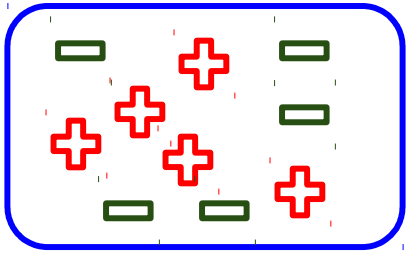
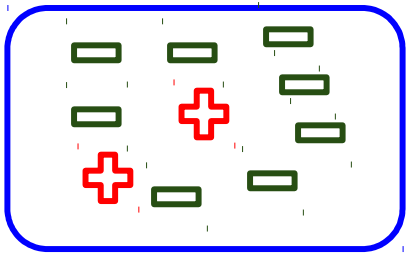
How to select best split point in Decision Trees?

Information Gain

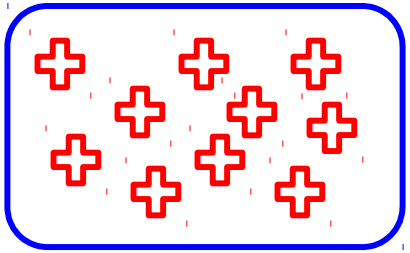
Information Gain



Node 1

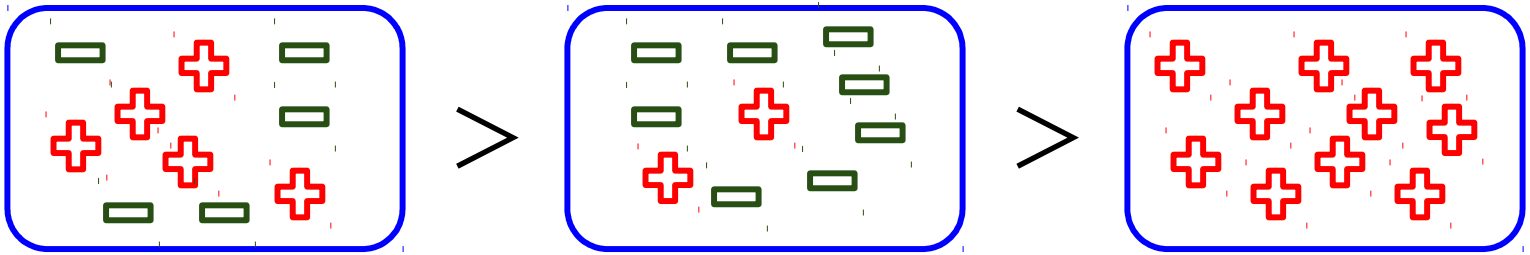


Node 2



Node 3

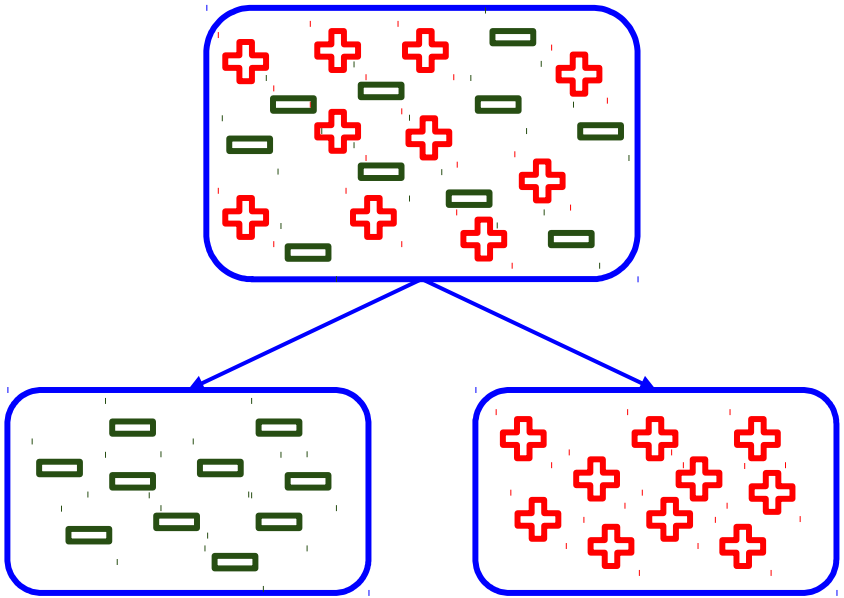
Information Gain

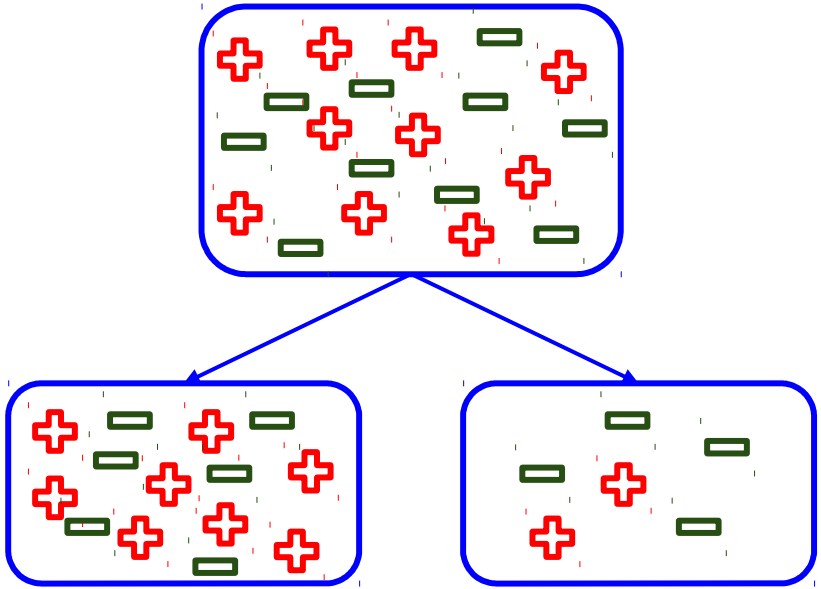


Node 1 | Node 2 | Node 3

Information required to describe the node

Information Gain





Analytics Vidhya
Learn everything about analytics

Information Gain

Information Gain = 1 - Entropy

Information Gain

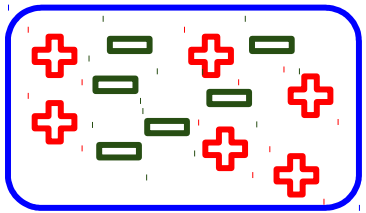
Entropy

$$- p_1 * \log_2 p_1 - p_2 * \log_2 p_2 - p_3 * \log_2 p_3 - ..... - p_n * \log_2 p_n$$

Information Gain

Entropy

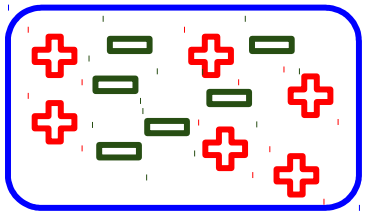
$$- p_1 * \log_2 p_1 - p_2 * \log_2 p_2 - p_3 * \log_2 p_3 - ..... - p_n * \log_2 p_n$$



Information Gain

Entropy

$$- p_1 * \log_2 p_1 - p_2 * \log_2 p_2 - p_3 * \log_2 p_3 - ..... - p_n * \log_2 p_n$$

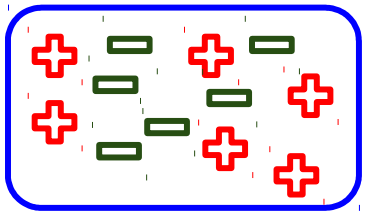


% Play = 0.50
% Not play = 0.50

Information Gain

Entropy

$$- p_1 * \log_2 p_1 - p_2 * \log_2 p_2 - p_3 * \log_2 p_3 - ..... - p_n * \log_2 p_n$$



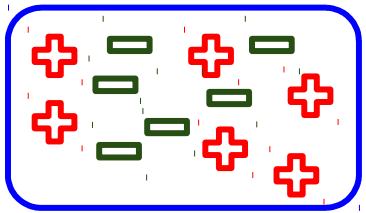
% Play = 0.50
% Not play = 0.50

Entropy = - (0.5) * log2(0.5) - (0.5) * log2(0.5)

Information Gain

Entropy

$$- p_1 * \log_2 p_1 - p_2 * \log_2 p_2 - p_3 * \log_2 p_3 - ..... - p_n * \log_2 p_n$$



% Play = 0.50
% Not play = 0.50

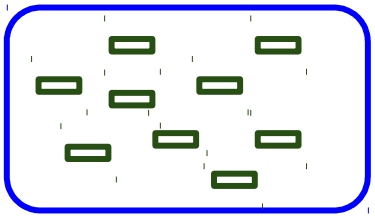
Entropy = - (0.5) * log2(0.5) - (0.5) * log2(0.5)

= 1

Information Gain

Entropy

$$- p_1 * \log_2 p_1 - p_2 * \log_2 p_2 - p_3 * \log_2 p_3 - ..... - p_n * \log_2 p_n$$



% Play = 0
% Not play = 1

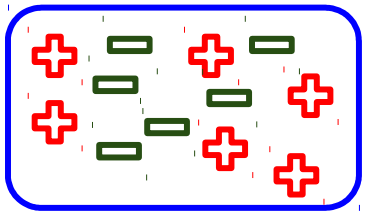
Entropy = - (0) * log2(0) - (1) * log2(1)

= 0

Information Gain

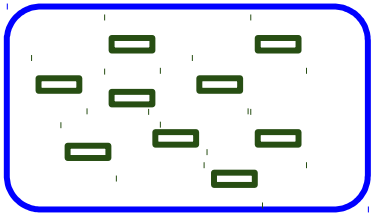
Entropy

$$- p_1 * \log_2 p_1 - p_2 * \log_2 p_2 - p_3 * \log_2 p_3 - ..... - p_n * \log_2 p_n$$



% Play = 0.50
% Not play = 0.50

Entropy = 1



% Play = 0
% Not play = 1

Entropy = 0

Properties of Entropy

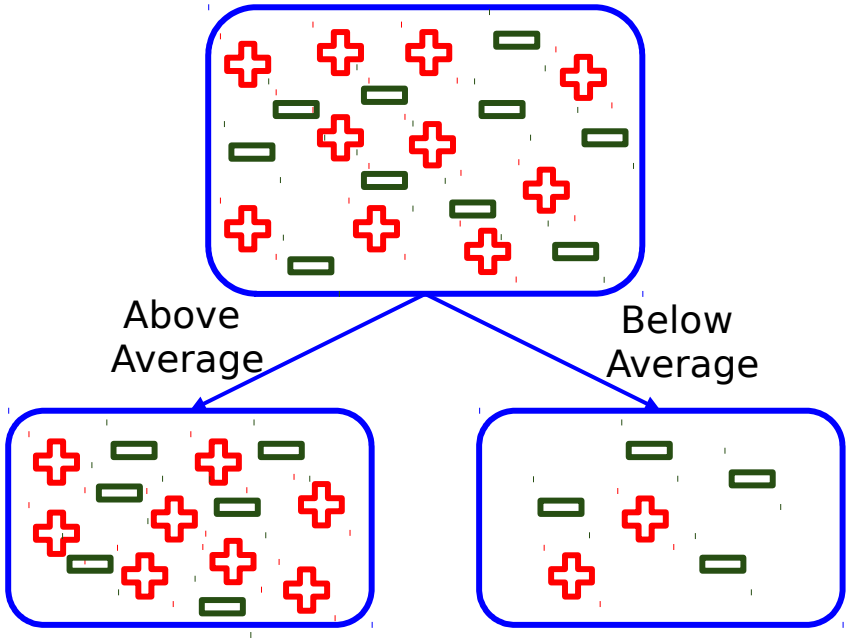
- Works only with categorical targets
- Lesser the Entropy, higher the homogeneity of nodes

Steps to calculate Entropy for a split

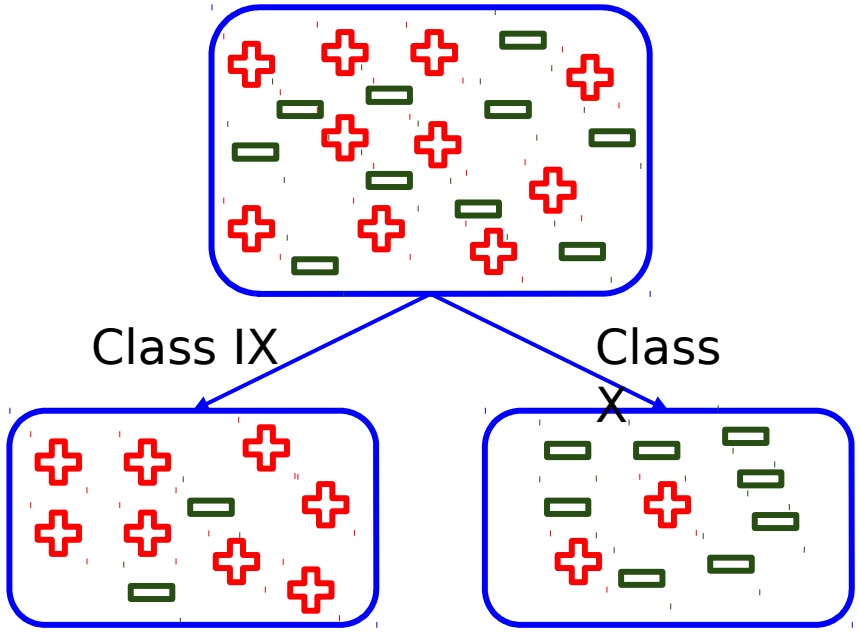
- Calculate the entropy of the parent node

- Calculate the entropy of each child node
- Calculate the weighted average entropy of the split

Steps to calculate Entropy for a split



Split on Performance in Class



Split on Class

Analytics Vidhya
Learn everything about analytics

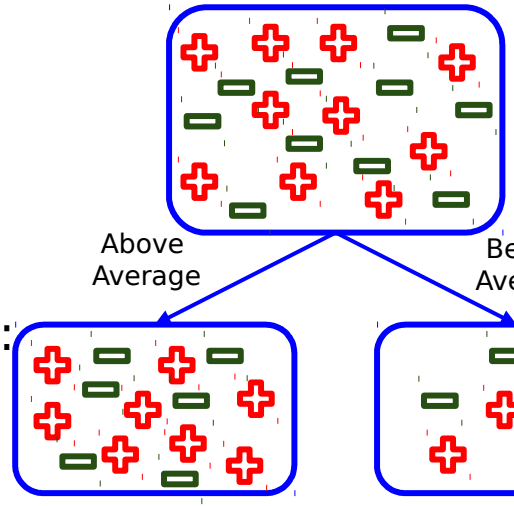
Steps to calculate Entropy for a split

Split on Performance in Class

Steps to calculate Entropy for a split

Split on Performance in Class

- Entropy for Parent node:
  $-(0.5)*\log_2(0.5) -(0.5)*\log_2(0.5) = 1$
- Entropy for sub-node Above Average:
  $-(0.57)*\log_2(0.57) -(0.43)*\log_2(0.43) = 0.98$
- Entropy for sub-node Below Average:
  $-(0.33)*\log_2(0.33) -(0.67)*\log_2(0.67) = 0.91$
- Weighted Entropy: Performance in Class:
  $(14/20)*0.98 + (6/20)*0.91 = 0.959$



Analytics Vidhya
Learn everything about analytics

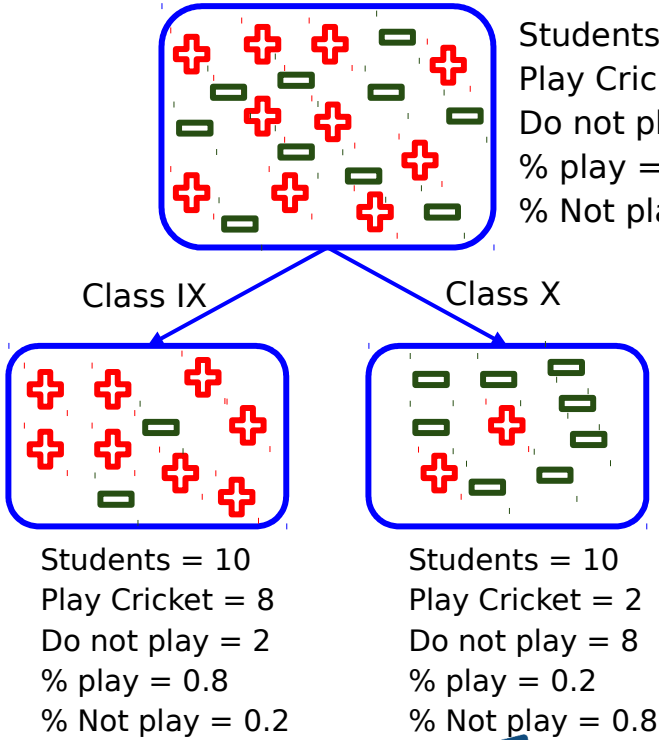
Steps to calculate Entropy for a split

Split on Class

Steps to calculate Entropy for a split

Split on Class

- Entropy for Parent node:
  $-(0.5)*\log_2(0,5) -(0.5)*\log_2(0.5) = 1$
- Entropy for sub-node Class IX:
  $-(0.8)*\log_2(0.8) -(0.2)*\log_2(0.2) = 0.722$
- Entropy for sub-node Class X:
  $-(0.2)*\log_2(0.2) -(0.8)*\log_2(0.8) = 0.722$
- Weighted Entropy: Class:
  $(10/20)*0.722 + (10/20)*0.722 = 0.722$



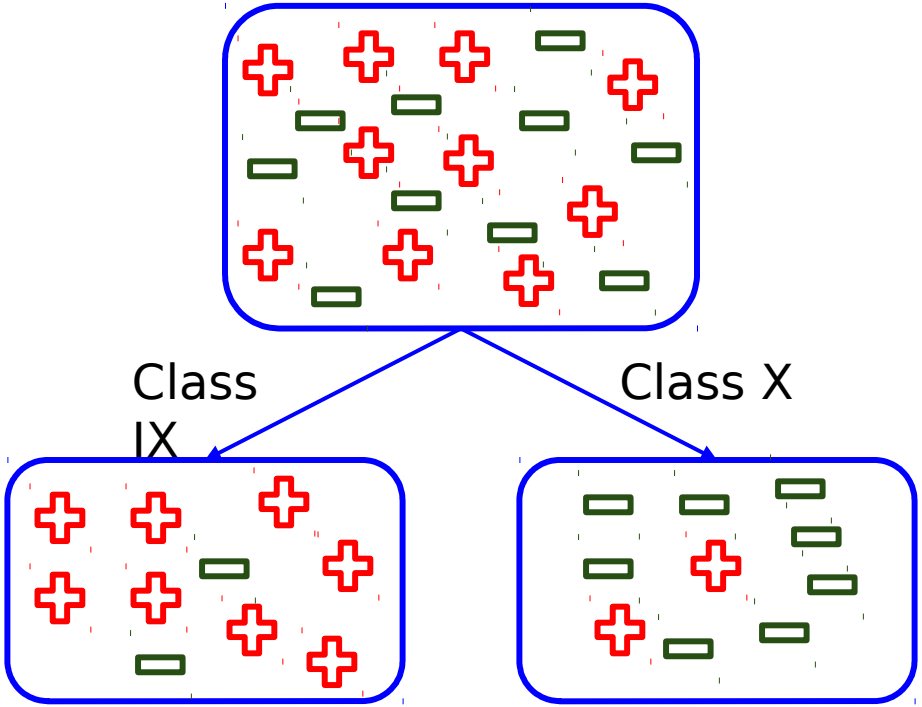
Steps to calculate Entropy for a split

Split	Entropy	Information Gain
Performance in Class	0.959	0.041
Class	0.722	0.278

Steps to calculate Entropy for a split

Split	Entropy	Information Gain
Performance in Class	0.959	0.041
Class	0.722	0.278

Steps to calculate Entropy for a split



Split on Class

Thank
You!

Analytics Vidhya
Learn everything about analytics