Tutorial 3

1. This is because there are 2^N ways of assigning the N 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 $\frac{2^N - 2}{2} = 2^{N-1} - 1$

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

   For the branch \$0 to \$15k, there is no need for further splitting, and we can form the leaf node **high risk**.

   For the branch \$15 to \$35k with associated partition {2,3,12,14}, we need to choose among the three attributes **credit history**, **debt**, **collateral** to perform splitting.

   The original entropy is 1 bit.

After splitting on **credit history**, the entropy becomes

$$\frac{2}{4}(-\frac{1}{2}\log_2\frac{1}{2}-\frac{1}{2}\log_2\frac{1}{2})+\frac{1}{4}(0)+\frac{1}{4}(0)=0.5 \text{ bit}$$

As a result

gain(**credit history**) = 1-0.5 = 0.5 bit

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

$$\frac{3}{4}(-\frac{2}{3}\log_2\frac{2}{3}-\frac{1}{3}\log_2\frac{1}{3})+\frac{1}{4}(0)=0.689 \text{ bit}$$

As a result,

gain(**debt**) = 1-0.689 = 0.311 bit

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

$$\frac{4}{4}(-\frac{2}{4}\log_2\frac{2}{4}-\frac{2}{4}\log_2\frac{2}{4})=1 \text{ bit}$$

As a result

gain(**collateral**) = 1-1 = 0 bit

As a result, we choose the attribute **credit history**, which split {2,3,12,14} into {2,3},{12},{14}.

For the partition {2,3}, we need to choose among the two attributes **debt** and **collateral** to perform splitting.

The original entropy is 1 bit.

After splitting on **debt**, the entropy becomes 0 bit.

As a result,

gain(**debt**) = 1-0 = 1 bit

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

$$\frac{2}{2}(-\frac{1}{2}\log_2\frac{1}{2}-\frac{1}{2}\log_2\frac{1}{2})=1 \text{ bit}$$

As a result,

gain(**collateral**) = 1-1 = 0 bit

As a result, we choose the attribute **debt**, which split {2,3} into {2},{3}.

For the branch over $35k with associated partition {5,6,8,9,10,13}, we need to choose among the three attributes **credit history**, **debt**, **collateral** to perform splitting.

The original entropy is 0.65bit.

After splitting on **credit history**, the entropy becomes 0 bit.

As a result

gain(**credit history**) = 0.65-0 = 0.65 bit

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

$$\frac{4}{6}(-\frac{3}{4}\log_2\frac{3}{4}-\frac{1}{4}\log_2\frac{1}{4})+\frac{2}{6}(0)=0.541 \text{ bit}$$

As a result,

gain(**debt**) = 0.65-0.541 = 0.109 bit

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

$$\frac{3}{6}(-\frac{2}{3}\log_2\frac{2}{3}-\frac{1}{3}\log_2\frac{1}{3})+\frac{3}{6}(0)=0.459 \text{ bit}$$

As a result

gain(**collateral**) = 0.65-0.459 = 0.191 bit

As a result, we choose the attribute **credit history**, which split {5,6,8,9,10,13} into
{5,6},{8},{9,10,13}.