

Date: \_\_\_\_\_  
M T W T F S S

Goal: Logistic Regression.  
Binary classification ( $y \in \{0, 1\}$ ) ①

1) Model Definition

$x = \sum_{i=0}^n x_i^i \rightarrow$  feature matrix.

$w = \sum_{i=0}^n w_i^i \rightarrow$  weight matrix.

$$Z = w^T x^i \rightarrow w_0 + w_1 x_1^i + w_2 x_2^i + w_3 x_3^i + \dots + w_n x_n^i$$

$$\hat{y}^{(i)} = \sigma(z^{(i)}) = \frac{1}{1 + e^{-z^{(i)}}} \rightarrow \text{sigmoid}$$

2) Probability of True Label

If  $y^{(i)} = 1 \rightarrow \text{Probability} = \hat{y}^{(i)}$

$y^{(i)} = 0 \rightarrow \text{Probability} = 1 - \hat{y}^{(i)}$

compact single Expression.

$$P(y^{(i)} | x^{(i)}, w) = \hat{y}^{(i)} y^{(i)} \cdot (1 - \hat{y}^{(i)})^{(1 - y^{(i)})}$$

Put  $y=1 \rightarrow \hat{y}^1 \cdot (1 - \hat{y})^0 = \hat{y}$

$y=0 \rightarrow \hat{y}^{(0)} \cdot (1 - \hat{y})^{1-0} = 1 - \hat{y}$

### 3) Likelihood of Entire Dataset

Assuming Independence:

$$L(w) = \prod_{i=1}^m (\hat{y}^{(i)})^{y^{(i)}} (1 - \hat{y}^{(i)})^{1 - y^{(i)}}$$

### 4) log likelihood:

Take natural log (turns product  $\rightarrow$  sum):

$$\ell(w) = \log L(w) = \sum_{i=1}^m (y^{(i)} \cdot \log \hat{y}^{(i)} + (1 - y^{(i)}) \log (1 - \hat{y}^{(i)}))$$

Goal maximize  $\ell(w)$ .

### 5) Cost Function: (Binary Cross-Entropy Loss)

Minimize negative average log likelihood.

$$J(w) = -\frac{1}{m} \ell(w) = -\frac{1}{m} \sum_{i=1}^m [y^{(i)} \cdot \log \hat{y}^{(i)} + (1 - y^{(i)}) \log (1 - \hat{y}^{(i)})]$$

### 6) Compute the Gradient:

Goal:  $\partial J / \partial w_j$  of each weight.

For one sample

$$J^{(i)} = -[y \cdot \log \hat{y} + (1 - y) \log (1 - \hat{y})], \hat{y} = \sigma(z), z = w^T x$$

Goal:  $\partial J^{(i)} / \partial w_j$

③

Required Chain Rule:

1) Derivative w.r.t  $\hat{y}$ :

$$\begin{aligned}\frac{\partial J^{(i)}}{\partial \hat{y}} &= - \left( \frac{y}{\hat{y}} - \frac{1-y}{1-\hat{y}} \right) = \frac{y(1-\hat{y}) - (1-y)\hat{y}}{\hat{y}(1-\hat{y})} \\ &= \frac{\hat{y} - y}{\hat{y}(1-\hat{y})}\end{aligned}$$

2) Derivative of Sigmoid.

$$\frac{d\hat{y}}{dz} = \frac{d\sigma}{dz} = \hat{y}(1-\hat{y})$$

3) Derivative of "z" w.r.t weight

$$\frac{\partial z}{\partial w_j} = x_j$$

4) Apply chain rule:

$$\begin{aligned}\frac{\partial J^{(i)}}{\partial w_j} &= \frac{\partial J^{(i)}}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial z}{\partial w_j} \\ &= \left( \frac{\hat{y} - y}{\hat{y}(1-\hat{y})} \right) \cdot \hat{y}(1-\hat{y}) \cdot x_j \\ &= (\hat{y} - y) x_j\end{aligned}$$

The  $\hat{y}(1-\hat{y})$  terms cancel perfectly!

### 7) Gradient for whole Dataset:

Average over all  $m$  examples

$$\frac{\partial J}{\partial w_j} = \frac{1}{m} \sum_{i=1}^m ((\hat{y}^{(i)} - y^{(i)}) x_j^{(i)})$$

In vectors form  $\nabla_w J(w) = \frac{1}{m} X^T (\hat{y} - y)$

$X \rightarrow$  bias column of 1s.

### 8) Gradient Descent Update:

Repeat many times:

For each weight:  $w \leftarrow w - \alpha \cdot \nabla_w J(w)$

$$w_j \leftarrow w_j - \alpha \cdot \frac{1}{m} \sum_{i=1}^m ((\hat{y}^{(i)} - y^{(i)}) \cdot x_j^{(i)})$$

### 9) Prediction on New Example:

$$z_{\text{new}} = w^T x_{\text{new}} \Rightarrow \hat{y}_{\text{new}} = \frac{1}{1 + e^{-z_{\text{new}}}}$$

Predict class 1 if  $\hat{y}_{\text{new}} \geq 0.5 \Rightarrow z_{\text{new}} \geq 0$

## Final Cheat Sheet

Linear Score

$$z = w^T x$$

Sigmoid function

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

predicted probability

$$\hat{y} = \sigma(w^T x)$$

Probability of True label  $(\hat{y})^y \cdot (1 - \hat{y})^{1-y}$

log likelihood  $\ell(w) = \sum (y \cdot \log \hat{y} + (1-y) \log(1-\hat{y}))$

Cost function  $\Rightarrow J(w) = -\frac{1}{m} \ell(w)$

Gradient

$$\frac{\partial J^{(i)}}{\partial w_j} = (\hat{y}^{(i)} - y^{(i)}) x_j^{(i)} = (\hat{y} - y) x$$

Full Gradient

$$(\cancel{w \leftarrow w - \alpha \cdot}) \quad \nabla J = \frac{1}{m} X^T (\hat{y} - y)$$

Weight update

$$w \leftarrow w - \alpha \nabla J$$

Decision Rule

Predict 1 if  $w^T x \geq 0$ , else 0