Zaki (2020) 624 635

Logistic Regression

Aim of log reg is to predict probs of response variable based on the indep variables

log reg = classification Tech that for a given point $x_i \in \mathbb{R}^d$ preds $P(G \mid x_i)$ for each class in the domain of Y

log reg: indep vars $X_1, X_2, \ldots, X_n$
Bin or Bernulli response $Y \in \{0,1\}$
Like linear Reg, Aug data to add X_0 value = 1 for bias weight
$\tilde{x}_i = (1, x_1, x_2, \ldots, x_d)^T \in \mathbb{R}^{d+1}$

In the Banary log reg there are only two outputs: So the prob mass function for $\tilde{X} = \tilde{x}$ is:

$$P(Y=1 \mid \tilde{X} = \tilde{x}) = \pi(\tilde{x})$$

means prob function

$$P(Y=0 \mid \tilde{X} = \tilde{x}) = 1 - \pi(\tilde{x})$$

So, $\pi(\tilde{x}) = Y = 1$ given $\tilde{X} = \tilde{x}$

Also denotes Expected value of Y given $\tilde{X} = \tilde{x}$

thus, in logistic regression the goal is to predict the prob

this is in conflict to say Bayes where the prob is observed & classif matter most

NOTE:
$P(Y=1 \mid \tilde{X} = \tilde{x})$ also = Exp Val of Y giv $\tilde{X} = \tilde{x}$

Since we want a probabilis it is not a correct to use the lnear reg model

This because we need output to be between 0 - 1

the sigmoid function achieves this

$$\Theta(z) = \frac{1}{1 + \exp\{-z\}} = \frac{\exp\{z\}}{1 + \exp\{z\}}$$

logistic reg model is defined as follow:

$$P(Y=1 \mid \hat{X} = \hat{x}) = \pi(\hat{x}) = \Theta(f(\hat{x})) = \Theta(\hat{w}^T \hat{X}) = \frac{\exp\{\hat{w}^T \hat{X}\}}{1 + \exp\{\hat{w}^T \hat{X}\}}$$

$$P(Y=0 \mid \hat{X} = \hat{x}) = 1 - P(Y=1 \mid \hat{X} = \hat{x}) = \Theta(-\hat{w}^T \hat{X}) = \frac{1}{1 + \exp\{\hat{w}^T \hat{X}\}}$$

here the probs for $Y=1$ & $Y=0$ are demonstrated using Different versions of the Sigmoid functions

the $Y=1$ uses the form $\frac{\exp(w^T x)}{1+\exp\{w^T x\}}$

this is because it shows how the prob of $y=1$ is related to the expo of the linear comb of $w^T x$

Exmpasis that as the ~~expo~~ (C increase so does the prob

Why?

- The term is in the numerator so over-all value of the fraction must increase as it does
- ~~term~~ remember it is expo so as $w^T x$ increase output does by allot

$y = (1)$ uses the form $\frac{1}{1 + \exp(w^T x)}$

this emphasizes ~~that as $w^T x$ increases, then~~ the prob is related to the inverse of the expo

As $w^T x$ increases the prob decreases

Perpendicular grows as $w^T x$ does

NOTE: both forms are the same and give the same result, this is just used for easy reading & interpretation

Combining the two outcome cases ($y=0$, $y=1$) the full logistic reg Model is given as:

$$P(Y=\tilde{Y}=\tilde{x}) = \Theta(\tilde{\omega}^T\tilde{x})^{\tilde{Y}} \cdot \Theta(-\tilde{\omega}^T\tilde{x})^{1-Y}$$

note $1-\Theta(z)$ same as $\Theta(-z)$

So $P(Y|\tilde{Y}=\tilde{x}) = \Theta(\tilde{\omega}^T\tilde{x}) = Y=1$

$P(Y|\tilde{Y}=\tilde{x}) = \Theta(-\tilde{\omega}^T\tilde{x}) = Y=0$

Log Odd Ratio

for occurance $Y=1$:

← just odds, not log

$$\text{odds}(Y=1 \mid \tilde{X}=\tilde{x}) = \frac{P(Y=1 \mid \tilde{X}=\tilde{x})}{P(Y=0 \mid \tilde{X}=\tilde{x})}$$

= Ratio of Probabilities

= General rule = the "Happen" class on top

$$= \frac{\exp\{\tilde{\omega}^T \tilde{x}\}}{1+\exp\{\tilde{\omega}^T \tilde{x}\}} \cdot \left(1+\exp\{\tilde{\omega}^T \tilde{x}\}\right)$$

$$= \exp\{\tilde{\omega}^T \tilde{x}\}$$

Next, we want to transform this into the log odds ratio. Why?

- odds are uncapped: 0 to ∞. large probs are hard to deal with
- log odds transforms to Neginf to Pos inf
- Makes it easier to compare coeffs
- A positive log odds = increased odds of $Y=1$
- A Negative log odds = decreases odds of $Y=1$

- Creates a linear odds ratio w/ preds
  - ratio of probs & sigmoid allow for inf & exponential values
- Allows for symetry of numbers used. w/ odds .2 & ½ represent same magnitude of change in diff directions (half & double) but are very diff numerically. log odds creates a 0 bounds & symetry: 2 = 0.7, ½ = -0.7

the logarithm of the odds ratio:

$$\ln(\text{odds}(Y=1 \mid \vec{X}=\vec{x})) = \ln\left(\frac{P(Y=1 \mid \vec{X}=\vec{x})}{P(Y=0 \mid \vec{X}=\vec{x})}\right)$$

$$= \ln(\exp\{\vec{\omega}^T \vec{x}\}) = \vec{\omega}^T \vec{x}$$

This is just the output of the linear Model

- log odds are linearly related to predictors
- coeffs (ω) rep change in log odds for one-unit Δ in pred
- log reg is modelling log-odds as linear preds of the predictors

the log odd ratio function is called logit

$$\text{logit}(z) = \ln\left(\frac{z}{1-z}\right)$$

where z = odd ratio calced by $\frac{p}{1-p}$

This is the inverse of the logistic function

$$\underbrace{\ln(\text{odds}(Y=1 \mid \vec{X}=\vec{x}))}_{\text{log function}} = \text{logit}(P(Y=1 \mid \vec{X}=\vec{x}))$$

Same as $\frac{1}{1+\exp(\vec{w}^T \vec{x})}$

logistic: $\frac{1}{1+\exp(\vec{w}^T \vec{x})}$

- takes in real number at input
- outputs value between 0 & 1
- Maps real number to Prob

logit:

- takes Prob or odd ratio as input
- Produces log odds
- Maps Prob to real number

Hence the two reverse each other

the logistic regression model is therefore based on the assumption that the log odds ratio for $Y=1$ given $\vec{X}=\vec{x}$ is a linear function of the independent

24.1.1 Maximum likelihood Estimation

Let D be train w/ n pomts x_i & labels

Let $\tilde{w} = (w_0, w_1, w_2, \ldots, w_d)^T$ be the Augment weight vector for est $\tilde{w}$

We will uses Max likelihood approach to learn the weight vector

likelihood = Prob of observing data given the est parametes $\hat{w}$

likelihood of observed resp is given as: $L(\hat{w}) = P(Y = \tilde{w}) =$

$$= \prod_{i=1}^{n} P(y_i \mid x_i) = \prod_{i=1}^{n} \theta(\tilde{w}^T \tilde{x}_i)^{y_i} \cdot \theta(-\tilde{w}^T \tilde{x}_i)^{1-y_i}$$

However, instead of trying to max the liklihood, we can max the log of likelihood to convert the product into a summation

$$\ln(L(\tilde{w})) = \sum y_i \cdot \ln(\theta(\tilde{w}^T \bar{x}_i)) + (1-y_i) \cdot \ln(\theta(-\tilde{w}^T \bar{x}))$$

the neg of the log-likelihood can be considered an error func:

cross-entrophy error function

$$E(\tilde{w}) = -\ln(L(\tilde{w})) =$$

$$= \sum y_i \cdot \ln\left(\frac{1}{\theta(\tilde{w}^T \bar{x}_i)}\right) + (1-y_i) \cdot \ln\left(\frac{1}{1-\theta(\tilde{w}^T \bar{x}_i)}\right)$$

the task of maxing the log-like is there the same as min the cross entrophy error

Typically, to solve the optimal weight vector $\tilde{w}$ we would Diff the log-like funct w/ resp to $\tilde{w}$ & set to 0 and solve for $\tilde{w}$

However, there results in no-closed form to solve for $\tilde{w}$. That is not $\tilde{w}$ term remains

instead use ~~it~~ iterative gradient Ascent to comp optimal value as log-likli function is concave

This method relies on the gradient of the log-lik which is obtained by partial Derivative w/ respect to $\hat{w}$

$$\text{Change in } \hat{w} = \nabla \hat{w} = \frac{\partial}{\partial \hat{w}} \{ \ln(L(\tilde{w})) \} =$$

$$\frac{\partial}{\partial \tilde{w}} \{ \underbrace{\sum y_i \cdot \ln(\theta(z_i)) + (1-y_i) \cdot \ln(\theta(-z))}_{\text{this is just the log-likli}} \}$$

$z_i = \hat{w}^T \tilde{x}$

To obtain $\nabla(\tilde{w})$ we need to use the chain rule to obtain deriv of $Ln(\tilde{\theta}(z_i))$ wi respect to $\tilde{v}$

for $y=1 \longrightarrow \tilde{\theta}(-z_i) \cdot x_i$

$y=2 \longrightarrow -\tilde{\theta}(z_i) \cdot \tilde{x}_i$

replace $Ln(\tilde{\theta}(z_i))$ for respective

$$\nabla(\tilde{w}) = \sum^{n} y_i \cdot \tilde{\theta}(-z_i) \cdot \tilde{x}_i - (1-y_i) \cdot \tilde{\theta}(z_i) \cdot \tilde{x}_i$$

$$= \sum^{n} \left(\tilde{\theta}(-z_i) + \tilde{\theta}(z_i)\right) \cdot x_i - \tilde{\theta}(z_i) \cdot x_i$$

$$= \sum^{n} \left(y_i - \tilde{\theta}(z_i)\right) \cdot \tilde{x}_i \quad \text{Since} = \textcircled{a}\, 1$$

$$= \sum^{n} \left(y_i - \tilde{\theta}(\tilde{w}^T \tilde{x})\right) \cdot \tilde{x}_i \longrightarrow \nabla(\tilde{w})$$

the sum of the class minus the linear output times the data point

from here the gradient ascent algm starts for some inital estimate of $\tilde{w} = \tilde{w}^0$

At each step t the method moves in the direction of the steepest ascent which given by the gradient vector $\nabla \tilde{w}$

Note: $\tilde{w}^t$ = est/start weight vector

$\Delta \tilde{w}^t$ = w^t plugged into part Deriv

$\Delta \tilde{w}^t$ is the direction & size to update

if $\tilde{w}^t$ is current we obtain the next by:

$$\tilde{w}^{t+1} = \tilde{w}^t + \eta \cdot \nabla(\tilde{w}^t)$$

$\eta > 0$ is user specified para called learning rate

if too large estimates will vary too much

too small & converges will be too costly

@ optimal $\tilde{w}$, $\nabla(\tilde{w}) = 0$

It is important to note / remember that each partial derivatives measures how sens the functoi is to changes in that specific variable ($\tilde{w}$)

A larger partial Deriv = a more rapid movement in a direction

note, when updating we are updating the whole weight vector

the gradient vector $\nabla \tilde{w}$ is the output of the partial derivatives for each weight

the weight vector is updated after each partial derivative stage

for computing optimal $\tilde{w}$

Stochastic Gradient Ascent

the previous gradient Ascent calc computes by considering all data points X_i

It is also called Batch GA

for large datasets, faster to compute gradient by considering one randomly chose point

the weight vector is updated at each partial Derivative stage

= SGA for compute optimal $\vec{w}$

$$\nabla(\vec{w}_i, \vec{x}_i) = (y_i - \theta(\vec{w}^T \vec{x}_i)) \cdot \vec{x}_i$$

SGA = MUCH Quicker & cheaper comp to Batch

Predictions

Once the model has be trained i.e. we know the optimal $\tilde{w}$ weight vector

we predict using

$$\hat{y} = \begin{cases} 1 & \text{if } \theta(\tilde{w}^T \tilde{z}) \geq 0.5 \\ 0 & \text{if } \theta(\tilde{w}^T \tilde{z}) < 0.5 \end{cases}$$