Linear Regression contd…

Aarti Singh

Machine Learning 10-315
Sept 28, 2020

Linear Regression

$$\hat{f}_n^L = \arg\min_{f \in \mathcal{F}_L} \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2$$

Err

$$f(X_i) = X_i\beta$$

parameter

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (X_i\beta - Y_i)^2$$

$$\hat{f}_n^L(X) = X\hat{\beta}$$

$$= \arg\min_{\beta} \frac{1}{n} (\mathbf{A}\beta - \mathbf{Y})^T (\mathbf{A}\beta - \mathbf{Y})$$

← J(β)

$$\mathbf{A} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} X_1^{(1)} & \dots & X_1^{(p)} \\ \vdots & \ddots & \vdots \\ X_n^{(1)} & \dots & X_n^{(p)} \end{bmatrix} \quad \mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_n \end{bmatrix}$$

Linear regression solution satisfies Normal Equations

$$(\mathbf{A}^T\mathbf{A})\hat{\beta} = \mathbf{A}^T\mathbf{Y}$$

p x p p x1 p x1

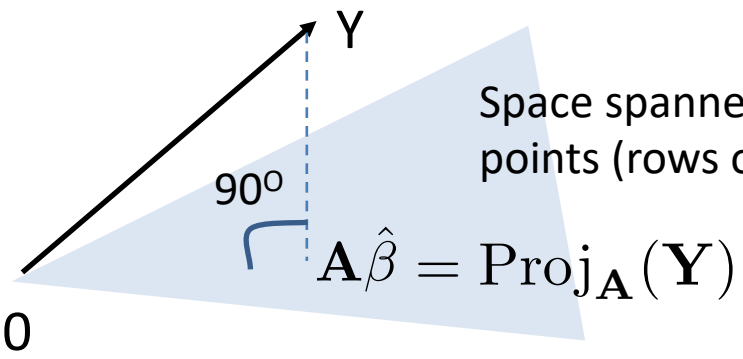
If $(\mathbf{A}^T\mathbf{A})$ is invertible,

$$\hat{\beta} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{Y} \qquad \hat{f}_n^L(X) = X\hat{\beta}$$

$\hat{f}_n(A) = A\hat{\beta} = \mathrm{Proj}_{\mathbf{A}}(\mathbf{Y})$

$\mathrm{Proj}_A = A(A^TA)^{-1}A^T$ projects onto

Predicted labels for training points



Linear regression solution satisfies Normal Equations

$$(\mathbf{A}^T\mathbf{A})\hat{\beta} = \mathbf{A}^T\mathbf{Y}$$

p x p p x1 p x1

If $(\mathbf{A}^T\mathbf{A})$ is invertible,

$$\hat{\beta} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{Y} \qquad \hat{f}_n^L(X) = X\hat{\beta}$$

Later: When is $(\mathbf{A}^T\mathbf{A})$ invertible ?

Recall: Full rank matrices are invertible. What is rank of $(\mathbf{A}^T\mathbf{A})$?

= p

Now: What if $(\mathbf{A}^T\mathbf{A})$ is invertible but expensive (p very large)?

Gradient Descent

Even when $(\mathbf{A}^T\mathbf{A})$ is invertible, might be computationally expensive if **A** is huge.

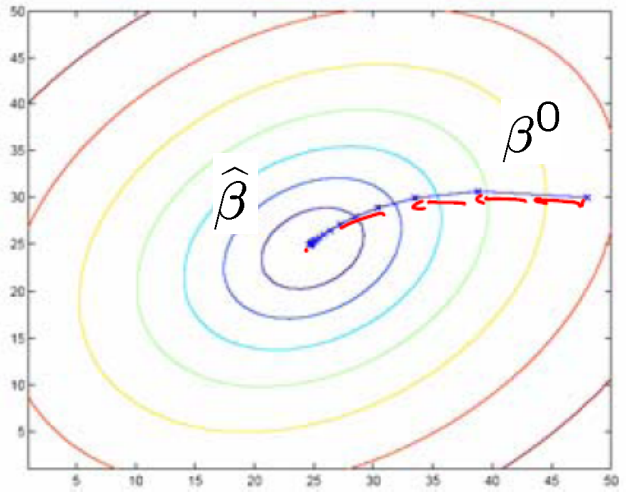
$$\widehat{\beta} = \arg\min_{\beta} \frac{1}{n}(\mathbf{A}\beta - \mathbf{Y})^T(\mathbf{A}\beta - \mathbf{Y}) = \arg\min_{\beta} J(\beta)$$

Since J(β) is convex, move along negative of gradient

Initialize: β^0

step size

Update:
$$\beta^{t+1} = \beta^t - \frac{\alpha}{2}\frac{\partial J(\beta)}{\partial \beta}\Big|_t$$
$$= \beta^t - \alpha\, \mathbf{A}^T(\underbrace{\mathbf{A}\beta^t - Y}_{0 \text{ if } \widehat{\beta} = \beta^t})$$



Stop: when some criterion met e.g. fixed # iterations, or $\frac{\partial J(\beta)}{\partial \beta}\Big|_{\beta^t} < \varepsilon$.

Linear regression solution satisfies Normal Equations

$$(\mathbf{A}^T\mathbf{A})\hat{\beta} = \mathbf{A}^T\mathbf{Y}$$

p x p p x1 p x1

When is $(\mathbf{A}^T\mathbf{A})$ invertible ?

Recall: Full rank matrices are invertible. What is rank of $(\mathbf{A}^T\mathbf{A})$?

null (M) = { v : Mv = 0 }

Null space argument

null space (A^TA) = null space (A) [Claim]

1) Consider v s.t. $A^TAv = 0$

$\Rightarrow v^TA^TAv = 0$

$\Rightarrow z^Tz = 0$

$z^Tz = \sum_i z_i^2 = 0$

$\Rightarrow z = 0$

$\equiv Av = 0$

2) Consider v s.t. $Av = 0$

$\Rightarrow A^TAv = 0$

rank = p - dim(null)

Linear regression solution satisfies Normal Equations

$$(\mathbf{A}^T\mathbf{A})\hat{\beta} = \mathbf{A}^T\mathbf{Y}$$

p x p p x1 p x1

* collinear rows or columns

or • $n < p$ high-dim setting not enough data

When is $(\mathbf{A}^T\mathbf{A})$ invertible ?

Recall: Full rank matrices are invertible. What is rank of $(\mathbf{A}^T\mathbf{A})$?

Null space argument

$\text{null}(A^TA) = \text{null}(A)$

$\Rightarrow \text{rank}(A^TA) = \text{rank}(A)$

same number of linearly independent rows & columns

How many ?

$$A_{n\times p} = \begin{bmatrix} x_1^{(1)} & \cdots & x_1^{(p)} \\ \vdots & & \\ x_n^{(1)} & \cdots & x_n^{(p)} \end{bmatrix}$$

$\text{rank}(A) \leq \min(n,p) \overset{?}{\neq} p$ (full rank)

*

Linear regression solution satisfies Normal Equations

$$(\mathbf{A}^T\mathbf{A})\hat{\beta} = \mathbf{A}^T\mathbf{Y}$$

p x p p x1 p x1

When is $(\mathbf{A}^T\mathbf{A})$ invertible ?

Recall: Full rank matrices are invertible. What is rank of $(\mathbf{A}^T\mathbf{A})$?

Rank$(\mathbf{A}^T\mathbf{A})$ = number of non-zero eigenvalues of $(\mathbf{A}^T\mathbf{A})$= number of non-zero singular values of **A** <= min(n,p) since **A** is n x p

So, rank$(\mathbf{A}^T\mathbf{A})$, r <= min(n,p) not invertible if r < p (e.g. n < p i.e. high-dimensional setting)

Linear regression solution satisfies Normal Equations

$$(\mathbf{A}^T\mathbf{A})\hat{\beta} = \mathbf{A}^T\mathbf{Y}$$

p x p p x 1 p x 1

→ r independent equations
p unknowns ($\hat{\beta}$)

When is $(\mathbf{A}^T\mathbf{A})$ invertible ?
Recall: Full rank matrices are invertible. What is rank of $(\mathbf{A}^T\mathbf{A})$?

If $\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^\top$, then normal equations $(\mathbf{S}\mathbf{V}^\top)\hat{\beta} = (\mathbf{U}^\top\mathbf{Y})$

S - r x r r x p p x 1 r x 1

r equations in p unknowns. Under-determined if r < p, hence no unique solution.

Regularized Linear Regression

Aarti Singh

Machine Learning 10-315
Sept 28, 2020

Regularized Least Squares

What if $(\mathbf{A}^T\mathbf{A})$ is not invertible ?

r equations , p unknowns – underdetermined system of linear equations
many feasible solutions

Need to constrain solution further

e.g. bias solution to "small" values of β (small changes in input don't translate to large changes in output)

$$\hat{\beta}_{\text{MAP}} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda\|\beta\|_2^2$$

Ridge Regression
(l2 penalty)

$$= \arg\min_{\beta} \; (\mathbf{A}\beta - \mathbf{Y})^T(\mathbf{A}\beta - \mathbf{Y}) \; + \lambda\|\beta\|_2^2 \qquad \lambda \geq 0$$

$$\hat{\beta}_{\text{MAP}} = (\mathbf{A}^\top\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^\top\mathbf{Y}$$

Regularized Least Squares

$$\hat{\beta}_{\text{MAP}} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda\|\beta\|_2^2$$

Ridge Regression (l2 penalty)

$$= \arg\min_{\beta} \; (\mathbf{A}\beta - \mathbf{Y})^T(\mathbf{A}\beta - \mathbf{Y}) + \lambda\|\beta\|_2^2 \qquad \lambda \geq 0$$

$\|\beta\|_2^2 = \beta^T\beta$

$$\frac{\partial J(\beta)}{\partial \beta} = 2A^TA\beta - 2A^TY + 2\lambda\beta$$

$$= 2(A^TA + \lambda I)\beta - 2A^TY \;\Big|\; \hat{\beta}_{MAP} = 0$$

$$\Rightarrow (A^TA + \lambda I)\hat{\beta}_{MAP} = A^TY$$

$$\Rightarrow \hat{\beta}_{MAP} = (A^TA + \lambda I)^{-1}A^TY \quad \text{if invertible}$$

all rows & col^m are independent

Is $(\mathbf{A}^\top\mathbf{A} + \lambda\mathbf{I})$ invertible ? Yes! $\lambda > 0$

$\text{eval}(A^TA + \lambda I) = \text{eval}(A^TA) + \lambda > 0$

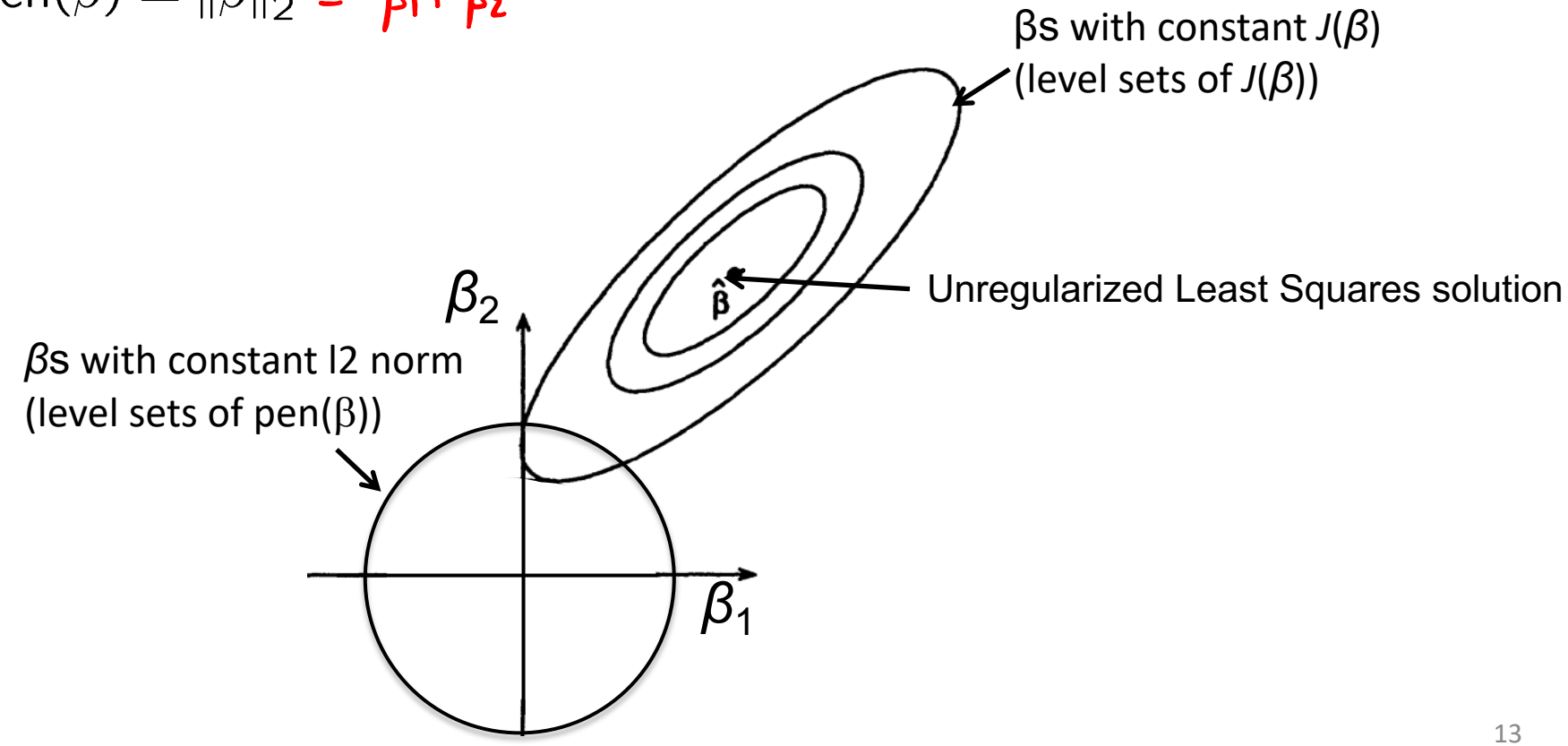
$(A^TA + \lambda I)V = A^TAV + \lambda V = \eta V + \lambda V = (\eta + \lambda)V$

Understanding regularized Least Squares

$$\min_{\beta}(\mathbf{A}\beta - \mathbf{Y})^T(\mathbf{A}\beta - \mathbf{Y}) + \lambda \text{pen}(\beta) = \min_{\beta} J(\beta) + \lambda \text{pen}(\beta)$$

Ridge Regression:

$\text{pen}(\beta) = \|\beta\|_2^2 = \beta_1^2 + \beta_2^2$



Regularized Least Squares

What if $(\mathbf{A}^T\mathbf{A})$ is not invertible ?

r equations , p unknowns – underdetermined system of linear equations
many feasible solutions

Need to constrain solution further

e.g. bias solution to "small" values of β (small changes in input don't translate to large changes in output)

$$\widehat{\beta}_{\text{MAP}} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda\|\beta\|_2^2 \quad = \sum_i \beta_i^2$$

Ridge Regression
(l2 penalty)

$$\widehat{\beta}_{\text{MAP}} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda\|\beta\|_1 \quad = \sum_i |\beta_i|$$

$\lambda \geq 0$

Lasso
(l1 penalty)

Many β can be zero – many inputs are irrelevant to prediction in high-dimensional settings (typically intercept term not penalized)

Regularized Least Squares

What if $(\mathbf{A}^T\mathbf{A})$ is not invertible ?

r equations , p unknowns – underdetermined system of linear equations
many feasible solutions

Need to constrain solution further

e.g. bias solution to "small" values of β (small changes in input don't translate to large changes in output)

$$\widehat{\beta}_{\text{MAP}} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda\|\beta\|_2^2$$

Ridge Regression
(l2 penalty)

$$\widehat{\beta}_{\text{MAP}} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda\|\beta\|_1$$

$\lambda \geq 0$

Lasso
(l1 penalty)

No closed form solution, but can optimize using sub-gradient descent (packages available)

Ridge Regression vs Lasso

$$\min_{\beta}(\mathbf{A}\beta - \mathbf{Y})^T(\mathbf{A}\beta - \mathbf{Y}) + \lambda \text{pen}(\beta) = \min_{\beta} J(\beta) + \lambda \text{pen}(\beta)$$

Ridge Regression:

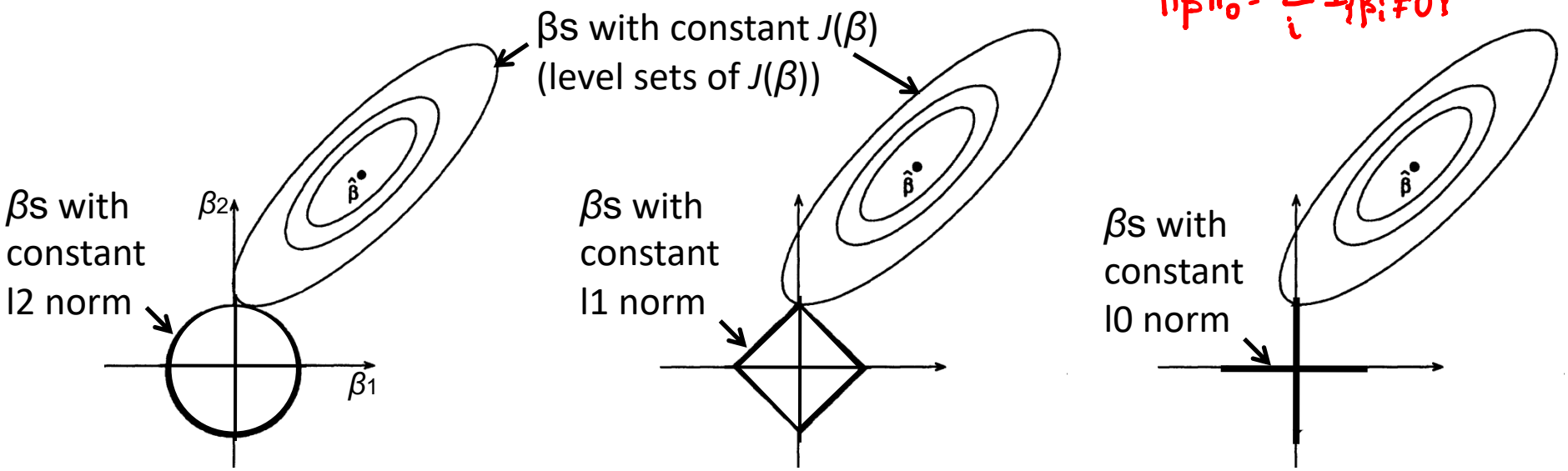
$\text{pen}(\beta) = \|\beta\|_2^2$

Lasso:

$\text{pen}(\beta) = \|\beta\|_1$

Ideally l0 penalty, but optimization becomes non-convex

$\|\beta\|_0 = \sum_i 1_{\{\beta_i \neq 0\}}$



Lasso (l1 penalty) results in sparse solutions – vector with more zero coordinates
Good for high-dimensional problems – don't have to store all coordinates, interpretable solution!

Matlab example

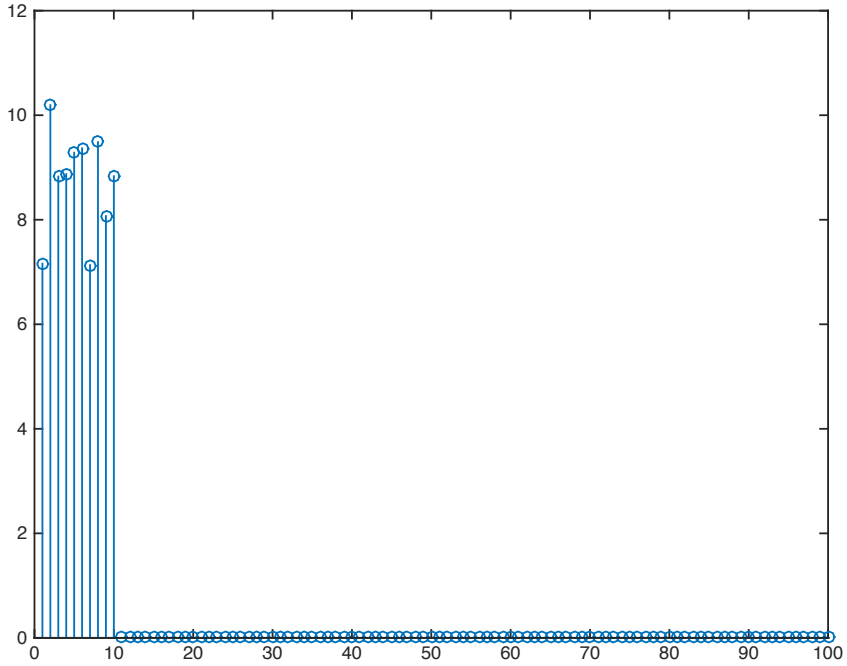
```
clear all
close all

n = 80;     % datapoints
p = 100;   % features
k = 10;     % non-zero features

rng(20);
X = randn(n,p);
weights = zeros(p,1);
weights(1:k) = randn(k,1)+10;
noise = randn(n,1) * 0.5;
Y = X*weights +  noise;

Xtest = randn(n,p);
noise = randn(n,1) * 0.5;
Ytest = Xtest*weights + noise;
```

```
lassoWeights = lasso(X,Y,'Lambda',1,
'Alpha', 1.0);
Ylasso = Xtest*lassoWeights;
norm(Ytest-Ylasso)

ridgeWeights = lasso(X,Y,'Lambda',1,
'Alpha', 0.0001);
Yridge = Xtest*ridgeWeights;
norm(Ytest-Yridge)

stem(lassoWeights)
pause
stem(ridgeWeights)
```

Matlab example

Test MSE = 33.7997

Lasso Coefficients



Test MSE = 185.9948

Ridge Coefficients