Fundamentals of Data Science

Semester B 20-21

Tutorial 5

1.a. The original Gini index is $1-(\frac{4}{9})^2-(\frac{5}{9})^2=0.494$

After splitting on a_1, the Gini index becomes

$$\frac{4}{9}[1-(\frac{3}{4})^2-(\frac{1}{4})^2]+\frac{5}{9}[1-(\frac{1}{5})^2-(\frac{4}{5})^2]=0.344$$

As a result, the change in Gini index is

$$\Delta G_{a_1} = 0.494 - 0.344 = 0.15$$

After splitting on a_2, the Gini index becomes

$$\frac{5}{9}[1-(\frac{2}{5})^2-(\frac{3}{5})^2]+\frac{4}{9}[1-(\frac{2}{4})^2-(\frac{2}{4})^2]=0.489$$

As a result,

$$\Delta G_{a_2} = 0.494 - 0.489 = 0.005$$

b. The original classification error rate is $1-\max(\frac{4}{9},\frac{5}{9})=\frac{4}{9}$

After splitting on a_1, the classification error rate becomes

$$\frac{4}{9}[1-\max(\frac{3}{4},\frac{1}{4})]+\frac{5}{9}[1-\max(\frac{1}{5},\frac{4}{5})]=\frac{2}{9}$$

As a result, the change in classification error rate is

$$\Delta E_{a_1} = \frac{4}{9}-\frac{2}{9}=\frac{2}{9}$$

After splitting on a_2, the classification error rate becomes

$$\frac{5}{9}[1-\max(\frac{2}{5},\frac{3}{5})]+\frac{4}{9}[1-\max(\frac{2}{4},\frac{2}{4})]=\frac{4}{9}$$

As a result,

$$\Delta E_{a_2} = \frac{4}{9}-\frac{4}{9}=0$$

c. We consider the different possible split points for a_3 as follows:

a_3	Class label	Split point	Entropy	Info gain
1	+	2.0	0.848	0.143
3	-	3.5	0.989	0.002
4	+	4.5	0.918	0.073
5 5	- -	5.5	0.984	0.007
6	+	6.5	0.973	0.018
7 7	+ -	7.5	0.889	0.102
8	-			

The best split for a_3 occurs when the split point is 2.0.