1. We consider the training examples shown in the following table for a binary classification problem.

Instance	a_1	a_2	a_3	Target Class	
1	Т	Т	1	+	
2	Т	Т	6	+	
3	Т	F	5	-	
4	F	F	4	+	
5	F	Т	7	-	
6	F	Т	3	-	
7	F	F	8	-	
8	Т	F	7	+	
9	F	Т	5	-	

a) What is the original entropy of this set of training instances?

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

b) What are the information gains when a_1 and a_2 are used for partitioning the training set respectively?

After splitting on a_1 , the entropy becomes

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

As a result,

gain
$$(a_1) = 0.991 - 0.762 = 0.229$$
 bit.

After splitting on a_2 , the entropy becomes

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

As a result,

gain
$$(a_2) = 0.991 - 0.984 = 0.007$$
 bit.

- 2. We again consider the training examples shown in Q.1
 - a) Calculate the respective changes in the Gini index value when a_1 and a_2 are used for partitioning the training set.

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

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

As a result, the change in Gini index is

$$\triangle G(a_1) = 0.494 - 0.344 = 0.15.$$

After splitting on a_2 , the Gini index becomes

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

As a result,

$$\triangle G(a_2) = 0.494 - 0.489 = 0.005.$$

b) Calculate the respective changes in the classification error when a_1 and a_2 are used for partitioning the training set.

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

After splitting on a_1 , the classification error becomes

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

As a result, the change in classification error is

$$\triangle E(a_1) = 4/9 - 2/9 = 2/9.$$

After splitting on a_2 , the classification error 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,

$$\triangle E(a_2) = 4/9 - 4/9 = 0.$$

c) For a_3 , which is a continuous attribute, compute the information gain for every possible split. What is the best threshold for splitting the set of attribute values?

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	0.984	0.007
5	-			
6	+	6.5	0.973	0.018
7	+	7.5	0.889	0.102
7	-			
8	_			

The best split for a_3 occurs when the split point is equal to 2.