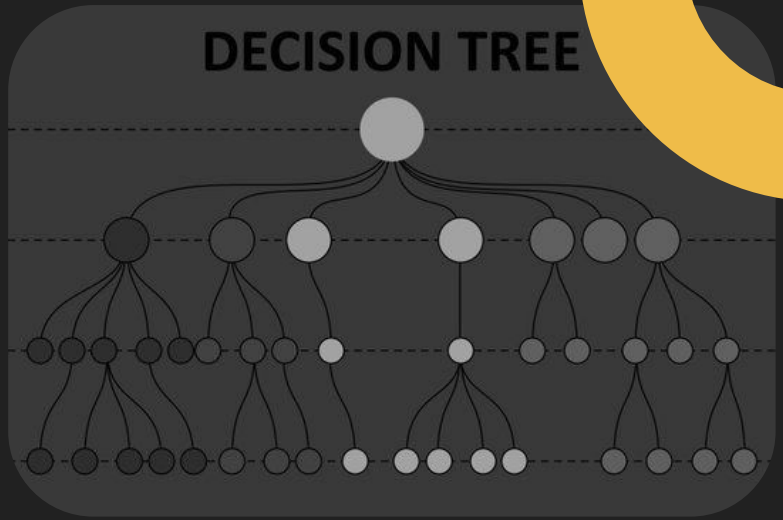
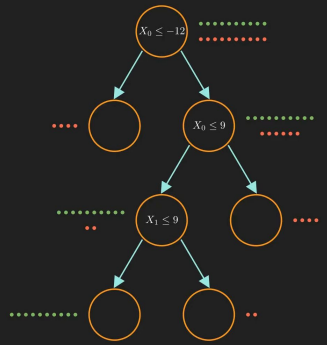
MACHINE LEARNING LAB 1





David O'Leary
&
Cillian Smith

ASSIGNMENT 0

Our initial assumptions about difficulty

MONK-1	**Easiest**	Only considers three attributes, which appear to have the simplest relationships.
MONK-2	**Hardest**	Considers all 6 attributes.
MONK-3	**Intermediate**	More complex than MONK-1 but still only considers 3 attributes.

ASSIGNMENT 1

- Entropy is a measure of uncertainty
- Information about predictability of a dataset

$$\text{Entropy}(S) = -\sum_i p_i \log_2 p_i$$

Dataset	Entropy
MONK-1	1.0
MONK-2	0.957117428264771
MONK-3	0.9998061328047111

ASSIGNMENT 2

Entropy, being a measure of uncertainty, is maximised when a distribution is uniform (all outcome are equally likely).

For a non-uniform distribution, where some outcomes are more likely to occur, the uncertainty in the outcome is decreased, resulting in a decreased entropy.

Fair Die

Unfair Die

$$Entropy = -\sum_i p_i log_2 p_i$$

P(6) = 0.5,
P(1,...,5) = 0.1

$$= -6 \cdot \frac{1}{6} \log_2 \frac{1}{6}$$

$$= 2.58 bits$$

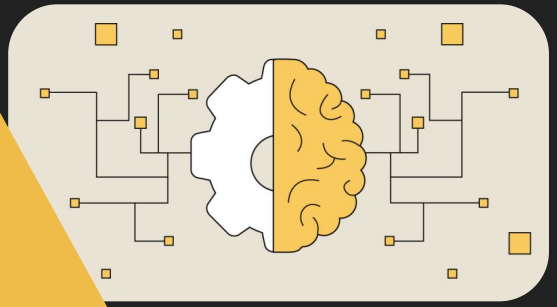
$$= -5 \cdot 0.1 \log_2 0.1 - 0.5 \cdot \log_2 0.5$$

$$= 2.16 bits$$

Entropy is maximised for a fair die, as the outcome is fair or completely random.

For an unfair die, there is less uncertainty about its outcome, as some result will be more likely than others.

ASSIGNMENT 3



- The average gain for each attribute was calculated for each MONK dataset.
- The larger the number, the higher the information gain.

Dataset	*a1*	*a2*	*a3*	*a4*	*a5*	*a6*
MONK-1	0.075272556	0.00583843	0.00470757	0.0263117	**0.28703075**	0.00075786
MONK-2	0.003756177	0.0024585	0.00105615	0.01566425	**0.01727718**	0.00624762
MONK-3	0.007120868	**0.29373617**	0.00083111	0.00289182	0.25591172	0.00707703

- Choose the attribute with the largest value for each dataset: a5 for MONK-1 and MONK-2; a2 for MONK-3

ASSIGNMENT 4

Dataset	*a1*	*a2*	*a3*	*a4*	*a5*	*a6*
MONK-1	0.075272556	0.00583843	0.00470757	0.0263117	**0.28703075**	0.00075786
MONK-2	0.003756177	0.0024585	0.00105615	0.01566425	**0.01727718**	0.00624762
MONK-3	0.007120868	**0.29373617**	0.00083111	0.00289182	0.25591172	0.00707703

When information gain is maximised, entropy will be as low as possible.

What is the motivation for using information gain as a heuristic for picking an attribute for splitting?

Measures the reduction in entropy of the system after splitting.

Maximising information gain means increasing predictivity of subsets.

Subsets are now more useful for classifying new data.

Entropy reduction implies highly homogenous subsets.

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{k \in \text{values}(A)} \frac{|S_k|}{|S|} \text{Entropy}(S_k)$$

ASSIGNMENT 5

Initial assumptions:

MONK-1 would be the least difficult for the decision tree to learn and MONK-2 would be the most difficult.

In reality:

MONK-3 was the easiest for the decision tree to learn and MONK-2 was the hardest.

	E_{train}	E_{test}
MONK-1	0.0	0.17129629629629628
MONK-2	0.0	0.30787037037037035
MONK-3	0.0	0.05555555555555558

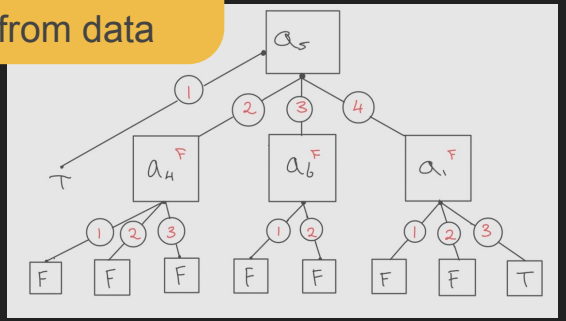
ASSIGNMENT 5

Attributes with the best information gain for a5 for the MONK-1 dataset

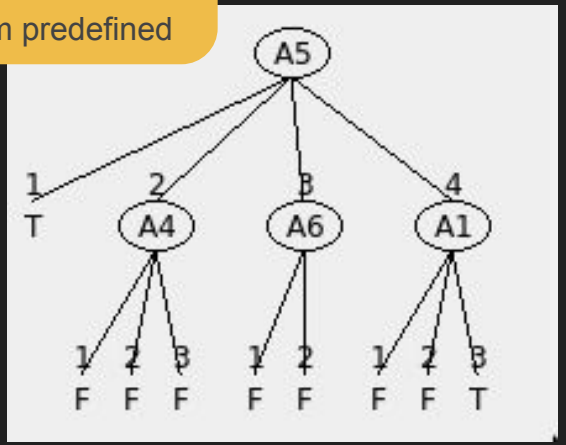
Value of a_5	Attribute with highest gain	Information gain
$a_5 = 1$	N/A (as a_5 results in True)	0.0
$a_5 = 2$	a_4	0.048892202
$a_5 = 3$	a_6	0.045108537
$a_5 = 4$	a_1	0.206290746

Using the most common function the majority classes for the subsets are shown

Drawn from data



Drawn from predefined



ASSIGNMENT 6

- Pruning is a technique used to reduce overfitting.
- In terms of bias-variance, it allows us to:
  - It reduces variance.
  - It increases bias.

It does so by removing redundant parameters that may cause deviations in the accuracy of the model.

However, over-pruning may lead to underfitting, causing an increase in the bias of the model.

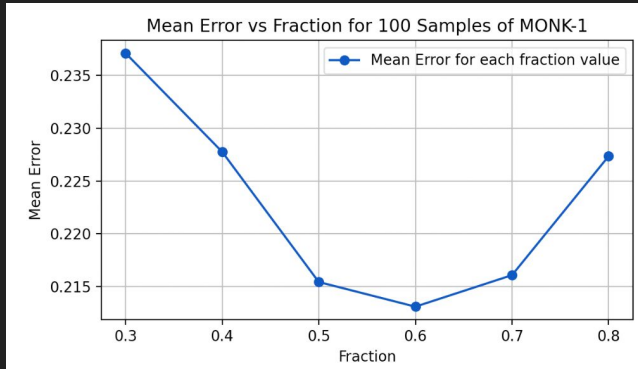
The goal of pruning is to find a balance between bias and variance to attain optimal performance of the model.

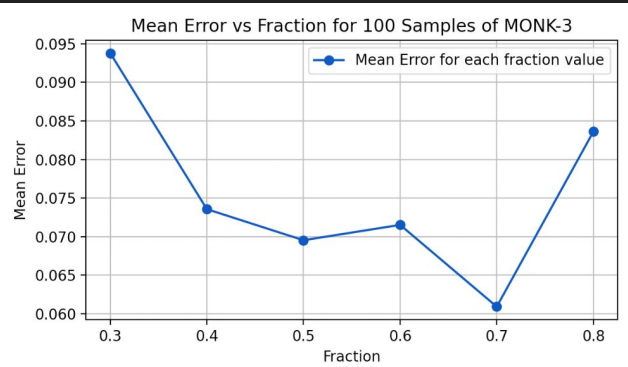
ASSIGNMENT 7

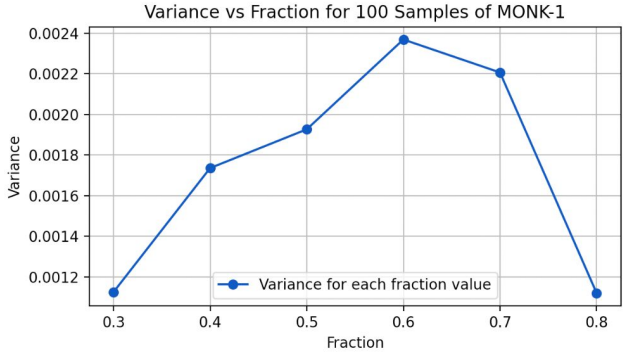
Optimised values of fraction:

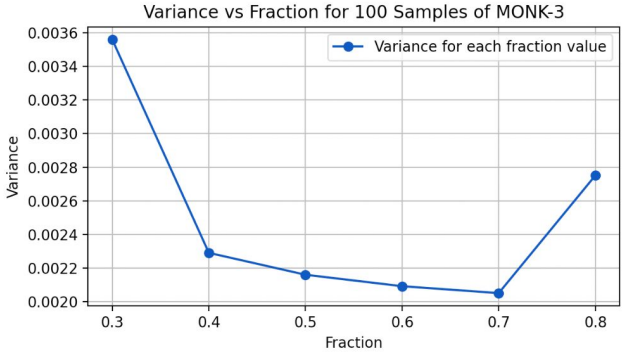
MONK-1: 0.6

MONK-3: 0.7









Mean Errors	0.3	0.4	0.5	0.6	0.7	0.8
MONK-1	0.237129629629629	0.227777777777777	0.215439814814814	0.213101851851851	0.216087962962962	0.227361111111110
MONK-3	0.093796296296296	0.073564814814814	0.069537037037037	0.071527777777777	0.060925925925925	0.083611111111111

Variance	0.3	0.4	0.5	0.6	0.7	0.8
MONK-1	0.00112518359174738	0.00173546161200482	0.00192611720341273	0.0023681625064084	0.0022059560583545	0.00112023224356043
MONK-3	0.00355995656149977	0.00229052138324257	0.00216022536752989	0.0020919824964321	0.0020509969378282	0.00275120027434842

THANK YOU FOR LISTENING!