Ridge Logistic Regression

The Logistic Regression model explains the binary qualitative response Y, The primary motivation for using **Ridge Logistic Regression** is closely similar to the Ridge Regression :

- Preventing Overfitting in high-dimensions (spare classification)
- Improve probability predictions
- Handling correlated Predictors(Features)

The Ridge solution for the logistic regression follows the same principles as in the Ridge Regression by adding a **Penalty Term** to the **Cross entropy loss** (cost function):

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} P_i^{y_i}(1-P_i)^{1-y_i}$$

- $P = \sigma(X\beta)$
  By taking the log of this **Likelihood Function** :

$$\log(\mathcal{L}(\beta)) = l(\beta) = y^\intercal \log(P) + (1-y)^\intercal \log(1-P)$$

- The $l(\beta)$ maximize the probability of the parameter β

$$l_{ridge}(\beta, \lambda) = l(\beta) - \frac{\lambda}{2}\|\beta\|_2^2$$

By taking the negative of $l(\beta)$ we obtain the **Cross entropy loss** function :

$$J(\beta;\lambda) = -l(\beta) + \frac{\lambda}{2}\|\beta\|_2^2$$

- This is the **regular logistic regression** with the **penalty term**
- It's a minimization problem

The **Ridge** adds a penalty term to the cross entropy loss, results in :

$$J(\beta;\lambda) = -l(\beta) + \frac{\lambda}{2}\|\beta\|_2^2$$

- Our goal is to minimize $J(\beta)$ while following the constraints the **ridge** introduce

Taking the gradient of $J(\beta)$ results in :

$$\nabla J(\beta;\lambda) = -\nabla l(\beta) + \nabla\left(\frac{\lambda}{2}\|\beta\|_2^2\right)$$

It's known that $\nabla l(\beta)$ is :

$$\nabla - l(\beta) = -\frac{1}{n}X^\intercal(y - \sigma(X\beta))$$

and $\nabla\left(\frac{\lambda}{2}\|\beta\|_2^2\right)$:

$$\nabla\left(\frac{\lambda}{2}\|\beta\|_2^2\right) = \lambda\beta$$

Results in :

$$\nabla J(\beta;\lambda) = -\frac{1}{n}X^\intercal(y - \sigma(X\beta)) + \lambda\beta$$

Logistic ridge regression estimator

Same as in the Logistic Regression the **logistic ridge regression** have no close-form solution unlike the Ridge Regression, using **iterative methods** to estimates the model coefficients :

- Gradient Descent
- **Newton's method**

- L-BFGS