COMPUTER SUBJECT:	BASIC ML CONCEPTS
TYPE:	GROUP WORK ASSIGNMENTS/DISCUSSION
IDENTIFICATION:	CHAPTER 6/MICL
COPYRIGHT:	*Michael Claudius*
LEVEL:	EASY
DURATION:	60 min
SIZE:	3 pages max!!
OBJECTIVE:	Understanding decision trees elements
REQUIREMENTS:	**ML Ch. 6**
COMMANDS:	

IDENTIFICATION: CHAPTER 6/MICL

ML Chapter 6 Assignments in Decision Trees

The following assignments are as usual to be solved in smaller groups (2-4 persons),

Assignment 1
What is a decision tree?

Assignment 2
N/A

Assignment 3
Give some examples where decision trees are applicable ?

Assignment 4
What is the approximate maximum depth of a Decision Tree trained (without restrictions) on a training set with one million instances?
What is the approximate maximum depth of a balanced Decision Tree trained (without restrictions) on a training set with one million instances?
Is it a good idea to utilize the maximum depth?

Assignment 5
Compare a node’s Gini impurity with its parent’s Gini-impurity?
Is it generally lower/greater, or always lower/greater?

Assignment 6
If a Decision Tree is overfitting the training set, is it a good idea to try decreasing max_depth?

Assignment 7
If a Decision Tree is underfitting the training set, is it a good idea to try scaling the input features?

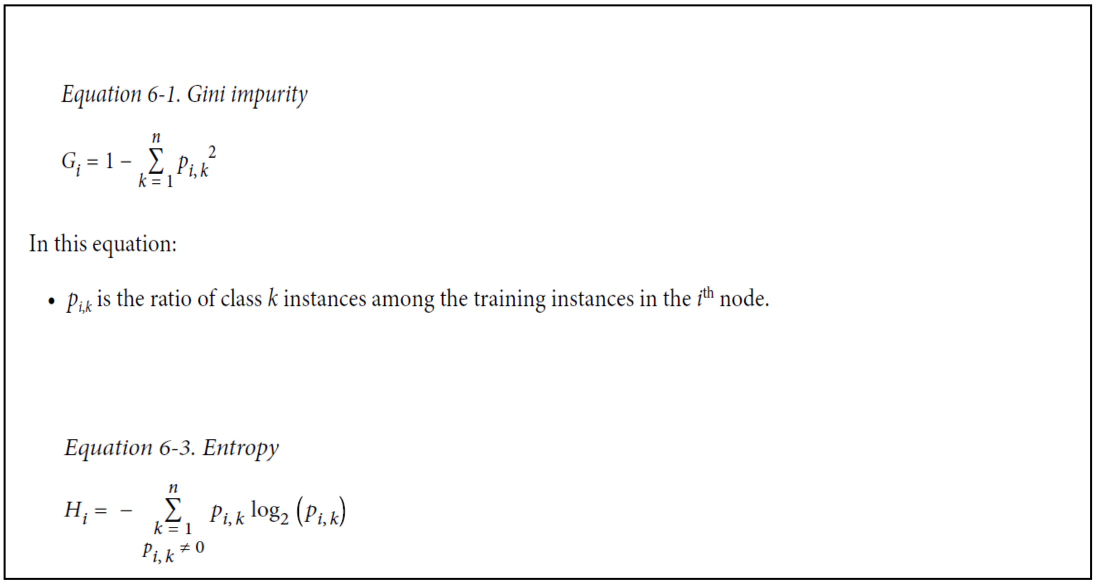
Assignment 8
If it takes one hour to train a Decision Tree on a training set containing 1 million instances, roughly how much time will it take to train another Decision Tree on a training set containing 10 million instances?
Tip, you should find a formula describing this……O(n x m x log(m))

Assignment 9
If your training set contains 100,000 instances, will setting presort=True speed up training?

Assignment 10

We shall now compare entropy vs impurity.

Take a look at formulas the figure below.



Equation 6-1. Gini impurity

$$G_i = 1 - \sum_{k=1}^{n} {p_{i,k}}^2$$

In this equation:

- $p_{i,k}$ is the ratio of class k instances among the training instances in the i^{th} node.

Equation 6-3. Entropy

$$H_i = - \sum_{\substack{k=1 \\ p_{i,k} \neq 0}}^{n} p_{i,k} \log_2 \left(p_{i,k}\right)$$

What is the Gini impurity function used for ?
What is entropy used for?

Look at the left leaf in the right subtree of the decision tree figure 6.1 below:

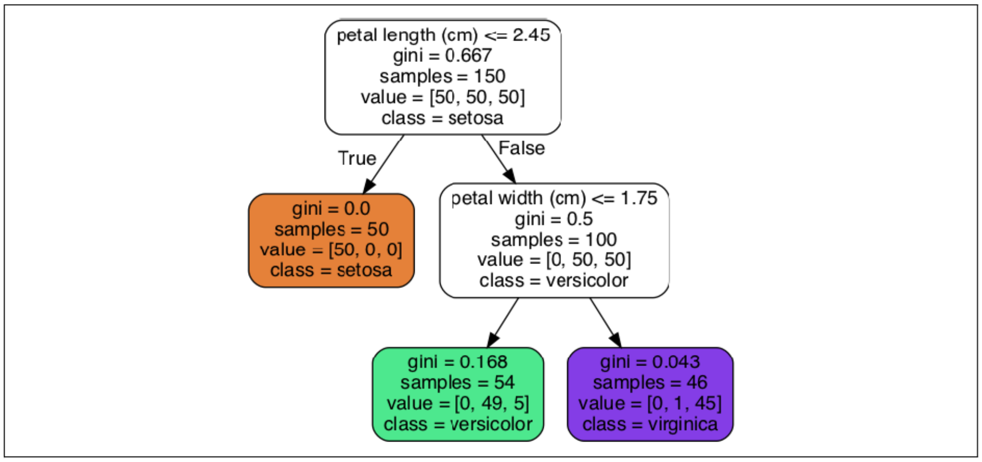
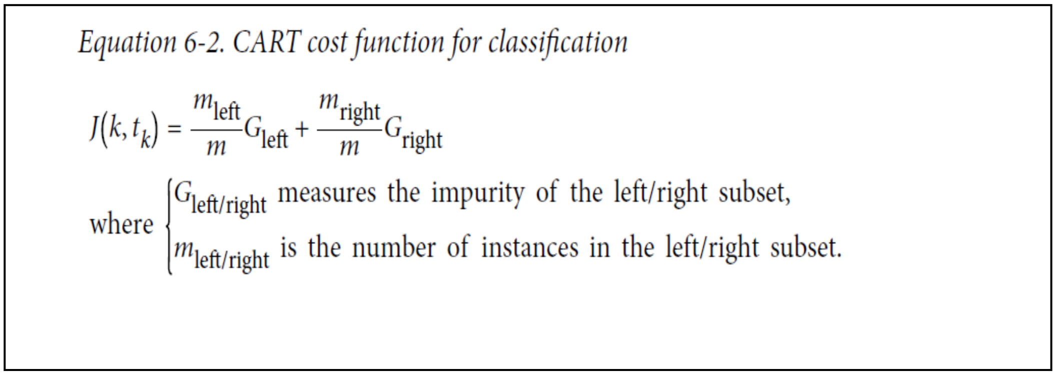


Figure 6-1. Iris Decision Tree

Verify Geni impurity $G_2 = 0.168$. $1 - (49/54)^2 - (5/54)^2 = 0.168$
Calculate the entropy (I suggest you to use $\log_{10}$ and not $\log_2$), $H_{2.} =$

Assignment 11

Take a look at the CART cost function, J, for a single training instance in equation, 6.2 below.



Equation 6-2. CART cost function for classification

$$J(k, t_k) = \frac{m_{\text{left}}}{m} G_{\text{left}} + \frac{m_{\text{right}}}{m} G_{\text{right}}$$

$$\text{where} \begin{cases} G_{\text{left/right}} \text{ measures the impurity of the left/right subset,} \\ m_{\text{left/right}} \text{ is the number of instances in the left/right subset.} \end{cases}$$

Calculate the cost, J, for the right subtree of the decision tree figure 6.1 below:

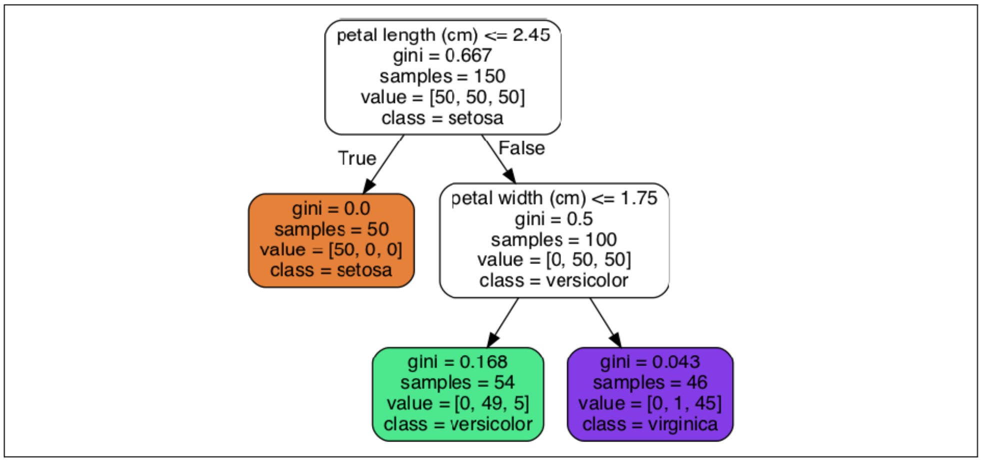


Figure 6-1. Iris Decision Tree

Got it ?

Assignment X deprecated

We shall now compare entropy vs impurity by another example

Take a look at formulas the figure below.
What is the Gini impurity function used for ?
What is entropy used for?