Problem Set 5
Due Monday, May 30, 2016 at 11:55pm

How to Submit

Create one .zip file (**not** .rar or something else) of your code and written answers and submit it via `ilearn.ucr.edu`. Your zip file should contain the following 3 files

- runq1.m
- ans2.{pdf/txt}

plus any other matlab functions your code depends on that you wrote.

Each file should include at the top (in comments if necessary)

- Your name
- Your UCR student ID number
- The date
- The course (CS 171)
- The assignment number (PS 5)

Note: Do not use any Matlab function that is in a toolbox for this problem set.

Problem 1. [20 pts]

In this problem, you are to use a decision tree to classify points. The training data are in the file bank-train.data. The first 19 columns are the features. The last column is the binary label: whether or not a customer responded positively to a phone offer to buy a certificate of deposit from a bank. The full details of the features can be found at `http://archive.ics.uci.edu/ml/datasets/Bank+Marketing`.

Some of the features are categorical, and some are numeric. The supplied code can handle either type of feature. Here are the supplied functions (all given as `.p` files, but they work just like `.m` files).

- `dt = learndt(X,Y,ftypes,scorefn)`. `X` and `Y` are the training set as you are familiar with from other assignments. `ftypes` is a vector reporting whether each feature is numeric or categorical. In particular, the corresponding value should be `0` if the feature is numeric and equal to the number of categories if the feature is categorical. For the bank dataset the vector should be

  $$\texttt{ftype} = \begin{bmatrix}0 & 12 & 4 & 8 & 3 & 3 & 3 & 2 & 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 & 0\end{bmatrix}$$

  The `scorefn` should be the function that will score a leaf. This function(s) you will write. The function should take in a vector of label fractions and return the score. For instance, if the input is `[ 0.5 0.25 0.25]` for the gini scoring, the function should return 0.625. To pass a function into another function, add `@` before the function's name. For instance, if you named your scoring function `giniscore`, then you would call `learndt` as `learndt(X,Y,ftypes,@giniscore)`.

- `drawdt(dt)`: `dt` is a decision tree (as learned from above). This function will draw it on the console.

- `predictdt(dt,X)` will return a vector of the categories for each row in `X`, classified according to the decision tree given by `dt`.

- `pdt = prunedt(dt,X,Y)` will returned the pruned verson of the decision tree `dt`, pruned according to the data given by `X` and `Y`.

You are to write a function called `runq1` that takes no parameters and returns a vector of the predicted values on the testing data given in the file `banktestX.data` (which is missing the last column). `runq1` should also return the learned decision tree: `[Y,dt] = runq1()`.

Problem 2. [30 pts]

Build and prune a decision tree by hand using the Gini scoring rule. Show all of your steps nicely and clearly. Draw the build tree, before and after pruning.

The data (all numeric features) are as below.

training

x_1	x_2	x_3	y
1	1	2	0
1	2	2	0
2	3	1	1
3	4	0	1
3	1	2	1
3	1	1	0
1	4	1	1
2	2	2	2
2	3	2	2
2	3	1	2

pruning

x_1	x_2	x_3	y
0	0	3	0
3	2	4	1
2	1	3	1
0	3	3	2
2	3	0	0
1	1	0	0
4	0	1	0
0	4	1	1
2	1	2	1
1	3	1	1