Week 5 Practical

Decision Tree – Measures of Node Impurity

Overview

In this Practical we will extend the discussion on various measures of node impurity. This is an important concept and easiest way to understand it well is to practice it.

For this practical you can use MS Excel or any other tool you find appropriate. There is no need for RStudio.

Objectives

- Review node impurity measures.
- Understand decision tree induction.

Sharing results

While tasks should be self-explanatory, some of the challenges might be more complex. Feel free to share your results or questions in the course discussion forum. Results to all the challenges will be available UPON REQUEST (discussion forum or email).

Task 1. Node impurity measures

Consider the training examples shown in Table 1. for a binary classification problem. This is an excerpt from the dataset we used in our previous practical.

Table 1. A sample of the Stroke prediction dataset.

Inst.	Gender	Age	Work Type	Avg. glucose level	BMI	Ever married	Stroke
1	Male	37	Private	79.56	25.0	Yes	0
2	Female	78	Self-employed	81.68	23.0	Yes	0
3	Male	74	Private	97.27	27.0	Yes	1
4	Male	64	Self-employed	111.98	19.0	No	1
5	Female	74	Private	231.61	37.0	No	1
6	Female	67	Private	82.09	14.0	Yes	0
7	Female	59	Private	57.47	31.0	No	0
8	Female	57	Private	82.62	29.0	No	0
9	Female	25	Self-employed	59.78	35.0	Yes	0
10	Male	58	Private	240.59	33.0	Yes	1

Questions:

1. What is the **entropy** of this collection of training examples?
2. What is the **information gain** for Gender and Work Type attributes, relative to these training examples?
    a. Which attribute (Gender or Work Type) would you select for split?
3. For BMI, which is a numerical attribute, compute the **information gain** for every possible split.
    a. How would you calculate the information gain for Age?
4. From our previous calculations, what would be the best split – Gender, Work Type, or BMI, according to the **information gain**?
5. What is the best split (between Gender and Work Type) according to the **classification error rate**?
6. What is the best split (between Gender and Work Type) according to the **Gini index**?