Decision Trees – Part 3

GINI Index

Decision Tree Algorithms

- Hunt’s Algorithm (one of the earliest)
- CART (Classification & Regression Tree)
- ID3 (Iterative Dichotomiser 3, Ross Quinlan)
- C4.5, C5.0 .. (Ross Quinlan)
- SLIQ (Supervised Learning In Quest,IBM)
- SPRINT(Scalable PaRallelizable INduction of decision Trees)

Measures of Node Impurity

- This is a measure of how inhomogeneous a node is.
- If all instances in a node are the same node impurity is 0.
- If there is a mix of classes, impurity is high.
- The decision tree algorithm is trying to find splits that lead to pure (homogenous) nodes.
- Or nodes with a low level of impurity.

Measures of Node Impurity

- All Decision Tree algorithms depend on a measure of node impurity.
- Three measures are
  - Gini Index
  - Entropy
  - Misclassification error
- The best split is the split that leads to the nodes with the lowest impurity.

Measure of Impurity: GINI

$$GINI(n) = 1 - \sum_{c} [P(c|n)]^2$$

- This calculates the GINI value of node n.
- P(c|n) is the probability of an instance in node n being of class c.
- This is also the relative frequency of class c in node n.

GINI

- For example, if a node n consists of 2 instances of class C1 and 4 instances of class C2 then
- P(C1|n) is 2/6
- P(C2|n) is 4/6
- P(C1|n) is also known as the relative frequencey of the class C1 in the node n.

GINI

$$GINI(n) = 1 - \sum_c [P(c|n)]^2$$

C1	**0**
C2	**6**
Gini=0.000	

C1	**1**
C2	**5**
Gini=0.278	

C1	**2**
C2	**4**
Gini=0.444	

C1	**3**
C2	**3**
Gini=0.500	

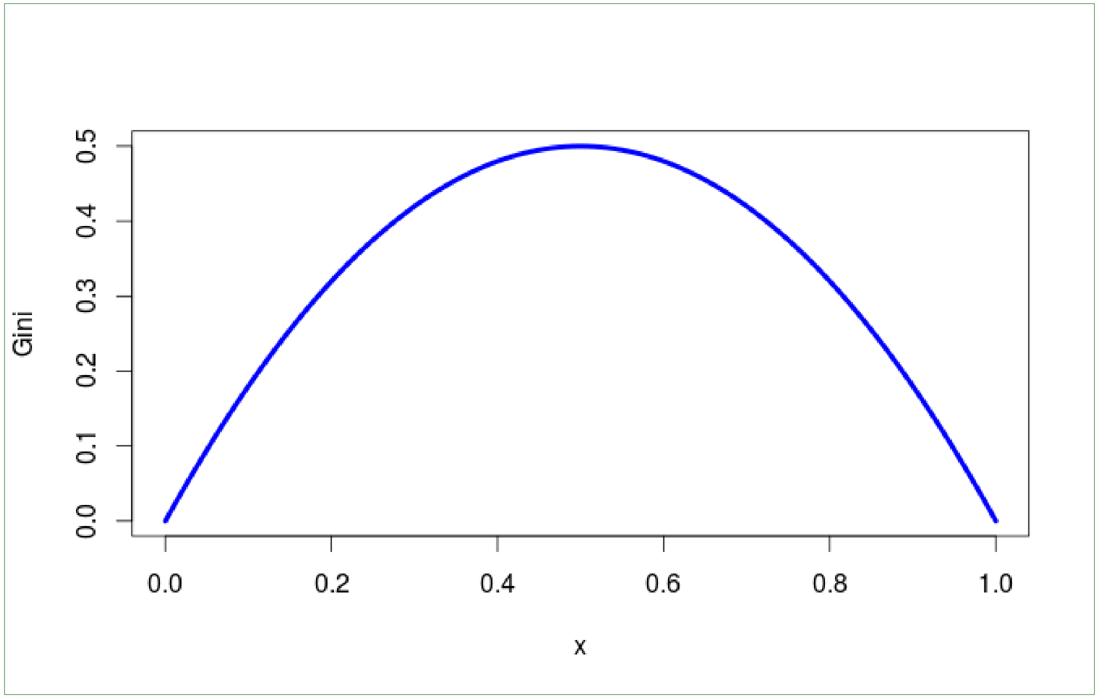
- GINI(n1) = 1 – $(0/6)^2$ – $(6/6)^2$ = 0
- GINI(n2) = 1 – $(1/6)^2$ – $(5/6)^2$ = 0.278
- GINI(n3) = 1 – $(2/6)^2$ – $(4/6)^2$ = 0.444
- GINI(n4) = 1 – $(3/6)^2$ – $(3/6)^2$ = 0.5

GINI

$$GINI(n) = 1 - \sum_{c} [P(c|n)]^2$$

- If n_c is the number of classes the maximum value of GINI is $(1 - 1/n_c)$
- This occurs when instances are equally distributed among all classes
- The minimum value of GINI is 0 (no matter how many classes there are)
- This occurs when all instances belong to one class.

GINI for two classes



- x = relative frequency of one of the classes
- (1-x = relative frequency of the other)

Computing GINI Index for a Split

- We know how to evaluate a node.
- But what the decision tree algorithm has to find is the best split of a node.
- Ths will depend on the GINI values of the two (or more) child nodes.
- One way to calculate this would be to take the average (mean) of the GINI values of the two child nodes (hint, not the right approach)

GINI value for a split – mean doesn't work

- Suppose we have a node of size 12.
- Which of the following splits are better
- 1/1 and 10/0 mean(0.5, 0) = 0.25
- 5/5 and 2/0 mean(0.5, 0) = 0.25
- The mean of GINI values gives the same result.
- Yet the first split is obviously the better split.
- The larger node is less impure (and more homogenous)

GINI value for a split

- This is defined as the weighted average of the GINI values of the children.
- For the split 1/1 and 10/0
  - GINI(split) = 2/12(0.5) + 10/12(0) = 1/12
- For the split 5/5 and 2/0
  - GINI(split) = 10/12(0.5) + 2/12(0) = 5/12
- Reflecting the fact that the first split is better.
- (Remember lower GINI values are better)

GINI

- Now we know how to evalute splits.
- The decision tree algorithm generates all possible splits and uses a measure such as GINI to pick to best one.