t-GLM: A generalization of logistic regression using information theory

Naveen Mathew Nathan S.

9/27/2019

Introduction

The definition of cross-entropy in statistics is $L(y,p) = \prod_{i=1}^{K} p_i^{y_i}$; $CE(y,p) = -log(L(y,p)) = -[\sum_{i=1}^{K} y_i * log(p_i)]$, where K is the number of classes. This simplifies to $L(y,p) = p^y * (1-p)^{1-y} \implies CE(y,p) = -[y * log(p) + (1-y) * log(1-p)]$. Therefore, the total likelihood (for a logistic regression model) can be written as $L(\mathbf{y}, \mathbf{p}) = \prod_{j=1}^{n} p_j^{y_j} * (1-p_j)^{1-y_j} \implies CE(\mathbf{y}, \mathbf{p}) = -[\sum_{j=1}^{n} y_j * log(p_j) + (1-y_j) * log(1-p_j)]$ assuming all the examples are independent. Interestingly, the logistic regression model simplifies to a link function of the form: $X_j\beta = \eta_j = h(p_j) = logit(p_j) = ln(\frac{p_j}{1-p_j})$. We know that the logit link does not fit all types of data. Is it possible to come up with a new link function that is universally better than logit link? The answer is yes.

New (inverse) link function

$$h(\eta, t) = \frac{1}{t} ln\left(\frac{2+t\eta}{2-t\eta}\right)$$

Equating with logistic regression $\implies ln(\frac{p}{1-p}) = \frac{1}{t}ln(\frac{2+t\eta}{2-t\eta}) \implies t * ln(\frac{p}{1-p}) = ln\left(\left(\frac{p}{1-p}\right)^t\right) = ln\left(\frac{2+t\eta}{2-t\eta}\right) - equation 1$

Assuming the quantity within the bracket on RHS of equation 1 is positive, applying componendo and dividendo: $\left(\frac{p}{1-p}\right)^t + 1 = \frac{4}{2-t\eta} \implies \eta = \frac{2\left[\left(\frac{p}{1-p}\right)^t - 1\right]}{t\left[\left(\frac{p}{1-p}\right)^t + 1\right]}$

Applying limit and L'Hospital rule: $\lim_{t\to 0} \eta = \lim_{t\to 0} \frac{2\left(\frac{p}{1-p}\right)^t \ln\left(\frac{p}{1-p}\right)}{\left(\frac{p}{1-p}\right)^t + 1 + t\left(\frac{p}{1-p}\right)^t \ln\left(\frac{p}{1-p}\right)} = \ln\left(\frac{p}{1-p}\right)$

Therefore, the new link function carries the same properties as logistic regression when t = 0. Also, we observe that the link is symmetric about t = 0: $h(\eta, -t) = \frac{1}{-t} ln(\frac{2-t\eta}{2+t\eta}) = \frac{1}{t} ln(\frac{2+t\eta}{2-t\eta}) = h(\eta, t)$

Therefore, a model with the new link function is guaranteed to perform at par with logistic regression for t = 0. By tuning t using cross validation it will perform better than logistic regression

The mathematics

Condition for being a proper link

Unlike the logit link that applies to the whole range of η , the set of parameters in the updated link is restricted. This is because in the above derivation we assumed that the term within the bracket on RHS of equation 1 is positive. Let us examine it carefully:

$$\frac{2+t\eta}{2-t\eta} > 0 \implies (2+t\eta)(2-t\eta) > 0 \text{ assuming } (2-t\eta) \neq 0$$

 $\implies -2 < t\eta < 2$ which may not be satisfied if we have random x on testing sets that has larger absolute value of η than the training set

Reasonable adjustment

Reasonable thresholds can be established to ensure that the mathematical inaccuracy can be avoided. For example, for any η we defined p = 1 if $t\eta \ge 2$ and p = 0 if $t\eta \le -2$

Link, inverse link, gradient of inverse link

Link function

From the previous derivation we observe that the link $\eta = \frac{2[\left(\frac{p}{1-p}\right)^t - 1]}{t[\left(\frac{p}{1-p}\right)^t + 1]}$

Inverse link

From the definition, the inverse link is given by $logit(p) = \frac{1}{t}ln(\frac{2+t\eta}{2-t\eta}) = ln\left(\left(\frac{2+t\eta}{2-t\eta}\right)^{\frac{1}{t}}\right) \implies \frac{p}{1-p} = ln\left(\frac{t}{t}\right)^{\frac{1}{t}}$

$$\left(\frac{2+t\eta}{2-t\eta}\right)^{\frac{1}{t}} \implies p = \frac{\left(\frac{2+t\eta}{2-t\eta}\right)^{\frac{1}{t}}}{\left(\frac{2+t\eta}{2-t\eta}\right)^{\frac{1}{t}}+1}$$

Gradient of link inverse with respect to η

$$\nabla_{\eta}g = \frac{1}{t} * \frac{2-t\eta}{2+t\eta} * \frac{(2-t\eta)*t-(2+t\eta)*(-t)}{(2-t\eta)^2} = \frac{4}{4-t^2\eta^2}$$
$$\nabla_{\eta}p: \text{ Let } x = \left(\frac{2+t\eta}{2-t\eta}\right)^{\frac{1}{t}}; dx = \frac{1}{t} \left(\frac{2+t\eta}{2-t\eta}\right)^{\frac{1}{t}-1} \frac{t*(2-t\eta)-t*(2+t\eta)}{(2-t\eta)^2} = -\frac{2}{t} \left(\frac{2+t\eta}{2-t\eta}\right)^{\frac{1-t}{t}} \frac{t^2\eta^2}{(2-t\eta)^2}$$

Putting things together: Newton method

$$L(\mathbf{x},\beta) = \prod_{i=1}^{n} p_i^{y_i} (1-p_i)^{1-y_i} \implies \hat{\beta} = argmax L(\mathbf{x},\beta)$$

Since log is a monotonous transformation, it does not change the actual value(s) of β for which the likelihood is maximized. Therefore, $l(\mathbf{x}, \beta) = log(L(\mathbf{x}, \beta)) \implies \hat{\beta} = argmax_{\beta}L(\mathbf{x}, \beta) = argmax_{\beta}l(\mathbf{x}, \beta)$

$$l(\mathbf{x},\beta) = log(\prod_{i=1}^{n} p_i^{y_i} (1-p_i)^{1-y_i}) = \sum_{i=1}^{n} [y_i log(\frac{p_i}{1-p_i}) + log(1-p_i)] = \sum_{i=1}^{n} [log(p_i) - (1-y_i)x\beta]$$

Differentiating with respect to beta gives $\nabla_{\beta}p$ which is related to $\nabla_{\eta}p$ that was calculated above. Further, the Hessian matrix can also be calculated for the loss with respect to β . Finally Newton's second order optimization update can be applied: $\beta := \beta - H^{-1}J$