WORKSHOP 5 – Decision trees (Solutions)

CMP9137M – MACHINE LEARNING

1. Overview

In this exercise, you are expected to build a decision tree for classification. Instead of using computer software, you are encouraged to calculate the information gains by hand. (you can use a calculator to compute entropy.)

2. Decision trees

A) Suppose a hypothetical UK rail service from Lincoln to Edinburgh is often subject to delays. The train service is run by three different train operating companies (TOC). Over the course of a year, a random sample of the services was taken. The following data was obtained

	Weather	Season	TOC	Day	Lateness
Case 1	Windy	Summer	RotRail	Weekday	On time
Case 2	Windy	Winter	GNAF	Weekday	Delayed
Case 3	Windy	Autumn	GNAF	Weekday	Delayed
Case 4	Calm	Summer	Virgo	Weekend	Delayed
Case 5	Windy	Winter	RotRail	Weekend	Delayed
Case 6	Calm	Summer	Virgo	Weekday	Delayed
Case 7	Calm	Spring	RotRail	Weekday	On time
Case 8	Windy	Autumn	GNAF	Weekend	Delayed
Case 9	Calm	Winter	Virgo	Weekend	Delayed
Case 10	Calm	Spring	Virgo	Weekday	Delayed
Case 11	Windy	Autumn	GNAF	Weekday	Delayed
Case 12	Windy	Spring	GNAF	Weekday	On time
Case 13	Windy	Summer	RotRail	Weekday	On time
Case 14	Calm	Autumn	RotRail	Weekday	On time
Case 15	Windy	Winter	RotRail	Weekday	Delayed
Case 16	Calm	Autumn	Virgo	Weekday	Delayed
Case 17	Windy	Summer	Virgo	Weekday	Delayed
Case 18	Windy	Spring	Virgo	Weekend	Delayed
Case 19	Calm	Winter	GNAF	Weekday	On time
Case 20	Calm	Spring	GNAF	Weekend	On time

Find the root (top) node selected using the maximum information gain tree building procedure to classify whether a train will be delayed or on time. Show that it selects according to which TOC is providing the service. You might find the following table a helpful starter

Weather	Delayed	On time
Calm	5	4
Windy	8	3

Season	Delayed	On time
Summer	3	2
Winter	4	1
Autumn	4	1
Spring	2	3

TOC	Delayed	On time
RotRail	2	4
GNAF	4	3
Virgo	7	0

Day	Delayed	On time
Weekday	8	6
Weekend	5	1

Solution: To find the root node, we need to calculate the information gains for each attribute. The attribute which has the highest information gain is chosen as the root. The formulae for information gain is $information\ gain = parent\ entropy - average\ children\ entropy$. For simplicity, let s denote each possible class: late, on time. Let M denote the classification based on all the data. Let M_i denote the classification based on just looking at the data corresponding to value i of an attribute A, for example, A is weather and i is calm. The information gain is given by

$$Gain(M, A) = Ent(M) - \sum_{i \in A} \frac{|M_i|}{|M|} Ent(M_i)$$

$$= Ent(M) - \sum_{i \in A} \frac{|M_i|}{|M|} \sum_{s} \frac{|M_i^s|}{|M_i|} log \frac{|M_i^s|}{|M_i|}$$

where $|M|$ denotes the number of elements in M, $Ent(M)$ represents the entropy of M. We can now calculate each attribute use this equation. Note we use log base 2.

The parent entropy is $Ent(M) = -\frac{13}{20} log_2 \left(\frac{13}{20}\right) - \frac{7}{20} log_2 \left(\frac{7}{20}\right) = 0.9341$

Weather

$Gain(M, weather) = 0.9341 - \frac{9}{20} \times \left(-\frac{5}{9} log \frac{5}{9} - \frac{4}{9} log \frac{4}{9}\right) - \frac{11}{20} \times \left(-\frac{8}{11} log \frac{8}{11} - \frac{3}{11} log \frac{3}{11}\right) = 0.0232$

Season

$$Gain(M, Season) = 0.9341 - \frac{5}{20} \times \left(-\frac{3}{5} log \frac{3}{5} - \frac{2}{5} log \frac{2}{5}\right) - \frac{5}{20} \times \left(-\frac{4}{5} log \frac{4}{5} - \frac{1}{5} log \frac{1}{5}\right) - \frac{5}{20} \times \left(-\frac{4}{5} log \frac{4}{5} - \frac{1}{5} log \frac{1}{5}\right) - \frac{5}{20} \times \left(-\frac{2}{5} log \frac{2}{5} - \frac{3}{5} log \frac{3}{5}\right) = 0.0877$$

TOC

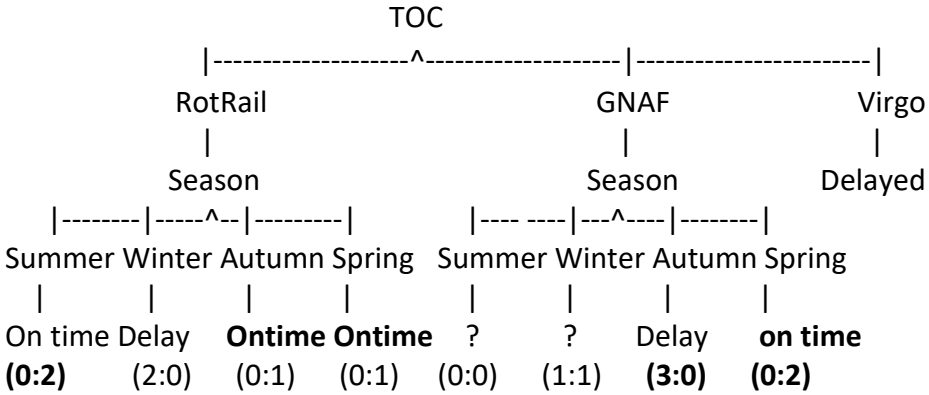
$$Gain(M, TOC) = 0.9341 - \frac{6}{20} \times \left(-\frac{2}{6} log \frac{2}{6} - \frac{4}{6} log \frac{4}{6}\right) - \frac{7}{20} \times \left(-\frac{4}{7} log \frac{4}{7} - \frac{3}{7} log \frac{3}{7}\right) - \frac{7}{20} \times \left(-\frac{7}{7} log \frac{7}{7} - \frac{0}{7} log \frac{0}{7}\right) = 0.3138$$

Day

$$Gain(M, Day) = 0.9341 - \frac{14}{20} \times \left(-\frac{8}{14} log \frac{8}{14} - \frac{6}{14} log \frac{6}{14}\right) - \frac{6}{20} \times \left(-\frac{5}{6} log \frac{5}{6} - \frac{1}{6} log \frac{1}{6}\right) = 0.0494$$

The largest information gain suggests to use TOC as the root of decision tree to classify if a train is late or on time.

The maximum information gain tree building procedure creates the following first two layers of the tree. Suppose the whole tree were pruned to this level (2 layers). Find the final decision tree by filling in the missing classification values and missing classification ratios below



Solution: See the filled values in the decision tree. However there is no classification decisions on `Summer' and `Winter', the two classes are draw. One possible way is to toss a coin to decide which class you want to choose.

B) Using your decision tree from question A, how would you classify

	Weather	Season	TOC	Day	Lateness
Example 1	Windy	Autumn	RotRail	Weekday	**On time**
Example 2	Calm	Summer	Virgo	Weekday	**Delayed**
Example 3	Calm	Spring	GNAF	Weekend	**On time**