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

Naveen Mathew Nathan S.<br>9/27/2019

## Introduction

The definition of cross-entropy in statistics is $L(y, p)=\prod_{i=1}^{K} p_{i}^{y_{i}} ; C E(y, p)=-\log (L(y, p))=-\left[\sum_{i=1}^{K} y_{i} *\right.$ $\left.\log \left(p_{i}\right)\right]$, where K is the number of classes. This simplifies to $L(y, p)=p^{y} *(1-p)^{1-y} \Longrightarrow C E(y, p)=$ $-[y * \log (p)+(1-y) * \log (1-p)]$. Therefore, the total likelihood (for a logistic regression model) can be written as $L(\mathbf{y}, \mathbf{p})=\prod_{j=1}^{n} p_{j}^{y_{j}} *\left(1-p_{j}\right)^{1-y_{j}} \Longrightarrow C E(\mathbf{y}, \mathbf{p})=-\left[\sum_{j=1}^{n} y_{j} * \log \left(p_{j}\right)+\left(1-y_{j}\right) * \log \left(1-p_{j}\right)\right]$ assuming all the examples are independent. Interestingly, the logistic regression model simplifies to a link function of the form: $X_{j} \beta=\eta_{j}=h\left(p_{j}\right)=\operatorname{logit}\left(p_{j}\right)=\ln \left(\frac{p_{j}}{1-p_{j}}\right)$. 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 $\Longrightarrow \ln \left(\frac{p}{1-p}\right)=\frac{1}{t} \ln \left(\frac{2+t \eta}{2-t \eta}\right) \Longrightarrow t * \ln \left(\frac{p}{1-p}\right)=\ln \left(\left(\frac{p}{1-p}\right)^{t}\right)=\ln \left(\frac{2+t \eta}{2-t \eta}\right)-$ equation 1
Assuming the quantity within the bracket on RHS of equation 1 is positive, applying componendo and dividendo: $\left(\frac{p}{1-p}\right)^{t}+1=\frac{4}{2-t \eta} \Longrightarrow \eta=\frac{2\left[\left(\frac{p}{1-p}\right)^{t}-1\right]}{t\left[\left(\frac{p}{1-p}\right)^{t}+1\right]}$
Applying limit and L'Hospital rule: $\lim _{t \rightarrow 0} \eta=\lim _{t \rightarrow 0} \frac{2\left(\frac{p}{1-p}\right)^{t} \ln \left(\frac{p}{1-p}\right)}{\left(\frac{p}{1-p}\right)^{t}+1+t\left(\frac{p}{1-p}\right)^{t} \ln \left(\frac{p}{1-p}\right)}=\ln \left(\frac{p}{1-p}\right)$
Therefore, the new link function carries the same properties as logistic regression when $t=0$. Also, we observe that the link is symmetric about $\mathrm{t}=0: h(\eta,-t)=\frac{1}{-t} \ln \left(\frac{2-t \eta}{2+t \eta}\right)=\frac{1}{t} \ln \left(\frac{2+t \eta}{2-t \eta}\right)=h(\eta, t)$
Therefore, a model with the new link function is guaranteed to perform at par with logistic regression for $\mathrm{t}=$ 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 $\eta$, 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:
$\frac{2+t \eta}{2-t \eta}>0 \Longrightarrow(2+t \eta)(2-t \eta)>0$ assuming $(2-t \eta) \neq 0$
$\Longrightarrow-2<t \eta<2$ which may not be satisfied if we have random x on testing sets that has larger absolute value of $\eta$ than the training set

## Reasonable adjustment

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

## Link, inverse link, gradient of inverse link

## Link function

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

## Inverse link

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

## Gradient of link inverse with respect to $\eta$

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

## Putting things together: Newton method

$L(\mathbf{x}, \beta)=\prod_{i=1}^{n} p_{i}^{y_{i}}\left(1-p_{i}\right)^{1-y_{i}} \Longrightarrow \hat{\beta}=\operatorname{argmax} L(\mathbf{x}, \beta)$
Since $\log$ is a monotonous transformation, it does not change the actual value(s) of $\beta$ for which the likelihood is maximized. Therefore, $l(\mathbf{x}, \beta)=\log (L(\mathbf{x}, \beta)) \Longrightarrow \hat{\beta}=\operatorname{argmax}_{\beta} L(\mathbf{x}, \beta)=\operatorname{argmax}_{\beta} l(\mathbf{x}, \beta)$
$l(\mathbf{x}, \beta)=\log \left(\prod_{i=1}^{n} p_{i}^{y_{i}}\left(1-p_{i}\right)^{1-y_{i}}\right)=\sum_{i=1}^{n}\left[y_{i} \log \left(\frac{p_{i}}{1-p_{i}}\right)+\log \left(1-p_{i}\right)\right]=\sum_{i=1}^{n}\left[\log \left(p_{i}\right)-\left(1-y_{i}\right) x \beta\right]$
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 $\beta$. Finally Newton's second order optimization update can be applied: $\beta:=\beta-H^{-1} J$

