Lecture 6 - *Ridge Regression*

OBJECTIVE: Here we learn a cost function for linear supervised learning that is more stable than the one in the previous lecture. We also introduce the very important notion of **regularization**.

Textbook: Pages 59–64.

All the answers so far are of the form

$$\widehat{\theta} = (XX^T)^{-1}X^TY$$

They require the inversion of XX^T. This can lead to problems if the system of equations is poorly conditioned. A solution is to add a small element to the diagonal:

$$\widehat{\theta} = (XX^T + \delta^2 I_d)^{-1}X^TY$$

This is the ridge regression estimate. It is the solution to the

following **regularised quadratic cost function**

$$C(\theta) = (Y - X\theta)^T(Y - X\theta) + \delta^2\theta^T\theta$$

⋆ Proof:

It is useful to visualise the quadratic optimisation function and the constraint region.

⋆

That is, we are solving the following **constrained optimisation** problem:

$$\min_{\theta \,:\, \theta^T\theta \le t} \left\{ (Y - X\theta)^T (Y - X\theta) \right\}$$

Large values of θ are penalised. We are **shrinking** θ towards zero. This can be used to carry out **feature weighting**. **An input $x_{i,d}$ weighted by a small θ_d will have less influence on the ouptut y_i.**

Spectral View of LS and Ridge Regression

Again, let $X \in \mathbb{R}^{n\times d}$ be factored as

$$X = U\Sigma V^T = \sum_{i=1}^{d} u_i \sigma_i v_i^T,$$

where we have assumed that the rank of X is d.

⋆ The least squares prediction is:

$$\widehat{Y}_{LS} = \sum_{i=1}^{d} u_i u_i^T Y$$

⋆ Likewise, for ridge regression we have:

$$\widehat{Y}_{ridge} = \sum_{i=1}^{d} \frac{\sigma_i^2}{\sigma_i^2 + \delta^2} u_i u_i^T Y$$

The filter factor

$$f_i = \frac{\sigma_i^2}{\sigma_i^2 + \delta^2}$$

penalises small values of σ^2 (they go to zero at a faster rate).

⋆

Also, by increasing δ^2 we are penalising the weights:

⋆

Small eigenvectors tend to be wobbly. The Ridge filter factor f_i gets rid of the wobbly eigenvectors. Therefore, the predictions tend to be more stable (smooth, regularised).

The smoothness parameter δ^2 is often estimated by cross-validation or Bayesian hierarchical methods.