BIG DATA MACHINE LEARNING

- RIDGE, LASSO, ELASTIC NET

Weipeng Li

03/23/2017

Reference

- Elements of Statistical Learning, *Trevor Hastie, Rober Tibshirani, Jeromy Friedman*
- Ridge Regression, LASSO and Elastic Net A talk given atNYC opendata meetup, www.nycopendata.com
- High-dimensional regression, Advanced Methods for Data Analysis, http://www.stat.cmu.edu/~ryantibs/advmethods/notes/highdim.pdf
- *Expression Arrays and the p>>n problem, https://pdfs.semanticscholar.org/6297/70429cb189494dca961e1900bda1a1b0099d.pdf*
- Ridge regression, Wessel van Wieringen, *http://www.few.vu.nl/~wvanwie/Courses/HighdimensionalDataAnalysis/WNvanWieringen_HDDA_Lecture4_RidgeRegression_20162017.pdf*

2

Linear Regression

n observations, each has one response variable and p predictors

$$Y = (y_1, \cdots, y_n)^T, \qquad n \times 1$$

$$X = (X_1, \cdots, X_p), \qquad n \times p$$

- We want to find a linear combination β of predictors $x = (x_1, \cdots, x_p)$ to
  - describe the actual relationship between y and $x_1, \cdots, x_p$
  - use $\hat{y} = x^T\beta$ to predict y
- Examples
  - find relationship between pressure and water boiling point
  - use GDP to predict interest rate (the accuracy of the prediction is important but the actual relationship may not matter)

3

Ordinary Least Square Estimate – Unbiased

Residual Sum of Square:

$$\begin{aligned} \mathrm{RSS}(\beta) &= \sum_{i=1}^{N}(y_i - f(x_i))^2 \\ &= \sum_{i=1}^{N}\Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\Big)^2. \end{aligned}$$

In a matrix form:

$$\mathrm{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta).$$

Differentiating with respect to β, we obtain:

$$\frac{\partial \mathrm{RSS}}{\partial \beta} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta)$$

Assuming (for the moment) that X has full column rank (each of the columns of the matrix are linearly independent), and hence $\mathbf{X}^T\mathbf{X}$ is invertible, we set the first derivative to zero, and get the unique solution to $\hat{\beta}$:

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0$$

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

OLS has the minimum mean square error among unbiased linear estimator (Gauss Markov Theorem) though a biased estimator may have smaller MSE than LSE

(Bias: difference between the expected model prediction and the true value)

4

Issues/Solution for Least Square Estimate

- Issues:
  - When multicollinearity exists, $\mathbf{X}^T\mathbf{X}$ is not invertible, least squares coefficients $\hat{\beta}$ have high variance and are poorly determined.
  - When p > n, the p × p matrix $\mathbf{X}^T\mathbf{X}$ has rank at most n, and is hence singular and cannot be inverted
- Solution:
  - Biased (Penalized) Estimator (sacrifice bias, reduce variance)

5

Ridge Regression – Shrink Coefficients

Add bias to the least square estimate of linear regression:

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

equivalent to:

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2,$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 \le t, \qquad \text{(the higher the λ, the lower the t)}$$

Ridge Regression Residual Sum of Squares in matrix form:

$$\text{RSS}(\lambda) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta^T\beta,$$

$$\hat{\beta}^{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y},$$

A wildly large positive coefficient on one variable can be canceled by a similarly large negative coefficient on its correlated cousin. By imposing a size constraint on the coefficients, this problem is alleviated.

6

Ridge Regression - Biased

$$
\begin{aligned}
E[\hat{\boldsymbol{\beta}}(\lambda)] &= E[(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}] \\
&= E\{[\mathbf{I}+\lambda(\mathbf{X}^T\mathbf{X})^{-1}]^{-1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}\} \\
&= E\{[\mathbf{I}+\lambda(\mathbf{X}^T\mathbf{X})^{-1}]^{-1}\hat{\boldsymbol{\beta}}\} \\
&= [\mathbf{I}+\lambda(\mathbf{X}^T\mathbf{X})^{-1}]^{-1}E(\hat{\boldsymbol{\beta}}) \\
&= [\mathbf{I}+\lambda(\mathbf{X}^T\mathbf{X})^{-1}]^{-1}\boldsymbol{\beta} \\
&\neq \boldsymbol{\beta}
\end{aligned}
$$

Unbiased when $\lambda = 0$

7

OLS Variance vs Ridge Variance

- OLS Variance

$$Var(\hat{\beta}) = \begin{bmatrix} \text{var}(\hat{\beta}_1) & \text{cov}(\hat{\beta}_1\hat{\beta}_2) & \text{cov}(\hat{\beta}_1\hat{\beta}_2) & \cdots & \text{cov}(\hat{\beta}_1\hat{\beta}_k) \\ \text{cov}(\hat{\beta}_2\hat{\beta}_1) & \text{var}(\hat{\beta}_2) & \text{cov}(\hat{\beta}_2\hat{\beta}_3) & \cdots & \text{cov}(\hat{\beta}_2\hat{\beta}_k) \\ \text{cov}(\hat{\beta}_3\hat{\beta}_1) & \text{cov}(\hat{\beta}_3\hat{\beta}_2) & \text{var}(\hat{\beta}_3) & \cdots & \text{cov}(\hat{\beta}_3\hat{\beta}_k) \\ & & \vdots & & \\ \text{cov}(\hat{\beta}_k\hat{\beta}_1) & \text{cov}(\hat{\beta}_k\hat{\beta}_2) & \text{cov}(\hat{\beta}_k\hat{\beta}_3) & \cdots & \text{var}(\hat{\beta}_k) \end{bmatrix}$$

$$Var(\hat{\beta}) = E[(\hat{\beta} - E[\hat{\beta}])(\hat{\beta} - E[\hat{\beta}])^T]$$
$$= E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T]$$

$$\hat{\beta} = \beta + (X^T X)^{-1} X^T U \text{ or } \hat{\beta} - \beta = (X^T X)^{-1} X^T U$$

$$\begin{aligned} Var(\hat{\beta}) &= E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T] \\ &= E[((X^T X)^{-1} X^T U)\ (U^T X (X^T X)^{-1})] \\ &= (X^T X)^{-1} X^T\ E[UU^T]\, X (X^T X)^{-1} \\ &= (X^T X)^{-1} X^T\ \sigma^2 I\, X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1} I \\ &= \sigma^2 (X^T X)^{-1} \end{aligned}$$

Variance: When repeated multiple times, how much predictions vary by different realizations of the model

OLS Variance vs Ridge Variance - Ridge Variance

Hereto define:

$$\mathbf{W}_\lambda = [\mathbf{I} + \lambda(\mathbf{X}^T\mathbf{X})^{-1}]^{-1}$$

Then note that:

$$\begin{aligned}
\mathbf{W}_\lambda \hat{\boldsymbol{\beta}} &= [\mathbf{I} + \lambda(\mathbf{X}^T\mathbf{X})^{-1}]^{-1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\
&= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\
&= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y} \\
&= \hat{\boldsymbol{\beta}}(\lambda)
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}[\hat{\boldsymbol{\beta}}(\lambda)] &= \mathrm{Var}[\mathbf{W}_\lambda\hat{\beta}] \\
&= \mathbf{W}_\lambda \mathrm{Var}[\hat{\beta}]\mathbf{W}_\lambda^{\mathrm{T}} \\
&= \sigma^2\mathbf{W}_\lambda(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{W}_\lambda^{\mathrm{T}}
\end{aligned}$$

$$\mathrm{Var}(\hat{\beta}) \succeq \mathrm{Var}[\hat{\beta}(\lambda)]$$

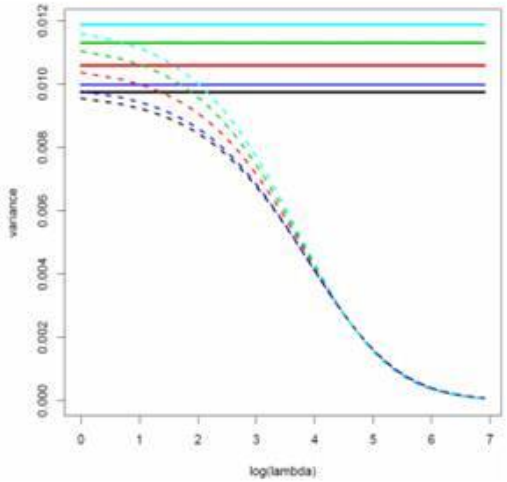
9

OLS Variance vs Ridge Variance - Orthonormal Case

In the orthonormal case, we have $\mathrm{Var}(\beta) = \sigma^2\mathbf{I}$ and

$$\begin{aligned}\mathrm{Var}[\hat{\beta}(\lambda)] &= \sigma^2\mathbf{W}_\lambda(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{W}_\lambda^T \\ &= \sigma^2[\mathbf{I}+\lambda\mathbf{I}]^{-1}\,\mathbf{I}\,\{[\mathbf{I}+\lambda\mathbf{I}]^{-1}\}^T \\ &= \sigma^2(1+\lambda)^{-2}\mathbf{I}\end{aligned}$$

As the penalty parameter is non-negative the former exceeds the latter.



10

Expected Prediction (Test) Error and λ Selection

Suppose β_0 is the true value and $y = x^T\beta_0 + \sigma\epsilon, \epsilon \sim \mathcal{N}(0,1)$

- Prediction error at x_0, the difference between the actual response and the model prediction

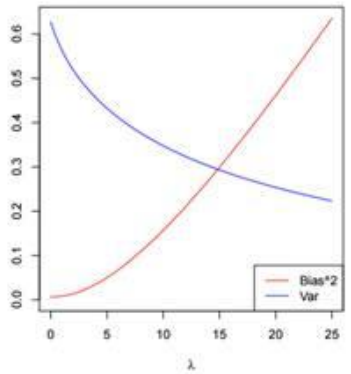
$$\mathrm{EPE}(x_0) = E[(y - x_0^T\hat{\beta})^2 | x = x_0]$$

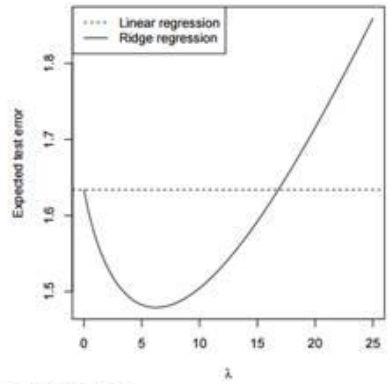
$$\mathrm{EPE}(x_0) = \sigma^2 + E(x_0^T\beta_0 - x_0^T\hat{\beta})^2$$

$$\mathrm{EPE}(x_0) = \sigma^2 + [\mathrm{Bias}^2(x_0^T\hat{\beta}) + \mathrm{Var}(x_0^T\hat{\beta})]$$

MSE

- A biased estimator may achieve a smaller prediction error than an un-biased estimator





- When λ reduces, bias reduces, variance increases
- λ is determined by cross validation from the smallest prediction error run

11

Ridge Regression Properties

- If two predictors are highly correlated among themselves, the estimated coefficients will be similar for them.
- If some variables are exactly identical, they will have same coefficients
- Ridge Regression does not zero coefficients

12

LASSO Regression – Feature Selection (Least Absolute Shrinkage and Selection Operator)

$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

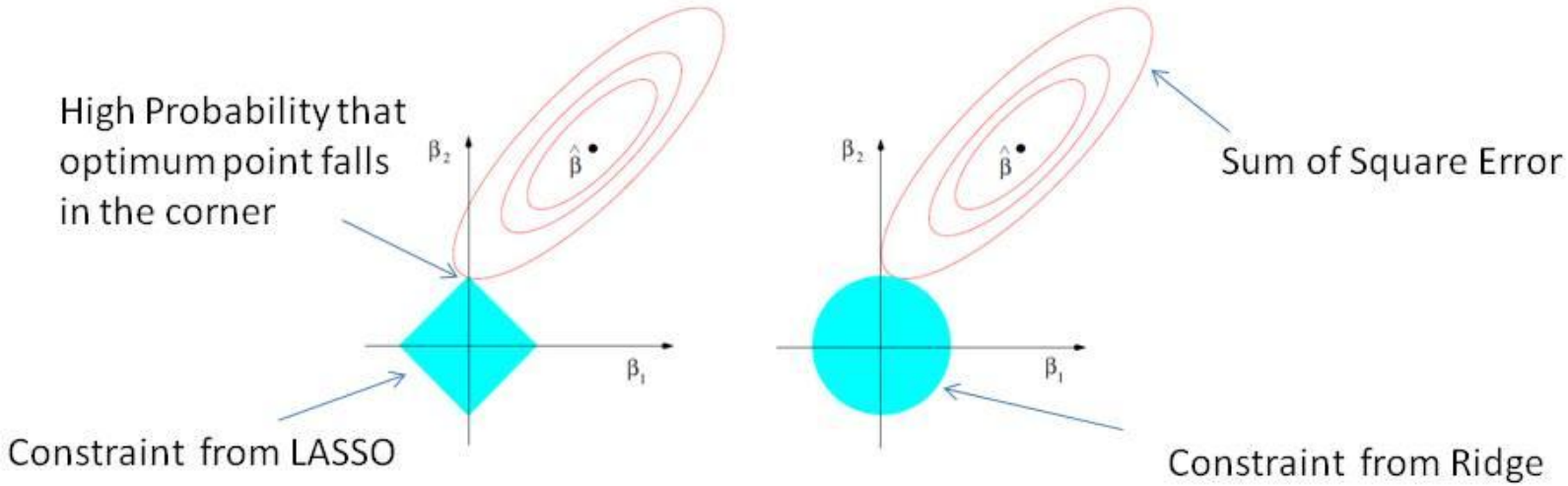
$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2$$

$$\text{subject to } \sum_{j=1}^{p} |\beta_j| \leq t.$$

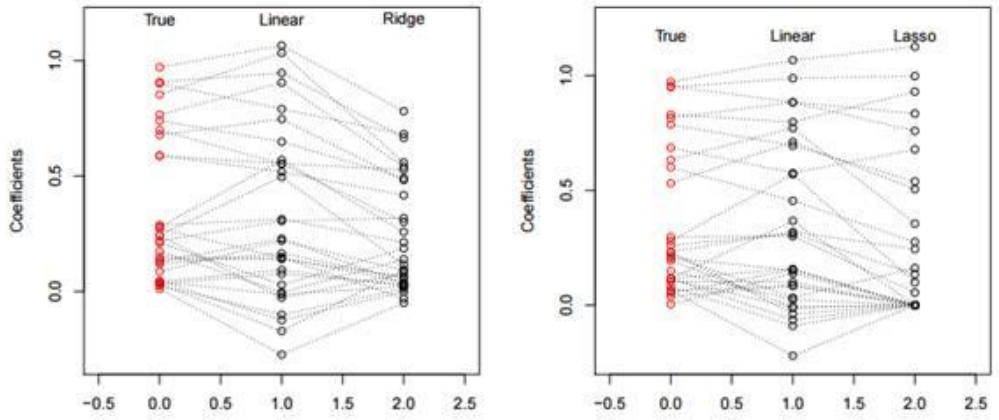
Solutions nonlinear in response variable, there is no closed form expression

13

Coefficients for Ridge and LASSO Regression (I)

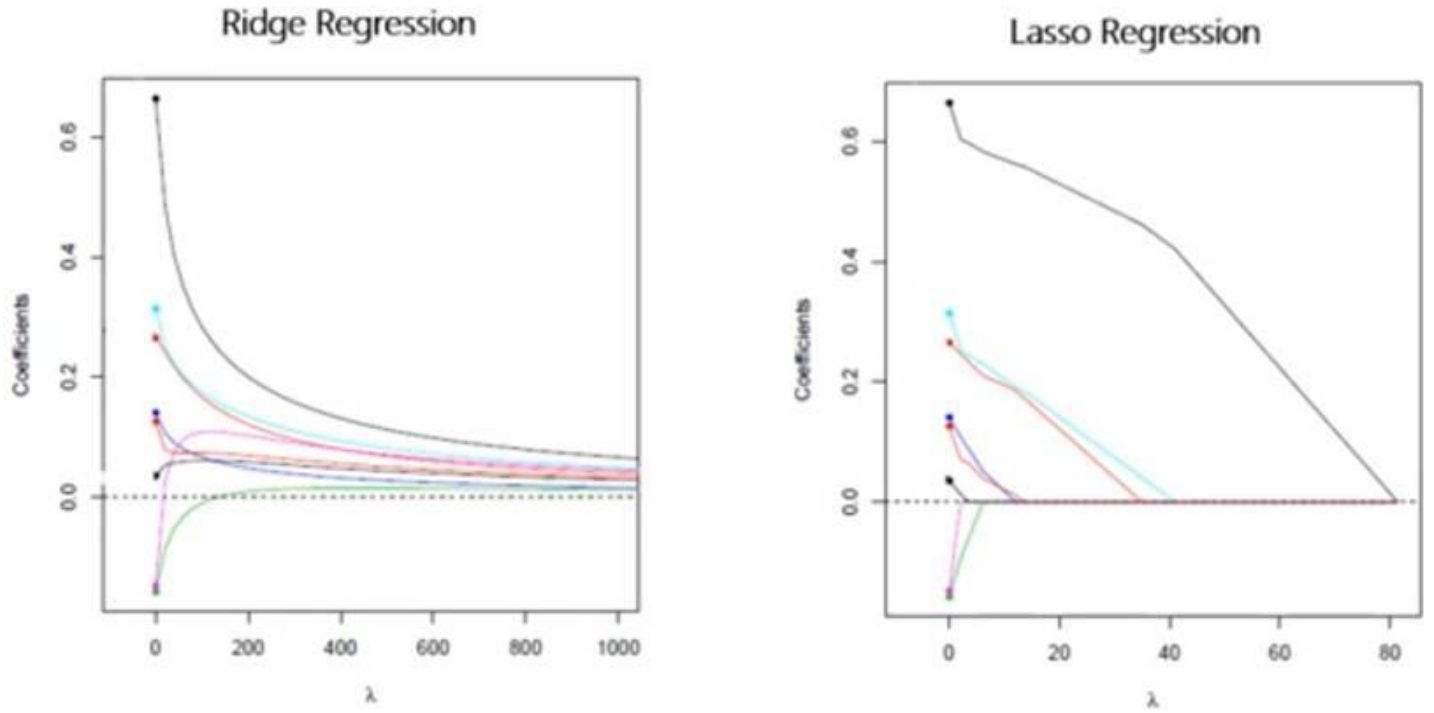


Minimize Sum of Square while meeting constraints



14

Coefficients for Ridge and LASSO Regression (II)



- When λ reduces to 0, they become OLS
- When λ increases, more regularization impact. Lasso zero coefficients eventually, Ridge Just reduces coefficients and saturates.

15

Issues/Solution for LASSO

Issues:

- If a group of predictors are highly correlated among themselves, LASSO tends to pick only one of them and shrink the other to zero
- For p > n problem, LASSO at most selects n features

Solution:

- Combine LASSO and Ridge

16

Elastic Net

The optimization problem for Naive Elastic Net is

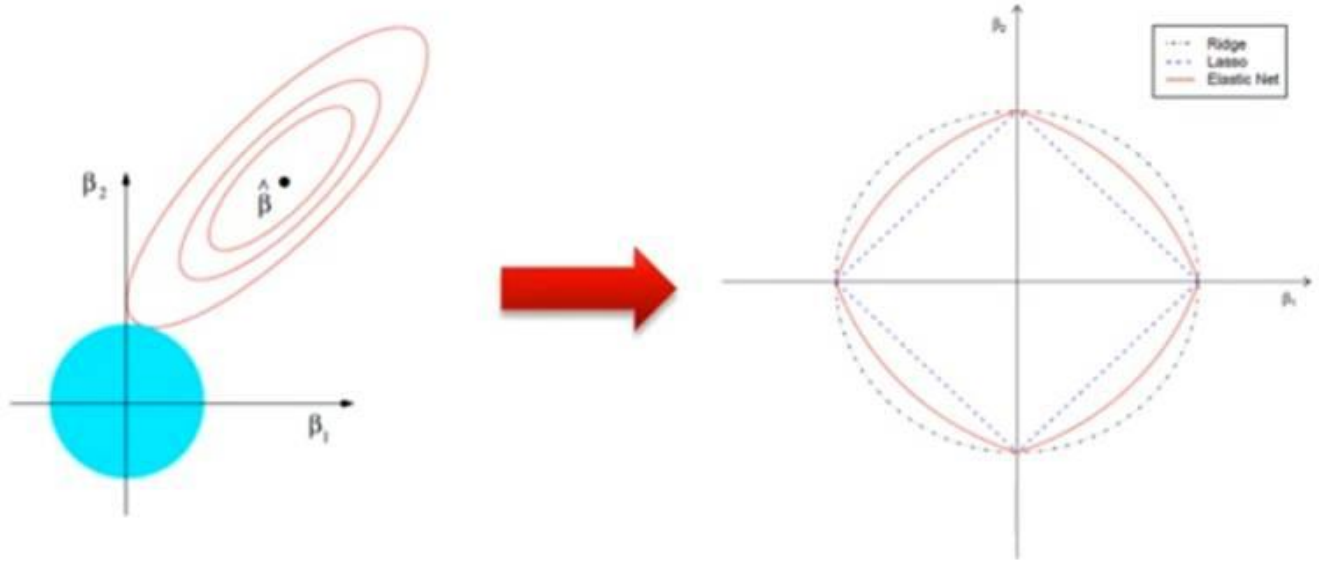
$$\hat{\boldsymbol{\beta}}(\text{Naive ENet}) = \arg\min_{\boldsymbol{\beta}} \quad \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda_1|\boldsymbol{\beta}|_1 + \lambda_2\|\boldsymbol{\beta}\|^2$$

- λ_1 and λ_2 are positive weights. Naive Elastic Net has a combined l_1 and l_2 penalty.
- $\lambda_1 \to 0$, Ridge regression; $\lambda_2 \to 0$, LASSO.
- Deficiency of the Naive Elastic Net:
  Empirical evidence shows the Naive Elastic Net does not perform satisfactorily.
  The reason is that there are two shrinkage procedures (Ridge and LASSO) in it.
  Double shrinkage introduces unnecessary bias.
- Re-scaling of Naive Elastic Net gives better performance, yielding the Elastic Net solution:

$$\hat{\boldsymbol{\beta}}(\text{ENet}) = (1 + \lambda_2) \cdot \hat{\boldsymbol{\beta}}(\text{Naive ENet})$$

17

Elastic Net - Constraints



18

Summary

- Ridge Regression:
  - Good for multicollinearity and grouped selection
  - Not good for variable selection
- LASSO
  - Good for variable selection
  - Not good for grouped selection or strongly correlated predictors
- Elastic Net
  - Combine strength of Ridge Regression and LASSO
- Regularization:
  - Trade bias for variance reduction
  - Better prediction accuracy

19

Appendix

OLS - Unbiased

$$\hat{\beta} = (X^T X)^{-1} X^T (X\beta + U) = (X^T X)^{-1} X^T X\beta + (X^T X)^{-1} X^T U$$
$$= I\beta + (X^T X)^{-1} X^T U$$
$$= \beta + (X^T X)^{-1} X^T U$$

$$E(\hat{\beta}) = E[\beta + (X^T X)^{-1} X^T U]$$
$$= E(\beta) + E[(X^T X)^{-1} X^T U]$$
$$= \beta + (X^T X)^{-1} X^T E(U)$$
$$= \beta$$

21

Bias Variance Decomposition

$$
\begin{aligned}
\text{MSE} &= E_{\mathbf{D}_N}[(\theta - \hat{\boldsymbol{\theta}})^2] = E_{\mathbf{D}_N}[(\theta - E[\hat{\boldsymbol{\theta}}] + E[\hat{\boldsymbol{\theta}}] - \hat{\boldsymbol{\theta}})^2] \\
&= E_{\mathbf{D}_N}[(\theta - E[\hat{\boldsymbol{\theta}}])^2] + E_{\mathbf{D}_N}[(E[\hat{\boldsymbol{\theta}}] - \hat{\boldsymbol{\theta}})^2] + E_{\mathbf{D}_N}[2(\theta - E[\hat{\boldsymbol{\theta}}])(E[\hat{\boldsymbol{\theta}}] - \hat{\boldsymbol{\theta}})] \\
&= E_{\mathbf{D}_N}[(\theta - E[\hat{\boldsymbol{\theta}}])^2] + E_{\mathbf{D}_N}[(E[\hat{\boldsymbol{\theta}}] - \hat{\boldsymbol{\theta}})^2] + 2(\theta - E[\hat{\boldsymbol{\theta}}])(E[\hat{\boldsymbol{\theta}}] - E[\hat{\boldsymbol{\theta}}]) \\
&= (E[\hat{\boldsymbol{\theta}}] - \theta)^2 + \text{Var}\left[\hat{\boldsymbol{\theta}}\right]
\end{aligned}
$$

22