

I 526/B659 Programming Assignment 1 - Due Tuesday September 23, 2014

Implement a fixed depth decision tree algorithm. In particular, the input to your algorithm will include the training data set and the maximum depth of the tree. For example, if the depth is set to one, you will learn a decision tree with one test node, which is also called a decision stump. Test your implementation, with depth=1, and 2 respectively, on the following data set as described below (train on the training data and test with the testing data set).

Data set information: This data set is extracted from the UCI zoo data set (zoo-train and zoo-test). Note that there are 16 features (the first 16 columns) and the class labels are in the last column. There are 7 classes (numerically specified as class 1 to 7). All features are binary except for feature 13, which is a categorical variable with possible values 0,2,4,5,6,8. Note that to create binary split, please use the one-vs-rest approach.

1. Start from depth = 1 and go to different depths (2,4,6,8...,16). For each depth, compute the error (the number of misclassifications) on the test set. Plot a learning curve with the depth of the tree on the x-axis and the accuracy on the y-axis.
2. Report the learned decision tree (depth 1 and depth 2) and report the confusion matrix for these two depths (A confusion matrix has the true label as rows and predicted labels in the columns. Each entry of the matrix is the number of examples. In a binary case, the top left corner is the number of negative examples correctly classified and the bottom right is the number of positives correctly classified).
3. Now, use Weka's default decision tree (J48) algorithm on this training set to learn a decision tree. Report the tree and the confusion matrix on the test set. Do not change the default parameters of Weka.
4. Repeat steps 2 and 3 with your "own" data set and report the confusion matrices.

The total points on this homework is 20.