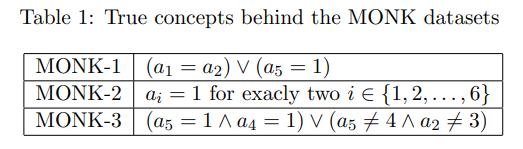
Lab1

Zhicheng Dong [zdo@kth.se](mailto:zdo@kth.se)

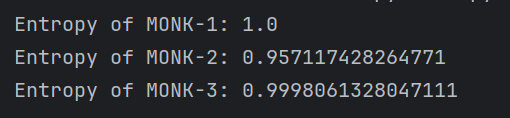
Qiaolan Luo [qiaolan@kth.se](mailto:qiaolan@kth.se)

**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.**



A decision tree checks the presence or absence of attributes and is better suited to partitioning data based on the presence of certain attribute values. Therefore, for MONK-1 and MONK-3, it can better deal with AND condition and OR condition. However, MONK-2 requires counting occurrences across multiple attributes, which makes MONK-2 the most difficult for a decision tree to learn.

**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.**



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

An entropy is a measurement of the unpredictability and is an expectation to the quantity of information under a given probability distribution. The lower entropy is, the uncertainty is lower and we are more likely to get certain information. According to Shannon’s formula, for a discrete random variable X, it’s entropy is:

Pi represents the probability of an outcome xi.

1. **Uniform distribution**

In a uniform distribution, the probability of each event is equal. The entropy will be:

Usually in a uniform distribution, the entropy is high because every event has the same probability and the uncertainty is at the largest.

Example:

Rolling a fair six-sided dice, each has an equal probability of 1/6. So, the entropy will be log6.

Flipping a fair coin, each side has an equal probability of ½. So, the entropy will be log2 = 1.

1. **Non-uniform distribution**

In a non-uniform distribution, the probabilities of events are not equal and some events are more likely to happen. So, the entropy will be lower compared to a uniform distribution of the same number of outcomes, because there is less uncertainty.

Example:

Suppose there is a biased six-sided dice, with the probability of rolling a 6 is 0.5, and the probabilities for 1, 2, 3, 4, and 5 are 0.1 each. The entropy now is 2.16, lower than a fair dice (log6≈2.585).

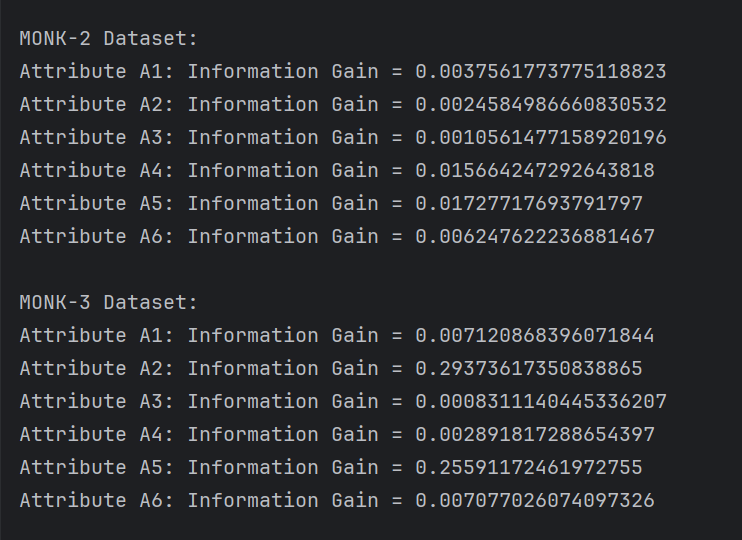
Consider a biased coin that comes up heads 95% of the time and tails 5% of the time. The entropy is 0.2864, much lower than a fair coin (1), because with a high probability of 95%, the uncertainty is much lower.

**Assignment 3: Use the function averageGain (defined in dtree.py) to calculate the expected information gain corresponding to each of the six attributes. Note that the attributes are represented as instances of the class Attribute (defined in monkdata.py) which you can access via m.attributes[0], ..., m.attributes[5]. Based on the results, which attribute should be used for splitting the examples at the root node?**

电脑屏幕的照片上有文字

描述已自动生成电脑屏幕的照片上有文字

中度可信度描述已自动生成

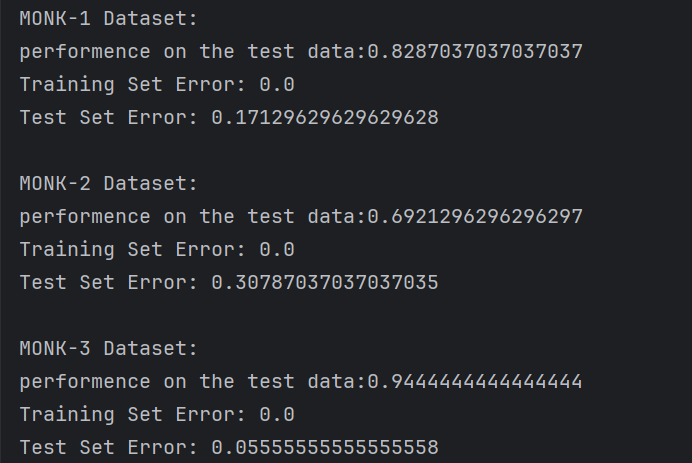


MONK-1: A5 MONK-2: A5 MONK-3: A2

**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 information gain is maximized, the entropy of the subsets is usually lower, which means the uncertainty in the subsets is lower, and the data tends to have the same classification label. Using information gain as a heuristic for splitting is motivated by the goal of reducing uncertainty or randomness in the data. By choosing an attribute that maximizes information gain, the decision tree can create subsets of data that are more homogeneous or pure in terms of the target variable. This means that after the split, the subsets are more organized and have a clearer structure, which is beneficial for building an effective and efficient decision tree.

**Assignment5: 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.**



In all three datasets, the decision trees achieved perfect classification on the training set (0% training error), indicating that the trees completely capture the patterns in the training data.

However, their performance on the test set varies significantly. Particularly, the performance on the MONK-2 dataset suggests that the model may be overly adapted to specific features of the training data and does not generalize well to new, unseen data.

MONK1：

许多的地图

描述已自动生成

MONK2：

一群鸟飞在空中

描述已自动生成

MONK3：

图片包含 规模, 线, 挂, 一群

描述已自动生成

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

Bias and variance are two sources of errors that affect the performance of a predictive model.

A decision tree with no pruning (fully grown) tends to have low bias because it can capture intricate details in the training data. However, it often has high variance because it can become overly complex and fit the noise in the training data. Pruning simplifies the decision tree by removing branches that do not contribute significantly to the model's performance on the validation or test data. This simplification reduces model complexity and, in turn, decreases variance.

In summary, pruning in decision trees is a strategy to find an optimal level of model complexity that minimizes the total prediction error by balancing bias and variance. It helps prevent overfitting (high variance) and ensures that the model captures meaningful patterns in the data (low bias).

**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 ∈ {0.3, 0.4, 0.5, 0.6, 0.7, 0.8}.**

**图表, 折线图

描述已自动生成**