Fundamentals of Data Science

Semester B 20-21

Tutorial 4

1 This is because there are 2^S ways of assigning the S attribute values to two classes. Since the left/right ordering of the classes is not important, and we exclude the two cases where all the attribute values are assigned to one of the classes, the resulting number of possible partitions is $\frac{2^S - 2}{2} = 2^{S-1} - 1$.

2.(a) The original entropy is $-\frac{4}{9}\log_2\frac{4}{9} - \frac{5}{9}\log_2\frac{5}{9} = 0.991$ bit

(b) After splitting on a_1, the entropy becomes

$$\frac{4}{9}(-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}) + \frac{5}{9}(-\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5}) = 0.762\,\text{bit}$$

As a result

$$gain(a_1) = 0.991 - 0.762 = 0.229\,\text{bit}$$

After splitting on a_2, the entropy becomes

$$\frac{5}{9}(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}) + \frac{4}{9}(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}) = 0.984\,\text{bit}$$

As a result,

$$gain(a_2) = 0.991 - 0.984 = 0.007\,\text{bit}$$

3. The attribute **AIR** is chosen as the first attribute for splitting the data set, as described in the lecture notes.

For the branch **High**, there is no need for further splitting, and we can form the leaf node **Large**.

For the branch **Low** with associated partition {1,3,4,6,9}, we need to choose among the two attributes **TEMP** and **HUMID** to perform splitting.

The original entropy is 0.971 bit.

After splitting on **TEMP**, the entropy becomes

$$\frac{1}{5}(0) + \frac{2}{5}(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}) + \frac{2}{5}(0) = 0.4 \text{ bit}$$

As a result

$$gain(\textbf{TEMP}) = 0.971 - 0.4 = 0.571\,\text{bit}$$

After splitting on **HUMID**, the entropy becomes

$$\frac{3}{5}(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}) + \frac{2}{5}(0) = 0.551\,\text{bit}$$

As a result,

$$gain(\textbf{HUMID}) = 0.971 - 0.551 = 0.42\,\text{bit}$$

As a result, we choose the attribute **TEMP**, which split {1,3,4,6,9} into {4},{3,6},{1,9}.

For the partition {4}, we can form the leaf node **Small**.

For the partition {1,9}, there is no need for further splitting, and we can form the leaf node **Large**.

For the partition {3,6}, we can use **HUMID** to split it into {6} and {3}. For {6}, we can form the leaf node **Small**, and for {3}, we can form the leaf node **Large**.