LASSO Regression

LASSO Regression

- L1 regularization adds a penalty equal to the **absolute value** of the magnitude of coefficients.

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \text{RSS} + \boxed{\lambda\sum_{j=1}^{p}|\beta_j|}$$

LASSO Regression

- L1 regularization adds a penalty equal to the **absolute value** of the magnitude of coefficients.
    - Limits the size of the coefficients.
    - Can yield sparse models where some coefficients can become zero.

LASSO Regression

- L1 regularization adds a penalty equal to the **absolute value** of the magnitude of coefficients.
    - Limits the size of the coefficients.
    - Can yield sparse models where some coefficients can become zero.

LASSO Regression

- LASSO can force some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large.
- Similar to subset selection, the LASSO performs variable selection.
- Models generated from the LASSO are generally much easier to interpret.

LASSO Regression

- LassoCV with sklearn operates on checking a number of alphas within a range, instead of providing the alphas directly.
- Let's explore the results of LASSO in Python and Scikit-Learn!

Elastic
NET

Elastic NET

- We’ve been able to perform Ridge and Lasso regression.
- We know Lasso is able to shrink coefficients to zero, but we haven’t taken a deeper dive into how or why that is.
- This ability becomes more clear when learning about **elastic net** which combines Lasso and Ridge together!

Elastic NET

- We can rewrite Lasso and Ridge:

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s$$

and

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq s,$$

Elastic NET

- There is some sum s which allows to rewrite the penalty as a requirement:

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s$$

and

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq s,$$

Elastic NET

- Start with a simple thought experiment:
  - A simple equation:
    - $\hat{y} = \beta_1 x_1 + \beta_2 x_2$
  - We know that regularization can be expressed as an additional requirement that RSS is subject to.

Elastic NET

- Start with a simple thought experiment:
  - A simple equation:
    - $\hat{y} = \beta_1 x_1 + \beta_2 x_2$
  - L1 constrains the sum of absolute values.
    - $\sum|\beta|$
  - L2 constrains the sum of squared values.
    - $\sum\beta^2$

Elastic NET

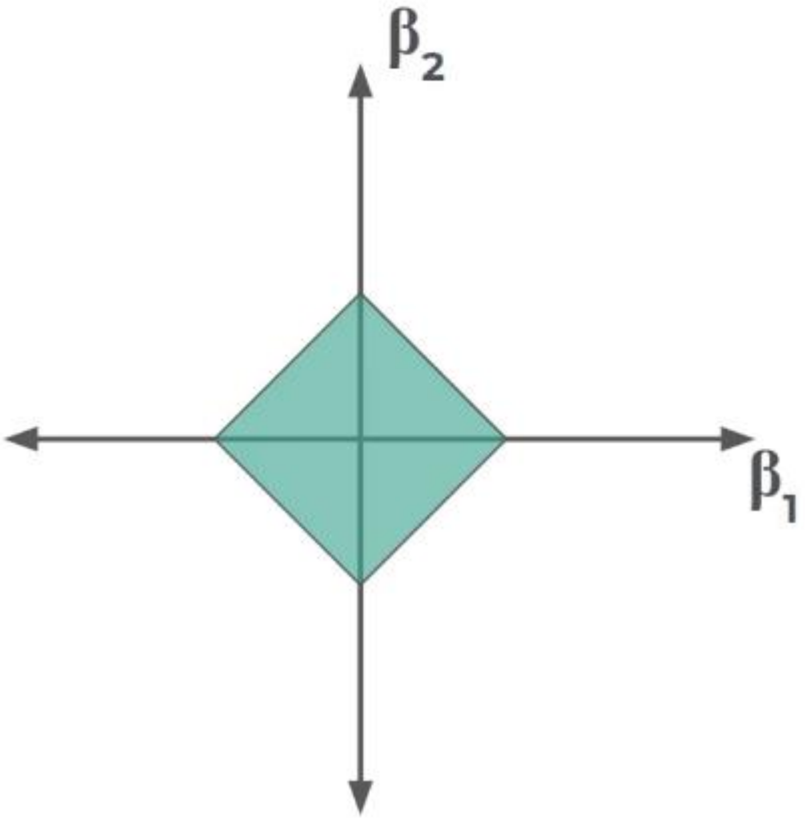
- Start with a simple thought experiment:
  - A simple equation:
    - $\hat{y} = \beta_1 x_1 + \beta_2 x_2$
  - There is some sum **s** that the penalty is less than.

Elastic NET

- For the case of only two features:
  - $\hat{y} = \beta_1 x_1 + \beta_2 x_2$
- Lasso Regression Penalty:
  - $|\beta_1| + |\beta_2| \leq s$
- Ridge Regression Penalty:
  - $\beta_1^2 + \beta_2^2 \leq s$

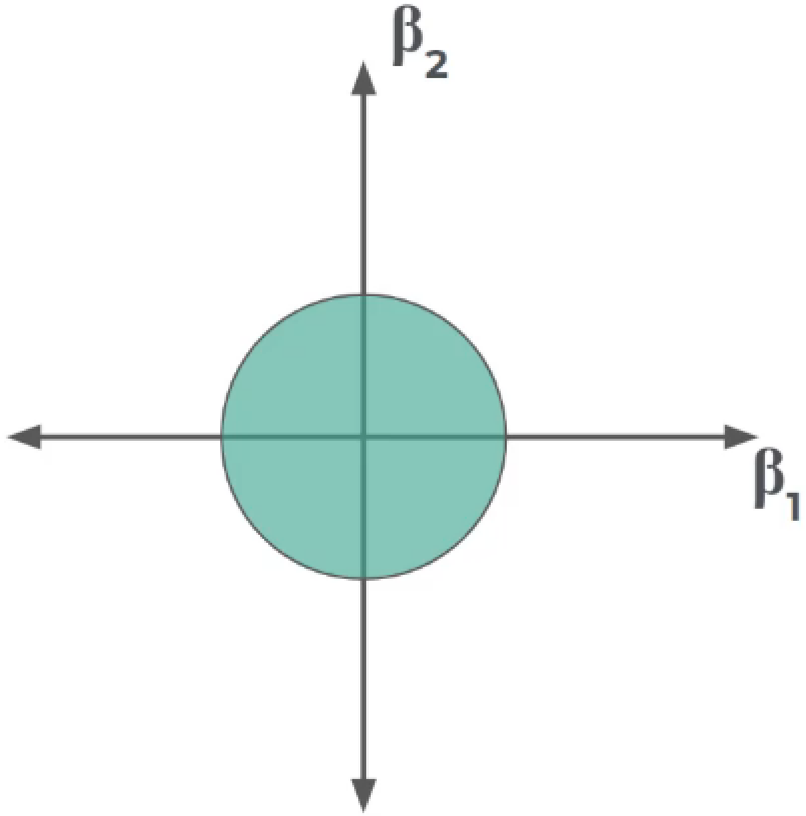
Elastic NET

- Let's plot Lasso: $|\beta_1| + |\beta_2| \leq s$



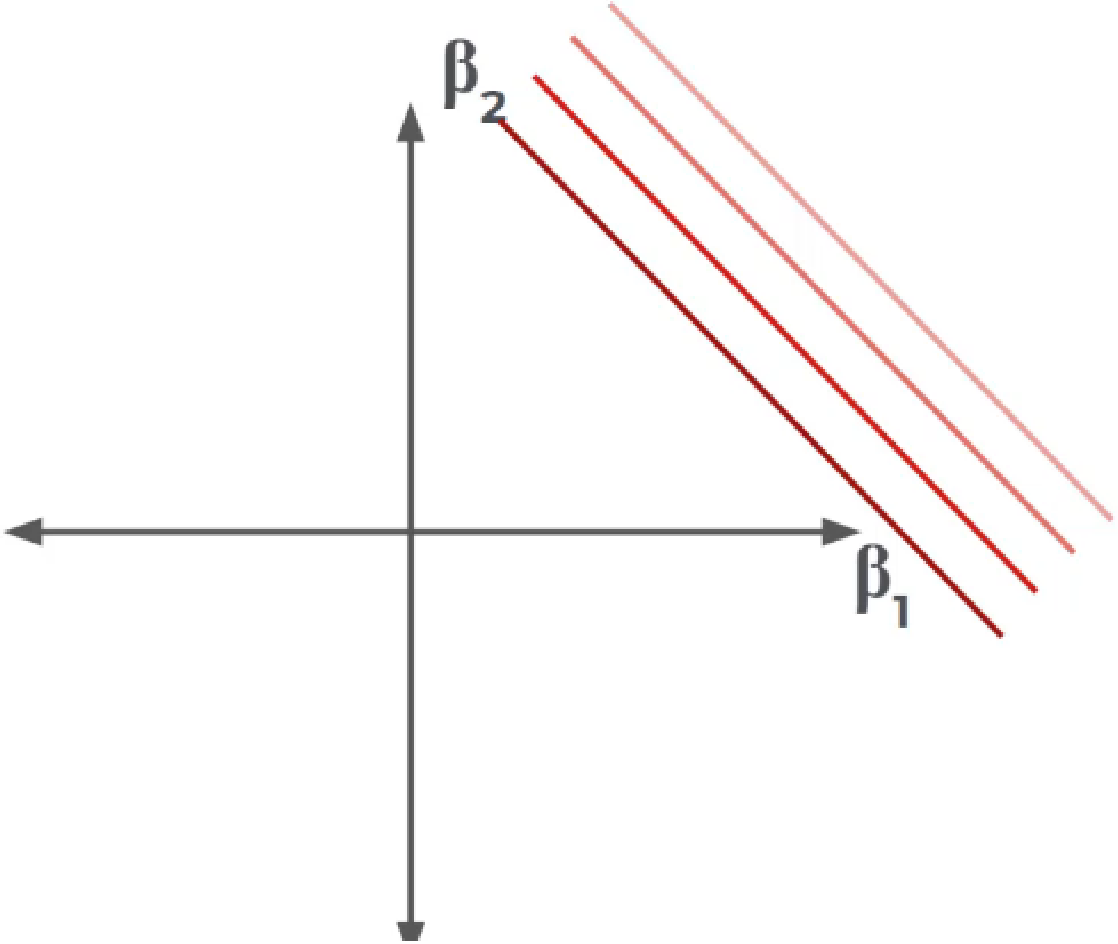
Elastic NET

- Let's plot Ridge: $\beta_1^2 + \beta_2^2 \leq s$



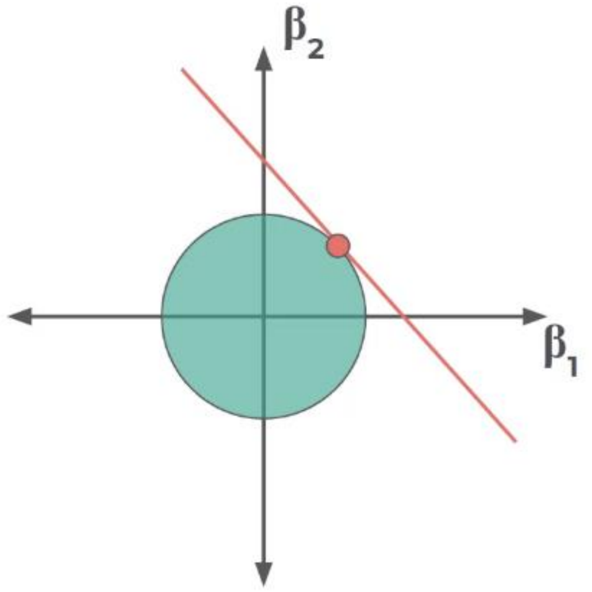
Elastic NET

- What would RSS look like?

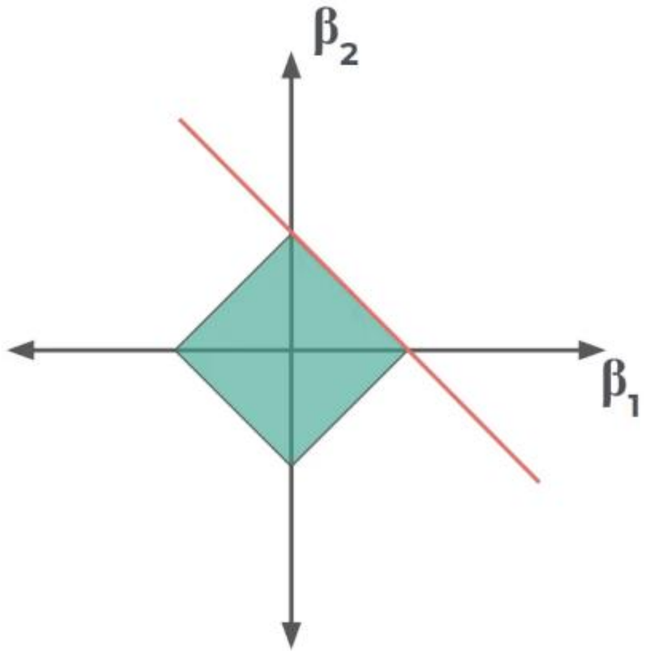


Elastic NET

Penalty for Ridge: $\beta_1^2 + \beta_2^2 \leq s$



Penalty for Lasso: $|\beta_1| + |\beta_2| \leq s$



Elastic NET

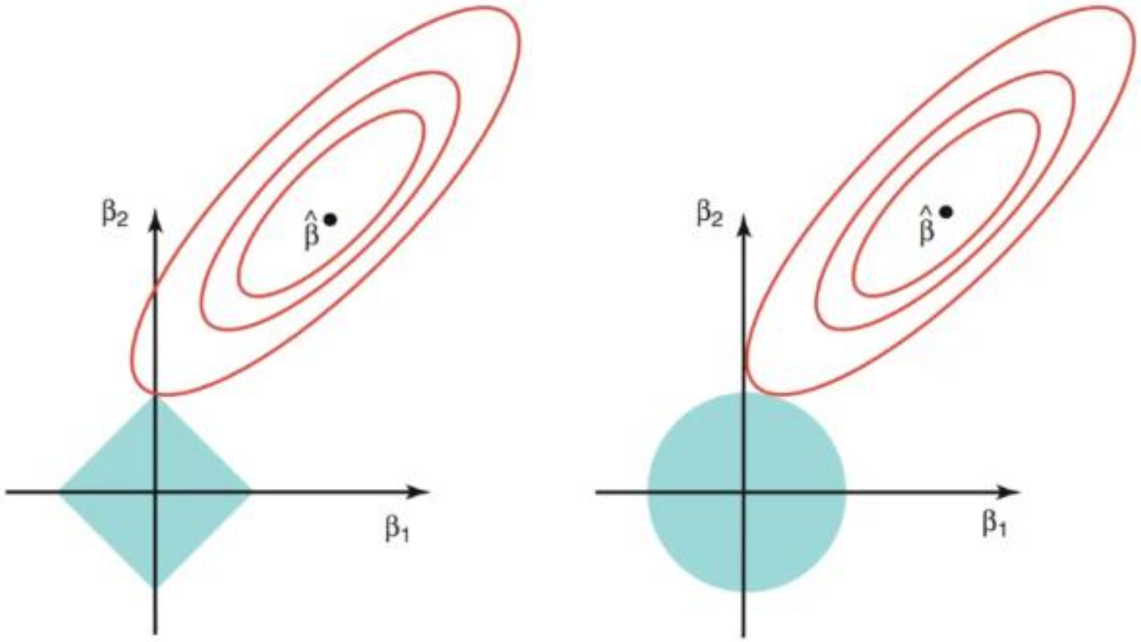
- Lasso:
  - A convex object that lies tangent to the boundary, is likely to encounter a corner of a hypercube, for which some components of β are identically zero.

Elastic NET

- Ridge: In the case of an n-sphere, the points on the boundary for which some of the components of β are zero are not distinguished from the others and the convex object is no more likely to contact a point at which some components of β are zero than one for which none of them are.

Elastic NET

- This is why Lasso is more likely to lead to coefficients as zero.
- This diagram is also commonly shown with contour RSS:



Elastic NET

- Elastic Net seeks to improve on both L1 and L2 Regularization by combining them:

$$\text{Error} = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda_1\sum_{j=1}^{p}\beta_j^2 + \lambda_2\sum_{j=1}^{p}|\beta_j|$$

Elastic NET

- Here we seek to minimize RSS and **both** the squared and absolute value terms:

$$\text{Error} = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda_1\sum_{j=1}^{p}\beta_j^2 + \lambda_2\sum_{j=1}^{p}|\beta_j|$$

Elastic NET

- Notice there are **two** distinct lambda values for each penalty:

$$\text{Error} = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda_1\sum_{j=1}^{p}\beta_j^2 + \lambda_2\sum_{j=1}^{p}|\beta_j|$$

Elastic NET

- We can alternatively express this as a ratio between L1 and L2:

$$\frac{\sum_{i=1}^{n}(y_i - x_i^J \hat{\beta})^2}{2n} + \lambda \left(\frac{1-\alpha}{2} \sum_{j=1}^{m} \hat{\beta}_j^2 + \alpha \sum_{j=1}^{m} |\hat{\beta}_j| \right)$$

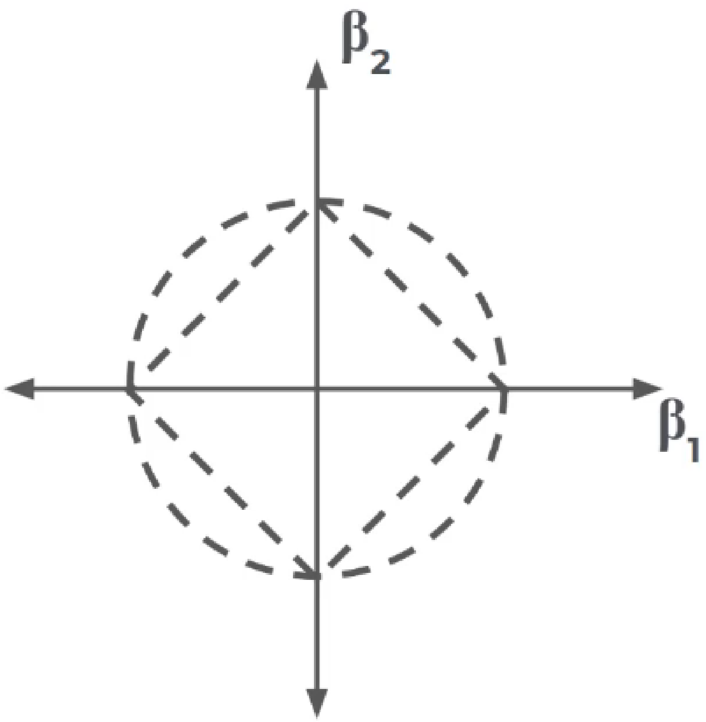
Elastic NET

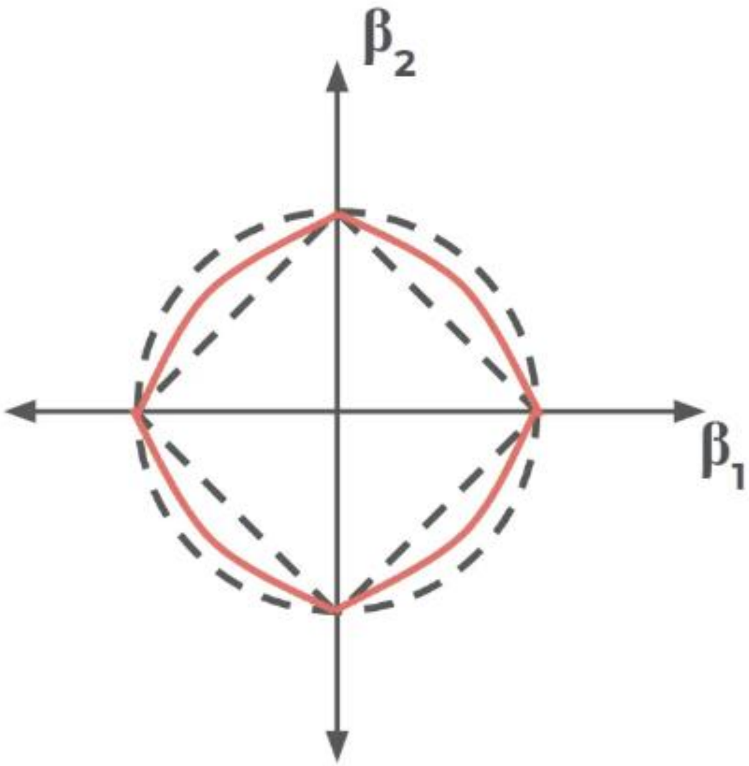
- We can also use simplified notation:

$$\hat{\beta} \equiv \underset{\beta}{\operatorname{argmin}}(||y - X\beta||^2 + \lambda_2||\beta||^2 + \lambda_1||\beta||_1)$$

Elastic NET

- Elastic Net Penalty Region:





Elastic NET

- Let’s See how Elastic Net in Python