11: Regularization in Multivariate Logistic Regression

Jacob Cluff

March 2019

1 LASSO Regression

$$J = -\frac{1}{N}\sum_{i=1}^{N} y_i \ln(\hat{p}_i) + (1-y_i)\ln(1-\hat{p}_i) + \frac{\lambda_1}{2N}\sum_{j=1}^{P}|w_j|$$

Using the chain rule, $\boldsymbol{\nabla} J = \frac{\partial J}{\partial \hat{p}}\frac{\partial \hat{p}}{\partial h}\frac{\partial h}{\partial w} + \frac{\lambda_1}{N}\ sign(w)$.

$$\begin{aligned}
\frac{\partial h}{\partial w_j} &= x_j \\
\frac{\partial \hat{p}}{\partial h} &= \frac{\partial}{\partial h}\left(1+e^{-h}\right)^{-1} \\
&= (-1)(-1)e^{-h}\left(1+e^{-h}\right)^{-2} \\
&= \frac{e^{-h}}{(1+e^{-h})^2} \\
&= \frac{1}{1+e^{-h}}\frac{e^{-h}}{1-e^{-h}} \\
&= \hat{p}(1-\hat{p}) \\
\frac{\partial J}{\partial \hat{p}} &= \frac{y_i}{\hat{p}_i} - \frac{1-y_i}{1-\hat{p}_i}
\end{aligned}$$

$$
\begin{aligned}
\boldsymbol{\nabla} J &= \frac{\partial J}{\partial \hat{p}} \frac{\partial \hat{p}}{\partial h} \frac{\partial h}{\partial w} + \frac{\lambda_1}{N}\ sign(w) \\
&= -\frac{1}{N} \sum_{i=1}^{N} \left(\frac{y_i}{\hat{p}_i} - \frac{1-y_i}{1-\hat{p}_i} \right) \hat{p}_i (1-\hat{p}_i) x_i j + \frac{\lambda_1}{N}\ sign(w) \\
&= -\frac{1}{N} \sum_{i=1}^{N} x_{ij} (y_i - y_i \hat{p}_i - \hat{p}_i + y_i \hat{p}_i) + \frac{\lambda_1}{N}\ sign(w) \\
&= -\frac{1}{N} \sum_{i=1}^{N} x_{ij} (y_i - \hat{p}_i) + \frac{\lambda_1}{N}\ sign(w) \\
&= \frac{1}{N} \sum_{i=1}^{N} x_{ij} (\hat{p}_i - y_i) + \frac{\lambda_1}{N}\ sign(w) \\
&= \frac{1}{N} \left[X^T (\hat{p} - y)) + \lambda_1\ sign(w) \right]
\end{aligned}
$$

For the probabilistic interpretation, let $F = e^{-J}$.

$$
\begin{aligned}
F &= \exp\left(\sum_{i=1}^{N} y_i \ln(\hat{p}_i) + (1-y_i)\ln(1-\hat{p}_i) - \lambda_1 \sum_{j=1}^{P} |w_j| \right) \\
&= \prod_{i=1}^{N} \exp(y_i \ln(\hat{p}_i) + (1-y_i)\ln(1-\hat{p}_i)) \prod_{j=1}^{P} \exp(-\lambda_1 |w_j|) \\
&= \prod_{i=1}^{N} \exp\Big(\ln(\hat{p}_i)^{y_i} + \ln(1-\hat{p}_i)^{(1-y_i)} \Big) \prod_{j=1}^{P} \exp(-\lambda_1 |w_j|) \\
&= \prod_{i=1}^{N} \hat{p}_i^{y_i} (1-\hat{p}_i)^{(1-y_i)} \prod_{j=1}^{P} \exp(-\lambda_1 |w_j|) \\
&= p(Y|X,W)\ p(W) \\
&\propto p(W|X,Y)
\end{aligned}
$$

The distribution $p(W)$ forms a Laplacian distribution.

2 Elastic Net Regression

$$J = -\frac{1}{N}\sum_{i=1}^{N} y_i \ln(\hat{p}_i) + (1-y_i)\ln(1-\hat{p}_i) + \frac{\lambda_2}{2N}\sum_{j=1}^{P} w_j^2 + \frac{\lambda_1}{2N}\sum_{j=1}^{P} |w_j|$$

Using the chain rule, $\boldsymbol{\nabla} J = \frac{\partial J}{\partial \hat{p}}\frac{\partial \hat{p}}{\partial h}\frac{\partial h}{\partial w} + \frac{\lambda_2}{N} w + \frac{\lambda_1}{N}\ sign(w)$.

$$\begin{aligned}
\frac{\partial h}{\partial w_j} &= x_j \\
\frac{\partial \hat{p}}{\partial h} &= \frac{\partial}{\partial h}\left(1+e^{-h}\right)^{-1} \\
&= (-1)(-1)e^{-h}\left(1+e^{-h}\right)^{-2} \\
&= \frac{e^{-h}}{(1+e^{-h})^2} \\
&= \frac{1}{1+e^{-h}}\frac{e^{-h}}{1-e^{-h}} \\
&= \hat{p}(1-\hat{p}) \\
\frac{\partial J}{\partial \hat{p}} &= \frac{y_i}{\hat{p}_i} - \frac{1-y_i}{1-\hat{p}_i}
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{\nabla} J &= \frac{\partial J}{\partial \hat{p}}\frac{\partial \hat{p}}{\partial h}\frac{\partial h}{\partial w} + \frac{\lambda_2}{N} w + \frac{\lambda_1}{N}\ sign(w) \\
&= -\frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i}{\hat{p}_i} - \frac{1-y_i}{1-\hat{p}_i}\right)\hat{p}_i(1-\hat{p}_i)x_ij + \frac{\lambda_2}{N} w + \frac{\lambda_1}{N}\ sign(w) \\
&= -\frac{1}{N}\sum_{i=1}^{N} x_{ij}(y_i - y_i\hat{p}_i - \hat{p}_i + y_i\hat{p}_i) + \frac{\lambda_2}{N} w + \frac{\lambda_1}{N}\ sign(w) \\
&= -\frac{1}{N}\sum_{i=1}^{N} x_{ij}(y_i - \hat{p}_i) + \frac{\lambda_2}{N} w + \frac{\lambda_1}{N}\ sign(w) \\
&= \frac{1}{N}\sum_{i=1}^{N} x_{ij}(\hat{p}_i - y_i) + \frac{\lambda_2}{N} w + \frac{\lambda_1}{N}\ sign(w) \\
&= \frac{1}{N}\left[X^T(\hat{p}-y)) + \lambda_2 w + \lambda_1\ sign(w)\right]
\end{aligned}$$

For the probabilistic interpretation, let $F = e^{-J}$.

$$
\begin{aligned}
F &= \exp\left(\sum_{i=1}^{N} y_i \ln(\hat{p}_i) + (1-y_i)\ln(1-\hat{p}_i) - \lambda_2 \sum_{j=1}^{P} w_j^2 - \lambda_1 \sum_{j=1}^{P} |w_j|\right) \\
&= \prod_{i=1}^{N} \exp(y_i \ln(\hat{p}_i) + (1-y_i)\ln(1-\hat{p}_i)) \prod_{j=1}^{P} \exp\left(-\lambda_2 w_j^2\right) \exp(-\lambda_1 |w_j|) \\
\\
&= \prod_{i=1}^{N} \exp\left(\ln(\hat{p}_i)^{y_i} + \ln(1-\hat{p}_i)^{(1-y_i)}\right) \prod_{j=1}^{P} \exp\left(-\lambda_2 w_j^2\right) \exp(-\lambda_1 |w_j|) \\
&= \prod_{i=1}^{N} \hat{p}_i^{y_i} (1-\hat{p}_i)^{(1-y_i)} \prod_{j=1}^{P} \exp\left(-\lambda_2 w_j^2\right) \exp(-\lambda_1 |w_j|) \\
&= p(Y|X,W)\ p(W) \\
&\propto p(W|X,Y)
\end{aligned}
$$

The distribution $p(W)$ forms the Laplaussian distribution[1]

3 Odds

The odds of are found by taking the probability –lets say A– and dividing it by the probability of everything but A.

$$
\begin{aligned}
odds &= \frac{\hat{p}}{1-\hat{p}} \\
&= \frac{\left(\frac{1}{1+e^{-h}}\right)}{1-\left(\frac{1}{1+e^{-h}}\right)} \\
&= \frac{1}{\left(1-\frac{1}{1+e^{-h}}\right)(1+e^{-h})} \\
&= \frac{1}{e^{-h}} \\
&= e^{h} \\
&= e^{w^T X}
\end{aligned}
$$

[1]Laplace × Gaussian. Pass it on so it catches on!