LASSO

Nate Wells

Math 243: Stat Learning

October 30th, 2020

Outline

In today's class, we will. . .

- Discuss LASSO as a method of penalized regression AND variable selection

Section 1

The LASSO

Metrics on R^p

How can we measure the distance of a point $x = (x_1, \dots, x_p) \in \mathbb{R}^p$ from the origin?

Metrics on R^p

How can we measure the distance of a point $x = (x_1, \dots, x_p) \in \mathbb{R}^p$ from the origin?

- A natural measurement is the Euclidean distance (i.e. the Pythagorean formula), or the ℓ_2 norm:

$$\|x\|_2 = \sqrt{x_1^2 + \cdots + x_p^2} = \sqrt{\sum_{i=1}^{p} x_i^2}$$

Metrics on R^p

How can we measure the distance of a point $x = (x_1, \ldots, x_p) \in \mathbb{R}^p$ from the origin?

- A natural measurement is the Euclidean distance (i.e. the Pythagorean formula), or the ℓ_2 norm:

$$\|x\|_2 = \sqrt{x_1^2 + \cdots + x_p^2} = \sqrt{\sum_{i=1}^{p} x_i^2}$$

- An alternative measurement is to use the sum of magnitudes of the coordinates (called the taxi-cab metric), or the ℓ_1 norm:

$$\|x\|_1 = |x_1| + \cdots + |x_p| = \sum_{i=1}^{p} |x_i|$$

Metrics on R^p

How can we measure the distance of a point $x = (x_1, \dots, x_p) \in \mathbb{R}^p$ from the origin?

- A natural measurement is the Euclidean distance (i.e. the Pythagorean formula), or the ℓ_2 norm:

$$\|x\|_2 = \sqrt{x_1^2 + \cdots + x_p^2} = \sqrt{\sum_{i=1}^{p} x_i^2}$$

- An alternative measurement is to use the sum of magnitudes of the coordinates (called the taxi-cab metric), or the ℓ_1 norm:

$$\|x\|_1 = |x_1| + \cdots + |x_p| = \sum_{i=1}^{p} |x_i|$$

- Sometimes, its useful to consider the ℓ_0 "norm" and ℓ_∞ norm

$$\|x\|_0 = \#(x_i \neq 0) \qquad \|x\|_\infty = \max |\beta_i|$$

LASSO

In ridge regression, we seek parameters β that minimize RSS plus the ℓ_2 norm of β:

$$\mathrm{RSS} + \lambda \sum_{i=1}^{p} \beta_i^2 = \mathrm{RSS} + \lambda \|\beta\|_2$$

LASSO

In ridge regression, we seek parameters β that minimize RSS plus the ℓ_2 norm of β:

$$\mathrm{RSS} + \lambda \sum_{i=1}^{p} \beta_i^2 = \mathrm{RSS} + \lambda \|\beta\|_2$$

Alternatively, we could seek parameters β that minimize RSS plus the ℓ_1 norm of β:

$$\mathrm{RSS} + \lambda \sum_{i=1}^{p} |\beta_i| = \mathrm{RSS} + \lambda \|\beta\|_1$$

This latter method is called the LASSO (least absolute shrinkage and selection operator)

LASSO

In ridge regression, we seek parameters β that minimize RSS plus the ℓ_2 norm of β:

$$\mathrm{RSS} + \lambda \sum_{i=1}^{p} \beta_i^2 = \mathrm{RSS} + \lambda \|\beta\|_2$$

Alternatively, we could seek parameters β that minimize RSS plus the ℓ_1 norm of β:

$$\mathrm{RSS} + \lambda \sum_{i=1}^{p} |\beta_i| = \mathrm{RSS} + \lambda \|\beta\|_1$$

This latter method is called the LASSO (least absolute shrinkage and selection operator)

- In addition to shrinking coefficients, it also happens to perform variable selection!

Alternative Formulations

Instead of thinking of Ridge Regression and LASSO as minimizing the sum of RSS and the shrinkage penalty, we can think of them as solving a restricted optimization problem:

Alternative Formulations

Instead of thinking of Ridge Regression and LASSO as minimizing the sum of RSS and the shrinkage penalty, we can think of them as solving a restricted optimization problem:

- For each $s \geq 0$, Ridge Regression seeks to minimize RSS subject to $\|\beta\|_2 \leq s$
- For each $s \geq 0$, LASSO seeks to minimize RSS subject to $\|\beta\|_1 \leq s$

Alternative Formulations

Instead of thinking of Ridge Regression and LASSO as minimizing the sum of RSS and the shrinkage penalty, we can think of them as solving a restricted optimization problem:

- For each $s \geq 0$, Ridge Regression seeks to minimize RSS subject to $\|\beta\|_2 \leq s$
- For each $s \geq 0$, LASSO seeks to minimize RSS subject to $\|\beta\|_1 \leq s$

The best subset algorithm also fits in this paradigm:

- For each $s \geq 0$, best s-subset seeks to minimize RSS subject to $\|\beta\|_0 \leq s$

Alternative Formulations

Instead of thinking of Ridge Regression and LASSO as minimizing the sum of RSS and the shrinkage penalty, we can think of them as solving a restricted optimization problem:

- For each $s \geq 0$, Ridge Regression seeks to minimize RSS subject to $\|\beta\|_2 \leq s$
- For each $s \geq 0$, LASSO seeks to minimize RSS subject to $\|\beta\|_1 \leq s$

The best subset algorithm also fits in this paradigm:

- For each $s \geq 0$, best s-subset seeks to minimize RSS subject to $\|\beta\|_0 \leq s$

Suppose q is 0, 1, or 2. For each $\lambda \geq 0$, there is exactly one $s \geq 0$ so that if β minimizes

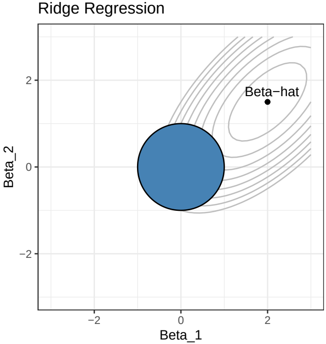
$$\mathrm{RSS} + \lambda \|\beta\|_q$$

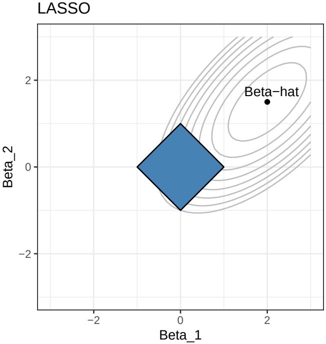
then β minimizes

$$\mathrm{RSS} \qquad \text{subject to } \|\beta\|_q \leq s$$

Variable Selection with LASSO

For LASSO, the solution to the optimization problem often lies on a vertex of the domain, which corresponds to a subspace where one or more parameters are 0.





Comparison of Penalized Regression Models

- Both ridge regression and LASSO can be implemented in R using `glmnet`. (Ridge regression uses `alpha = 0`, while LASSO uses `alpha = 1`)

Comparison of Penalized Regression Models

- Both ridge regression and LASSO can be implemented in R using `glmnet`. (Ridge regression uses `alpha = 0`, while LASSO uses `alpha = 1`)
- Both Ridge Regression and LASSO can be fit in about the same amount of time as ordinary least squares

Comparison of Penalized Regression Models

- Both ridge regression and LASSO can be implemented in R using `glmnet`. (Ridge regression uses `alpha = 0`, while LASSO uses `alpha = 1`)
- Both Ridge Regression and LASSO can be fit in about the same amount of time as ordinary least squares
- Both ridge regression and LASSO trade slightly increased bias for greatly reduced variance, compared to the full model.

Comparison of Penalized Regression Models

- Both ridge regression and LASSO can be implemented in R using `glmnet`. (Ridge regression uses `alpha = 0`, while LASSO uses `alpha = 1`)
- Both Ridge Regression and LASSO can be fit in about the same amount of time as ordinary least squares
- Both ridge regression and LASSO trade slightly increased bias for greatly reduced variance, compared to the full model.
- LASSO performs variable selection in addition to coefficient shrinkage

Comparison of Penalized Regression Models

- Both ridge regression and LASSO can be implemented in R using `glmnet`. (Ridge regression uses `alpha = 0`, while LASSO uses `alpha = 1`)
- Both Ridge Regression and LASSO can be fit in about the same amount of time as ordinary least squares
- Both ridge regression and LASSO trade slightly increased bias for greatly reduced variance, compared to the full model.
- LASSO performs variable selection in addition to coefficient shrinkage
- In Ridge Regression, correlated predictors tend to have similar coefficients. The same is not true of LASSO.

Comparison of Penalized Regression Models

- Both ridge regression and LASSO can be implemented in R using `glmnet`. (Ridge regression uses `alpha = 0`, while LASSO uses `alpha = 1`)
- Both Ridge Regression and LASSO can be fit in about the same amount of time as ordinary least squares
- Both ridge regression and LASSO trade slightly increased bias for greatly reduced variance, compared to the full model.
- LASSO performs variable selection in addition to coefficient shrinkage
- In Ridge Regression, correlated predictors tend to have similar coefficients. The same is not true of LASSO.
- In general, LASSO tends to outperform Ridge Regression in cases where some of the coefficients are nearly or truly 0.

Comparison of Penalized Regression Models

- Both ridge regression and LASSO can be implemented in R using `glmnet`. (Ridge regression uses `alpha = 0`, while LASSO uses `alpha = 1`)
- Both Ridge Regression and LASSO can be fit in about the same amount of time as ordinary least squares
- Both ridge regression and LASSO trade slightly increased bias for greatly reduced variance, compared to the full model.
- LASSO performs variable selection in addition to coefficient shrinkage
- In Ridge Regression, correlated predictors tend to have similar coefficients. The same is not true of LASSO.
- In general, LASSO tends to outperform Ridge Regression in cases where some of the coefficients are nearly or truly 0.
- Conversely, Ridge Regression outperforms LASSO when all coefficients are significant (but variance is still a liability for MSE)