CS 194 - 172 HW 2

Collaborators: Zaid Ahmad

Oct 4th 2021

Stefan Bielmeier

1a) Prove (1) $\dfrac{e^{\beta_0+\beta_1 \cdot x}}{1+e^{\beta_0+\beta_1 \cdot x}}$ = (2) $\dfrac{1}{1+e^{-\beta_0-\beta_1 \cdot x}}$

$\Rightarrow$ take (1): $\dfrac{e^{\beta_0+\beta_1 \cdot x} \cdot e^{-\beta_0-\beta_1 \cdot x}}{\left(1+e^{\beta_0+\beta_1 \cdot x}\right)\left(e^{-\beta_0-\beta_1 \cdot x}\right)} = \dfrac{e^{\beta_0+\beta_1 \cdot x-\beta_0-\beta_1 \cdot x}}{e^{-\beta_0-\beta_1 \cdot x}+e^{\beta_0+\beta_1 \cdot x-\beta_0-\beta_1 \cdot x}}$

$$= \frac{e^0}{e^{-\beta_0-\beta_1 x}+e^0} = \frac{1}{1+e^{-\beta_0-\beta_1 \cdot x}} = (2) \checkmark\checkmark$$

$1 \hat{=} 1 \hat{=} 1$ number one, sorry for any confusion.

1b) $g(z) = \dfrac{1}{1+e^{-z}} \Rightarrow g'(z) = \dfrac{0 \cdot (1+e^{-z}) - (1+e^{-z})' \cdot 1}{(1+e^{-z})^2}$

$$= \frac{-(0-e^{-z})}{(1+e^{-z})^2} = \frac{e^{-z}}{(1+e^{-z})^2} = \underbrace{\frac{1}{(1+e^{-z})}}_{g(z) \checkmark} \cdot \underbrace{\left(\frac{e^{-z}}{(1+e^{-z})}\right)}_{\text{get it to } 1-g(z)!}$$

$$= \frac{1}{(1+e^{-z})} \cdot \left(\frac{1+e^{-z}-1}{1+e^{-z}}\right) = \frac{1}{1+e^{-z}} \left(\frac{1+e^{-z}}{1+e^{-z}} - \frac{1}{1+e^{-z}}\right) = \frac{1}{1+e^{-z}} \cdot \left(1 - \frac{1}{1+e^{-z}}\right)$$

$$= g(z) \cdot (1-g(z)) \checkmark\checkmark = g'(z)$$

SB

1c) $f(x; \beta_0, \beta_1) = \frac{1}{1+e^{-\beta_0-\beta_1 \cdot x}} = \frac{1}{1+e^{-(\beta_0+\beta_1 \cdot x)}}$

$\Rightarrow$ with $z = \beta_0 + \beta_1 \cdot x$

$\Rightarrow f(x; \beta_0, \beta_1) = g(\beta_0 + \beta_1 \cdot x)$ with $g(z) = \frac{1}{1+e^{-z}}$

1d) We will go from general to specific to answer this question.
Probability of observing $\underbrace{\text{a 1 given}}_{\text{data}}$ a specific datapoint x & β_0, β_1 (model)

$\Rightarrow P(Y=1 \mid x; \beta_0, \beta_1) = \frac{1}{1+e^{-(\beta_0+\beta_1 \cdot x)}}$, using rule of total probability & knowing only 2 outcomes,

$P(Y=0 \mid x; \beta_0, \beta_1) = 1 - \frac{1}{1+e^{-(\beta_0+\beta_1 \cdot x)}} \Leftarrow 0 \,\&\, 1$

$\Rightarrow$ for any given data point (y,x) $P(Y=y \mid x, \beta_0, \beta_1) = \left(\frac{1}{1+e^{-(\beta_0+\beta_1 \cdot x)}}\right)^{y} \cdot \left(\frac{1}{1+e^{(\beta_0+\beta_1 \cdot x)}}\right)^{(1-y)}$

(Bernoulli probability mass function)

$\Rightarrow$ Now, for the equivalent of a sequence in Markov model, a dataset with n observations, the likelihood of a logistic regression is:

$\Rightarrow L(\beta_0, \beta_1) = \prod_{i=1}^{n} P(Y=y_i \mid X=x_i, \beta_0, \beta_1)$, using $\frac{1}{1+e^{-(\beta_0+\beta_1 x)}} = g(\beta_0+\beta_1 \cdot x)$

$$= \prod_{i=1}^{n} \left(g(\beta_0+\beta_1 \cdot x_i)\right)^{y_i} \cdot \left(1-g(\beta_0+\beta_1 x_i)\right)^{1-y_i}$$

$$\Rightarrow \log L(\beta_0, \beta_1) = \sum_{i=1}^{n} \log\left(g(\beta_0+\beta_1 \cdot x_i)^{y_i} \cdot (1-g(\beta_0+\beta_1 \cdot x_i))^{1-y_i} \right)$$

$$= \sum_{i=1}^{n} \log\left(g(\beta_0+\beta_1 \cdot x_i)^{y_i}\right) + \log\left((1-g(\beta_0+\beta_1 \cdot x_i))^{1-y_i}\right)$$

$$= \sum_{i=1}^{n} y_i \log[g(\beta_0+\beta_1 \cdot x_i)] + (1-y_i) \cdot \log[1-g(\beta_0+\beta_1 \cdot x_i)]$$

1e) The log function is a monotonically increasing function, as $f(x) = \log x$, and $f'(x) = \frac{1}{x}$ if log is to the base e ~~Because~~

SB
Oct 4th
2021
CS194-172

With $x > 0$ as per definition of the log, $\frac{1}{x} > 0$!
That means if we maximize x,* $\log_e x$ will also maximize. ⇒ OK to use, and more convenient to compute.

*which we do, max $L(\beta_0, \beta_1)$

1f) There are 2 ways to calculate the gradients – one extensive, one simpler by substitution. I'll do one each. $\underline{\log = \ln}$

$$\underline{\frac{\partial \log L(\beta_0, \beta_1)}{\partial \beta_0}} = \sum_{i=1}^{n} y_i \cdot \underbrace{\frac{1}{\cancel{g(\beta_0+\beta_1 x_i)}}}_{\frac{\partial \log(g)}{\partial g}} \cdot \underbrace{\cancel{g(\beta_0+\beta_1 \cdot x_i)} \cdot (1-g(\beta_0+\beta_1 \cdot x_i))}_{\frac{\partial g(\beta_0+\beta_1 \cdot x_i)}{\partial(\beta_0+\beta_1 \cdot x_i)}} \cdot \underbrace{1}_{\frac{\partial \beta_0+\beta_1 \cdot x_i}{\partial \beta_0}}$$

$$+ (1-y_i) \cdot \frac{1}{\cancel{1-g(\beta_0+\beta_1 x_i)}} \cdot (-1) \cdot g(\beta_0+\beta_1 \cdot x_i) \cdot \cancel{(1-g(\beta_0+\beta_1 \cdot x_i))} \cdot 1$$

$$= \sum_{i=1}^{n} y_i - \cancel{y_i\, g(\beta_0+\beta_1 \cdot x_i)} - g(\beta_0+\beta_1 \cdot x_i) + \cancel{y_i\, g(\beta_0+\beta_1 \cdot x_i)}$$

$$= \underline{\underline{\sum_{i=1}^{n} y_i - g(\beta_0+\beta_1 \cdot x_i)}} = \sum_{i=1}^{n} y_i - \frac{1}{1+e^{-(\beta_0+\beta_1 \cdot x_i)}}$$

for $\underline{\frac{\partial \log L(\beta_0, \beta_1)}{\partial \beta_1}}$ we use: $f = \log(~~\ldots~~)$ & $h = \beta_0+\beta_1 \cdot x_i$ with $f' = \frac{\partial f}{\partial g(h)}$ & $h' = \frac{\partial h}{\partial \beta_1}$

$$= \sum_{i=1}^{n} y_i \cdot f'(g(h)) \cdot g'(h) \cdot h' + (1-y_i) \cdot f'(1-g(h)) \cdot (1-g(h))' \cdot h'$$

$$= \sum_{i=1}^{n} y_i \cdot \frac{1}{\cancel{g(h)}} \cdot \cancel{g(h)}\,(1-g(h)) \cdot \overbrace{\frac{\partial h}{\partial \beta_1}}^{=x_i} + (1-y_i) \frac{1}{\cancel{1-g(h)}} (-1) \cdot g(h) \cdot \cancel{(1-g(h))} \cdot x_i$$

$$= \sum_{i=1}^{n} (y_i - \cancel{y_i\, g(h)} - g(h) + \cancel{y_i\, g(h)}) \cdot x_i = \sum_{i=1}^{n} (y_i - g(h)) \cdot x_i$$

$$= \sum_{i=1}^{n} \left(y_i - \frac{1}{1+e^{-(\beta_0+\beta_1 \cdot x_i)}}\right) \cdot x_i \quad ✓✓$$

1g) Initialize $\beta_0 = 0$, $\beta_1 = 0$;

SB

Do:

$$\beta_0^{new} = \beta_0^{old} + \alpha \cdot \sum_{i=1}^{n} y_i - g(\beta_0^{old} + \beta_1^{old} \cdot x_i)$$

$$\beta_1^{new} = \beta_1^{old} + \alpha \sum_{i=1}^{n} (y_i - g(\beta_0^{old} + \beta_1^{old} \cdot x_i)) \cdot x_i$$

Until: $\|\beta_0^{new} - \beta_0^{old}\| <$ threshold and $\|\beta_1^{new} - \beta_1^{old}\| <$ threshold

α being step size of gradient ascent.

Homework 2 - ct'd

SB
CompGen
Oct 8th 2021

2a) Prove that $FWER \leq \alpha$ for rejecting all H_0^j at $p_j \leq \frac{\alpha}{m}$ (after Bonferri correction)

$FWER = P(\text{rejecting} \geq 1 H_0^j \text{ falsely})$ for all $H_1, \dots, H_m$

$= P(\bigcup_{j=1}^{m} \text{rejecting } H_0^j \text{ falsely})$

knowing that $P(A \cup B) \leq P(A) + P(B)$, no matter if A and B are independent

$\Rightarrow FWER = P(\bigcup_{j=1}^{m} \text{rejecting } H_0^j \text{ falsely}) \leq \sum_{j=1}^{m} P(\text{rejecting } H_0^j \text{ falsely})$

$\Rightarrow FWER(\alpha) \leq \sum_{j=1}^{m} \alpha$, with Bonferri correction follows

$\Rightarrow FWER(\alpha) \leq \sum_{j=1}^{m} \frac{\alpha}{m} = m \cdot \frac{\alpha}{m} = \alpha$

$\Rightarrow \underline{FWER(\alpha) \leq \alpha}$

Bonferri correction controls $FWER(\alpha)$ at $\leq \alpha$! for a family of hypothesis. The p-values need not be independent because of beer use of inequality - we do not require to know if events / Hypothesis are independent. (which is probably helpful in variant analysis because of LD)