
```
% IDS/ACM/CS 158: Fundamentals of Statistical Learning
% PS5, Problem 5: Classification Trees for Stock Market Data
% Author: Michael Li, mlli@caltech.edu
```

```
%-----
clear;
```

```
% Stock market Data
train = readmatrix('stock_market_train.csv');
test = readmatrix('stock_market_test.csv');
T0 = fitctree(train(:,1:end-1), train(:,end));
view(T0, 'Mode', 'graph')
```

```
train_preds = predict(T0, train(:,1:end-1));
test_preds = predict(T0, test(:, 1:end-1));
```

```
train_err = mean(train(:,end) ~= train_preds);
test_err = mean(test(:,end) ~= test_preds);
```

```
fprintf("Training Error for T_0: %s\n", train_err);
fprintf("Test Error for T_0: %s\n", test_err);
```

```
% The training error is .118
% the test error is .476
```

```
% loop over each alpha and run loocv
alphas = linspace(0, .04, 41);
best = [];
lowest_err = 1;
```

```
for a = alphas
    err = 0;
    tree = [];
    % for each datapoint leave it out and test error
    for i = 1:length(train)
        x = repmat(train, 1);
        x(i,:) = [];
        test_x = train(i,:);

        tree = fitctree(x(:,1:end-1), x(:,end));
        tree = prune(tree, 'Alpha', a);
        pred = predict(tree, test_x(1, 1:end-1));
        err = err + (pred ~= test_x(1, end));
    end

    err = err / length(train);

    % if error is lower than previous, replace
    if err < lowest_err
        best = a;
        lowest_err = err;
    end
end
```

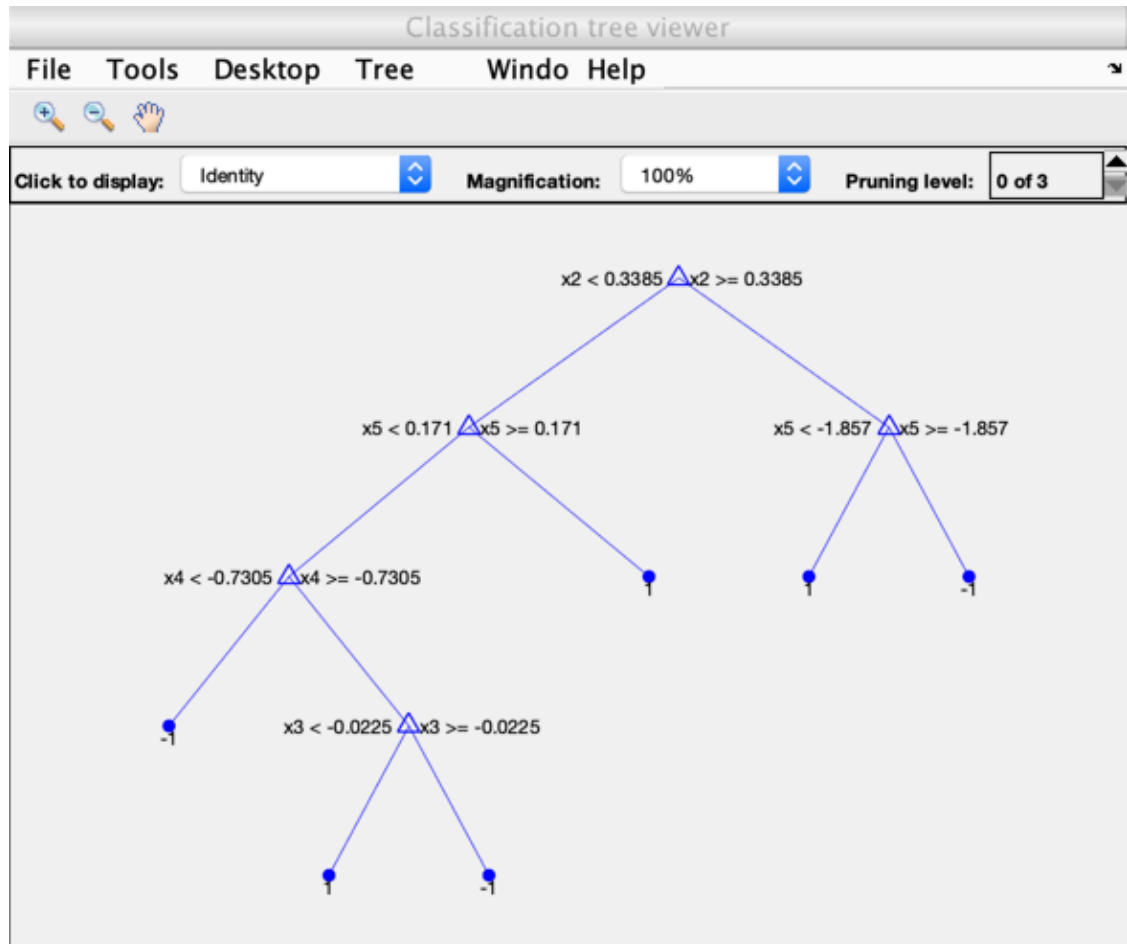
```
% refit best alpha on all data
T_best = fitctree(train(:,1:end-1), train(:,end));
T_best = prune(T_best, 'Alpha', best);
view(T_best, 'Mode', 'graph');

train_preds = predict(T_best, train(:,1:end-1));
test_preds = predict(T_best, test(:, 1:end-1));

train_err = mean(train(:,end) ~= train_preds);
test_err = mean(test(:,end) ~= test_preds);
fprintf("\nBest Alpha: %s\n", best);
fprintf("Training Error for T_best: %s\n", train_err);
fprintf("Test Error for T_best: %s\n", test_err);
% Optimal value of pruning parameter is .01
% The training error for T is .401
% the test error for T is .584

Training Error for T_0: 1.180000e-01
Test Error for T_0: 4.760000e-01

Best Alpha: 1.000000e-02
Training Error for T_best: 4.010000e-01
Test Error for T_best: 5.840000e-01
```

Published with MATLAB® R2019a