Shannon Entropy

$$Entropy - H(p) = -\sum_i p_i * log_2(p_i)$$

1. Average Amount of information that we get from one sample drawn from a given Probability Distribution.
2. It tells about how unpredictable the Probability Distribution is or how uncertain the events are.
3. Sharing 1bit of information reduces the recipient uncertainty by a factor of 2.

   Consider the following Probability Distribution of weather on a day.

   P(Sunny) =0.5

   P(Rain) = 0.5

   If we get to know tomorrow it would be sunny,this information has reduced our uncertainty by a factor of 2. In other words,there were two likely options but with the given information. it is just one. So a 1bit useful information is shared.

   Likewise , consider for 8 possible equally likely events,if we get to know the next likely event,the uncertainty is reduced by a factor of 8. Hence 3 bits of useful information is sent.

$$2^3 = 8$$

$$log_2(8) = 3$$

   In cases where it is not equally likely,

   P(Sunny) = 0.75

   P(Rain) = 0.25

   Knowing it will rain tomorrow,the uncertainty is reduced by a factor of 8, or it is the inverse of the events probability.

$$Bits(Rainy) = log_2(1/p) = log_2(1/0.25) = log_2(4) = 2bits$$

$$Bits(Sunny) = log_2(1/p) = log_2(1/0.75) = -log_2(0.75) = 0.41bits$$

If we calculate , on average how much information we get knowing a particular event or in other terms,how uncertain the events are.

$$Entropy = 0.75 * -log(0.75) + .25 * -log(0.25) = 0.81$$

$$Entropy = 0.75 * 0.41 + .25 * 2 = 0.81$$

$$Entropy - H(p) = -\sum_i p_i * log_2(p_i)$$

Splitting Criteria in Decision Tree

Information Gain of a node for a Dataset(D) with a Categorical feature(Xj) :

$$Gain(D, x_j) = H(D) - \sum_{v \in Values(x_j)} \frac{|D_v|}{|D|} H(D_v)$$

$$\text{where H(D) - Entropy of the parent node}$$

$$H(D_v) - \text{Entropy of the child nodes.}$$

Higher the Information Gain,better is the split.

Same formula works for the Continuous Feature,the only difference would be

$X(j) <= threshod, X(j) > threshold$ will be the representation of the child nodes for binary trees

1. Entropy

$$Entropy - H(p) = -\sum_i p_i * log_2(p_i)$$

Range - [0,1]

2. Gini Impurity

Total Gini Impurity = Weighted average of Gini Impurities for the Leaves.

$$\text{Gini Impurity for a Leaf} = 1 - \sum_i (p(i)^2)$$

Lower the Gini Impurity,better is the split.

Range - [0,0.5]

References

1. Splitting Criteria in Decision Tree
2. Entropy
3. The StatQuest Illustrated Guide to Machine Learning!!!