Lasso Regularization

[View all machine learning examples](https://www.mathworks.com/solutions/machine-learning/examples.html)

This example demonstrates the use of lasso for feature selection by looking at a dataset and identifying predictors of diabetes in a population. The dataset contains 10 predictors. The goal is to identify important predictors and discard those that are unnecessary.

View the [complete set of data and functions](http://www.stanford.edu/~hastie/Papers/LARS/diabetes.data) for this demonstration.

Download Data

filename = 'diabetes.txt'; urlwrite('http://www.stanford.edu/~hastie/Papers/LARS/diabetes.data',filename);

Import data

Once the file is saved, you can import data into MATLAB as a table using the [Import Tool](https://www.mathworks.com/help/matlab/import_export/import-data-interactively.html) with default options. Alternatively, you can use the following code which can be auto generated from the Import Tool:

formatSpec = '%f%f%f%f%f%f%f%f%f%f%f%[^\n\r]';

fileID = fopen(filename,'r');

dataArray = textscan(fileID, formatSpec, 'Delimiter', '\t', 'HeaderLines' ,1, 'ReturnOnError', false);

fclose(fileID);

diabetes = table(dataArray{1:end-1}, 'VariableNames', {'AGE','SEX','BMI','BP','S1','S2','S3','S4','S5','S6','Y'});

clearvars filename delimiter startRow formatSpec fileID dataArray ans;

% Delete the file

delete diabetes.txt

Read the Predictors and Response Variables from the Table

predNames = diabetes.Properties.VariableNames(1:end-1);

X = diabetes{:,1:end-1};

y = diabetes{:,end};

Perform Lasso Regularization

[beta, FitInfo] = lasso(X,y,'Standardize',true,'CV',10,'PredictorNames',predNames);

lassoPlot(beta,FitInfo,'PlotType','Lambda','XScale','log');

hlplot = get(gca,'Children');

% Generating colors for each line in the plot

colors = hsv(numel(hlplot));

for ii = 1:numel(hlplot)

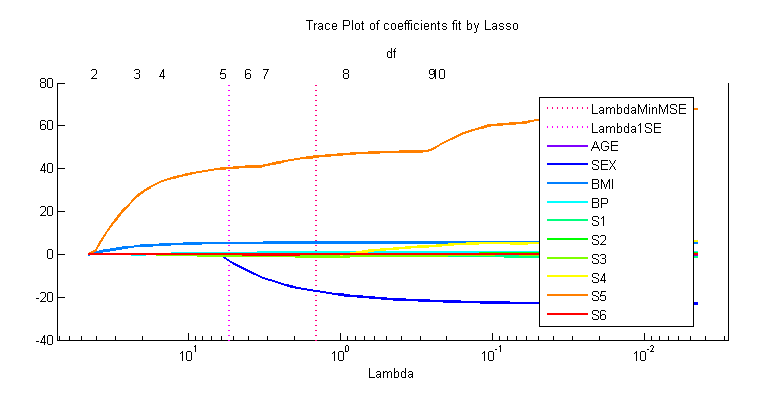
set(hlplot(ii),'color',colors(ii,:));

end

set(hlplot,'LineWidth',2)

set(gcf,'Units','Normalized','Position',[0.2 0.4 0.5 0.35])

legend('Location','Best')



Larger values of lambda appear on the left side of the graph, which means that there is increased regularization. As the lambda value increases, the number of nonzero predictors also increases.

Important Predictors

As a rule of thumb, one standard-error value is often used for choosing a smaller model with a good fit.

lam = FitInfo.Index1SE;

isImportant = beta(:,lam) ~= 0;

disp(predNames(isImportant))

'BMI' 'BP' 'S3' 'S5'

Fit a Linear Model with the Terms for Comparison

mdlFull = fitlm(X,y,'Intercept',false);

disp(mdlFull)

Linear regression model:

y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10

Estimated Coefficients:

Estimate SE tStat pValue

x1 0.022296 0.22256 0.10018 0.92025

x2 -26.073 5.9561 -4.3775 1.5074e-05

x3 5.3537 0.73462 7.2877 1.5112e-12

x4 1.0178 0.2304 4.4175 1.2635e-05

x5 1.2636 0.33044 3.8239 0.00015068

x6 -1.2849 0.3468 -3.7051 0.00023877

x7 -3.0683 0.37189 -8.2505 1.9259e-15

x8 -5.508 5.5883 -0.98565 0.32486

x9 5.5034 9.4293 0.58365 0.55976

x10 0.12339 0.2788 0.44256 0.6583

Number of observations: 442, Error degrees of freedom: 432

Root Mean Squared Error: 55.6

Compare the MSE for regularized and unregularized models.

disp(['Lasso MSE: ', num2str(FitInfo.MSE(lam))])

disp(['Full MSE: ', num2str(mdlFull.MSE)])

Lasso MSE: 3176.5163

Full MSE: 3092.896

The mean squared error (MSE) of the fit using only the important predictors as determined by lasso, is quite close to the error from the linear model that uses all the predictors. Lasso is often used to prevent overfitting or remove redundant predictors to improve model accuracy.