Assignment Sheet 4

Assignment 4.1

Explain in your own words how a decision tree looks like, how a decision tree is used for classification, and how a decision tree is learned. Try to restrict to the most important points and be as clear as possible.

Assignment 4.2

Explain how a possible decision tree for the following problem description may look like. Discuss on choices for features in the decision tree and possible challenges.
Children should be classified based on their behavior into three classes of Good, Neutral and Bad. Each child is evaluated daily on the following points: how many **chores** they fulfilled (All, Most, Few, None), how **noisy** they are (a grade between 1 and 10), whether they made a **mess** (Yes or No) and how they **treated other** children (Nice, Mean, No Interaction). Additionally, for each child you have information about them, meaning their name, birth date, supervisor and the income of the parents.
It has been found out, that the classification of a child mostly depends on their age, their noisiness and how they treated others. Additionally, some supervisors grade children worse than others.

Assignment 4.3

Given are the following training examples for the target concept "Enjoy Sport" using attributes: Sky (Sunny, Cloudy, Rainy), Temperature (Warm, Cold), Humidity (High, Normal), Wind (Strong, Weak), Water (Warm, Cold) and Forecast (Same, Change).

$$
\begin{aligned}
x_1 &= \langle \texttt{Sunny, Warm, Normal, Strong, Warm, Same} \rangle & y_1 &= + \\
x_2 &= \langle \texttt{Sunny, Warm, High, Strong, Warm, Same} \rangle & y_2 &= + \\
x_3 &= \langle \texttt{Rainy, Cold, High, Strong, Warm, Change} \rangle & y_3 &= - \\
x_4 &= \langle \texttt{Sunny, Warm, High, Strong, Cool, Change} \rangle & y_4 &= +
\end{aligned}
$$

Build a decision tree using the ID3 algorithm. Write down all Information Gain values. Compare your result tree with the version space.

Assignment 4.4

Shortly discuss overfitting for decision trees. Especially consider how decision trees might overfit and how it can be potentially prevented. Relate your discussion to the terms of Bias and Variance! Additionally, read up on **two** alternative methods for deciding on feature splits for (classification) decision trees.