DD2421 Machine Learning

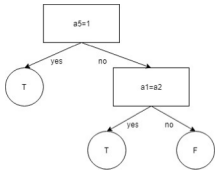
Lab 1: Decision Trees

Daiki Shirafuji
daikish@kth.se
Takuya Nishi
nishi@kth.se

1

Assignment 0: Each one of the datasets has properties which makes hard to learn. Motivate which of the three problems is most dicult for a decision tree algorithm to learn.

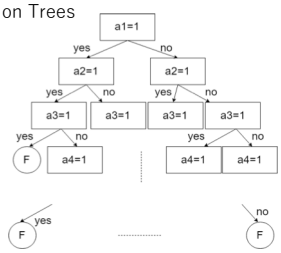
- MONK1 Decision Trees



2

Assignment 0: Each one of the datasets has properties which makes hard to learn. Motivate which of the three problems is most dicult for a decision tree algorithm to learn.

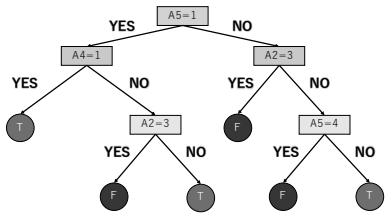
- MONK2 Decision Trees



3

Assignment 0: Each one of the datasets has properties which makes hard to learn. Motivate which of the three problems is most dicult for a decision tree algorithm to learn.

- MONK3 Decision Trees



4

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.

$$\text{Entropy} = \sum_i -p_i \log_2 p_i$$

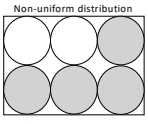
Dataset	Entropy
MONK-1	1.0
MONK-2	0.9571
MONK-3	0.9998

5

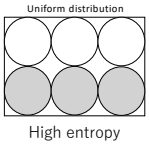
Assignment 2: Explain entropy for a uniform distribution and a non-uniform distribution, present some example distributions with high and low entropy.

Example of entropy:

$\text{Ent(B)} = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} \cong 0.92$ < $\text{Ent(B)} = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$



Low entropy



High entropy

The model of a higher entropy is more unpredictable

6

Assignment 3: Use the function averageGain (defined in dtree.py) 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?

In Monk1, a5 should be used for splitting the examples at the root node.
In Monk2, a5 should be used for splitting the examples at the root node.
In Monk3, a2 should be used for splitting the examples at the root node.

Dataset	a1	a2	a3	a4	a5	a6
MONK-1	0.07527	0.005838	0.004708	0.02631	**0.2870**	0.0007579
MONK-2	0.003756	0.002458	0.001056	0.01566	**0.01728**	0.006248
MONK-3	0.007121	**0.2937**	0.0008311	0.002892	0.2559	0.007077

7

Assignment 4: For splitting we choose the attribute that maximizes the information gain, Eq.3. 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? Think about reduction in entropy after the split and what the entropy implies.

- When the information gain is maximized, **the entropy of the subsets is minimized** because of the following formula.

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

- **The highest information gain** in all attributes is the most useful for estimating the appropriate model.

8

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? Explain the results you get for the training and test datasets.

- Error rates are shown in the following table.
- OUR Assumption: MONK-1 is the easiest and MONK-2 is the most difficult.
  - **MONK-2 is certainty the most difficult dataset.** Only this dataset resulted in over **30%** error rate.
  - MONK-1 did **NOT** result in the best score, but **MONK-3** did. The reason why MONK-1 resulted bad score on test dataset may be the overfitting.

	ERROR for training dataset	ERROR for test dataset
MONK-1	0.0	0.1713
MONK-2	0.0	0.3079
MONK-3	0.0	0.05556

9

Assignment 6: Explain pruning from a bias variance trade-off perspective.

- Our Decision Trees Model:
  Can classify training dataset perfectly, but does not work well on test dataset (b.c. **Low Bias & Higher Variance**).
  -> Model may be overfitting
- Therefore, we have to avoid overfitting with "Pruning". Pruning is to identify and remove the irrelevant branches, which may be outliers in order to increase the accuracy.
- After pruning the decision trees, **the complexity of a model become decreasing, so the model get higher Bias and lower Variance**.
- However, if we remove branches a lot, the accuracy will decrease.

10

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 validation by optimizing the parameter fraction. Plot the classification error on the test sets as a function of the parameter fraction ∈ {0.3, 0.4, 0.5, 0.6, 0.7, 0.8}.

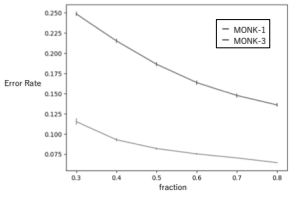
- Generally, it can be said that the error rate of the model with pruning is improved somehow because the model can avoid overfitting.
- The results are shown in the following table (the number of runs is 3,000).
- MONK-1: the error rate certainty ***decreases*** by about 4%.
- However, the error rate ***increases*** by about 4% on MONK-2, the error rate ***increases*** about 1%, and on MONK-3 because the model is not overfitted, thus this case pruned necessary branches.

	ERROR without pruning	ERROR with pruning
MONK-1	0.1713	**0.1354** (best fraction=0.8)
MONK-2	0.3079	0.3466 (best fraction=0.8)
MONK-3	0.05556	**0.06729** (best fraction=0.8)

11

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 validation by optimizing the parameter fraction. Plot the classification error on the test sets as a function of the parameter fraction ∈ {0.3, 0.4, 0.5, 0.6, 0.7, 0.8}.

- The mean error rate improves as the parameter fraction increases.
- The variance is not so large (even the largest one is just 0.003856).



12

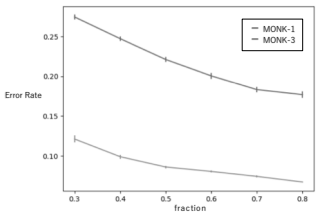
APPENDIX: Test dataset without training dataset

- In this assignments, the test datasets include the training dataset, thus we separate the training data from the test datasets.
- Error rates with/without pruning are shown in the following table.

	ERROR without pruning	ERROR with pruning
MONK-1	0.2403	0.1770 (fraction=0.8)
MONK-2	0.5057	0.4534 (fraction=0.8)
MONK-3	0.05807	0.06729 (fraction=0.8)

13

APPENDIX: Test dataset without training dataset



14