Course 2 Section 4.11 - CLASSIFICATION TREES

Jiaying Wu

17/10/2020

These are easy to think about when the response variable has only two categories, say y = 0 or 1 (a binary variable) then:

- Gini is $2p(1-p)$, and
- Entropy is $-p(log_e p - (1-p)log_e 1 - p$

where p is the proportion of cases in a subset belonging to class 1. Everything else in the fitting procedure remains the same.

Q1.What is the value of Gini when $p = 0.5$? $p = 1$? $p = 0.5$?

When p = 0.5,

$$2p(1-p) = 2 * 0.5 * (1 - 0.5) = 0.5$$

When p = 1,

$$2p(1-p) = 2 * 1 * (1 - 1) = 0$$

When p = 1,

$$2p(1-p) = 2 * 0 * (1 - 0) = 0$$

Q2.Which value would indicate a subset of cases that are all one class?

If we think of p as "the proportion of cases in a subset belonging to class 1." Then p=0 would indicate a subset of cases that are all one class, while p = 1 means the opposite.

Q3.Should a higher or lower value of Gini indicate a subset is more pure, that is mostly one class?

A lower Gini indicate a subset is more pure.