Linear models for regression

Michela Mulas

A recap on linear models

- We have a training set of data, from which we observe the outcome and feature measurements for a set of objects.
- Using this data we build a prediction model, or learner. This model will enable us to predict the outcome for new unseen objects.
- A linear regression model assumes that the regression function is linear in the inputs $X_1, \dots, X_p$.

A recap on linear models

The input matrix $\mathbf{X}$ of dimension $N \times (p+1)$ has the form:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{N,1} & x_{N,2} & \cdots & x_{N,p} \end{pmatrix}$$

The output vector $\mathbf{y}$ is

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{pmatrix}$$

A recap on linear models

The estimated of the parameters β is $\hat{\beta}$.

The fitted values at the training inputs are:

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{p} x_{ij} \hat{\beta}_j$$

where:

$$\hat{\mathbf{y}} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \dots \\ \hat{y}_N \end{pmatrix}$$

A recap on linear models

In matrix form, the **least squares estimation** of $\hat{\beta}$ is:

$$\hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

The predicted values at an input vector x_0 are given by $f(y_0) = f(1 : x_0^T)\hat{\beta}$.

The fitted values at the training input are:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \underbrace{\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T}_{\text{Hat matrix}}\mathbf{y}$$

A recap on linear models

Assume now that:

- y_i are uncorrelated and have constant variance σ^2.
- x_i are fixed (non random).

The variance-covariance matrix of the least squares parameter estimates is derived from the least squares estimates and is given by:

$$Var(\hat{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2 \quad \text{and}$$

$$\hat{\sigma}^2 = \frac{1}{N-p-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

- $\hat{\sigma}^2$ is the un unbiased estimate of the variance σ^2 (due to $N-p-1$ in the denominator).

A recap on linear models

Prediction accuracy can sometimes be **improved by shrinking** or setting some coefficients to zero.

By doing so we sacrifice a little bit of bias to reduce the variance of the predicted values, and hence may improve the overall prediction accuracy.

Ridge regression shrinks the coefficients by imposing a penalty on their size.

Lasso regression is a shrinkage method that replaces the ridge penalty.

A recap on linear models

The **ridge coefficients** minimize a penalized residual sum of squares:

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\text{argmin}}\left\{\sum_{i=1}^{N}\Big(y_i - \beta_0 - \sum_{j=1}^{p}x_{ij}\beta_j\Big)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\right\}$$

- $\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage.
  The larger the value of λ, the greater the amount of shrinkage.
- The coefficients are shrunk toward zero (and each other).
- The idea of penalizing by the sum-of-squares of the parameters is also used in neural networks, where it is known as *w*eight decay.

A recap on linear models

An equivalent way to write the ridge problem is:

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2$$
$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 \leq t$$

- This makes explicit the size constraint on the parameters.
- There is a one- to-one correspondence between the parameters λ and t.

A recap on linear models

Notes on Ridge

- The ridge solutions are not equivariant under scaling of the inputs, and so one normally standardizes the inputs before solving ridge.
- The intercept β_0 has been left out of the penalty term: Penalization of the intercept would make the procedure depend on the origin chosen for Y.

A recap on linear models

Notes on Ridge

- In matrix form, we have:

$$RSS(\lambda) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta^T\beta$$
$$\hat{\beta}^{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

  I is the $p \times p$ identity matrix.
- With the choice of quadratic penalty $\beta^T\beta$, the ridge regression solution is again a linear function of **y**.
- The solution adds a positive constant to the diagonal of $\mathbf{X}^T\mathbf{X}$ before inversion.
- This makes the problem nonsingular, even if $\mathbf{X}^T\mathbf{X}$ is not of full rank, and was the main motivation for ridge regression when it was first introduced in statistics.

A recap on linear models

The **lasso** estimate is defined as:

$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2$$
$$\text{subject to } \sum_{j=1}^{p} |\beta_j| \leq t$$

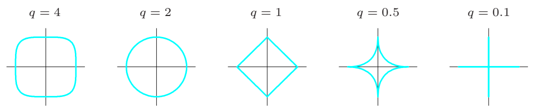
We can also write the lasso problem in the equivalent **Lagrangian form**

$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

A recap on linear models

We can generalize ridge regression and the lasso:

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\} \quad \text{for } q \geq 0$$



A recap on linear models

The **elastic net model** propose a different compromise between the ridge and lasso:

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda_1 \sum_{j=1}^{p} \beta_j^2 + \lambda_2 \sum_{j=1}^{p} |\beta_j| \right\}$$

The **elastic net penalty** can be written as:

$$\lambda \sum_{j=1}^{p} \left(\alpha\beta_j^2 + (1-\alpha)|\beta_j| \right)$$