Lab 1
Decision Trees

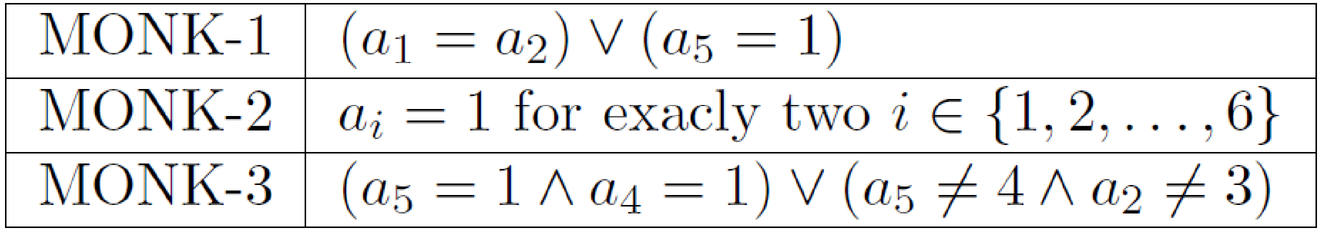
Annika Grothaus & Bastien Massion

17-09-2021

DD2421 – Machine Learning

Assignment 0

Each one of the datasets has properties which makes them hard to learn. Motivate which of the three problems is most difficult for a decision tree algorithm to learn.



MONK-1	$(a_1 = a_2) \vee (a_5 = 1)$
MONK-2	$a_i = 1$ for exacly two $i \in \{1, 2, \ldots, 6\}$
MONK-3	$(a_5 = 1 \wedge a_4 = 1) \vee (a_5 \neq 4 \wedge a_2 \neq 3)$

- MONK-2 is the most difficult one
  - In the form above, boolean statement depends on an unknown combination of attributes and not specific elements to test
  - Can be rewritten in a closed form as a boolean statement with 14 "or" (15 = C^6_2 possibilities/substatements) and each of them have 6 inner conditions (5 "and")

Assignment 1

The file dtree.py defines a function entropy which calculates the entropy of a dataset. Import this file along with the monks datasets and use it to calculate the entropy of the training datasets.

Dataset	Entropy
MONK-1	1.0
MONK-2	~0.9571
MONK-3	~0.9998

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

Assignment 2

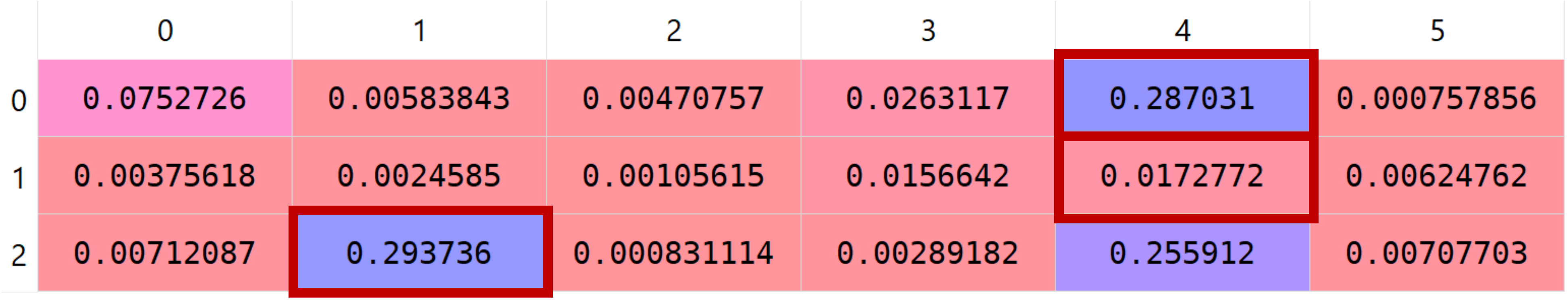
Explain entropy for a uniform distribution and a non-uniform distribution, present some example distributions with high and low entropy.

- Uniform distribution
  - Entropy is maximal : maximal uncertainty/unpredictability
- Non-uniform distribution
  - Entropy is smaller : less uncertainty/unpredictability

- Examples of High and Low Entropy
  - Which number/sign in a card deck is drawn?
    - High entropy: Original card deck
    - Low Entropy: Card deck, which has most numbered cards removed (and known)

Assignment 3

Use the function averageGain to calculate the expected information gain corresponding to each of the six attributes. Based on the results, which attribute should be used for splitting the examples at the root node?



	0	1	2	3	4	5
0	0.0752726	0.00583843	0.00470757	0.0263117	0.287031	0.000757856
1	0.00375618	0.0024585	0.00105615	0.0156642	0.0172772	0.00624762
2	0.00712087	0.293736	0.000831114	0.00289182	0.255912	0.00707703

- For MONK-1, we should use a_5 as root node
- For MONK-2, we should use a_5 as root node
- For MONK-3, we should use a_2 as root node

Assignment 4

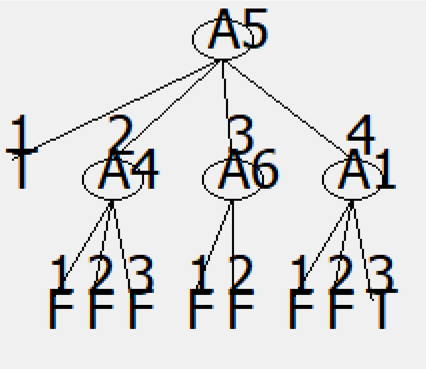
For splitting we choose the attribute that maximizes the information gain. Looking at Eq.3 how does the entropy of the subsets, Sk, look like when the information gain is maximized? How can we motivate using the information gain as a heuristic for picking an attribute for splitting?

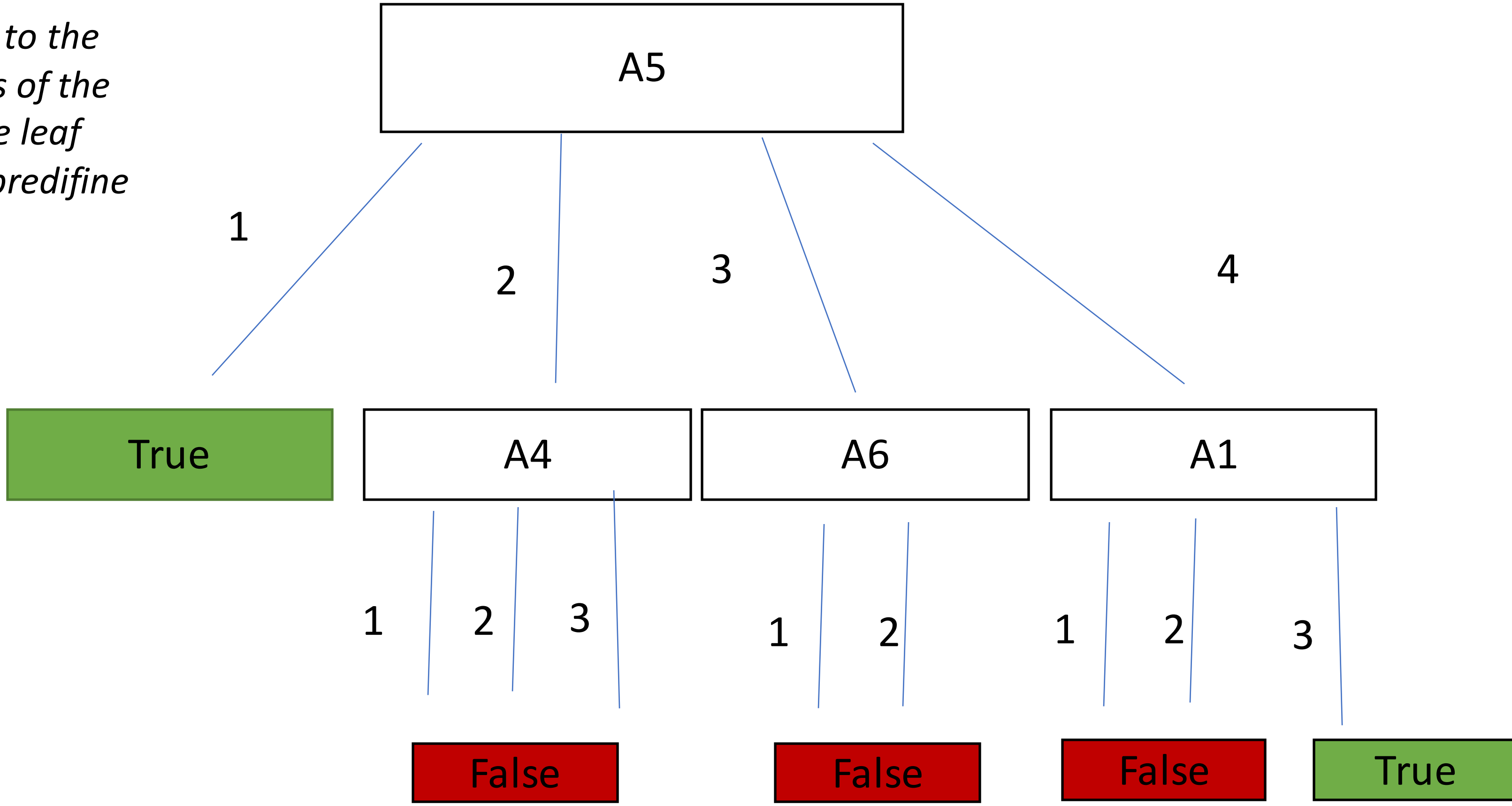
$$\mathrm{Gain}(S, A) = \mathrm{Entropy}(S) - \sum_{k \in \mathrm{values}(A)} \frac{|S_k|}{|S|} \mathrm{Entropy}(S_k)$$

- For the highest information gain we need to minimize the weighted average of the entropy of the subsets (= expected entropy after splitting)
  - We pay respect to the number of samples in S_k
  - When the split results in subset(s) with a lot of elements the entropy of these subset(s) should be low
  - When the split results in subset(s) with few elements the entropy of these subset(s) can be higher without having too much impact on the information gain
- The higher the information gain, the lower the expected new entropy, the lower the new expected uncertainty ➤ higher predictability

Assignment 5

For the monk1 *data draw the decision tree up to the first two levels and assign the majority of class of the subsets that resulted from the two splits to the leaf nodes. Compare your results with that of the predifine routine for ID3.*



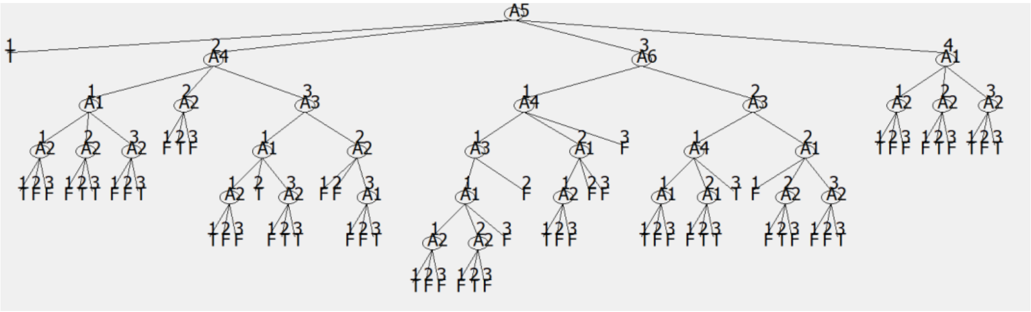


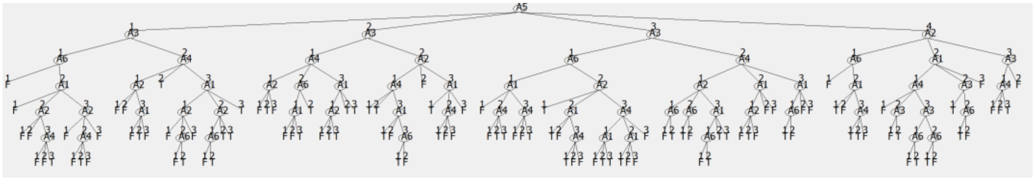
Assignment 5

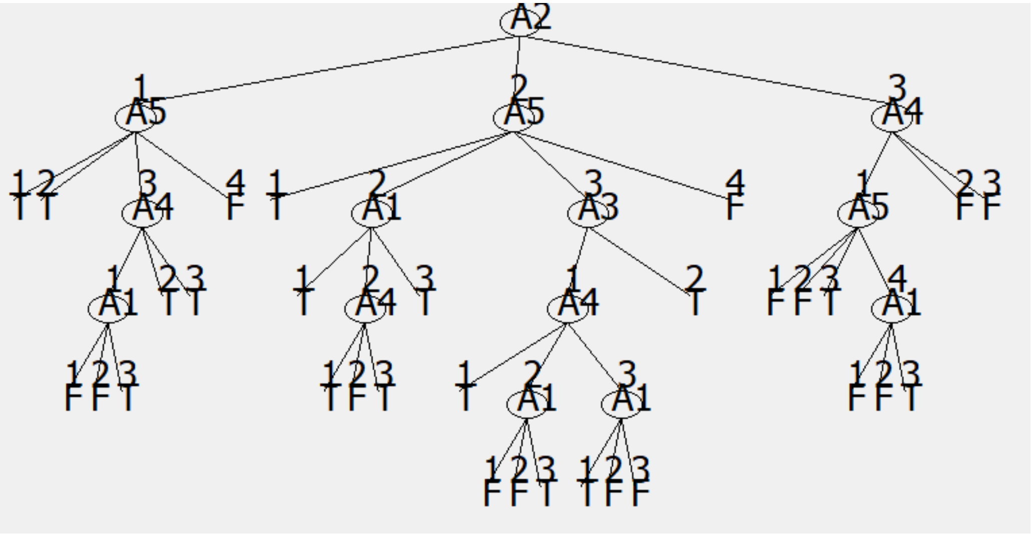
Compute the train and test set errors for the three Monk datasets for the full trees. Were your assumptions about the datasets correct?

Dataset	Training accuracy	Testing accuracy	Accuracy on new data	Number of leaves
MONK-1	1.0	0.8287	0.7597	~62
MONK-2	1.0	0.6921	0.4942	~107
MONK-3	1.0	0.9444	0.9225	~31

- Monk-2 shows the smallest accuracy when classifying samples from the testing set ➤Hard to learn : assumption correct
- 100% accuracy on the training set: logical because we stop the growing of the full tree when we reach perfect labelling of training set



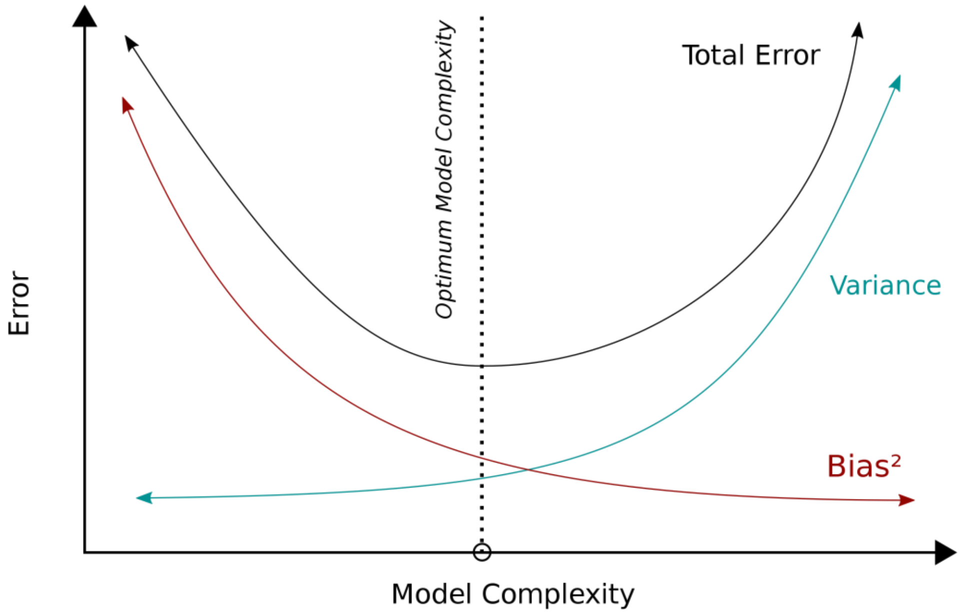




Assignment 6

Explain pruning from a bias variance trade-off perspective.

- Non-pruned tree is a "too complex" model
  - Low bias but high variance (overfitting)
- Pruning the tree simplifies the model
  - increases the bias but decreases the variance: bias-variance trade-off
- Note: minimizing error in least-square sense is equal to minimize the variance + the bias squared



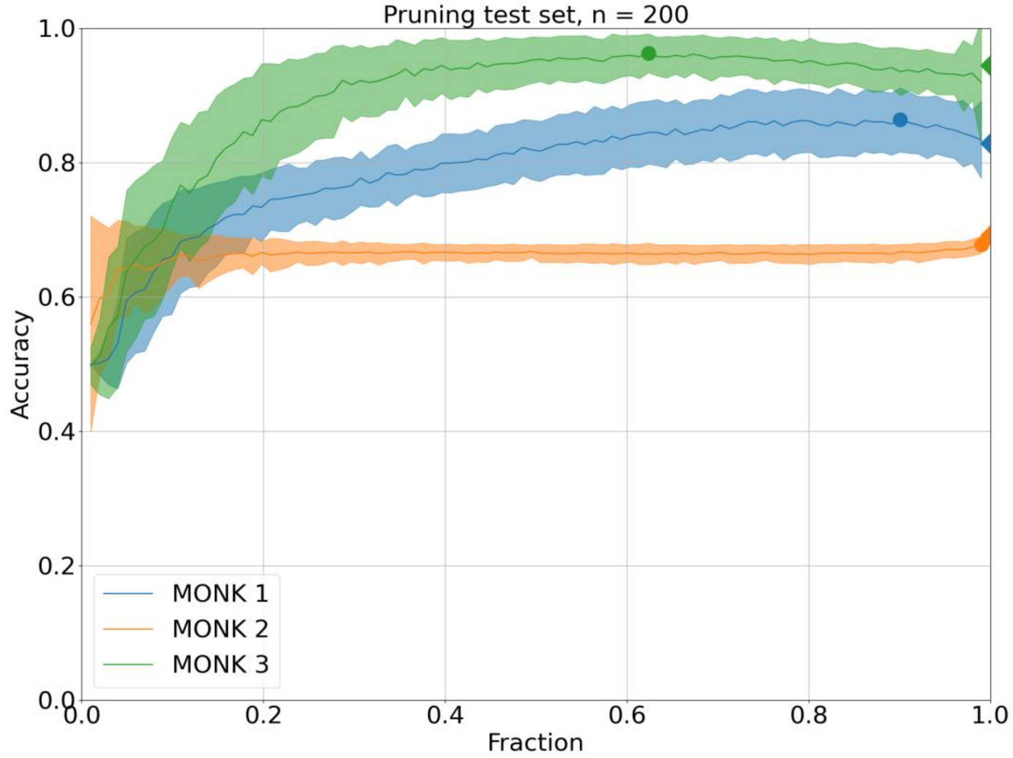
$$\mathrm{E}_D\left[\left(y-\hat{f}\left(x;D\right)\right)^2\right] = \left(\mathrm{Bias}_D\left[\hat{f}\left(x;D\right)\right]\right)^2 + \mathrm{Var}_D\left[\hat{f}\left(x;D\right)\right] + \sigma^2$$

Assignment 7

Evaluate the effect pruning has on the test error for the monk1 and monk3 datasets, in particular determine the optimal partition into training and pruning by optimizing the parameter fraction. Plot the classification error on the test sets as a function of the parameter fraction.

- Pruning can improve the accuracy for MONK-1 and MONK-3 if the parameter *fraction* is well-chosen

Dataset	Accuracy without pruning	Max mean accuracy with pruning	Optimal fraction	Good range for fraction
MONK-1	0.8287	0.8637	0.90	[0.7,0.95]
MONK-2	0.6921	0.6777	"1.0"	"1.0"
MONK-3	0.9444	0.9629	0.62	[0.5,0.7]


Pruning test set, n = 200
Accuracy
Fraction
MONK 1
MONK 2
MONK 3