DECISION TREES

Gini Impurity

Now

for a set 's'

If $[y_1, y_2, \cdots y_n]$ are the labels.

and let there be $k \leq n$ different kinds of labels

so let $p(j)$ be the probability that if you randomly select a label from 's' it will belong to class 'j'.

$$p(j) = \frac{\{y_i, y_i = j\}_{i=1 \text{ to } n}}{\{y_i\}_{i=1 \text{ to } n}}$$

$$p(j) = \frac{\text{No: of labels of class 'j'}}{\text{Total no: of labels.}}$$

Now suppose we do an experiment of picking 2 things with replacement from 's'

The probability that both will be of class 'j' is $p(j)^2$

The probability that both will be of same class is $p(1)^2 + p(2)^2 + \cdots p(j)^2$

$$= \sum_{i=1}^{j} p(i)^2$$

Thus the probability of 2 diff groups is $1 - \sum_{i=1}^{j} p(i)^2$

This should be minimum as we don't want various outputs

and This is defined to be gini impurity of set S

$$G(s) = 1 - \sum_{i=1}^{j} p(i)^2$$

This is same as $G(s) = \sum_{i=1}^{j} p(i)(1-p(i))$

as $\sum_{i=1}^{j} p(i)(1-p(i))$

$= \sum_{i=1}^{j} p(i) - p(i)^2$

$= \sum_{i=1}^{j} p(i) - \sum_{i=1}^{j} p(i)^2$

$\sum_{i=1}^{j} p(i)$ = probability of picking any class = 1

$= 1 - \sum_{i=1}^{j} p(i)^2$