IAML Tutorial 2

1. Train delay based on weather. [see sheet for data]

Find the root (top) node selected using maximum information gain tree building procedure to classify whether a train will be delayed or on time.

Show that it selects to which TOC is providing the service.

Table

	Delayed	On time
Calm	5	4
Windy	8	3
Summer	3	2
Winter	4	1
Autumn	4	1
Spring	2	3

	Delayed	On-time
RotRail	2	4
GNAF	4	3
Virgo	7	0
Weekday	8	6
Weekend	5	1

Let s denote each possible class (late, on time). Let M be the classification based on all the data.

Let M_i be the classification based on just looking at the data corresponding to value i of some attribute A.
e.g. (A = weather, i = calm)

Then Information Gain is given by

$$\begin{aligned} Gain(M, A) &= Ent(M) - \sum_{i \in A} \frac{|M_i|}{|M|} Ent(M_i) \\ &= Ent(M) + \sum_{i \in A} \frac{|M_i|}{|M|} \sum_{s} \frac{|M_i^s|}{|M_i|} \log \frac{|M_i^s|}{|M_i|} \\ &= Ent(M) + \sum_{i \in A} \frac{1}{|M|} \left[\left(\sum_{s} |M_i^s| \log |M_i^s| \right) - |M_i| \log |M_i| \right] \end{aligned}$$

Because $Ent(M)$ and $|M|$ are fixed, maximising the information gain is equivalent to maximising

$$\boxed{\sum_{i \in A} \left(\sum_{s} |M_i^s| \log |M_i^s| \right) - |M_i| \log |M_i|}$$

Then calculate this for each attribute

Weather → $[(5\log 5 + 4\log 4) - 9\log 9 + (8\log 8 + 3\log 3) - 11\log 11]$

$= -12.63$ nats $= -18.22$ bits

Season → $[(3\log 3 + 2\log 2) - 5\log 5 + (4\log 4 + 1\log 1) - 5\log 5 + (4\log 4 + 1\log 1) - 5\log 5 + (2\log 2 + 3\log 3) - 5\log 5] = -11.73$ nats $= -16.92$ bits

TOC → $2\log 2 + 4\log 4 - 6\log 6 + 4\log 4 + 3\log 3 - 7\log 7 + 7\log 7 + 0\log 0 - 7\log 7 = -8.60$ nats $= -12.40$ bits

Day → $8\log 8 + 6\log 6 - 14\log 14 + 5\log 5 + 1\log 1 - 6\log 6$

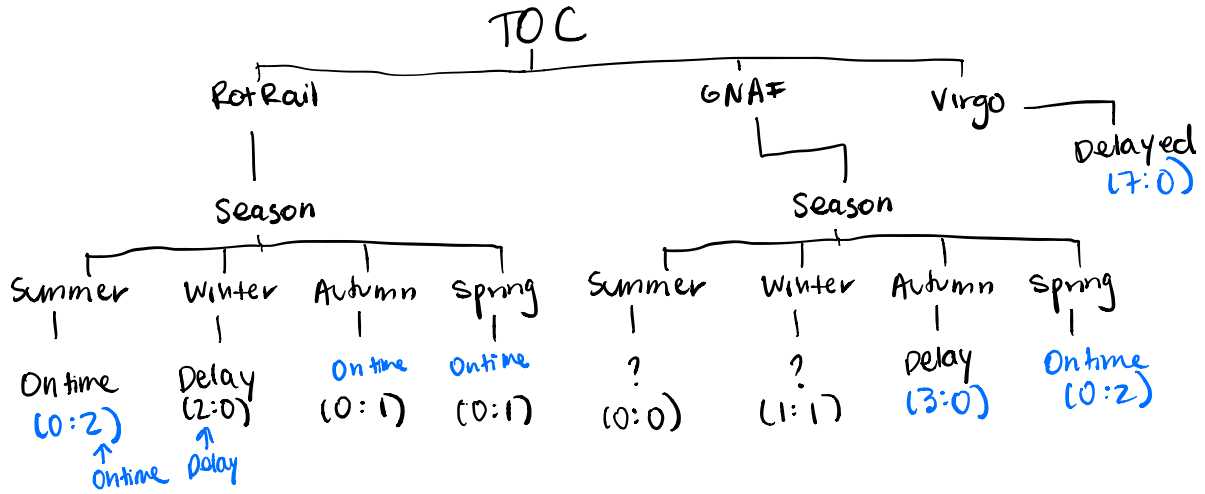
$= -12.26$ nats $= -17.69$ nats

* to get information gain, divide these values by $|M| = 20$ and add the entropy of the split of the whole dataset on the delayed and on time categories (13/20 and 7/20 respectively) = 0.65 nats

The largest information gain comes from choosing to classify according to the train operating company.

The maximum IG 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 the values:



2. Classify from your DT in (1).

	Weather	Season	TOC	Day	
Example 1	Windy	Autumn	RotRail	Weekday	Ontime
Example 2	Calm	Summer	Virgo	Weekday	Delayed
Example 3	Calm	Spring	ONAF	Weekend	Ontime

3. A training set consists of one-dimensional examples from two classes.

Training examples from $x_1 = \{0.5, 0.1, 0.2, 0.4, 0.3, 0.2, 0.2, 0.1, 0.35, 0.25\}$
Training examples from $x_2 = \{0.9, 0.8, 0.75, 1.0\}$

Fit a 1-D Gaussian using Maximum Likelihood to each of these two classes.

$\sigma_1^2 = 0.0149 \quad \sigma_2^2 = 0.0092$

Find $p(y=1 \mid x=0.6)$?

$$\mu_1 = \frac{0.5+0.1+0.2+0.4+0.3+0.2+0.2+0.1+0.35+0.25}{10} = 0.26$$

$$\mu_2 = \frac{0.9+0.8+0.75+1.0}{4} = 0.8625$$

$p_1 = 10/14 = 0.7143 \quad p_2 = 4/14 = 0.2857 \rightarrow$ class prior probabilities

Now, the probability that a point x belongs to class 1 is

$$p(c_1 \mid x) = \frac{p(x \mid c_1)\, p(c_1)}{p(x \mid c_1)p(c_1) + p(x \mid c_2)p(c_2)}$$

where $p(x \mid c_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(\frac{(x-\mu_k)^2}{2\sigma_k^2}\right)$

$$p(x = 0.6 \mid y=1) = \frac{1}{\sqrt{2\pi(0.0149)}} \exp\left(\frac{(0.6-0.26)^2}{2(0.0149)}\right) = 0.0675$$

$$p(x = 0.6 \mid y=2) = \frac{1}{\sqrt{2\pi(0.0092)}} \exp\left(\frac{(0.6-0.8625)^2}{2(0.0092)}\right) = 0.0983$$

$$P(y=1|x) = \frac{0.0675 \times 0.7143}{0.0675 \times 0.7143 + 0.2857 \times 0.0983}$$

$$= 0.6319$$

$$P(y=2|x) = 1 - 0.6319 = 0.3681$$

Probability that $x = 0.6$ belongs to class 1 is 0.63.

* Note that μ_2 is nearer to $x = 0.6$ than $\mu_1 = 0.26$, but $\sigma_1^2 = 0.0149$ is broader than $\sigma_2^2 = 0.0092$.

4. Spam Detection

* expected performance on validation set ~ generalisation error.

  but small validation set ⇒ large variance relative to the true generalisation error.