b)

	mis	b	r	
$[0,-1]^T$	$2\ 2$	T	T	V
$[2,-3]^T$	$1\ 1$	−	−	V
$[3,-2]^T$	$2\ 2$ ×	F	−	
$[1,0]^T$	$1\ 1$ ×	T		
$[3,2]^T$	$2\ 2$	T	T	V
$[2,3]^T$	$3\ 3$ ×	T	F	

△ Activity 23 $\quad \sigma(z) = \max\{0, z\}$

* 1. $h_1 = \sigma(2+4+2) = 8$

$h_2 = \sigma(-2+12-8) = 2$

$y_1 = \sigma(4+(-1)) = 3$

$y_2 = \sigma(-4+1) = 0$

* 2, a) $\hat{y}_i = \begin{cases} \sum_{j=0}^{p} w_j x_j^i, & \text{if } \sum_{j=0}^{p} w_j x_j^i > 0 \\ 0, & \text{if } \sum_{j=0}^{p} w_j x_j^i < 0 \end{cases}$

b)

i) chain rule $\dfrac{\partial f(w)}{\partial w_j} = \sum_{i=1}^{n} \underbrace{\dfrac{\partial f^i(w)}{\partial \hat{y}_i}}_{①} \cdot \underbrace{\dfrac{\partial \hat{y}_i}{\partial w_j}}_{②}$ (whole dataset)

$= \begin{cases} \sum_{i=1}^{n} (\hat{y}^i - y^i) \cdot x_j^i, & \sum_{j=0}^{p} w_j x_j^i > 0 \\ \sum_{i=1}^{n} (\hat{y}_i - y_i) \cdot 0, & \sum_{j=0}^{p} w_j x_j^i < 0 \end{cases}$

$\dfrac{\partial f^i(w)}{\partial \hat{y}_i} = \dfrac{\frac{1}{2}\,\partial (\hat{y}_i - y_i)^2}{\partial \hat{y}_i} = \dfrac{1}{2} \dfrac{\partial\, \hat{y}_i^2 - 2 y_i \hat{y}_i + y_i^2}{\partial \hat{y}_i} = \dfrac{1}{2}(2\hat{y}_i - 2y_i)$ ①

ii) ① Initialize w_j^0

② Iterate for $t = 0, 1, 2, \ldots$

③ SGD choose $i_t \in \{1, 2, \ldots N\}$ at random

④ forward Net compute $\hat{y}_j^{i_t}$ from $x_j^{i_t}, w_j^t$

⑤ Backward update $w_j^{t+1} = \begin{cases} w_j^t - \alpha_t \cdot (\hat{y}_j^{i_t} - y_j^{i_t}) \cdot x_j^{i_t} \\ w_j^t \end{cases}$ (single data)

c) $\dfrac{\partial f(w)}{\partial w_j} = \sum_{i=1}^{n} \dfrac{\partial f^i(w)}{\partial \hat{y}_i} \cdot \dfrac{\partial \hat{y}_i}{\partial w_j}$

$\cancel{\hat{y}_i = \dfrac{1}{1+e^{-(\sum_{i=1} w_j x_j^i)}}}$

$\dfrac{\partial \frac{1}{2}(\hat{y}_i - y_i)^2 + \lambda \sum_{j=1}^{p} w_j^2}{\partial \hat{y}_i} = (\hat{y}_i - y_i)$

$\hat{y}_i = \sum_{j=0}^{p} w_j x_j^i$

$\rightarrow x_j^i \quad x_j^i$

$\dfrac{\partial f(w)}{\partial w_j} = (\hat{y}_i - y_i) \cdot \sigma(\sum w_j x_j^i)(1 - \sigma(\sum w_j x_j^i))$

$\Rightarrow \hat{y}_i = \sigma\big(\underbrace{\sum w_j x_j}_{g(w)}\big)$

$\sigma(g(w)) = \dfrac{1}{1 + e^{-g(w)}}$

$\dfrac{\partial \hat{y}_i}{\partial g(w)} = \sigma(g(w))(1 - \sigma(g(w)))$

$\dfrac{\partial g(w)}{\partial w_j} = x_j^i$

$\dfrac{\partial f(w)}{\partial w_j} = \dfrac{\partial f(w)}{\partial \sigma} \dfrac{\partial \sigma}{\partial w}$

$\Rightarrow (\hat{y}_i - y_i) \cdot \hat{y}_i \cdot (1 - \hat{y}_i) x_j^i$