Stat 435 Intro to Statistical Machine Learning

Week 6: Additional exercises

Richard Li

May 9, 2017

How do they compare

$\| \vec{u} \|_2^2 = \sum u_i^2$

$= u^T u$

Assuming all $\boldsymbol{X}$ and $\boldsymbol{Y}$ are centered,

$$\boldsymbol{\beta}_1 = \text{argmin } \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$$

$$\boldsymbol{\beta}_2 = \text{argmin } \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2$$

$$\boldsymbol{\beta}_3 = \text{argmin } \|\boldsymbol{Y} - a\boldsymbol{X}\boldsymbol{\beta}\|_2^2$$

$$\boldsymbol{\beta}_4 = \text{argmin } \|\boldsymbol{Y} - a\boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2$$

$$\boldsymbol{\beta}_5 = \text{argmin } \|\boldsymbol{Y} - a\boldsymbol{X}\boldsymbol{\beta}\|_2^2 + a^2\lambda\|\boldsymbol{\beta}\|_2^2$$

$$\{\beta_0, \boldsymbol{\beta}_6\} = \text{argmin } \|(\boldsymbol{Y} + 1) - \beta_0 - (\boldsymbol{X} + 2)\boldsymbol{\beta}\|_2^2$$

$$\{\beta_0, \boldsymbol{\beta}_7\} = \text{argmin } \|(\boldsymbol{Y} + 1) - \beta_0 - (\boldsymbol{X} + 2)\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2$$

And what about the resulted RSS?

Least square: shifting by a constant

$$\boldsymbol{\beta}_1 = \text{argmin } \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$$
$$\{\beta_0, \boldsymbol{\beta}_6\} = \text{argmin } \|(\boldsymbol{Y} + 1) - \beta_0 - (\boldsymbol{X} + 2)\boldsymbol{\beta}\|_2^2$$

- $\boldsymbol{\beta}_6 = \boldsymbol{\beta}_1$
- We can calculate β_0 from $\boldsymbol{\beta}_6$, $\bar{\boldsymbol{X}}$, and $\bar{\boldsymbol{Y}}$
- RSS is also the same

Least square: changing scales

$$\beta_1 = \text{argmin}\ \|\boldsymbol{Y} - \boldsymbol{X}\beta\|_2^2$$

$$\beta_3 = \text{argmin}\ \|\boldsymbol{Y} - a\boldsymbol{X}\beta\|_2^2$$

- $\beta_3 = \beta_1/a$
- since $\boldsymbol{Y} - \boldsymbol{X}\beta_1 = \boldsymbol{Y} - a\boldsymbol{X}(\beta_1/a)$
- RSS is also the same

Ridge: shifting by a constant

$$\boldsymbol{\beta}_2 = \text{argmin } \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2$$

$$\{\beta_0, \boldsymbol{\beta}_7\} = \text{argmin } \|(\boldsymbol{Y} + 1) - \beta_0 - (\boldsymbol{X} + 2)\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2$$

- $\boldsymbol{\beta}_7 = \boldsymbol{\beta}_2$
- Notice β_0 is not penalized, so similar to the argument before
- We can calculate β_0 from $\boldsymbol{\beta}_7$, $\bar{\boldsymbol{X}}$, and $\bar{\boldsymbol{Y}}$
- RSS is also the same

Ridge: compare with least square

$$\beta_1 = \text{argmin } \|\boldsymbol{Y} - \boldsymbol{X}\beta\|_2^2$$
$$\beta_2 = \text{argmin } \|\boldsymbol{Y} - \boldsymbol{X}\beta\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2$$

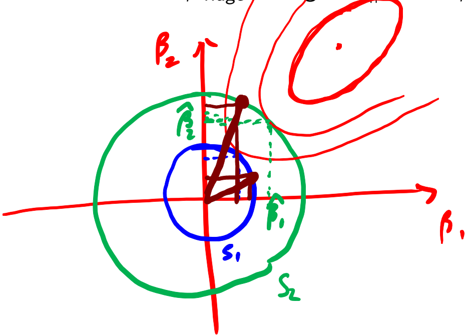
- For any single element: $|\beta_{2j}|$? $|\beta_{1j}|$, **not sure in general!**
  - In some special cases, we know the relationship for sure, e.g., in class you derived the case when $\boldsymbol{X}$ is identity matrix.
- For the sum of squares $\|\boldsymbol{\beta}_2\|_2^2 < \|\boldsymbol{\beta}_1\|_2^2$
  - which implies when $p = 1$, $|\beta_2| < |\beta_1|$.

Geometric intuition / alternative view

There is a one-to-one relationship between λ and s in

$$\hat{\boldsymbol{\beta}}_{ridge} = \text{argmin } \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2$$

$$\hat{\boldsymbol{\beta}}_{ridge} = \text{argmin } \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \quad s.t. \ \|\boldsymbol{\beta}\|_2^2 \leq s$$



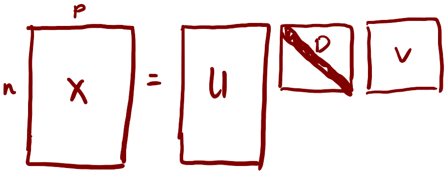
$\lambda_1 > \lambda_2$

$S_2 < S_2$

Proof 1: using SVD

In order to show this is true, we need to review something called **singular value decomposition (SVD)**.

- For any real matrix $\boldsymbol{X} \in R^{n \times p}$, $n \geq p$, we can write
- $\boldsymbol{X} = \boldsymbol{UDV}^T$, where
  1. $\boldsymbol{U} \in R^{n \times p}$ and $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}_p$,
  2. $\boldsymbol{D} \in R^{p \times p}$ and is diagonal,
  3. $\boldsymbol{V} \in R^{p \times p}$ and $\boldsymbol{V}^T = \boldsymbol{V}^{-1}$.



Proof 1: Plug in SVD

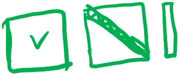
$X = UDV^T \qquad U^TU = I, \quad V^T = V^{-1}$

$$\hat{\beta}_{ridge} = (X^TX + \lambda I)^{-1} X^TY$$
$$= (VD\cancel{U^TU}DV^T + \lambda I)^{-1} VDU^TY$$
$$= (VD^2V^T + \lambda I)^{-1} VDU^TY$$
$$= (VD^2V^T + V(\lambda I)V^T)^{-1} VDU^TY$$
$$= (V(D^2 + \lambda I)V^T)^{-1} VDU^TY$$
$$= V(D^2+\lambda I)^{-1} \cancel{V^TV} DU^TY$$
$$= V(D^2+\lambda I)^{-1} D U^TY \implies \text{Let } U^TY \in R^{p\times 1} = a$$
$$= \sum_{j=1}^{p} v_j \left(\frac{d_j}{d_j^2+\lambda}\right) a_j$$

$V^TV = I$

$\lambda I = \lambda V V^T$
$= V(\lambda I)V^T$

$(V\tilde{D}V^T)(V\tilde{D}^{-1}V^T)$
$= V\tilde{D}\tilde{D}^{-1}V^T$
$= VV^T = I$



Proof 1: Plug in SVD

$$\|\hat{\beta}_{ridge}\|_2^2 = \hat{\beta}_{ridge}^T \hat{\beta}_{ridge} = Y^T U D (D^2+\lambda I)^{-1} \cancel{V^T V} (D^2+\lambda I)^{-1} D U^T Y$$

$$= a^T \, \mathrm{diag}(\cdots) \, a$$

$$= \sum_{j=1}^{p} a_j^2 \cdot \frac{d_j^2}{(d_j^2+\lambda)^2}$$

$$\lambda \uparrow \quad \|\hat{\beta}\|_2^2 \downarrow$$

Proof 2: By definition

Consider the general case with $\lambda_1 < \lambda_2$ *(in the previous definition, $\lambda_1 = 0$, $\lambda_2 = \lambda$)*

$$\beta_1 = \text{argmin } \|\boldsymbol{Y} - \boldsymbol{X}\beta\|_2^2 + \lambda_1\|\beta\|_2^2$$

$$\beta_2 = \text{argmin } \|\boldsymbol{Y} - \boldsymbol{X}\beta\|_2^2 + \lambda_2\|\beta\|_2^2$$

Now by definition,

$$\|\boldsymbol{Y} - \boldsymbol{X}\beta_1\|_2^2 + \lambda_1\|\beta_1\|_2^2 \leq \|\boldsymbol{Y} - \boldsymbol{X}\beta_2\|_2^2 + \lambda_1\|\beta_2\|_2^2$$

$$\|\boldsymbol{Y} - \boldsymbol{X}\beta_2\|_2^2 + \lambda_2\|\beta_2\|_2^2 \leq \|\boldsymbol{Y} - \boldsymbol{X}\beta_1\|_2^2 + \lambda_2\|\beta_1\|_2^2$$

Adding the two inequalities together,

$$\lambda_1\|\beta_1\|_2^2 + \lambda_2\|\beta_2\|_2^2 \leq \lambda_1\|\beta_2\|_2^2 + \lambda_2\|\beta_1\|_2^2$$

$$(\lambda_1 - \lambda_2)\|\beta_1\|_2^2 \leq (\lambda_1 - \lambda_2)\|\beta_2\|_2^2$$

$$\|\beta_1\|_2^2 \geq \|\beta_2\|_2^2$$

Proof 2: By definition

Now plug in $\|\boldsymbol{\beta}_1\|^2 > \|\boldsymbol{\beta}_2\|^2$ back in to

$$\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_1\|_2^2 + \lambda_1\|\boldsymbol{\beta}_1\|_2^2 \leq \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_2\|_2^2 + \lambda_1\|\boldsymbol{\beta}_2\|_2^2$$

$$RSS_1 \leq RSS_2 + \lambda_1 \|\beta_2\|_2^2 - \lambda_1 \|\beta_1\|_2^2$$

$$\leq RSS_2$$

Summary so far

So far we have shown that for ridge regression

- When we increase λ, $\|\beta\|_2^2$ decreases.
- But we cannot guarantee every element in $|\beta|$ decreases.
- When we increase λ, RSS increases.

Now this fact allows us to compare more complicated scenarios...

Ridge: changing scale and λ accordingly

$$\beta_2 = \text{argmin } \|\boldsymbol{Y} - \boldsymbol{X}\beta\|_2^2 + \lambda\|\beta\|_2^2$$
$$\beta_5 = \text{argmin } \|\boldsymbol{Y} - a\boldsymbol{X}\beta\|_2^2 + a^2\lambda\|\beta\|_2^2$$

By change of variable, if we let $\boldsymbol{\alpha} = a \times \boldsymbol{\beta}$ in the second problem,

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{Y} - a\boldsymbol{X}\beta\|_2^2 + a^2\lambda\|\boldsymbol{\beta}\|_2^2 = \min_{\boldsymbol{\alpha}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_2^2$$

Notice we know the minimizer of RHS is β_2, thus $\boldsymbol{\alpha} = \beta_2$, or

$$\beta_5 = \frac{1}{a}\beta_2$$

And RSS is the same.

Ridge: changing scale and λ fixed

$$\boldsymbol{\beta}_2 = \text{argmin } \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2$$

$$\boldsymbol{\beta}_5 = \text{argmin } \|\boldsymbol{Y} - a\boldsymbol{X}\boldsymbol{\beta}\|_2^2 + a^2\lambda\|\boldsymbol{\beta}\|_2^2$$

$$\boldsymbol{\beta}_4 = \text{argmin } \|\boldsymbol{Y} - a\boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2$$

Consider $a > 1$, using $\boldsymbol{\beta}_5$ as a middle step,

- $\|\boldsymbol{\beta}_4\|_2^2 > \|\boldsymbol{\beta}_5\|_2^2 = \frac{1}{a^2}\|\boldsymbol{\beta}_2\|_2^2$
- which implies if $p = 1$, $|\beta_4| > |\beta_5| = \frac{1}{a}|\beta_2|$

($a < 1$ is just the opposite.)

$$\boldsymbol{\beta}_1 = \text{argmin } \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$$
$$\boldsymbol{\beta}_2 = \text{argmin } \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2$$
$$\boldsymbol{\beta}_3 = \text{argmin } \|\boldsymbol{Y} - a\boldsymbol{X}\boldsymbol{\beta}\|_2^2$$
$$\boldsymbol{\beta}_4 = \text{argmin } \|\boldsymbol{Y} - a\boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2$$
$$\boldsymbol{\beta}_5 = \text{argmin } \|\boldsymbol{Y} - a\boldsymbol{X}\boldsymbol{\beta}\|_2^2 + a^2\lambda\|\boldsymbol{\beta}\|_2^2$$
$$\{\beta_0, \boldsymbol{\beta}_6\} = \text{argmin } \|(\boldsymbol{Y} + 1) - \beta_0 - (\boldsymbol{X} + 2)\boldsymbol{\beta}\|_2^2$$
$$\{\beta_0, \boldsymbol{\beta}_7\} = \text{argmin } \|(\boldsymbol{Y} + 1) - \beta_0 - (\boldsymbol{X} + 2)\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2$$

So, in the simple case of $p = 1$, and $a > 1$

$$|\beta_7| = |\beta_2| = a|\beta_5| < a|\beta_4| < a|\beta_3| = |\beta_1| = |\beta_6|$$

If we replace $|\cdot|$ with $\|\cdot\|_2^2$, it holds for $p > 1$ case.

How about RSS?

We already know $RSS_1 = RSS_3 = RSS_6$:

$$\beta_1 = \text{argmin } \|\boldsymbol{Y} - \boldsymbol{X}\beta\|_2^2$$

$$\beta_3 = \text{argmin } \|\boldsymbol{Y} - a\boldsymbol{X}\beta\|_2^2$$

$$\{\beta_0, \beta_6\} = \text{argmin } \|(\boldsymbol{Y} + 1) - \beta_0 - (\boldsymbol{X} + 2)\beta\|_2^2$$

Remember RSS increases as λ increases, we have $RSS_3 < RSS_4 < RSS_5$

$$\beta_3 = \text{argmin } \|\boldsymbol{Y} - a\boldsymbol{X}\beta\|_2^2$$

$$\beta_4 = \text{argmin } \|\boldsymbol{Y} - a\boldsymbol{X}\beta\|_2^2 + \lambda\|\beta\|_2^2$$

$$\beta_5 = \text{argmin } \|\boldsymbol{Y} - a\boldsymbol{X}\beta\|_2^2 + a^2\lambda\|\beta\|_2^2$$

And we also have shown $RSS_2 = RSS_5 = RSS_7$

$$\beta_2 = \text{argmin } \|\boldsymbol{Y} - \boldsymbol{X}\beta\|_2^2 + \lambda\|\beta\|_2^2$$

$$\beta_5 = \text{argmin } \|\boldsymbol{Y} - a\boldsymbol{X}\beta\|_2^2 + a^2\lambda\|\beta\|_2^2$$

$$\{\beta_0, \beta_7\} = \text{argmin } \|(\boldsymbol{Y} + 1) - \beta_0 - (\boldsymbol{X} + 2)\beta\|_2^2 + \lambda\|\beta\|_2^2$$

Questions/exercises for you

- Derive using SVD that RSS can be written as

$$\boldsymbol{Y}^T\boldsymbol{Y} + \sum_{j=1}^{p}((\frac{d_j^2}{d_j^2+\lambda} - 1)^2 - 1)\boldsymbol{a}_j^2$$

  where $\boldsymbol{a} = \boldsymbol{U}^T\boldsymbol{Y}$, and show this is an increasing function of $\lambda > 0$.

- Can you say something similar if we change all ridge regression into Lasso?

Derivation of lasso estimator when $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}$

To derive the lasso estimator, we first notice that

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^{(lasso)} &= \text{argmin }_{\boldsymbol{\beta}}\{(\boldsymbol{Y} - \boldsymbol{X}\beta)^T(\boldsymbol{Y} - \boldsymbol{X}\beta) + \lambda\sum_{j=1}^{p}|\beta_j|\} \\
&= \text{argmin }_{\boldsymbol{\beta}}\{\boldsymbol{Y}^T\boldsymbol{Y} - 2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{Y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + \lambda\sum_{j=1}^{p}|\beta_j|\} \\
&= \text{argmin }_{\boldsymbol{\beta}}\{\boldsymbol{Y}^T\boldsymbol{Y} - 2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{Y} + \boldsymbol{\beta}^T\boldsymbol{\beta} + \lambda\sum_{j=1}^{p}|\beta_j|\} \\
&= \text{argmin }_{\boldsymbol{\beta}}\{\sum_{i=1}^{n}y_i^2 + \sum_{j=1}^{p}(\beta_j^2 + \lambda|\beta_j| - 2\beta_j\sum_{i=1}^{n}x_{ij}y_i)\}
\end{aligned}$$

Derivation of lasso estimator when $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}$

The above objective function allows us to optimize each element of $\boldsymbol{\beta}$ separately. Similar to what we have derived in class, we can see the minimum is achieved at

$$\hat{\beta}_j = \max\{\sum_{i=1}^{n} x_{ij}y_i - \frac{\lambda}{2}, 0\} \quad \text{if} \quad \hat{\beta}_j > 0$$

$$\hat{\beta}_j = \min\{\sum_{i=1}^{n} x_{ij}y_i + \frac{\lambda}{2}, 0\} \quad \text{if} \quad \hat{\beta}_j < 0$$

which gives us the lasso solution

$$\hat{\boldsymbol{\beta}}^{(lasso)} = \begin{cases} \sum_{i=1}^{n} x_{ij}y_i - \frac{\lambda}{2} & \text{if } \sum_{i=1}^{n} x_{ij}y_i > \frac{\lambda}{2} \\ \sum_{i=1}^{n} x_{ij}y_i + \frac{\lambda}{2} & \text{if } \sum_{i=1}^{n} x_{ij}y_i < -\frac{\lambda}{2} \\ 0 & \text{otherwise} \end{cases}$$