Data Mining

Classification II

1 Decition Tree - DT

- Select a dataset with available label (for example Iris dataset)

- Divide the original dataset into two subsets: one for training (80%) and one for testing (20%).

- Build a DT for the training subset and test the built model for data from the testing subset. Note: Try the "tree" package from sklearn in Python or the function fitctree() in Matlab.

- Calculate the error of classification.

2 Random Forests

- Select a high dimensional dataset with avaiable label.

- Create K = 100 training set (using cross-validation or bagging technique), and build 1 testing set.

- Build a DT for each training set.

- Classify data from the testing set using one DT and all DTs and calculate the error of classification.

- Conclude the obtained results.