GRADIENT & HESSIAN FOR BINARY LOGISTIC REGRESSION

$$MLE = \prod_{\forall y=1} p(y^{(i)}=1 \mid x^{(i)}) \prod_{\forall y=0} p(y^{(i)}=0 \mid x^{(i)})$$

$$\ln(MLE) =$$

$$L(w) = \sum_{\forall y=1} \ln(p(\cdot)) + \sum_{\forall y=0} \ln(p(\cdot))$$

$$y^{(+)} = 1$$

$$y^{(-)} = 0$$

$$= \sum_{\forall y} y^{(i)} \ln(p(y^{(i)}=1 \ldots)) + (1-y^{(i)}) \ln(p(y^{(i)}=0 \ldots))$$

BUT $p(y^{(i)}=1) = g(x^{(i)}) = \frac{1}{1-\exp(-w^T x^{(i)})}$

$p(y^{(i)}=0) = 1 - g(x^{(i)})$.

$$L(w) = \sum_i y^{(i)} \ln(g(x^{(i)})) + (1-y^{(i)}) \ln(1-g(x^{(i)}))$$

- Want to maximize iteratively

$$\frac{\partial}{\partial w_j} L(w) = \sum_i y^{(i)} \frac{\partial}{\partial w_j} \ln(g) + (1-y^{(i)}) \frac{\partial}{\partial w_j} \ln(1-g(\ldots))$$

$\vdots$ (SIMPLIFIES!) (SEE VIDEO)

$$= \sum_i (y^{(i)} - g(x^{(i)})) x_j^{(i)}$$

$$w := w + \eta \sum_i (y^{(i)} - \hat{y}^{(i)}) x^{(i)}$$

This is gradient descent update equation

$$\frac{\partial}{\partial w_j} \ell(w) = \sum_i \left(y^{(i)} - g(x^{(i)})\right) x_j^{(i)}$$

PLUG IN

$$H[k,j] = \frac{\partial}{\partial w_k}\frac{\partial}{\partial w_j}\ell = \frac{\partial}{\partial w_k}\left(\sum_i \left(y^{(i)} - g(x^{(i)})\right) x_j^{(i)}\right)$$

$$= \sum_i \frac{\partial}{\partial w_k} \overset{0}{y^{(i)}} x_j^{(i)} - \frac{\partial}{\partial w_k} g(x^{(i)}) x_j^{(i)}$$

LEFT OVER TERM

$$= -\sum_i \frac{\partial}{\partial w_k} g(x^{(i)}) x_j^{(i)}$$

Already know $\frac{\partial}{\partial w_k} g(w^T x^{(i)})$ side calculation

$$= g(x^{(i)})\left(1 - g(x^{(i)})\right) \frac{\partial}{\partial w_k}\left(w^T x^{(i)}\right)$$

$$= g(x^{(i)})\left(1 - g(x^{(i)})\right) x_k^{(i)}$$

PLUG IN

$$\therefore = -\sum_i g(x^{(i)})\left(1 - g(x^{(i)})\right) x_k^{(i)} x_j^{(i)}$$

THAT'S THE HESSIAN!

$H(k,j) =$ This is a valid equation for the Hessian, but we want to represent it using linear algebra

$$\boxed{= -\sum_i \underset{\text{SIGMOID}}{g(x^{(i)})}\left(1 - g(x^{(i)})\right) x_k^{(i)} x_j^{(i)}}$$

USE LINEAR ALG.

$$= \begin{bmatrix} \sum_i g^{(i)}(1-g^{(i)})(x_1^{(i)})^2 & \sum_i g(1-g) x_1^{(i)} x_2^{(i)} & \cdots & \sum g(1-g) x_1^{(i)} x_M^{(i)} \\ & \ddots & & \\ \sum_i g(1-g) x_N^{(i)} x_1^{(i)} & \cdots & & \sum g(1-g)(x_N^{(i)})^2 \end{bmatrix}$$

$$= \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ x^{(1)} & x^{(2)} & \cdots & x^{M} \\ \downarrow & \downarrow & & \downarrow \end{bmatrix} \begin{bmatrix} g^1(1-g^1) & & & 0 \\ 0 & g^2(1-g^2) & & \\ \vdots & & \ddots & \\ 0 & & & g^N(1-g^N) \end{bmatrix} \begin{bmatrix} — & x^{(1)} & — \\ — & x^{(2)} & — \\ & \vdots & \\ & x^{(M)} & \end{bmatrix}$$

$$g^1(1-g^1) x_1^{(1)} + g^2(1-g^2) x_1^{(2)} \cdots = \sum_i g(1-g) x_1^{(i)} x_1^{(i)} \checkmark$$

X transpose

DIAG[g(1-g)]

X

New Update:

$$w := w + \eta\, H^{-1} \nabla l(w)$$

$$:= w + \eta \underbrace{\underset{N \times N}{H[l(w))]^{-1}} \; \underset{(N \times N)}{\nabla l(w)}}_{N \times 1}$$

where

$$H[l(w)] = X^T \cdot \underset{i}{\mathrm{DIAG}}\left(g(x^{(i)})(1-g(x^{(i)}))\right) \cdot X$$

HESSIAN FOR BINARY LR

AS LINEAR ALGEBRA OPERATIONS