Lasso regression

Nathanaël Carraz Rakotonirina

Mathématiques Informatique et Statistique Appliquées (MISA)
Université d'Antananarivo

Model

We want the parameters to no just be small as in ridge regression, but to be exactly zero. l_1 norm is used for the penalty. This way, $\hat{\boldsymbol{w}}$ is sparse. This approach is called lasso or least absolute shrinkage and selection operator.
The loss function becomes

$$J(\boldsymbol{w}) = \mathrm{RSS}(\boldsymbol{w}) + \lambda ||\boldsymbol{w}||_1$$

where λ controls the strength of the regularizer

Probabilistic interpretation

We use MAP estimation with a Laplace prior on the weights $p(\boldsymbol{w}; \lambda) = \prod_{i=1}^{D} Lap(w_i; 0, 1/\lambda)$. The MAP estimate corresponds to minimizing the previous objective function

$$J(\boldsymbol{w}) = ||\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}||_2^2 + \lambda||\boldsymbol{w}||_1$$

Optimization

There is no analytical solution as in linear and ridge regression. However, the objective function is still convex.

Explore further

Solve Lasso using:

- LARS (least angle regression and shrinkage)
- Gradient descent
- Coordinate descent

Combine lasso and ridge to obtain elastic net whose objective is:

$$J(\boldsymbol{w}) = ||\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}||_2^2 + \lambda_2||\boldsymbol{w}||_2^2 + \lambda_1||\boldsymbol{w}||_1$$