What is Sparsity and the Lasso? (ISLR 6.2.2)

Objectives:
- compute sparsity and ℓ_1 norms
- formulation of Lasso and comparison w/ ridge

Ridge: first example of regularization
to reduce model complexity,
reduce variance at cost of adding bias.

$$R(\vec{\beta}) = \lambda \|\vec{\beta}\|^2 = \lambda \sum_{i=1}^{d} \beta_i^2 = \ell_2 \text{ norm of } \vec{\beta}$$

Equivalent to prior $\|\vec{\beta}\|$ small, $\vec{\beta} \sim N(0, \frac{1}{\lambda})$

Lasso: a new regularization using ℓ_1 norm
prior that $\vec{\beta}$ is sparse

Defn The sparsity of a vector $\vec{\beta}$ is the number of entries in $\vec{\beta}$ that are nonzero.

$$\text{sparsity} = \sum_{i=1}^{d} \mathbb{1}_{\{\beta_i \neq 0\}} \qquad \text{indicator}$$

$$= \|\vec{\beta}\|_0$$

We call this the ℓ_0 "norm" (not a real norm)

ex/ $\vec{\beta} = [1, -2, 0]^T$

$$\|\vec{\beta}\|_0 = 1 + 1 + 0 = 2$$

Defn A sparse vector has many zeros.

$\iff \|\vec{\beta}\|_0$ is small

Ridge : $\min_{\vec{\beta}} \|X\vec{\beta} - \vec{y}\|^2 + \lambda \|\vec{\beta}\|^2$ ← made $\|\vec{\beta}\|$ small

To get sparsity in outputs

$$\min_{\vec{\beta}} \|X\vec{\beta} - \vec{y}\|^2 + \lambda \|\vec{\beta}\|_0 \quad , \quad \lambda > 0$$

Problem: $\|\cdot\|_0$ is not convex, optimization in NP-hard.

Solution: Change the problem, use different penalty that gives sparsity.

Lasso: least absolute shrinkage + selection operator

Tibshirani, Donoho, Mallat

Idea is to replace $\|\cdot\|_0$ with $\|\cdot\|_1$

the "closest" true norm, which is convex.

$$\vec{\beta}_{lasso} = \arg\min_{\vec{\beta}} \|X\vec{\beta} - \vec{y}\|^2 + \lambda \|\vec{\beta}\|_1$$

$$\text{where } \|\vec{\beta}\|_1 = \sum_{i=1}^{d} |\beta_i|$$

ex/ $\vec{\beta} = [1, -2, 0]^T$

$$\|\vec{\beta}\|_1 = |1| + |-2| + |0| = 3$$

difference:
cares about
magnitude of entries

$\Rightarrow$ shrinkage

How to find $\vec{\beta}_{lasso}$?

- gradient descent with soft-thresholding (ISTA / FISTA)
- will cover after gradient descent

Takeaway: efficient algorithm for sparsity