Data Mining & Machine Learning

Regularization Methods for Regression

Francisco Javier Arceo
Data Scientist

NYU Polytechnic School of Engineering
Commonwealth Bank of Australia

March 23, 2015

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Motivation
Framework
Similarities

Ridge Regression

$$\hat{\beta}_{ridge} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \qquad (1)$$

$$\hat{\boldsymbol{\beta}}_{ridge} = \underset{\beta}{\operatorname{argmin}}(\boldsymbol{y} - \boldsymbol{x}^T \boldsymbol{\beta})^2 + \boldsymbol{\lambda}\boldsymbol{\beta}^2 \qquad (2)$$

▶ First published by the Russian Mathematician Andrey
  Nikolayevich Tikhonov (1906-1993) in 1943

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Motivation
Framework
Similarities

Ridge Regression

$$\hat{\beta}_{ridge} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \qquad (1)$$

$$\hat{\boldsymbol{\beta}}_{ridge} = \underset{\beta}{\operatorname{argmin}} (\boldsymbol{y} - \boldsymbol{x}^T \boldsymbol{\beta})^2 + \boldsymbol{\lambda} \boldsymbol{\beta}^2 \qquad (2)$$

▶ First published by the Russian Mathematician Andrey Nikolayevich Tikhonov (1906-1993) in 1943

▶ Originally called Tikhonov regularization

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Motivation
Framework
Similarities

Ridge Regression

$$\hat{\beta}_{ridge} = \operatorname*{argmin}_{\beta} \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \qquad (1)$$

$$\hat{\boldsymbol{\beta}}_{ridge} = \operatorname*{argmin}_{\beta}(\boldsymbol{y} - \boldsymbol{x}^T\boldsymbol{\beta})^2 + \boldsymbol{\lambda}\boldsymbol{\beta}^2 \qquad (2)$$

▶ First published by the Russian Mathematician Andrey Nikolayevich Tikhonov (1906-1993) in 1943

▶ Originally called Tikhonov regularization

▶ First introduced to the statistical literature in 1970 by Hoerl and Kennard (*Technometrics*)

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Motivation
Framework
Similarities

Ridge Regression

$$\hat{\beta}_{ridge} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \qquad (1)$$

$$\hat{\boldsymbol{\beta}}_{ridge} = \underset{\beta}{\operatorname{argmin}}(\boldsymbol{y} - \boldsymbol{x}^T\boldsymbol{\beta})^2 + \boldsymbol{\lambda}\boldsymbol{\beta}^2 \qquad (2)$$

- ▶ First published by the Russian Mathematician Andrey Nikolayevich Tikhonov (1906-1993) in 1943
- ▶ Originally called Tikhonov regularization
- ▶ First introduced to the statistical literature in 1970 by Hoerl and Kennard (*Technometrics*)
- ▶ Method to deal with non-invertible **(X'X)** matrices

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Motivation
Framework
Similarities

Ridge Regression

$$\hat{\beta}_{j,ridge} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(y_i - \bar{y})}{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2} \frac{1}{1 + \lambda} = \frac{\hat{\beta}_{j,ols}}{1 + \lambda} \qquad (3)$$

$$\hat{\boldsymbol{\beta}}_{ridge} = (\boldsymbol{X'X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X'y} \qquad (4)$$

▶ Equation (3) is a special case of an orthonormal design matrix

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Motivation
Framework
Similarities

Ridge Regression

$$\hat{\beta}_{j,ridge} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(y_i - \bar{y})}{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2} \frac{1}{1+\lambda} = \frac{\hat{\beta}_{j,ols}}{1+\lambda} \qquad (3)$$

$$\hat{\beta}_{ridge} = (X'X + \lambda I)^{-1}X'y \qquad (4)$$

▶ Equation (3) is a special case of an orthonormal design matrix

▶ λ is a penalty term that is added to the diagonal elements of $(X'X)$ and I is an (nxn) matrix

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Motivation
Framework
Similarities

Ridge Regression

$$\hat{\beta}_{j,ridge} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(y_i - \bar{y})}{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}\frac{1}{1 + \lambda} = \frac{\hat{\beta}_{j,ols}}{1 + \lambda} \tag{3}$$

$$\hat{\beta}_{ridge} = (X'X + \lambda I)^{-1}X'y \tag{4}$$

- Equation (3) is a special case of an orthonormal design matrix
- λ is a penalty term that is added to the diagonal elements of $(X'X)$ and I is an (nxn) matrix
- Question: what happens when $\lambda{=}0$?

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Motivation
Framework
Similarities

Ridge Regression

$$\hat{\beta}_{j,ridge} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(y_i - \bar{y})}{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2} \frac{1}{1+\lambda} = \frac{\hat{\beta}_{j,ols}}{1+\lambda} \qquad (3)$$

$$\hat{\boldsymbol{\beta}}_{ridge} = (\boldsymbol{X'X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X'y} \qquad (4)$$

- ▶ Equation (3) is a special case of an orthonormal design matrix
- ▶ λ is a penalty term that is added to the diagonal elements of $(\boldsymbol{X'X})$ and $\boldsymbol{I}$ is an (nxn) matrix
- ▶ Question: what happens when $\lambda=0$?
- ▶ Back to Ordinary Least-Squares solution!

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Motivation
Framework
Similarities

Ridge Regression

$$\hat{\boldsymbol{\beta}}_{ridge} = \underset{\beta}{\operatorname{argmin}}(\boldsymbol{y} - \boldsymbol{x}^T\boldsymbol{\beta})^2 + \boldsymbol{\lambda}\boldsymbol{\beta}^2 \tag{5}$$

$$\hat{\boldsymbol{\beta}}_{ridge} = \underset{\beta}{\operatorname{argmin}}(\sum_{i=1}^{n}(y_i - x_i^T\boldsymbol{\beta})^2) \ \ s.t. ||\boldsymbol{\beta}||_2 \leq t \tag{6}$$

▶ Equation (5) is equivalent to solving the constrained optimization problem in equation (6)

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Motivation
Framework
Similarities

Ridge Regression

$$\hat{\boldsymbol{\beta}}_{ridge} = \underset{\beta}{\operatorname{argmin}}(\mathbf{y} - \mathbf{x}^T\boldsymbol{\beta})^2 + \lambda\boldsymbol{\beta}^2 \tag{5}$$

$$\hat{\boldsymbol{\beta}}_{ridge} = \underset{\beta}{\operatorname{argmin}}(\sum_{i=1}^{n}(y_i - x_i^T\boldsymbol{\beta})^2) \ s.t. ||\boldsymbol{\beta}||_2 \leq t \tag{6}$$

- Equation (5) is equivalent to solving the constrained optimization problem in equation (6)
- Note that the $||\boldsymbol{\beta}||_2$ is the 2-norm of the vector $\boldsymbol{\beta}$ and $t \geq 0$

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Motivation
Framework
Similarities

Ridge Regression

$$\hat{\boldsymbol{\beta}}_{ridge} = \underset{\beta}{\operatorname{argmin}}(\boldsymbol{y} - \boldsymbol{x}^T\boldsymbol{\beta})^2 + \lambda\boldsymbol{\beta}^2 \qquad (5)$$

$$\hat{\boldsymbol{\beta}}_{ridge} = \underset{\beta}{\operatorname{argmin}}(\sum_{i=1}^{n}(y_i - x_i^T\boldsymbol{\beta})^2) \ s.t.||\boldsymbol{\beta}||_2 \leq t \qquad (6)$$

▶ Equation (5) is equivalent to solving the constrained optimization problem in equation (6)

▶ Note that the $||\boldsymbol{\beta}||_2$ is the 2-norm of the vector $\boldsymbol{\beta}$ and $t \geq 0$

▶ What happens when we change the norm? What about the L1 norm?

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Similarities
Motivation
Framework

LASSO Regression

$$\hat{\boldsymbol{\beta}}_{ridge} = \underset{\beta}{\operatorname{argmin}}(\sum_{i=1}^{n}(y_i - x_i^T\boldsymbol{\beta})^2) \ s.t.||\boldsymbol{\beta}||_2 \leq t \qquad (7)$$

$$\hat{\boldsymbol{\beta}}_{lasso} = \underset{\beta}{\operatorname{argmin}}(\sum_{i=1}^{n}(y_i - x_i^T\boldsymbol{\beta})^2) \ s.t.||\boldsymbol{\beta}||_1 \leq t \qquad (8)$$

▶ $||\boldsymbol{\beta}||_1 = \sum_{j=1}^{p}|\beta_j|$, $||\boldsymbol{\beta}||_2 = \sum_{j=1}^{p}\beta_j^2$, and $p :=$ the number of features

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Similarities
Motivation
Framework

LASSO Regression

$$\hat{\beta}_{ridge} = \underset{\beta}{\operatorname{argmin}}(\sum_{i=1}^{n}(y_i - x_i^T\beta)^2) \ s.t. ||\beta||_2 \leq t \qquad (7)$$

$$\hat{\beta}_{lasso} = \underset{\beta}{\operatorname{argmin}}(\sum_{i=1}^{n}(y_i - x_i^T\beta)^2) \ s.t. ||\beta||_1 \leq t \qquad (8)$$

- $||\beta||_1 = \sum_{j=1}^{p} |\beta_j|$, $||\beta||_2 = \sum_{j=1}^{p} \beta_j^2$, and $p :=$ the number of features
- $||\beta||_1$ is the unit-norm of the vector β and $t \geq 0$

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Similarities
Motivation
Framework

LASSO Regression

$$\hat{\beta}_{ridge} = \underset{\beta}{\operatorname{argmin}}(\sum_{i=1}^{n}(y_i - x_i^T\beta)^2) \ \ s.t. ||\beta||_2 \leq t \qquad (7)$$

$$\hat{\beta}_{lasso} = \underset{\beta}{\operatorname{argmin}}(\sum_{i=1}^{n}(y_i - x_i^T\beta)^2) \ \ s.t. ||\beta||_1 \leq t \qquad (8)$$

▶ $||\beta||_1 = \sum_{j=1}^{p}|\beta_j|$, $||\beta||_2 = \sum_{j=1}^{p}\beta_j^2$, and $p :=$ the number of features

▶ $||\beta||_1$ is the unit-norm of the vector β and $t \geq 0$

▶ Difference is in the constraint of the minimization problem

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Similarities
Motivation
Framework

LASSO Regression

$$\hat{\boldsymbol{\beta}}_{lasso} = \underset{\beta}{\text{argmin}}(\sum_{i=1}^{n}(y_i - x_i^T\boldsymbol{\beta})^2) \ \ s.t. ||\boldsymbol{\beta}||_1 \leq t \qquad (9)$$

$$\hat{\boldsymbol{\beta}}_{lasso} = \underset{\beta}{\text{argmin}}(\boldsymbol{y} - \boldsymbol{x}^T\boldsymbol{\beta})^2 + \boldsymbol{\lambda}|\boldsymbol{\beta}| \qquad (10)$$

▶ Equation (9) and (10) lead to a minimization similar to equation (2) from Ridge, except the penalization to β is linear instead of quadratic

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Similarities
Motivation
Framework

LASSO Regression

$$\hat{\boldsymbol{\beta}}_{lasso} = \underset{\beta}{\operatorname{argmin}}(\sum_{i=1}^{n}(y_i - x_i^T\boldsymbol{\beta})^2) \ s.t. ||\boldsymbol{\beta}||_1 \leq t \qquad (9)$$

$$\hat{\boldsymbol{\beta}}_{lasso} = \underset{\beta}{\operatorname{argmin}}(\boldsymbol{y} - \boldsymbol{x}^T\boldsymbol{\beta})^2 + \boldsymbol{\lambda}|\boldsymbol{\beta}| \qquad (10)$$

▶ Equation (9) and (10) lead to a minimization similar to
  equation (2) from Ridge, except the penalization to β is linear
  instead of quadratic

▶ What's the benefit of using the L1-norm versus the L2-norm?

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Similarities
Motivation
Framework

LASSO Regression

$$\hat{\boldsymbol{\beta}}_{lasso} = \underset{\beta}{\operatorname{argmin}}(\sum_{i=1}^{n}(y_i - x_i^T\boldsymbol{\beta})^2) \ s.t.||\boldsymbol{\beta}||_1 \leq t \qquad (9)$$

$$\hat{\boldsymbol{\beta}}_{lasso} = \underset{\beta}{\operatorname{argmin}}(\boldsymbol{y} - \boldsymbol{x}^T\boldsymbol{\beta})^2 + \boldsymbol{\lambda}|\boldsymbol{\beta}| \qquad (10)$$

▶ Equation (9) and (10) lead to a minimization similar to equation (2) from Ridge, except the penalization to β is linear instead of quadratic

▶ What's the benefit of using the L1-norm versus the L2-norm?

▶ Sparsity!

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Similarities
Motivation
Framework

LASSO Regression

$$\hat{\boldsymbol{\beta}}_{lasso} = \underset{\beta}{\operatorname{argmin}}(\sum_{i=1}^{n}(y_i - x_i^T\boldsymbol{\beta})^2) \ s.t. ||\boldsymbol{\beta}||_1 \leq t \qquad (9)$$

$$\hat{\boldsymbol{\beta}}_{lasso} = \underset{\beta}{\operatorname{argmin}}(\boldsymbol{y} - \boldsymbol{x}^T\boldsymbol{\beta})^2 + \boldsymbol{\lambda}|\boldsymbol{\beta}| \qquad (10)$$

▶ Equation (9) and (10) lead to a minimization similar to equation (2) from Ridge, except the penalization to β is linear instead of quadratic

▶ What's the benefit of using the L1-norm versus the L2-norm?

▶ Sparsity!

▶ Using LASSO results in models with less parameters (features)

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Similarities
Motivation
Framework

LASSO Regression

$$\hat{\beta}_{lasso}(\lambda) = S(\hat{\beta}, \lambda) \equiv sign(\hat{\beta})(|\hat{\beta}| - \lambda)_{+} \qquad (11)$$

$$S(\hat{\beta}, \lambda) = \begin{cases} \hat{\beta} - \lambda, & \text{if } \hat{\beta} > 0 \text{ and } \lambda < |\hat{\beta}| \\ \hat{\beta} + \lambda, & \text{if } \hat{\beta} < 0 \text{ and } \lambda < |\hat{\beta}| \\ 0, & \text{if } \lambda > |\hat{\beta}|. \end{cases}$$

$$\hat{\beta}_{j,lasso}^{t} \leftarrow S(\hat{\beta}_{j,lasso}^{t-1} + \sum_{i=1}^{n} x_{ij}(y_i - \hat{y}_i^{t-1}), \lambda) \qquad (12)$$

▶ Equation (11) is called the soft-threshold operator

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Similarities
Motivation
Framework

LASSO Regression

$$\hat{\beta}_{lasso}(\lambda) = S(\hat{\beta}, \lambda) \equiv sign(\hat{\beta})(|\hat{\beta}| - \lambda)_+ \tag{11}$$

$$S(\hat{\beta}, \lambda) = \begin{cases} \hat{\beta} - \lambda, & if \, \hat{\beta} > 0 \text{ and } \lambda < |\hat{\beta}| \\ \hat{\beta} + \lambda, & if \, \hat{\beta} < 0 \text{ and } \lambda < |\hat{\beta}| \\ 0, & if \, \lambda > |\hat{\beta}|. \end{cases}$$

$$\hat{\beta}_{j,lasso}^t \leftarrow S(\hat{\beta}_{j,lasso}^{t-1} + \sum_{i=1}^n x_{ij}(y_i - \hat{y}_i^{t-1}), \lambda) \tag{12}$$

▶ Equation (11) is called the soft-threshold operator
▶ Apply soft-threshold operator as we cycle through our features at the tth iteration (12) updating the residual along the way

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Similarities
Motivation
Framework

LASSO Regression

$$\hat{\beta}_{lasso}(\lambda) = S(\hat{\beta}, \lambda) \equiv sign(\hat{\beta})(|\hat{\beta}| - \lambda)_+ \tag{11}$$

$$S(\hat{\beta}, \lambda) = \begin{cases} \hat{\beta} - \lambda, & \text{if } \hat{\beta} > 0 \text{ and } \lambda < |\hat{\beta}| \\ \hat{\beta} + \lambda, & \text{if } \hat{\beta} < 0 \text{ and } \lambda < |\hat{\beta}| \\ 0, & \text{if } \lambda > |\hat{\beta}|. \end{cases}$$

$$\hat{\beta}_{j,lasso}^t \leftarrow S(\hat{\beta}_{j,lasso}^{t-1} + \sum_{i=1}^{n} x_{ij}(y_i - \hat{y}_i^{t-1}), \lambda) \tag{12}$$

▶ Equation (11) is called the soft-threshold operator
▶ Apply soft-threshold operator as we cycle through our features at the tth iteration (12) updating the residual along the way
▶ LASSO does variable selection by setting select variables to 0!

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Important Facts
Empirical Example: Kaggle's Amazon Competition
Amazon Data:LASSO Model
Amazon Data: Ridge Model
Amazon Data: Performance

LASSO and Ridge Regression

▶ How do we choose λ? **Cross Validation**

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Important Facts
Empirical Example: Kaggle's Amazon Competition
Amazon Data:LASSO Model
Amazon Data: Ridge Model
Amazon Data: Performance

LASSO and Ridge Regression

- ▶ How do we choose λ? **Cross Validation**
- ▶ When do we use Ridge vs LASSO? When we want a model with less coefficients (predictors) use LASSO

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Important Facts
Empirical Example: Kaggle's Amazon Competition
Amazon Data:LASSO Model
Amazon Data: Ridge Model
Amazon Data: Performance

LASSO and Ridge Regression

- ▶ How do we choose λ? **Cross Validation**
- ▶ When do we use Ridge vs LASSO? When we want a model with less coefficients (predictors) use LASSO
- ▶ LASSO can be used when p $>>$ n

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Important Facts
Empirical Example: Kaggle's Amazon Competition
Amazon Data:LASSO Model
Amazon Data: Ridge Model
Amazon Data: Performance

LASSO and Ridge Regression

- ▶ How do we choose λ? **Cross Validation**
- ▶ When do we use Ridge vs LASSO? When we want a model with less coefficients (predictors) use LASSO
- ▶ LASSO can be used when $p >> n$
- ▶ Both methods reduce the size of the coefficients but Ridge shrinks to 0 continuously; LASSO shrinks discretely

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Important Facts
Empirical Example: Kaggle's Amazon Competition
Amazon Data:LASSO Model
Amazon Data: Ridge Model
Amazon Data: Performance

LASSO and Ridge Regression

- ▶ How do we choose λ? **Cross Validation**
- ▶ When do we use Ridge vs LASSO? When we want a model with less coefficients (predictors) use LASSO
- ▶ LASSO can be used when $p >> n$
- ▶ Both methods reduce the size of the coefficients but Ridge shrinks to 0 continuously; LASSO shrinks discretely
- ▶ Both biased in the statistical sense but asymptotically unbiased and can be shown that they select the true features with probability $\rightarrow 1$

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Important Facts
Empirical Example: Kaggle's Amazon Competition
Amazon Data:LASSO Model
Amazon Data: Ridge Model
Amazon Data: Performance

LASSO and Ridge Regression

▶ Ridge and Lasso can be developed using a Bayesian framework

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Important Facts
Empirical Example: Kaggle's Amazon Competition
Amazon Data:LASSO Model
Amazon Data: Ridge Model
Amazon Data: Performance

LASSO and Ridge Regression

- ▶ Ridge and Lasso can be developed using a Bayesian framework
- ▶ Equivalent to setting a Gaussian Prior and Laplace prior on each coefficient for Ridge and Lasso, respectively

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Important Facts
Empirical Example: Kaggle's Amazon Competition
Amazon Data:LASSO Model
Amazon Data: Ridge Model
Amazon Data: Performance

LASSO and Ridge Regression

- ▶ Ridge and Lasso can be developed using a Bayesian framework
- ▶ Equivalent to setting a Gaussian Prior and Laplace prior on each coefficient for Ridge and Lasso, respectively
- ▶ Both are convex optimization problems where Ridge has a closed form solution and LASSO must be solved iteratively

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Important Facts
Empirical Example: Kaggle's Amazon Competition
Amazon Data:LASSO Model
Amazon Data: Ridge Model
Amazon Data: Performance

LASSO and Ridge Regression

- ▶ Ridge and Lasso can be developed using a Bayesian framework
- ▶ Equivalent to setting a Gaussian Prior and Laplace prior on each coefficient for Ridge and Lasso, respectively
- ▶ Both are convex optimization problems where Ridge has a closed form solution and LASSO must be solved iteratively
- ▶ L1 and L2-Norm can be extended to L-P Norm

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Important Facts
Empirical Example: Kaggle's Amazon Competition
Amazon Data:LASSO Model
Amazon Data: Ridge Model
Amazon Data: Performance

LASSO and Ridge Regression

- ▶ Ridge and Lasso can be developed using a Bayesian framework
- ▶ Equivalent to setting a Gaussian Prior and Laplace prior on each coefficient for Ridge and Lasso, respectively
- ▶ Both are convex optimization problems where Ridge has a closed form solution and LASSO must be solved iteratively
- ▶ L1 and L2-Norm can be extended to L-P Norm
- ▶ P-norms < 1 yield sparser models

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Important Facts
Empirical Example: Kaggle's Amazon Competition
Amazon Data:LASSO Model
Amazon Data: Ridge Model
Amazon Data: Performance

LASSO and Ridge Regression

- Ridge and Lasso can be developed using a Bayesian framework
- Equivalent to setting a Gaussian Prior and Laplace prior on each coefficient for Ridge and Lasso, respectively
- Both are convex optimization problems where Ridge has a closed form solution and LASSO must be solved iteratively
- L1 and L2-Norm can be extended to L-P Norm
- P-norms < 1 yield sparser models
- Many different ways to regularize parameters, extensive literature exists

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Important Facts
Empirical Example: Kaggle's Amazon Competition
Amazon Data:LASSO Model
Amazon Data: Ridge Model
Amazon Data: Performance

LASSO and Ridge Regression

- ▶ Ridge and Lasso can be developed using a Bayesian framework
- ▶ Equivalent to setting a Gaussian Prior and Laplace prior on each coefficient for Ridge and Lasso, respectively
- ▶ Both are convex optimization problems where Ridge has a closed form solution and LASSO must be solved iteratively
- ▶ L1 and L2-Norm can be extended to L-P Norm
- ▶ P-norms < 1 yield sparser models
- ▶ Many different ways to regularize parameters, extensive literature exists
- ▶ Useful in banking, tech, insurance, marketing, and other fields

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Important Facts
Empirical Example: Kaggle's Amazon Competition
Amazon Data:LASSO Model
Amazon Data: Ridge Model
Amazon Data: Performance

LASSO and Ridge Regression

▶ Amazon Kaggle competition: Employee Access Challenge

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Important Facts
Empirical Example: Kaggle's Amazon Competition
Amazon Data:LASSO Model
Amazon Data: Ridge Model
Amazon Data: Performance

LASSO and Ridge Regression

- ▶ Amazon Kaggle competition: Employee Access Challenge
- ▶ Training data on 32,768 employee with

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Important Facts
Empirical Example: Kaggle's Amazon Competition
Amazon Data:LASSO Model
Amazon Data: Ridge Model
Amazon Data: Performance

LASSO and Ridge Regression

- ▶ Amazon Kaggle competition: Employee Access Challenge
- ▶ Training data on 32,768 employee with
- ▶ Outcome/Label is binary outcome of whether an employee received access (Average Training Response = 0.94)

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Important Facts
Empirical Example: Kaggle's Amazon Competition
Amazon Data:LASSO Model
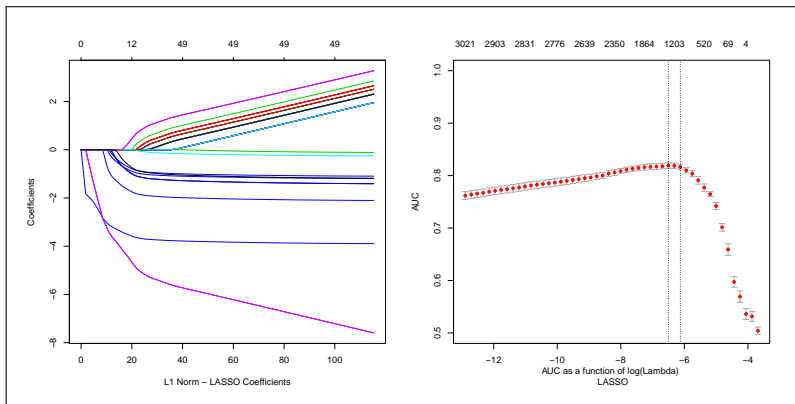Amazon Data: Ridge Model
Amazon Data: Performance

LASSO and Ridge Regression

- ▶ Amazon Kaggle competition: Employee Access Challenge
- ▶ Training data on 32,768 employee with
- ▶ Outcome/Label is binary outcome of whether an employee received access (Average Training Response = 0.94)
- ▶ 1,687 teams with most submissions performing reasonably well

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Important Facts
Empirical Example: Kaggle's Amazon Competition
Amazon Data:LASSO Model
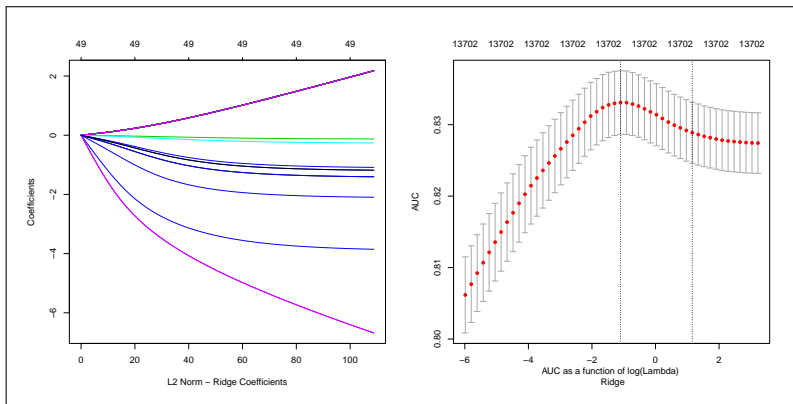Amazon Data: Ridge Model
Amazon Data: Performance

LASSO and Ridge Regression

- ▶ Amazon Kaggle competition: Employee Access Challenge
- ▶ Training data on 32,768 employee with
- ▶ Outcome/Label is binary outcome of whether an employee received access (Average Training Response $= 0.94$)
- ▶ 1,687 teams with most submissions performing reasonably well
- ▶ Features are all categorical variables and highly sparse

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Important Facts
Empirical Example: Kaggle's Amazon Competition
Amazon Data:LASSO Model
Amazon Data: Ridge Model
Amazon Data: Performance

LASSO and Ridge Regression

- ▶ Amazon Kaggle competition: Employee Access Challenge
- ▶ Training data on 32,768 employee with
- ▶ Outcome/Label is binary outcome of whether an employee received access (Average Training Response = 0.94)
- ▶ 1,687 teams with most submissions performing reasonably well
- ▶ Features are all categorical variables and highly sparse
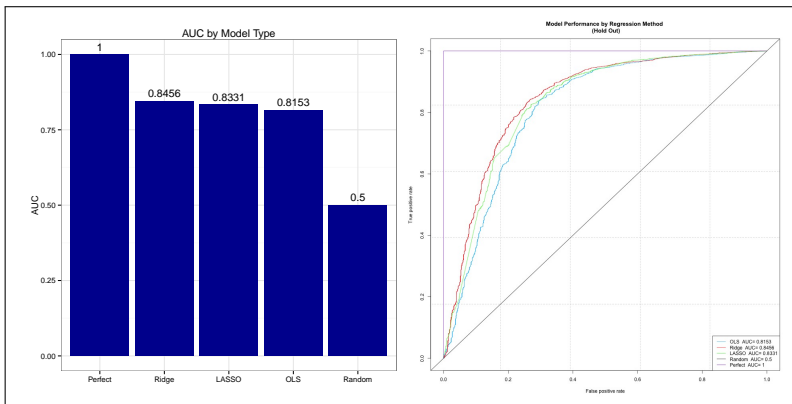- ▶ Binary representation of all 10 features leads to a very sparse (32,769 x 15,618) matrix

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Important Facts
Empirical Example: Kaggle's Amazon Competition
Amazon Data:LASSO Model
Amazon Data: Ridge Model
Amazon Data: Performance

LASSO Model Performance

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Important Facts
Empirical Example: Kaggle's Amazon Competition
Amazon Data:LASSO Model
Amazon Data: Ridge Model
Amazon Data: Performance

Ridge Model Performance

Regularization Methods for Regression
L2 Norm: Ridge
L1 Norm: Lasso
Regularization Methods
Some Resources

Important Facts
Empirical Example: Kaggle's Amazon Competition
Amazon Data:LASSO Model
Amazon Data: Ridge Model
Amazon Data: Performance

Comparison of Ridge, LASSO, and OLS

Useful Links below

Thank you.

- ▶ My Email (Not as Useful)

Useful Links below

Thank you.

- ▶ My Email (Not as Useful)
- ▶ GitHub (Moderately Useful)

Useful Links below

Thank you.

- ► My Email (Not as Useful)
- ► GitHub (Moderately Useful)
- ► GLMNET in R (Very Useful)

Useful Links below

Thank you.

- ▶ My Email (Not as Useful)
- ▶ GitHub (Moderately Useful)
- ▶ GLMNET in R (Very Useful)
- ▶ SKLearn (Also Very Useful)

Useful Links below

Thank you.

- ▶ My Email (Not as Useful)
- ▶ GitHub (Moderately Useful)
- ▶ GLMNET in R (Very Useful)
- ▶ SKLearn (Also Very Useful)
- ▶ Original LASSO Publication (REALLY Useful)

Useful Links below

Thank you.

- ▶ My Email (Not as Useful)
- ▶ GitHub (Moderately Useful)
- ▶ GLMNET in R (Very Useful)
- ▶ SKLearn (Also Very Useful)
- ▶ Original LASSO Publication (REALLY Useful)
- ▶ Original Ridge Publication (Somewhat useful)

Useful Links below

Thank you.

- ▶ My Email (Not as Useful)
- ▶ GitHub (Moderately Useful)
- ▶ GLMNET in R (Very Useful)
- ▶ SKLearn (Also Very Useful)
- ▶ Original LASSO Publication (REALLY Useful)
- ▶ Original Ridge Publication (Somewhat useful)
- ▶ Slides from Stanford (Most Useful)

Useful Links below

Thank you.

- ▶ My Email (Not as Useful)
- ▶ GitHub (Moderately Useful)
- ▶ GLMNET in R (Very Useful)
- ▶ SKLearn (Also Very Useful)
- ▶ Original LASSO Publication (REALLY Useful)
- ▶ Original Ridge Publication (Somewhat useful)
- ▶ Slides from Stanford (Most Useful)
- ▶ Sweet paper on asymptotic properties of Ridge and LASSO