INF264 project 1

Report

We (Tobias Eilertsen, tei006 and Ingrid Liabakk Eriksen, ier008) have been working togheter on this assignment. We have mostly done pair programming, but also worked by our self. Tobias has done most of the programming, but made sure that what we have done is understood by both and that both agree.

We decided to use Python becuse thats the language we've used most and it has numpy.

Task 1.1

Implemented the decision tree. We splitted the x-feature on the average of that column. We split the data into test and training data by randomly selected rows with a given seed and a percentage. The Tree is represented as Node classes, with fields representing child nodes and some data used in building and predicting.

Task 1.2

Added gini as an option for impurity measure

Task 1.3

Split training data into training and pruning/validation data. Each node has a field for storing the majority label in subtree. Tries to change node to leaf node and comparing accuracy. If the leaf accuracy is worse, then we restore the tree to its previous state.

The pruning happens recursively in the learn method, after the subtree is built. We're pruning from the leaf node to the top, but it would possibly give better performance to prune from the root.

Task 1.4

Performance with different settings when testing for accuracy on test data (30% of all data)

	Pruning	No pruning
entropy	97.93	98.44
Gini	69.25	85.78

No pruning and entropy gives best result, which we think may be because the test data is very similar to the training data, and it doesnt overfit.

Estimate: it takes 0.002 sec to predict 400 new data points

Task 1.5

sklearning Decision tree class:

98.19% accuracy, 0.015 sec

Our tree class

Time used to build the tree and prune tree and predict all test cases: 1.2 sec Time used without pruning: 0.19 Pruning is slow in our tree becuse we check accuracy of the whole tree for every node, and becuase it prunes from the bottom up