Assignment 3 - Individual

Machine Learning

MSc Business Analytics 2020/2021

Students should give the mathematical formulae for entropy and information gain at least once in their work. Entropy of root node

$$\begin{aligned} H(\text{root node}) &= -\frac{13}{25}\log_2\left(\frac{13}{25}\right) - \frac{12}{25}\log_2\left(\frac{12}{25}\right) \\ &= \mathbf{0.999}\,(3\,\text{s.f}) \end{aligned} \tag{1}$$

calculations for the first split

If we split by 'Day':
weekday: 12 'no', 8 'yes'
weekend: 1 'no', 4 'yes'

$$\begin{aligned} \text{Information gain} &= (0.9988) \\ &+ \frac{20}{25}\left(\frac{12}{20}\log_2\left(\frac{12}{20}\right) + \frac{8}{20}(\log_2\left(\frac{8}{20}\right)\right) \\ &+ \frac{5}{25}\left(\frac{1}{5}\log_2\left(\frac{1}{5}\right) + \frac{4}{5}(\log_2\left(\frac{4}{5}\right)\right) \\ &= \mathbf{0.0777}\,(3\,\text{s.f}) \end{aligned} \tag{2}$$

If we split by 'Weather':
rainy: 1 'no', 7 'yes'
sunny: 12 'no', 5 'yes'

$$\begin{aligned} \text{Information gain} &= (0.9988) \\ &+ \frac{8}{25}\left(\frac{1}{8}\log_2\left(\frac{1}{8}\right) + \frac{7}{8}(\log_2\left(\frac{7}{8}\right)\right) \\ &+ \frac{17}{25}\left(\frac{12}{17}\log_2\left(\frac{12}{17}\right) + \frac{5}{17}(\log_2\left(\frac{5}{17}\right)\right) \\ &= \mathbf{0.0231}\,(3\,\text{s.f}) \end{aligned} \tag{3}$$

If we split by 'Time':
8 am: 5 'no', 6 'yes'

1 pm: 8 'no', 6 'yes'

$$\begin{aligned}\text{Information gain} &= (0.9988) \\ &+ \frac{11}{25}\left(\frac{5}{11}\log_2\left(\frac{5}{11}\right) + \frac{6}{11}(\log_2\left(\frac{6}{11}\right)\right) \\ &+ \frac{14}{25}\left(\frac{8}{14}\log_2\left(\frac{8}{14}\right) + \frac{6}{14}(\log_2\left(\frac{6}{14}\right)\right) \\ &= \mathbf{0.00974}\,(3\,\text{s.f})\end{aligned} \tag{4}$$

Decision: split by **'Weather'**

calculations for the second split - 'Sunny node'

$$\begin{aligned}H(\text{Sunny node}) &= -\frac{12}{17}\log_2\left(\frac{12}{17}\right) - \frac{5}{17}\log_2\left(\frac{5}{17}\right) \\ &= \mathbf{0.874}\,(3\,\text{s.f})\end{aligned} \tag{5}$$

If we split by 'Day':
weekday: 12 'no', 3 'yes'
weekend: 0 'no', 2 'yes'

$$\begin{aligned}\text{Information gain} &= (0.87398) \\ &+ \frac{15}{17}\left(\frac{12}{15}\log_2\left(\frac{12}{15}\right) + \frac{3}{15}(\log_2\left(\frac{3}{15}\right)\right) \\ &= \mathbf{0.237}\,(3\,\text{s.f})\end{aligned} \tag{6}$$

If we split by 'Time':
8 am: 4 'no', 3 'yes'
1 pm: 8 'no', 2 'yes'

$$\begin{aligned}\text{Information gain} &= (0.87398) \\ &+ \frac{7}{17}\left(\frac{4}{7}\log_2\left(\frac{4}{7}\right) + \frac{3}{7}(\log_2\left(\frac{3}{7}\right)\right) \\ &+ \frac{10}{17}\left(\frac{8}{10}\log_2\left(\frac{8}{10}\right) + \frac{2}{10}(\log_2\left(\frac{2}{10}\right)\right) \\ &= \mathbf{0.0436}\,(3\,\text{s.f})\end{aligned} \tag{7}$$

Decision: split Sunny node by **'Day'**

calculations for the second split - 'Rainy node'

$$\begin{aligned}H(\text{Rainy node}) &= -\frac{1}{8}\log_2\left(\frac{1}{8}\right) - \frac{7}{8}\log_2\left(\frac{7}{8}\right) \\ &= \mathbf{0.544}\,(3\,\text{s.f})\end{aligned} \tag{8}$$

If we split by 'Day':
weekday: 0 'no', 5 'yes'
weekend: 1 'no', 2 'yes'

$$\begin{aligned}\text{Information gain} &= (0.5436) \\ &\quad + \frac{3}{8}\left(\frac{1}{3}\log_2\left(\frac{1}{3}\right) + \frac{2}{3}(\log_2\left(\frac{2}{3}\right)\right) \\ &= \mathbf{0.199}\,(3\,\text{s.f})\end{aligned} \tag{9}$$

If we split by 'Time':
8 am: 1 'no', 3 'yes'
1 pm: 0 'no', 4 'yes'

$$\begin{aligned}\text{Information gain} &= (0.5436) \\ &\quad + \frac{4}{8}\left(\frac{1}{4}\log_2\left(\frac{1}{4}\right) + \frac{3}{4}(\log_2\left(\frac{3}{4}\right)\right) \\ &= \mathbf{0.138}\,(3\,\text{s.f})\end{aligned} \tag{10}$$

Decision: split Rainy node by **'Day'**

Decisions for third split
Do not split 'Sunny-weekend' further as maximum purity reached.
Do not split 'Rainy-weekday' further as maximum purity reached.
Split 'Sunny-weekday' by the last available variable, time
Split 'Rainy-weekend' by the last available variable, time

Applying tree to trainset
Produce Table 1.
Students should be careful of using a majority rule in this context. Exercising good statistical judgement is key to creating a good model. This exercise is a test of students' logical reasoning abilities. Here 2 in 25 needs to be classified randomly (8%). In the test set, this is 2 in 15 or 13%. It is clear that a majority rule will not result in a reliable solution and may be over-conservative for a traffic situation. In these circumstances, it is best to decouple the uncertainties from the model. Especially, it becomes important to see which part of prediction (TP, TN, FP, FN) is being affected the most by the model's uncertainty limitations. Students are free to quantify the errors or uncertainties in any reasonable way, a minimum/maximum uncertainty method (simple counting method) is given here. Part marks are awarded to students who apply a majority rule without realising the implications this has in the context of the question, especially for a traffic situation.
On the training set, we obtain 12 TNs, 8 TPs, 0 FPs, 3 FNs, and 2 cases contributing to stochastic errors. The confusion matrix is given as shown in Table 2. Part marks are awarded for values falling in the uncertainty range.

Day	Weather	Time	Traffic	Prediction	Class
weekday	sunny	1 pm	N	N	TN
weekday	rainy	1 pm	Y	Y	TP
weekday	sunny	8 am	N	N	TN
weekday	sunny	1 pm	N	N	TN
weekday	rainy	1 pm	Y	Y	TP
weekday	sunny	8 am	N	N	TN
weekend	rainy	8 am	Y	Y/N	*
weekend	sunny	1 pm	Y	Y	TP
weekday	sunny	8 am	Y	N	FN
weekday	sunny	1 pm	N	N	TN
weekday	sunny	1 pm	N	N	TN
weekend	rainy	1 pm	Y	Y	TP
weekday	rainy	1 pm	Y	Y	TP
weekday	sunny	8 am	N	N	TN
weekday	sunny	1 pm	N	N	TN
weekend	sunny	1 pm	Y	Y	TP
weekday	rainy	8 am	Y	Y	TP
weekday	sunny	1 pm	N	N	TN
weekday	sunny	8 am	N	N	TN
weekday	sunny	1 pm	N	N	TN
weekday	sunny	8 am	Y	N	FN
weekend	rainy	8 am	N	Y/N	*
weekday	rainy	1 pm	N	N	TN
weekday	rainy	8 am	Y	Y	TP
weekday	sunny	8 pm	Y	N	FN

Table 1: Use of tree on train set

The confusion matrix now allows us to see that there is one unsure prediction affecting all TP, TN, FP, FN equally.

	Y_{actual}	N_{actual}
$Y_{predicted}$	8^{+1}_{-0}	0^{+1}_{-0}
$N_{predicted}$	3^{+1}_{-0}	12^{+1}_{-0}

Table 2: confusion matrix train set

$$\text{misclassification} = \frac{\text{wrongly classified instances}}{\text{total instances}} = \frac{3^{+1}_{-0} + 0^{+1}_{-0}}{25} = \frac{3^{+2}_{-0}}{25} \qquad (11)$$

It is now clear that with a majority rule, we may make an error of $(5-3=2)/3$ = 67% on the the number of instances misclassified. This figure cannot be neglected and is a fact that needs to be known to people using the model.

$$\text{specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}} = \frac{12^{+1}_{-0}}{12^{+1}_{-0}+0^{+1}_{-0}} = \frac{12^{+1}_{-0}}{12^{+2}_{-0}} \tag{12}$$

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}} = \frac{8^{+1}_{-0}}{8^{+1}_{-0}+3^{+1}_{-0}} = \frac{8^{+1}_{-0}}{11^{+2}_{-0}} \tag{13}$$

Provided mathematical formulae (for misclassification, sensitivity and specificity) and workings are provided, part marks may be awarded for reasonable values in the range of the uncertainties.

Applying tree to test set
Produce Table 3.

Day	Weather	Time	Traffic	Prediction	Class
weekend	rainy	8 am	N	Y/N	*
weekday	sunny	8 am	Y	N	FN
weekend	sunny	1 pm	Y	Y	TP
weekday	sunny	8 am	N	N	TN
weekend	sunny	1 pm	Y	Y	TP
weekday	rainy	8 am	N	Y	FP
weekday	sunny	8 am	Y	N	FN
weekday	sunny	1 pm	Y	N	FN
weekday	sunny	1 pm	N	N	TN
weekday	sunny	1 pm	N	N	TN
weekday	rainy	8 am	Y	*	*
weekend	sunny	8 am	Y	N	FN
weekday	sunny	1 pm	N	N	TN
weekday	rainy	1 pm	Y	Y	TP
weekday	sunny	1 pm	N	N	TN

Table 3: Use of tree on test set

	Y_{actual}	N_{actual}
$Y_{predicted}$	3^{+1}_{-0}	1^{+1}_{-0}
$N_{predicted}$	4^{+1}_{-0}	5^{+1}_{-0}

Table 4: confusion matrix train set

$$\text{misclassification} = \frac{\text{wrongly classified instances}}{\text{total instances}} = \frac{4^{+1}_{-0} + 1^{+1}_{-0}}{15} = \frac{5^{+2}_{-0}}{15} \tag{14}$$

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} = \frac{5^{+1}_{-0}}{5^{+1}_{-0} + 1^{+1}_{-0}} = \frac{5^{+1}_{-0}}{6^{+2}_{-0}} \tag{15}$$

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{3^{+1}_{-0}}{8^{+1}_{-0} + 4^{+1}_{-0}} = \frac{3^{+1}_{-0}}{7^{+2}_{-0}} \tag{16}$$