Decision Trees

Applied ML in Engineering - Exercise 06

TU Berlin, Summer Term 2023

Prof. Dr.-Ing. Merten Stender – merten.stender@tu-berlin.de

Students are asked to implement a basic decision tree model from scratch using object-oriented programming in Python. The exemplary data set, also used in the lecture slides, is provided in the file `decision_tree_dataset.txt` where the first two columns provide the coordinates and the last column provides the targets (0 and 1 for the binary classification problem). Remember that the left child of a split will always carry all data points that are $\leq$ the splitting threshold, and the right child will contain all data that are $>$ the splitting condition.

Problem 1

Implement various helper functions that will be required for the generation of a decision tree:

- `def entropy(y)` where `y` is a vector of labels (integers). Function returns the information entropy $H(x)$.
- `def information_gain(y_parent, index_split)` where `y_parent` gives the labels of the parent node `index_split` is a binary vector, `1` indicating samples assigned to the left child node, and `0` indicating samples assigned to the right child node. Returns the information gain that given split.
- `def best_split(X, y)` where `X` is the data, and `y` is the distribution of corresponding labels at the parent node. Returns `split_dim` , e.g. 0 for the first feature dimension, and `split_val` , e.g. 1.5 for the decision rule $x_1 \leq 1.5$, indicating the best split w.r.t. information gain given the current data and labels.
- `def create_split(X, split_dim, split_val` where `X` is the data at the parent node and the splitting definition as in the previous point. Returns a boolean vector `split_idx` that indicates assignment to the left child by `1` and to the right child by `0` .

Problem 2

Validate your implementation using the data set provided (or a self-generated data set) and manual computation of the entropy and information gain to check your code. Define the main for testing as `if __name__ == __main__:` . Implement a plotting routine for labeling the data points and showing the decision boundary (given by `split_dim` and `split_val` , and also returning the entropy and information gain values in the title of the figure.