# Decision trees versus naive Bayes in mail classification

Jori Bomanson (81819F)
jori.bomanson@aalto.fi

Sami J. Lehtinen (44814P)
sjl@iki.fi

October 14, 2011

Our goal in this exercise project is to compare a naive Bayes classifier with decision trees. We will compare learning speed and classification accuracy, among other characteristics.

Decision trees are easy to visualize graphically and following a graph of the tree allows to see exactly what steps the classifier takes when it processes a sample.

The implementation of the decision tree is done with Numpy. This approach was chosen because we felt that Python would be better suited in manipulating the necessary data structures.

The predictions were created with a decision tree classifier, chosen manually from a few runs based on the accuracy of the classifier (ratio of correctly classified samples from the validation set)[1]. The tree was prepruned with threshold values in impurity and the number of columns and rows left after splitting.

We did some manual classification checks based on the classifier output of the test set. Looking into the test data with `R` and checking the flags set in the data, there were no obvious errors with a dozen or so rows. Better validation procedures are necessary.

## References

[1] Ethem Alpaydin, *Introduction to Machine Learning*. Massachusetts Institute of Technology, Cambridge, Massachusetts, 1st edition, 2004. 415 pages. ISBN 0-262-01211-1.

[1] In the project, we will perform cross-validation to choose the best classifiers with the training data.