Deriving the gradient for logistic regression

VERY OPTIONAL

Log-likelihood function

- Goal: choose coefficients **w** maximizing likelihood:

$$\ell(\mathbf{w}) = \prod_{i=1}^{N} P(y_i \mid \mathbf{x}_i, \mathbf{w})$$

- Math simplified by using log-likelihood – taking (natural) log:

$$\ell\ell(\mathbf{w}) = \ln \prod_{i=1}^{N} P(y_i \mid \mathbf{x}_i, \mathbf{w})$$

natural
log

The log trick, often used in ML…

- Products become sums:

$\ln a \cdot b = \ln a + \ln b$ | $\ln \frac{a}{b} = \ln a - \ln b$

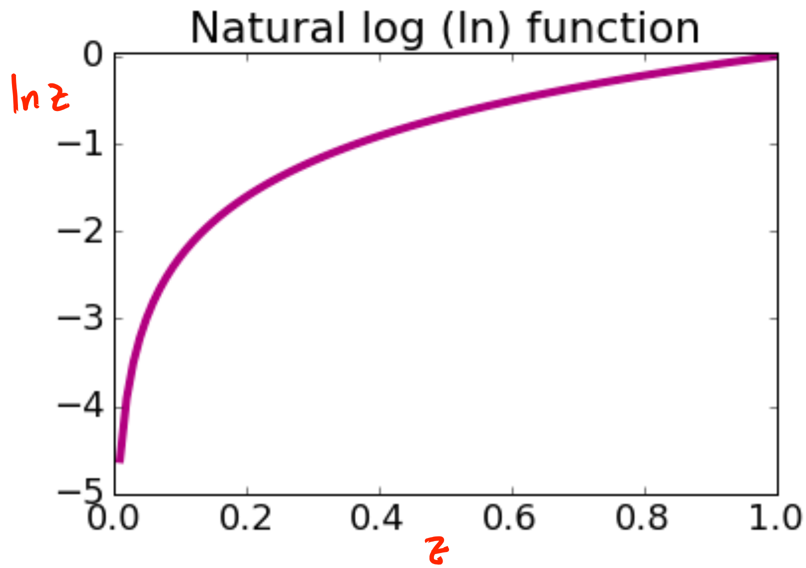
- Doesn't change maximum!
  - If $\hat{\mathbf{w}}$ maximizes f($\mathbf{w}$):

$$\hat{w} = \underbrace{\arg\max_{w}}_{\text{the } w \text{ that makes } f(w) \text{ largest}} f(w)$$

  - Then $\hat{\mathbf{w}}_{\ln}$ maximizes ln(f($\mathbf{w}$)):

$$\hat{w}_{\ln} = \arg\max_{w} \ln(f(w))$$

$$\hat{w} = \hat{w}_{\ln}$$



Expressing the log-likelihood

Using log to turn products into sums

$$\ln \prod_{i=1}^{N} f_i = \sum_{i=1}^{N} \ln f_i$$

- The log of the product of likelihoods becomes the sum of the logs:

$$\ell\ell(\mathbf{w}) = \ln \prod_{i=1}^{N} P(y_i \mid \mathbf{x}_i, \mathbf{w})$$

$$= \sum_{i=1}^{N} \ln P(y_i \mid x_i, w)$$

Rewriting log-likelihood

- For simpler math, we'll rewrite likelihood with indicators:

$$\ell\ell(\mathbf{w}) = \sum_{i=1}^{N} \ln P(y_i \mid \mathbf{x}_i, \mathbf{w})$$

$$= \sum_{i=1}^{N} [\mathbb{1}[y_i = +1] \ln P(y = +1 \mid \mathbf{x}_i, \mathbf{w}) + \mathbb{1}[y_i = -1] \ln P(y = -1 \mid \mathbf{x}_i, \mathbf{w})]$$

if $y_i = +1$ ✓ 0

if $y_i = -1$ 0 ✓

Deriving probability that y=-1 given x

VERY OPTIONAL

Logistic regression model: $P(y=-1|\mathbf{x},\mathbf{w})$

- Probability model predicts y=+1:

$$P(y=+1|\mathbf{x},\mathbf{w}) = \frac{1}{1+e^{-\mathbf{w}^\top h(\mathbf{x})}}$$

- Probability model predicts y=-1:

$$P(y=-1|x,w) = 1 - P(y=+1|x,w) = 1 - \frac{1}{1+e^{-w^T h(x)}}$$

$$= \frac{1+e^{-w^T h(x)} - 1}{1+e^{-w^T h(x)}} = \frac{e^{-w^T h(x)}}{1+e^{-w^T h(x)}}$$

Rewriting the log-likelihood

VERY OPTIONAL

Plugging in logistic function for 1 data point

$$P(y=+1 \mid \mathbf{x},\mathbf{w}) = \frac{1}{1+e^{-\mathbf{w}^\top h(\mathbf{x})}} \qquad P(y=-1 \mid \mathbf{x},\mathbf{w}) = \frac{e^{-\mathbf{w}^\top h(\mathbf{x})}}{1+e^{-\mathbf{w}^\top h(\mathbf{x})}}$$

$$\ell\ell(\mathbf{w}) = \mathbb{1}[y_i=+1]\ln P(y=+1 \mid \mathbf{x}_i,\mathbf{w}) + \mathbb{1}[y_i=-1]\ln P(y=-1 \mid \mathbf{x}_i,\mathbf{w})$$

$$= \mathbb{1}[y_i=+1] \ln \frac{1}{1+e^{-w^T h(x_i)}} + (1-\mathbb{1}[y_i=+1]) \ln \frac{e^{-w^T h(x_i)}}{1+e^{-w^T h(x_i)}}$$

$$= -\mathbb{1}[y_i=+1] \ln(1+e^{-w^T h(x_i)}) + (1-\mathbb{1}[y_i=+1])[-w^T h(x_i) - \ln(1+e^{-w^T h(x_i)})]$$

$$\underbrace{= -(1-\mathbb{1}[y_i=+1])\, w^T h(x_i) - \ln(1+e^{-w^T h(x_i)})}_{\text{Simpler form}}$$

$$\ln e^a = a$$

$$\mathbb{1}[y_i=-1] = 1-\mathbb{1}[y_i=+1]$$

$$\ln \frac{1}{1+e^{-w^T h(x_i)}} = -\ln(1+e^{-w^T h(x_i)})$$

$$\ln \frac{e^{-w^T h(x_i)}}{1+e^{-w^T h(x_i)}} = \underbrace{\ln e^{-w^T h(x_i)}}_{-w^T h(x_i)} - \ln(1+e^{-w^T h(x_i)})$$

Deriving gradient of log-likelihood

Gradient for 1 data point

$$\ell\ell(\mathbf{w}) = -(1-\mathbb{1}[y_i=+1])\mathbf{w}^\top h(\mathbf{x}_i) - \ln\left(1+e^{-\mathbf{w}^\top h(\mathbf{x}_i)}\right)$$

$$\frac{\partial \ell\ell}{\partial w_j} = -(1-\mathbb{1}[y_i=+1])\frac{\partial}{\partial w_j} w^T h(x_i) - \frac{\partial}{\partial w_j}\ln\left(1+e^{-w^T h(x_i)}\right)$$

$$= -(1-\mathbb{1}[y_i=+1])\, h_j(x_i) + h_j(x_i)\, P(y=-1|x_i,w)$$

$$= h_j(x_i)\left[\mathbb{1}[y_i=+1] - P(y=+1|x_i,w)\right]$$

$$\frac{\partial}{\partial w_j} w^T h(x_i) = h_j(x_i)$$

$$\frac{\partial}{\partial w_j}\ln\left(1+e^{-w^T h(x_i)}\right)$$

$$= -h_j(x_i)\underbrace{\frac{e^{-w^T h(x_i)}}{1+e^{-w^T h(x_i)}}}_{P(y=-1|x_i,w)}$$

Finally, gradient for all data points

- Gradient for one data point:

$$h_j(\mathbf{x}_i)\Big(\mathbb{1}[y_i = +1] - P(y = +1 \mid \mathbf{x}_i, \mathbf{w})\Big)$$

- Adding over data points:

$$\frac{\partial \ell\ell}{\partial w_j} = \sum_{i=1}^{N} h_j(x_i)\Big(\mathbb{1}[y_i = +1] - \underline{P(y=+1 \mid x_i, w)}\Big)$$