STA304XS - Assignment 2: Machine Learning

Jing Yeh
yhxjin001@myuct.ac.za

Saurav Sathnarayan
sthsau001@myuct.ac.za

2025-10-22

Abstract

Keywords:

Contents

0.1 Question 1 - BAYESIAN INTERPRETATION

We start with the logistic regression model:

$$\Pr(Y_i = 1 \mid x_i) = \operatorname{logit}^{-1}(x_i^\top \beta) = \frac{1}{1 + \exp(-x_i^\top \beta)}.$$

The log-likelihood for all n observations is

$$l(\beta) = \sum_{i=1}^{n} \left[y_i x_i^\top \beta - \log(1 + e^{x_i^\top \beta}) \right].$$

Assume independent Normal priors for the coefficients:

$$\beta_j \sim N(0, \tau^2), \quad j = 1, \ldots, p.$$

Hence, the prior density is

$$\pi(\beta) = \prod_{j=1}^{p} \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{\beta_j^2}{2\tau^2} \right)$$

Taking logs, we obtain the log-prior:

$$\log \pi(\beta) = -\frac{p}{2} \log(2\pi\tau^2) - \frac{1}{2\tau^2} \sum_{j=1}^{p} \beta_j^2$$

Now, by Bayes' theorem,

$$\pi(\beta \mid y) = \frac{\pi(y \mid \beta)\, \pi(\beta)}{\pi(y)}$$

Taking logs of both sides gives

$$\log \pi(\beta \mid y) = \log \pi(y \mid \beta) + \log \pi(\beta) - \log \pi(y)$$

The term $\log \pi(y)$ is a normalising constant that does not depend on β, so when maximising over β, it can be ignored. Therefore,

$$\log \pi(\beta \mid y) \propto \log \pi(y \mid \beta) + \log \pi(\beta)$$

Substituting the expressions for $\log p(y \mid \beta)$ and $\log p(\beta)$, we have

$$\log p(\beta \mid y) \propto l(\beta) - \frac{1}{2\tau^2} \sum_{j=1}^{p} \beta_j^2$$

This is the expression for the log-posterior up to a constant.

To obtain the maximum a posteriori (MAP) estimate, we maximise $\log p(\beta \mid y)$ with respect to β. Equivalently, we minimise the negative log-posterior:

$$\widehat{\beta}_{MAP} = \arg\min_{\beta} \left[-l(\beta) + \frac{1}{2\tau^2} \sum_{j=1}^{p} \beta_j^2 \right]$$

If we define $\lambda = \dfrac{1}{2\tau^2}$, then the optimisation problem becomes

$$\boxed{\widehat{\beta}_{MAP} = \arg\min_{\beta} \left[-l(\beta) + \lambda \|\beta\|_2^2 \right]}$$

This shows that the MAP estimator under a Normal prior is equivalent to the ridge-regularised logistic regression estimator, where the penalty parameter λ corresponds to the precision of the prior.

(b)

0.2 Question 2 - DERIVING RIDGE-IWLS

(a)

In IWLS, the weight for observation i is given by

$$w_i^{(t)} = \frac{1}{\mathrm{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 .$$

For logistic regression, $\mathrm{Var}(Y_i) = p_i(1-p_i)$ and $\frac{\partial \mu_i}{\partial \eta_i} = p_i(1-p_i)$, so

$$w_i^{(t)} = \frac{\left(p_i^{(t)}(1-p_i^{(t)}) \right)^2}{p_i^{(t)}(1-p_i^{(t)})} = p_i^{(t)}(1-p_i^{(t)}).$$

These are exactly the diagonal entries of the weight matrix:

$$W^{(t)} = \mathrm{diag}\left(p_1^{(t)}(1-p_1^{(t)}), \ldots, p_n^{(t)}(1-p_n^{(t)}) \right).$$

and $p^{(t)} = (p_1^{(t)}, \ldots, p_n^{(t)})^\top$ are the predicted probabilities at $\beta^{(t)}$.
and for $z^{(t)}$

$$z_i^{(t)} = \eta_i^{(t)} + \frac{y_i - \mu_i^{(t)}}{\frac{\partial \mu_i}{\partial \eta_i}}$$

where

$$\eta_i^{(t)} = x_i^\top \beta^{(t)}, \quad \mu_i^{(t)} = \sigma(\eta_i^{(t)}) = p_i^{(t)}, \quad \frac{\partial \mu_i}{\partial \eta_i} = p_i^{(t)}(1-p_i^{(t)}).$$

Substituting the derivative for logistic regression gives

$$z_i^{(t)} = x_i^\top \beta^{(t)} + \frac{y_i - p_i^{(t)}}{p_i^{(t)}(1-p_i^{(t)})}.$$

In matrix form, for all n observations:

$$z^{(t)} = X\beta^{(t)} + (W^{(t)})^{-1}(y - p^{(t)}),$$

From our IWLS formula, we have:

$$(\mathbf{X}^T\mathbf{W}^{(t)}\mathbf{X})^{-1}\beta^{(t+1)} = \mathbf{X}^T\mathbf{W}^{(t)}z^{(t)}$$

Thus,

$$\beta^{(t+1)} = (\mathbf{X}^T\mathbf{W}^{(t)}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{(t)}z^{(t)}$$

(b)

First we use the taylor expansion around $l(\beta)$, which is:

$$l(\beta) \approx l(\beta^{(t)}) + (\beta - \beta^{(t)})^\top \nabla l(\beta^{(t)}) + \frac{1}{2}(\beta - \beta^{(t)})^\top \nabla^2 f(\beta^{(t)})(\beta - \beta^{(t)})$$

where

$$\nabla l(\beta^{(t)})$$

is the gradient vector.

$l(\beta) \approx l(\beta^{(t)}) + (\beta - \beta^{(t)})^\top \nabla l(\beta^{(t)}) + \frac{1}{2}(\beta - \beta^{(t)})^\top H^{(t)}(\beta - \beta^{(t)})$,

where $H^{(t)} = \nabla^2 l(\beta^{(t)})$ is the Hessian.

For GLMs, the negative Hessian is often written as:

$-H^{(t)} = X^\top W^{(t)} X$,

where $W^{(t)}$ is the diagonal weight matrix.

The penalized objective is:

$f(\beta) = -l(\beta) + \lambda\|\beta\|_2^2$.

The gradient of the penalized objective is:

$\nabla f(\beta) = -\nabla l(\beta) + 2\lambda\beta$,

and the Hessian is:

$\nabla^2 f(\beta) = -\nabla^2 l(\beta) + 2\lambda I = X^\top W^{(t)} X + 2\lambda I$.

The Newton-Raphson update for minimizing $f(\beta)$ is:

$\beta^{(t+1)} = \beta^{(t)} - [\nabla^2 f(\beta^{(t)})]^{-1} \nabla f(\beta^{(t)})$.

Plug in the gradient and Hessian:

$\beta^{(t+1)} = \beta^{(t)} - (X^\top W^{(t)} X + 2\lambda I)^{-1}(-\nabla l(\beta^{(t)}) + 2\lambda\beta^{(t)})$

$= \beta^{(t)} + (X^\top W^{(t)} X + 2\lambda I)^{-1}(\nabla l(\beta^{(t)}) - 2\lambda\beta^{(t)})$.

and the gradient satisfies:

$\nabla l(\beta^{(t)}) = X^\top(y - \mu^{(t)}) = X^\top W^{(t)}(z^{(t)} - X\beta^{(t)})$.

Plugging this into the update:

$\beta^{(t+1)} = \beta^{(t)} + (X^\top W^{(t)} X + 2\lambda I)^{-1}(X^\top W^{(t)}(z^{(t)} - X\beta^{(t)}) - 2\lambda\beta^{(t)})$

$$= (X^\top W^{(t)} X + 2\lambda I)^{-1}(X^\top W^{(t)} z^{(t)}).$$
$$\beta^{(t+1)} = (X^\top W^{(t)} X + 2\lambda I)^{-1} X^\top W^{(t)} z^{(t)}.$$

0.3 Question 3 - IMPLEMENTATION AND EVALUATION

1 (a)

```
## $coefficients
##                      [,1]
## 1              0.05766608
## Measurement1   0.16427287
## Measurement2   0.67327480
## Measurement3  -1.24407501
##
## $iterations
## [1] 9
##
## $converged
## [1] TRUE
```