Logistic Model

$$\Pr(y = 1|x) = \frac{1}{1 + \exp(-g(x))}$$

$$where\ g(x) = x^T W_1 x + x^T W_2 + b$$

$$= (x_1 \quad \dots \quad x_d)\begin{pmatrix} w_{1_{11}} & \cdots & w_{1_{1d}} \\ \vdots & \ddots & \vdots \\ w_{1_{d1}} & \cdots & w_{1_{dd}} \end{pmatrix}\begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} + (x_1 \quad \dots \quad x_d)\begin{pmatrix} w_{2_1} \\ \vdots \\ w_{2_d} \end{pmatrix} + b$$

Parameters and Dimensions

$$W_1: d \times d\ (matrix)$$

$$W_2: d \times 1\ (column\ vector)$$

$$b: 1 \times 1\ (scalar)$$

$$Total\ d^2 + d + 1\ number\ of\ trainable\ parameters.$$

Maximum Likelihood

$$\Pr(D; W_1, W_2, b) = \prod_{i|y_i=1} \Pr(x_i|W_1, W_2, b) \times \prod_{i|y_i=0} (1 - \Pr(x_i|W_1, W_2, b))$$

$$\log \Pr(D; W_1, W_2, b) = \sum_{i|y_i=1} \log \Pr(x_i|W_1, W_2, b) + \sum_{i|y_i=0} \log(1 - \Pr(x_i|W_1, W_2, b))$$

$$\boldsymbol{Objective}:\ argmax_{W_1, W_2, b} \log \Pr(D; W_1, W_2, b)$$

Partial Derivatives of Log Likelihood with respect to Parameters

$$\frac{\partial \log \Pr(D; W_1, W_2, b)}{\partial W_1}$$
$$= \sum_{i|y_i=1} \big(1 - \sigma(x^T W_1 x + x^T W_2 + b)\big) * \frac{\partial g}{\partial W_1} + \sum_{i|y_i=0} \sigma(x^T W_1 x + x^T W_2 + b) * \frac{\partial g}{\partial W_1}$$
$$= \sum_{i|y_i=1} \big(1 - \sigma(x^T W_1 x + x^T W_2 + b)\big) x x^T + \sum_{i|y_i=0} \sigma(x^T W_1 x + x^T W_2 + b) x x^T$$

$$\frac{\partial \log \Pr(D; W_1, W_2, b)}{\partial W_2}$$
$$= \sum_{i|y_i=1} \big(1 - \sigma(x^T W_1 x + x^T W_2 + b)\big) * \frac{\partial g}{\partial W_2} + \sum_{i|y_i=0} \sigma(x^T W_1 x + x^T W_2 + b) * \frac{\partial g}{\partial W_2}$$
$$= \sum_{i|y_i=1} \big(1 - \sigma(x^T W_1 x + x^T W_2 + b)\big) x + \sum_{i|y_i=0} \sigma(x^T W_1 x + x^T W_2 + b) x$$

$$\frac{\partial \log \Pr(D; W_1, W_2, b)}{\partial b}$$
$$= \sum_{i|y_i=1} \big(1 - \sigma(x^T W_1 x + x^T W_2 + b)\big) * \frac{\partial g}{\partial b} + \sum_{i|y_i=0} \sigma(x^T W_1 x + x^T W_2 + b) * \frac{\partial g}{\partial b}$$
$$= \sum_{i|y_i=1} \big(1 - \sigma(x^T W_1 x + x^T W_2 + b)\big) + \sum_{i|y_i=0} \sigma(x^T W_1 x + x^T W_2 + b)$$

Gradient ascent pseudocode

$Let\ \gamma\ be\ the\ step\ size.$

$Let\ J(\theta)\ be\ the\ \log likelihood\ above.$

$$J(\theta) = \log \Pr(D; W_1, W_2, b) = \sum_{i|y_i=1} \log \Pr(x_i|W_1, W_2, b) + \sum_{i|y_i=0} \log(1 - \Pr(x_i|W_1, W_2, b))$$

$Initialize\ parameters\ W_1, W_2, b.$

$While\ not\ converged, repeat:$

$$W_{1_{new}} \leftarrow W_{1_{old}} + \gamma * \frac{\partial J}{\partial W_1}$$

$$where\ \frac{\partial J}{\partial W_1} = \sum_{i|y_i=1} \big(1 - \sigma(x^T W_1 x + x^T W_2 + b)\big) x x^T + \sum_{i|y_i=0} \sigma(x^T W_1 x + x^T W_2 + b) x x^T$$

$$W_{2_{new}} \leftarrow W_{2_{old}} + \gamma * \frac{\partial J}{\partial W_2}$$

$$where\ \frac{\partial J}{\partial W_2} = \sum_{i|y_i=1} \big(1 - \sigma(x^T W_1 x + x^T W_2 + b)\big) x + \sum_{i|y_i=0} \sigma(x^T W_1 x + x^T W_2 + b) x$$

$$b_{new} \leftarrow b_{old} + \gamma * \frac{\partial J}{\partial b}$$

$$where\ \frac{\partial J}{\partial b} = \sum_{i|y_i=1} \big(1 - \sigma(x^T W_1 x + x^T W_2 + b)\big) + \sum_{i|y_i=0} \sigma(x^T W_1 x + x^T W_2 + b)$$