, 2 x 2, 3 x 0

$$E(YearSal \leq 52800, cr = 1) = -\frac{0}{4}\left(\log_2 \frac{0}{4}\right) = 0.0$$

$$E(YearSal \leq 52800, cr = 2) = -\frac{2}{4}\left(\log_2 \frac{2}{4}\right) = 0.5$$

$$E(YearSal \leq 52800, cr = 3) = -\frac{2}{4}\left(\log_2 \frac{2}{4}\right) = 0.5$$

$$E(YearSal \leq 52800, yes) = 0.0 + 0.5 + 0.5 = 1.0$$

$$E(YearSal > 52800, cr = 1) = -\frac{4}{6}\left(\log_2 \frac{4}{6}\right) = 0.389975$$

$$E(YearSal > 52800, cr = 2) = -\frac{2}{6}\left(\log_2 \frac{2}{6}\right) = 0.528321$$

$$E(YearSal > 52800, cr = 3) = -\frac{0}{6}\left(\log_2 \frac{0}{0}\right) = 0.0$$

$$\begin{aligned} E(YearSal \leq 52800, no) &= 0.389975 + 0.528321 + 0.0 \\ &= 0.918296 \end{aligned}$$

Information gain

$$\mathrm{IG}(ds, YearSal \leq 52800) = E(ds, YearSal \leq 52800) - \left[\frac{4}{10} \times E(YearSal \leq 52800, yes) + \frac{6}{10} \times E(YearSal \leq 52800, no)\right]$$

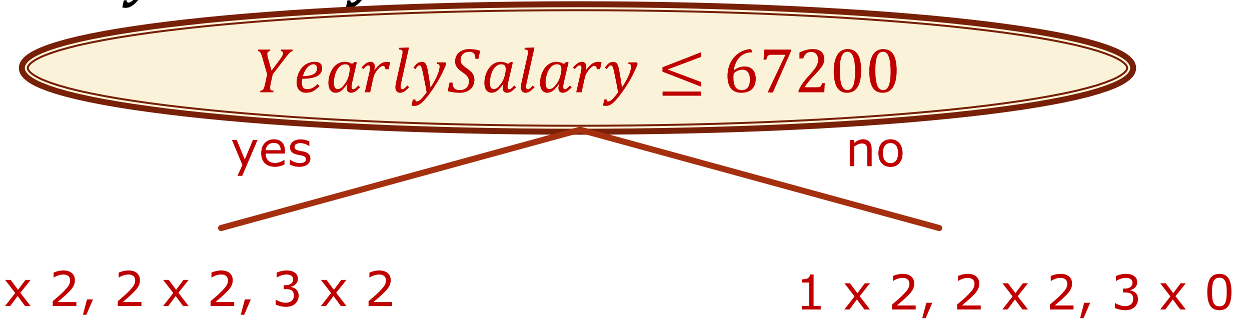
$$= 1.52193 - \left[\left(\frac{4}{10} \times 1.0\right) + \left(\frac{6}{10} \times 0.918296\right)\right]$$

$$= 1.52193 - [0.4 + 0.5509776]$$

$$IG(ds, YearSal \leq 52800 = 0.570952$$

Information gain

- Feature: $YearlySalary \leq 67200$



$YearlySalary \leq 67200$

yes

no

1 x 2, 2 x 2, 3 x 2

1 x 2, 2 x 2, 3 x 0

$$E(YearSal \leq 67200, cr = 1) = -\frac{2}{6}\left(\log_2 \frac{2}{6}\right) = 0.528321$$

$$E(YearSal \leq 67200, cr = 2) = -\frac{2}{6}\left(\log_2 \frac{2}{6}\right) = 0.528321$$

$$E(YearSal \leq 67200, cr = 3) = -\frac{2}{6}\left(\log_2 \frac{2}{6}\right) = 0.528321$$

$$\begin{aligned} E(YearSal \leq 67200, yes) = {} & 0.528321 + \\ & 0.528321 + \\ & 0.528321 = 1.584963 \end{aligned}$$

$$E(YearSal > 67200, cr = 1) = -\frac{2}{4}\left(\log_2 \frac{2}{4}\right) = 0.5$$

$$E(YearSal > 67200, cr = 2) = -\frac{2}{4}\left(\log_2 \frac{2}{4}\right) = 0.5$$

$$E(YearSal > 67200, cr = 1) = -\frac{0}{6}\left(\log_2 \frac{0}{0}\right) = 0.0$$

$$\begin{aligned} E(YearSal \leq 67200, no) &= 0.5 + 0.51 + 0.0 \\ &= 1.0 \end{aligned}$$

Information gain

$$\text{IG}(ds, YearSal \leq 67200) = E(ds, YearSal \leq 67200) - \left[\frac{6}{10} \times E(YearSal \leq 67200, yes) + \frac{4}{10} \times E(YearSal \leq 67200, no)\right]$$

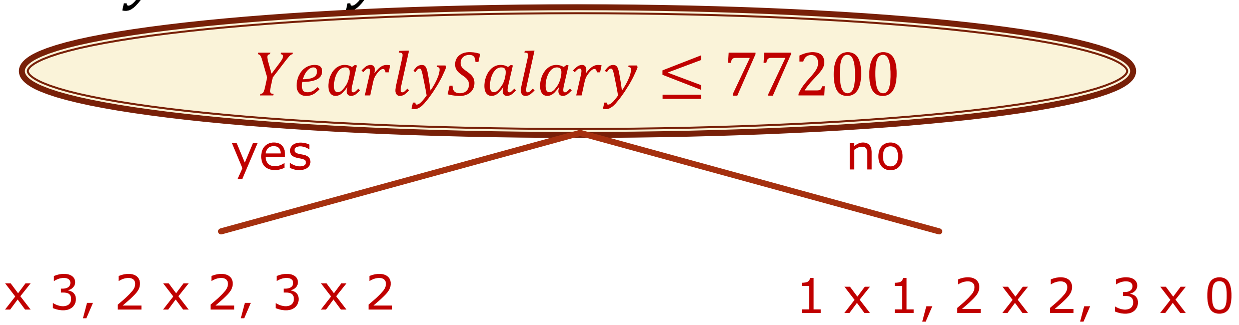
$$= 1.52193 - \left[\left(\frac{4}{10} \times 1.584963\right) + \left(\frac{6}{10} \times 1.0\right)\right]$$

$$= 1.52193 - [0.6339852 + 0.6]$$

$$IG(ds, YearSal \leq 67200 = 0.170952$$

Information gain

- Feature: $YearlySalary \leq 77200$



$YearlySalary \leq 77200$

yes

no

1 x 3, 2 x 2, 3 x 2

1 x 1, 2 x 2, 3 x 0

$$E(YearSal \leq 77200, cr = 1) = -\frac{3}{7}\left(\log_2 \frac{3}{7}\right) = 0.523882$$

$$E(YearSal > 77200, cr = 1) = -\frac{1}{3}\left(\log_2 \frac{1}{3}\right) = 0.528321$$

$$E(YearSal \leq 77200, cr = 2) = -\frac{2}{7}\left(\log_2 \frac{2}{7}\right) = 0.516387$$

$$E(YearSal > 77200, cr = 2) = -\frac{2}{3}\left(\log_2 \frac{2}{3}\right) = 0.389975$$

$$E(YearSal \leq 77200, cr = 3) = -\frac{2}{7}\left(\log_2 \frac{2}{7}\right) = 0.516387$$

$$E(YearSal > 77200, cr = 3) = -\frac{0}{3}\left(\log_2 \frac{0}{3}\right) = 0.0$$

$$\begin{aligned} E(YearSal \leq 77200, yes) = {} & 0.523882 + \\ & 0.516387 + \\ & 0.516387 = 1.556657 \end{aligned}$$

$$\begin{aligned} E(YearSal \leq 77200, no) &= 0.528321 + 0.389975 + 0.0 \\ &= 0.918296 \end{aligned}$$

Information gain

$$\text{IG}(ds, YearSal \leq 77200) = E(ds, YearSal \leq 77200) - \left[\frac{7}{10} \times E(YearSal \leq 77200, yes) + \frac{3}{10} \times E(YearSal \leq 77200, no)\right]$$

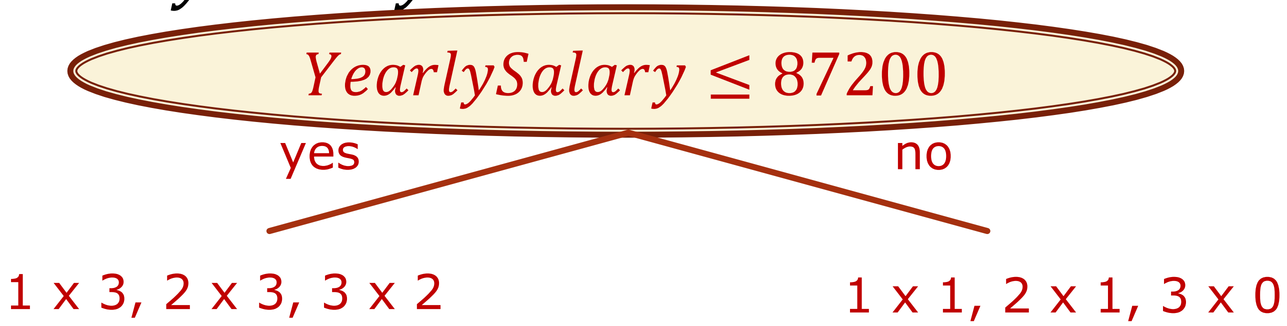
$$= 1.52193 - \left[\left(\frac{7}{10} \times 1.556657\right) + \left(\frac{3}{10} \times 0.918296\right)\right]$$

$$= 1.52193 - [1.0896592 + 0.2754888]$$

$$IG(ds, YearSal \leq 77200) = 0.156782$$

Information gain

- Feature: $YearlySalary \leq 82000$



$YearlySalary \leq 87200$

yes

no

1 x 3, 2 x 3, 3 x 2

1 x 1, 2 x 1, 3 x 0

$$E(YearSal \leq 82000, cr = 1) = -\frac{3}{8}\left(\log_2 \frac{3}{8}\right) = 0.530639$$

$$E(YearSal > 82000, cr = 1) = -\frac{1}{2}\left(\log_2 \frac{1}{2}\right) = 0.5$$

$$E(YearSal \leq 82000, cr = 1) = -\frac{3}{8}\left(\log_2 \frac{3}{8}\right) = 0.530639$$

$$E(YearSal > 82000, cr = 2) = -\frac{1}{2}\left(\log_2 \frac{1}{2}\right) = 0.5$$

$$E(YearSal \leq 82000, cr = 1) = -\frac{2}{8}\left(\log_2 \frac{2}{8}\right) = 0.5$$

$$E(YearSal > 82000, cr = 3) = -\frac{0}{2}\left(\log_2 \frac{0}{2}\right) = 0.0$$

$$\begin{aligned} E(YearSal \leq 82000, yes) = 0.530639 +{}& \\ 0.530639 +{}& \\ 0.5 = 1.561278& \end{aligned}$$

$$\begin{aligned} E(YearSal \leq 82000, no) &= 0.5 + 0.5 + 0.0 \\ &= 1.0 \end{aligned}$$

Information gain

$$\mathrm{IG}(ds, YearSal \leq 82000) = E(ds, YearSal \leq 82000) - \left[\frac{8}{10} \times E(YearSal \leq 82000, yes) + \frac{2}{10} \times E(YearSal \leq 82000, no)\right]$$

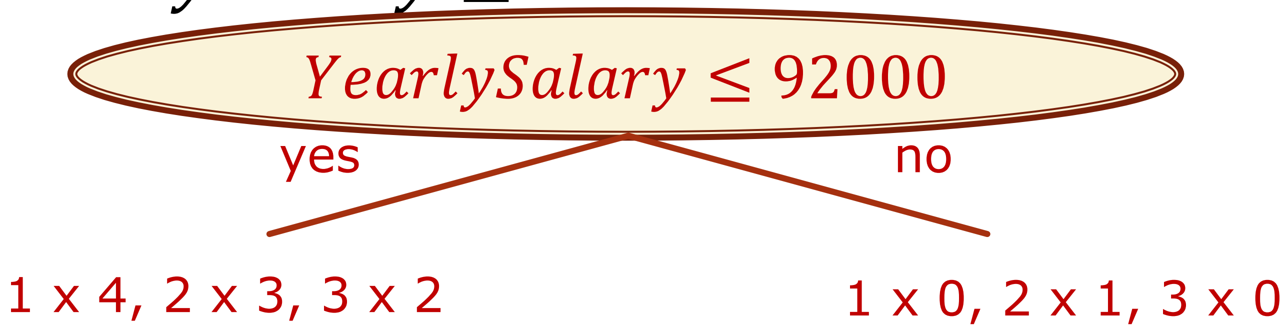
$$= 1.52193 - \left[\left(\frac{8}{10} \times 1.561278\right) + \left(\frac{2}{10} \times 1.0\right)\right]$$

$$= 1.52193 - [1.2490224 + 0.2]$$

$$IG(ds, YearSal \leq 82000) = 0.072908$$

Information gain

- Feature: $YearlySalary \leq 92000$



$$E(YearSal \leq 92000, yes)_{cr=1} = -\frac{4}{9}\left(\log_2 \frac{4}{9}\right) = 0.519967 \quad E(YearSal \leq 92000, no)_{cr=1} = -\frac{0}{2}\left(\log_2 \frac{0}{2}\right) = 0.0$$

$$E(YearSal \leq 92000, yes)_{cr=2} = -\frac{3}{9}\left(\log_2 \frac{3}{9}\right) = 0.528321 \quad E(YearSal \leq 92000, no)_{cr=2} = -\frac{1}{2}\left(\log_2 \frac{1}{2}\right) = 0.0$$

$$E(YearSal \leq 92000, yes)_{cr=3} = -\frac{2}{9}\left(\log_2 \frac{2}{9}\right) = 0.482206 \quad E(YearSal \leq 92000, no)_{cr=3} = -\frac{0}{2}\left(\log_2 \frac{0}{2}\right) = 0.0$$

$$\begin{aligned} E(YearSal \leq 92000, yes) = {} & 0.519967 + \\ & 0.528321 + \\ & 0.482206 = 1.530493 \end{aligned}$$

$$\begin{aligned} E(YearSal \leq 92000, no) & = 0.0 + 0.0 + 0.0 \\ & = 0.0 \end{aligned}$$

Information gain

$$\mathrm{IG}(ds, YearSal \leq 92000) = E(ds, YearSal \leq 92000) - \left[\frac{9}{10} \times E(YearSal \leq 92000, yes) + \frac{1}{10} \times E(YearSal \leq 92000, no)\right]$$

$$= 1.52193 - \left[\left(\frac{9}{10} \times 1.530493\right) + \left(\frac{1}{10} \times 0.0\right)\right]$$

$$= 1.52193 - [1.3774437 + 0.0]$$

$$= 0.144486$$

Information gain

- Summary:

$$IG(ds, Gender) = 0.32193$$

$$IG(ds, SelfEmployed) = 0.27549$$

$$IG(ds, YearSal \leq 18000) = 0.268995$$

$$\boldsymbol{IG(ds, YearSal \leq 33000) = 0.72193}$$

Feature with the highest gain.

$$IG(ds, YearSal \leq 45000) = 0.55678$$

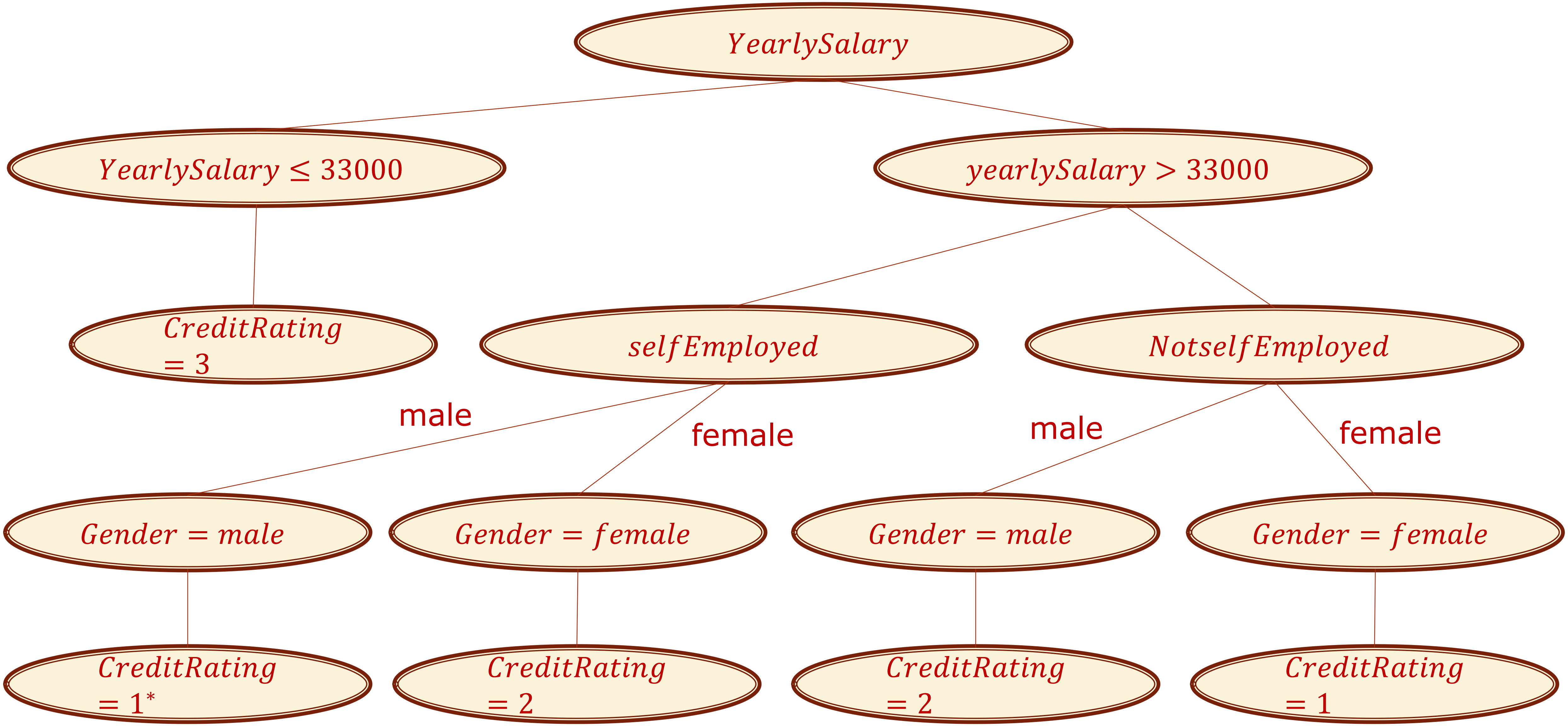
$$IG(ds, YearSal \leq 52800 = 0.570952$$

$$IG(ds, YearSal \leq 67200 = 0.170952$$

$$IG(ds, YearSal \leq 77200) = 0.156782$$

$$IG(ds, YearSal \leq 82000) = 0.072908$$

Decision Tree



$YearlySalary$

$YearlySalary \leq 33000$

$yearlySalary > 33000$

$CreditRating = 3$

$selfEmployed$

$NotselfEmployed$

male

female

male

female

$Gender = male$

$Gender = female$

$Gender = male$

$Gender = female$

$CreditRating = 1^*$

$CreditRating = 2$

$CreditRating = 2$

$CreditRating = 1$

Gini index

Gini Index

- The Gini index is the name of the cost function used to evaluate splits in the dataset. It is a measure of statistical dispersion intended to represent the income inequality or wealth inequality within a nation or any other group of people. [https://en.wikipedia.org/wiki/Gini_coefficient, 7 Feb 2021]
- In decision tree, a split in the dataset involves one input attribute and one value for that attribute. It can be used to divide training patterns into two groups of rows.

Gini Index

- A Gini index gives an idea of how good a split is by how mixed the classes are in the two groups created by the split.
- A perfect separation results in a Gini index of 0, whereas the worst case split that results in 50/50 classes in each group result in a Gini index of 0.5 (for a 2-class problem).

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

Gini index

- Example: In a class consisting of 10 students, there are 6 males students and 4 female students.

$$dataset\text{:}\ \{male, male, female, male, female, female, female, male, male, male\}$$

- The Gini index is then,

$$G_{male} = 1 - \left[\left(\frac{6}{10}\right)^2\right]$$

$$G_{male} = 1 - 0.36$$

$$G_{male} = 0.384$$

$$G_{female} = 1 - \left[\left(\frac{4}{10}\right)^2\right]$$

$$G_{female} = 1 - 0.16$$

$$G_{female} = 0.336$$

Gini index

- Gini index for the dataset is then equal the sum of $(Gini_{female}\times$ $weight\ by\ the\ size\ of\ the\ group\ relative\ to\ the\ sample)$ and $(Gini_{male}\times$

$$Gini_{Dataset} = Gini_{female} \times \left(\frac{4}{10}\right) + Gini_{male} \times \left(\frac{6}{10}\right)$$

$$Gini_{Dataset} = 0.336 \times \left(\frac{4}{10}\right) + 0.384 \times \left(\frac{6}{10}\right) = 0.3648$$

Gini index

- In case of a discrete-valued attribute, the subset that gives the minimum gini index for that chosen is selected as a splitting attribute. In the case of continuous-valued attributes, the strategy is to select each pair of adjacent values as a possible split-point and point with smaller gini index chosen as the splitting point.
- The attribute with minimum (lowest) Gini index is chosen as the splitting attribute.

Gain Ratio

- The gain ratio is a variation of information gain.
- Information gain is biased for the attribute with many outcomes. It means it prefers the attribute with a large number of distinct values.
- To overcome this bias behaviour, gain ratio takes into account the number and size of branches when choosing an attribute.
- In each level of decision tree, the feature having the highest gain ratio will be the decision rule (splitting node).

Gain ratio

$$IG(S,A) = E_{(s)} - \sum_{v \in values(A)}^{c} \frac{|S_v|}{|S|} E_{(S_v)}$$

$$SplitInfo(S,A) = - \sum_{v \in values(A)}^{c} \frac{|S_v|}{|S|} E_{(S_v)}$$

$$GainRatio(S,A) = \frac{IG(S,A)}{SplitInfo(S,A)}$$

- In each level of decision tree, the feature having the highest gain ratio will be the decision rule (splitting node).

	IG		SplitDecision	GainRatio
IG(ds,Gender):	0.32193		1.2	0.268275
IG(ds,SelfEmployed):	0.27549		1.24644	0.221021
IG(ds,YearSal<=18000):	0.26900		1.252935	0.214692
IG(ds,YearSal<=33000):	0.72193		0.90515	**0.797581**
IG(ds,YearSal<=45000):	0.55678		0.96515	0.576884
IG(ds,YearSal<=52800):	0.57095		0.95098	0.600383
IG(ds,YearSal<=67200):	0.17095		1.35098	0.126539
IG(ds,YearSal<=77200):	0.15678		1.36515	0.114846
IG(ds,YearSal<=82000):	0.07291		1.44902	0.050315
IG(ds,YearSal<=92000):	0.14449		1.37744	0.104895

Variance

- Reduction in variance is a method of splitting the node when the target variable is continuous.
- The formula of variance is as follow:

$$Variance = \frac{\sum(X-\mu)^2}{N}$$

Where:

X is the data point

μ is the mean of the data points, and

N is the total number of data points.

Reduction in Variance

- For each split, individually calculate the variance of each child node.
- Calculate the variance of each split as the weighted average variance of child nodes.
- Select the split with the lowest variance.

- https://www.python-course.eu/Decision_Trees.php

References

- https://www.python-course.eu/Decision_Trees.php
- https://machinelearningmastery.com/implement-decision-tree-algorithm-scratch-python/
- https://www.datacamp.com/community/tutorials/decision-tree-classification-python
- https://www.analyticsvidhya.com/blog/2020/10/all-about-decision-tree-from-scratch-with-python-implementation/
- https://towardsdatascience.com/decision-tree-from-scratch-in-python-46e99dfea775