1 Experiments and Analysis

First, “implement” the following methods:

• Decision trees (with pruning). I suggest using either reduced error pruning (this is the form of pruning we covered in class) or minimal cost complexity pruning (it’s implemented in scikit-learn; see here). For the split criterion, you can use information gain or the Gini index or some other good criterion. You might even try a few different ones as part of your analysis (explained further below).

• Random forests (do not use pruning). What forest size (number of trees) should you use? Well, you should experiment with different forest sizes and see what happens. For the number of random features d' when selecting the split feature at each decision node, I suggest starting at $\sqrt{d}$ (where d is the number of features) and experimenting by going up and down from there. What should be the size n' of the random sample (sampled with replacement) used to learn each tree? When you are doing an experiment that varies other parameters (like d'), please set n' to be equal to the original sample size; this way, each decision tree in the forest will be trained using a bootstrap sample.

• Neural networks. Any number of layers is fine, as long at there is at least one hidden layer; I suggest going with 3 layers (i.e. the input layer, 1 hidden layer, and the output layer). To get good results, check out the data preprocessing suggestions in Appendix A. There are a number of hyperparameters that you could play with when experimenting with neural networks. For example, one hyperparameter is the number of nodes in the hidden layer, another is the choice of nonlinearity, yet another is the regularization parameter (which modulates how much you penalize the size of the weights), and yet another is choice of optimization algorithm (common choices are stochastic gradient descent and Adam).

What to test on

You’ll be analyzing the performance of these methods on a binary classification problem. This problem comes from adult dataset from the UCI repository:

https://archive.ics.uci.edu/ml/datasets/adult