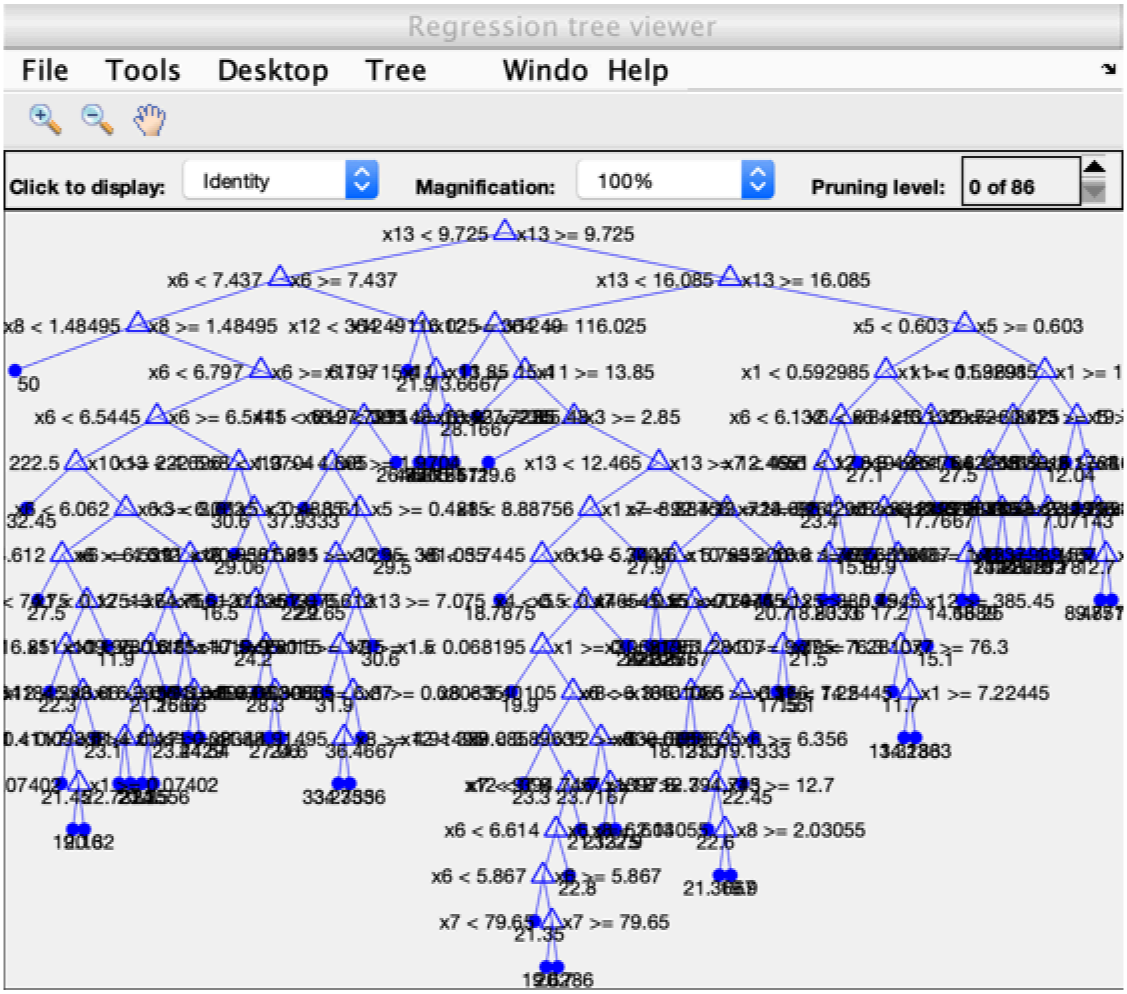
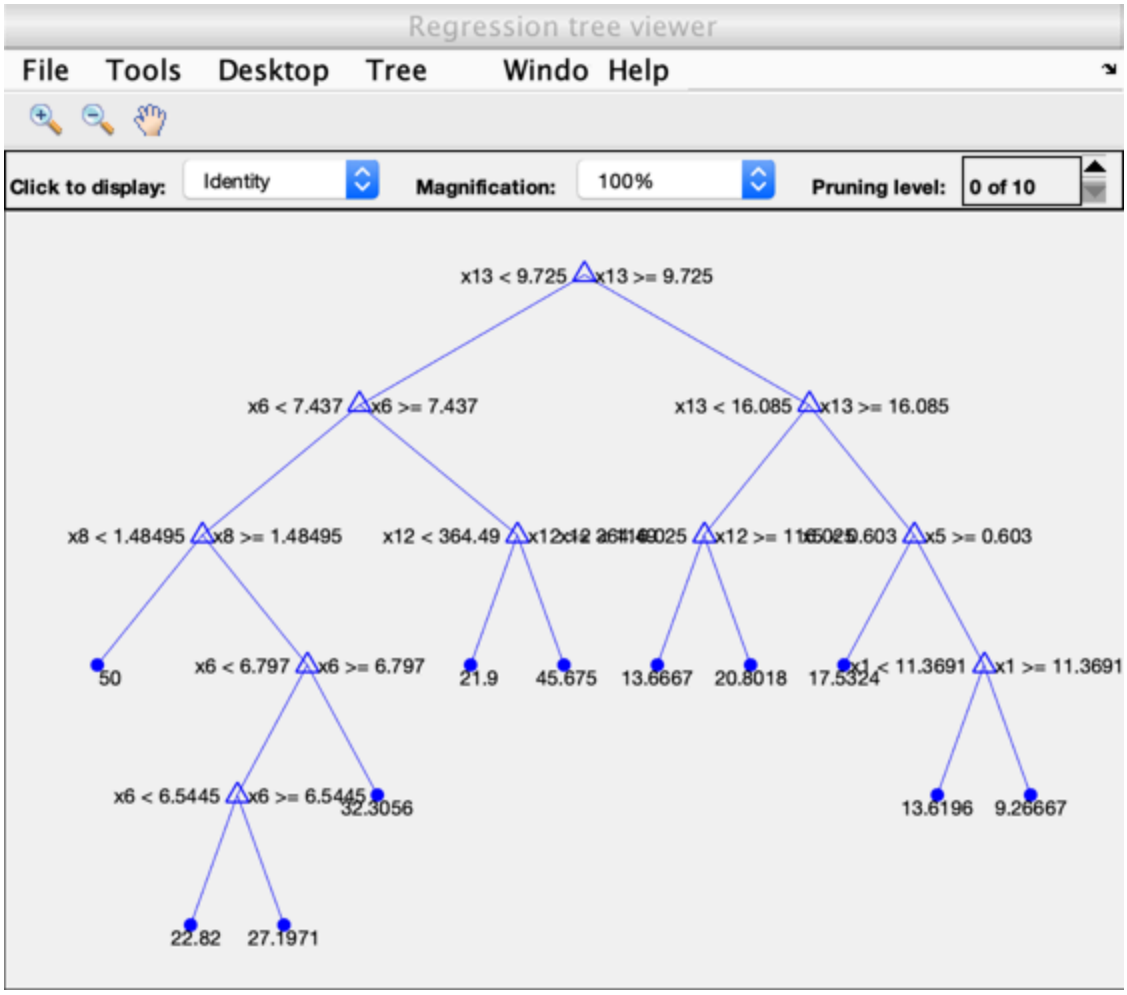
```
% IDS/ACM/CS 158: Fundamentals of Statistical Learning
% PS5, Problem 4: Regression Trees for Boston Housing Data
% Author: Michael Li, mlli@caltech.edu
%--------------------------------------------------------------------------
clear;

% Boston housing Data
train = readmatrix('Boston_train.csv');
test = readmatrix('Boston_test.csv');
TO = fitrtree(train(:,1:end-1), train(:,end));
view(TO, 'Mode', 'graph')

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

train_err = norm(train(:,end) - train_preds)^2 / length(train);
test_err = norm(test(:,end) - test_preds)^2 / length(test);

fprintf("Training Error for T_0: %s\n", train_err);
fprintf("Test Error for T_0: %s\n", test_err);
% The training error for T0 is 2.1707
% the test error is for T0 12.9867

alphas = linspace(0, 2, 21);
best = [];
lowest_err = 10^10;

% loop over each alpha and run loocv
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 = fitrtree(x(:,1:end-1), x(:,end));
        tree = prune(tree, 'Alpha', a);
        pred = predict(tree, test_x(1, 1:end-1));
        err = err + (test_x(1, end) - pred)^2;
    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 = fitrtree(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 = norm(train(:,end) - train_preds)^2 / length(train);
test_err = norm(test(:,end) - test_preds)^2 / length(test);
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 .7
% The training error for T is 10.21203
% the test error for T is 9.607979
```

```
Training Error for T_0: 2.170719e+00
Test Error for T_0: 1.298670e+01

Best Alpha: 7.000000e-01
Training Error for T_best: 1.021203e+01
Test Error for T_best: 9.607979e+00
```


Regression tree viewer
File
Tools
Desktop
Tree
Windo
Help
Click to display:
Identity
Magnification:
100%
Pruning level:
0 of 86
x13 < 9.725
x13 >= 9.725
x6 < 7.437
x6 >= 7.437
x13 < 16.085
x13 >= 16.085
x8 < 1.48495
x8 >= 1.48495
x5 < 0.603
x5 >= 0.603
50
x6 < 6.797
x1 < 0.592985
x6 < 6.5445
x13 < 12.465
x6 < 6.614
x8 >= 2.03055
x6 < 5.867
x7 < 79.65
x7 >= 79.65



Published with MATLAB® R2019a