Ridge regression

Nathanaël Carraz Rakotonirina

MISA
Université d'Antananarivo

Model

To avoid overfitting, a common strategy is to constrain the weights to be smaller. Weights that become too large in magnitude are penalized. This is called **weight decay** or L^2 **regularization**.

The loss function is updated as follows

$$J(\boldsymbol{w}) = RSS(\boldsymbol{w}) + \lambda ||\boldsymbol{w}||_2^2$$

where λ controls the strength of the regularizer

Probabilistic interpretation

Instead of MLE, we use MAP estimation with a zero-mean Gaussian prior on the weights $p(\boldsymbol{w}) = \prod_{i=1}^{D} \mathcal{N}(w_i; 0, \tau^2)$. The MAP estimate corresponds to minimizing

$$J(\boldsymbol{w}) = \frac{1}{2\sigma^2}||\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}||_2^2 + \frac{1}{2\tau^2}||\boldsymbol{w}||_2^2$$

where $\lambda = \frac{\sigma^2}{\tau^2}$

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

Solving the MAP estimate

The gradient is given by

$$\nabla_{\boldsymbol{w}} J(\boldsymbol{w}) = 2(\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}^\top \boldsymbol{y} + \lambda \boldsymbol{w})$$

we set the gradient to zero to obtain

$$\hat{\boldsymbol{w}} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_D)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

Numerical issues

As in standard linear regression, computing directly the inverse is slow and could be numerically unstable. There are other ways.

- Convert it to standard least squares then use QR decomposition
- SVD