GRADIENT OF A LINEAR CLASSIFIER WITH CROSS-ENTROPY LOSS

$$\underbrace{\begin{bmatrix} \longrightarrow w_1^T \longrightarrow \\ \longrightarrow w_2^T \longrightarrow \\ \vdots \\ \longrightarrow w_m^T \longrightarrow \\ \vdots \\ \longrightarrow w_{y_i}^T \longrightarrow \end{bmatrix}}_{W} \begin{bmatrix} \\ \\ x_i \\ \\ \\ \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \\ \vdots \\ f_{y_i} \\ \vdots \end{bmatrix}$$

← ground-truth class

$f_m = w_m^T x_i$ $\qquad$ w_m: m-th row of W

$$\begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \\ \vdots \\ f_{y_i} \\ \vdots \end{bmatrix} \xrightarrow[\text{softmax}(\cdot)]{} \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_m \\ \vdots \\ p_{y_i} \\ \vdots \end{bmatrix} \qquad p_m = \frac{e^{f_m}}{\sum e^{f_j}}$$

← probability of ground-truth class

cross-entropy loss for training sample (x_i, y_i):

$$= L_i = -\log\left(\frac{e^{f_{y_i}}}{\sum e^{f_j}}\right) = -\log\left(p_{y_i}\right)$$

Want $\nabla_W L_i = \begin{bmatrix} — & \nabla_{w_1} L_i & \longrightarrow \\ — & \nabla_{w_2} L_i & — \\ & \vdots & \\ — & \nabla_{w_m} L_i & — \\ & \vdots & \\ — & \nabla_{w_{y_i}} L_i & — \end{bmatrix}$

Note: $\nabla_W L_i$ is of same dimension as W

$$\frac{\partial L_i}{\partial w_m} = \frac{\partial}{\partial w_m}\left[-\log(p_{y_i})\right]$$

$$= \underbrace{\frac{\partial\left[-\log(p_{y_i})\right]}{\partial p_{y_i}}}_{①} \underbrace{\frac{\partial p_{y_i}}{\partial f_m}}_{②} \underbrace{\frac{\partial f_m}{\partial w_m}}_{③}$$ chain rule

① : $\frac{\partial}{p_{y_i}}\left[-\log(p_{y_i})\right] = \frac{-1}{p_{y_i}}$

③ : $\frac{\partial f_m}{\partial w_m} = \frac{\partial}{w_m}\left[w_m^T x_i\right] = x_i$

② : $\dfrac{\partial p_{y_i}}{\partial f_m} = \dfrac{\partial}{\partial f_m}\left[\dfrac{e^{f_{y_i}}}{\sum e^{f_j}}\right]$

(i) $m \neq y_i$ i.e. f_m and f_{y_i} are different variables

Recall : $\dfrac{\partial}{\partial x}\left[\dfrac{g(x)}{h(x)}\right] = \dfrac{g'(x)\,h(x) - g(x)\,h'(x)}{[h(x)]^2}$ quotient rule

$$\frac{\partial p_{y_i}}{\partial f_m} = \frac{\partial}{\partial f_m}\left[\frac{e^{f_{y_i}}}{\sum e^{f_j}}\right]$$

$$= \frac{(e^{f_{y_i}})' \sum e^{f_j} - e^{f_{y_i}} \left[\sum e^{f_j}\right]'}{\left[\sum e^{f_j}\right]^2}$$

$$= \frac{0 \cdot \sum e^{f_j} - e^{f_{y_i}} e^{f_m}}{\left[\sum e^{f_j}\right]^2}$$

Note : $\dfrac{\partial [e^{f_{y_i}}]}{\partial f_m} = 0$ $\because f_m, f_{y_i}$ are different variables.

$$= \frac{-e^{f_{y_i}} e^{f_m}}{\left[\sum e^{f_j}\right]^2}$$

$$\frac{\partial p_{y_i}}{\partial f_m} = -p_{y_i}\, p_m \quad (\text{when } m \neq y_i)$$

(ii) $m = y_i$ i.e. f_m, f_{y_i} are the same variable

$$\frac{\partial p_{y_i}}{\partial f_m} = \frac{\partial}{\partial f_m}\left[\frac{e^{f_{y_i}}}{\sum e^{f_j}}\right]$$

$$= \frac{\left(e^{f_{y_i}}\right)' \sum e^{f_j} - e^{f_{y_i}}\left[\sum e^{f_j}\right]'}{\left[\sum e^{f_j}\right]^2}$$

$$= \frac{e^{f_m} \sum e^{f_j} - e^{f_{y_i}} e^{f_m}}{\left[\sum e^{f_j}\right]^2}$$

$$= \frac{e^{f_m}}{\sum e^{f_j}} - \frac{e^{f_{y_i}} e^{f_m}}{\sum e^{f_j} \sum e^{f_j}}$$

$$= p_m - p_{y_i} p_m$$

$$\frac{\partial p_{y_i}}{\partial f_m} = p_{y_i}(1 - p_{y_i}) \qquad \left(\text{when } m = y_i\right)$$

$$\frac{\partial L_i}{\partial w_m} = \begin{cases} \dfrac{-1}{p_{y_i}} \left(-p_{y_i} p_m\right) x_i & m \neq y_i \\ & \left(\text{see } p.2\right) \\ \dfrac{-1}{p_{y_i}} \, p_{y_i}(1-p_{y_i})\, x_i & m = y_i \end{cases}$$

$$= \begin{cases} p_m x_i & m \neq y_i \\ (p_{y_i} - 1)\, x_i & m = y_i \end{cases}$$

$$\nabla_w L_i = \begin{bmatrix} \text{---} & p_1 x_i & \text{---} \\ \text{---} & p_2 x_i & \text{---} \\ & \vdots & \\ \text{---} & p_m x_i & \text{---} \\ & \vdots & \\ \text{---} & (p_{y_i} - 1) x_i & \text{---} \\ & \vdots & \end{bmatrix}$$

* Remember this is the gradient matrix for only one training sample (hence the index i)

Gradient descent:

$$W' = W - \gamma \nabla_w L_i$$

$\because p_m x_i \geq 0$

$\therefore w_m$ will decrease, so as to decrease f_m

$\because (p_{y_i} - 1) x_i \leq 0$

$\therefore w_{y_i}$ will increase, so as to