约定：特征：$x\ (n, m)$　　(n：批量数　m：特征数)

标签：$y\ (n, 1)$ 或 $y\ (n, c)$　(c：类别数)

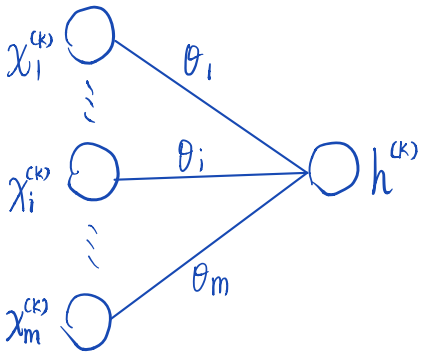
y → 二分类　y → 多分类　one-hot（"独热"编码）

可学习参数：$\theta\ (m, 1)$ 或 $\theta\ (m, c)$

Linear Regression (线性回归)

假设： $x\theta$

损失： 均方差(MSE)



$$h^{(k)} = \sum_{i=1}^{m} x_i^{(k)} \cdot \theta_i$$

$$L^{(k)} = \frac{1}{2} \cdot (h^{(k)} - y^{(k)})^2$$

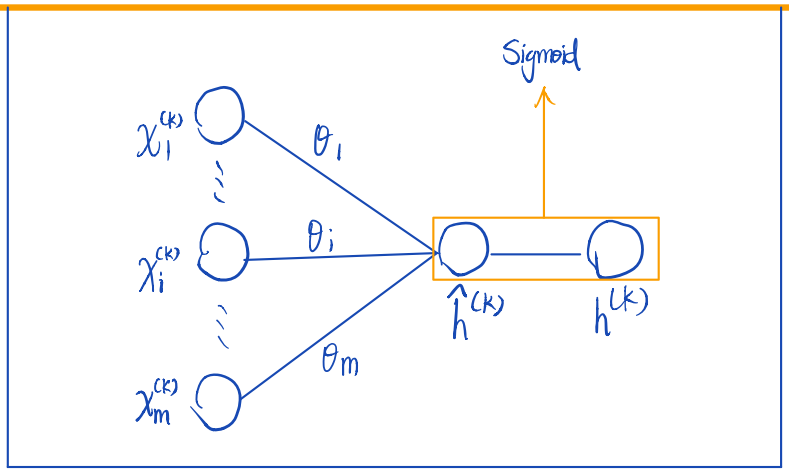
$$\frac{\partial L^{(k)}}{\partial \theta_i} = \frac{\partial L^{(k)}}{\partial h^{(k)}} \cdot \frac{\partial h^{(k)}}{\partial \theta_i} = (h^{(k)} - y^{(k)}) \cdot x_i^{(k)}$$

$$= x^T \cdot (h - y)$$

Logistic Regression (逻辑回归) (二分类)

① 假设：sigmoid(xθ)

损失：交叉熵 (CE)



$$\hat{h}^{(k)} = \sum_{i=1}^{m} x_i^{(k)} \cdot \theta_i$$

$$h^{(k)} = \text{Sigmoid}(\hat{h}^{(k)}) = \frac{1}{1+e^{-\hat{h}^{(k)}}}$$

$$L^{(k)} = -\log\left(h^{(k)^{y^{(k)}}} \cdot (1-h^{(k)})^{(1-y^{(k)})}\right) = -\left[y^{(k)}\log h^{(k)} + (1-y^{(k)})\log(1-h^{(k)})\right]$$

似然(最大化)

$$\frac{\partial L^{(k)}}{\partial \theta_i} = \frac{\partial L^{(k)}}{\partial h^{(k)}} \cdot \frac{\partial h^{(k)}}{\partial \hat{h}^{(k)}} \cdot \frac{\partial \hat{h}^{(k)}}{\partial \theta_i}$$

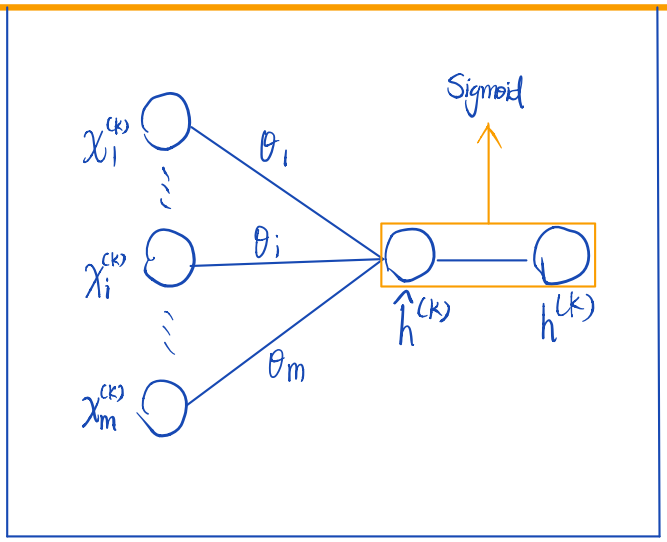
$$= -\left(\frac{y^{(k)}}{h^{(k)}} - \frac{1-y^{(k)}}{1-h^{(k)}}\right) \cdot \left(h^{(k)} \cdot (1-h^{(k)})\right) \cdot x_i$$

$$= \left(h^{(k)} - y^{(k)}\right) \cdot x_i^{(k)}$$

$$= x^T \cdot (h - y)$$

② 假设：sigmoid(xθ)

损失：均方差（mse）



$$\hat{h}^{(k)} = \sum_{i=1}^{m} x_i^{(k)} \cdot \theta_i$$

$$h^{(k)} = \text{Sigmoid}(\hat{h}^{(k)}) = \frac{1}{1+e^{-\hat{h}^{(k)}}}$$

$$L^{(k)} = \frac{1}{2}(h^{(k)} - y^{(k)})^2$$

$$\frac{\partial L^{(k)}}{\partial \theta_i} = \frac{\partial L^{(k)}}{\partial h^{(k)}} \cdot \frac{\partial h^{(k)}}{\partial \hat{h}^{(k)}} \cdot \frac{\partial \hat{h}^{(k)}}{\partial \theta_i}$$

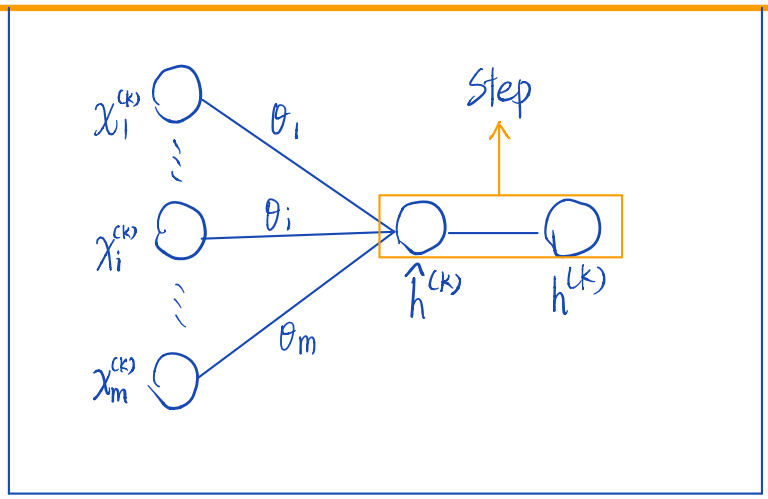
$$= (h^{(k)} - y^{(k)}) \cdot \left(h^{(k)} \cdot (1 - h^{(k)})\right) \cdot x_i$$

$$= x^T \cdot [(h-y) \times h \times (1-h)]$$

感知机 (Perceptron) (二分类)

假设：$\begin{cases} 1 & x\theta \geq 0 \\ 0 & x\theta < 0 \end{cases}$

损失：$\sum_{v \in M_0} x^{(v)}\theta - \sum_{s \in M_1} x^{(s)}\theta$ ("本0判1" − "本1判0")



$$\hat{h}^{(k)} = \sum_{i=1}^{m} x_i^{(k)} \cdot \theta_i$$

$$h^{(k)} = \begin{cases} 1 & \sum_{i=1}^{m} x_i^{(k)}\theta_i \geq 0 \\ 0 & \sum_{i=1}^{m} x_i^{(k)}\theta_i < 0 \end{cases}$$

$$L^{(k)} = (1-y^{(k)}) \cdot h^{(k)} \cdot \hat{h}^{(k)} - y^{(k)} \cdot (1-h^{(k)}) \cdot \hat{h}^{(k)}$$

"本0判1" ← (1−y)·h·ĥ　　y·(1−h)·ĥ →"本1判0"

$$= (h^{(k)} - y^{(k)}) \cdot \hat{h}^{(k)}$$

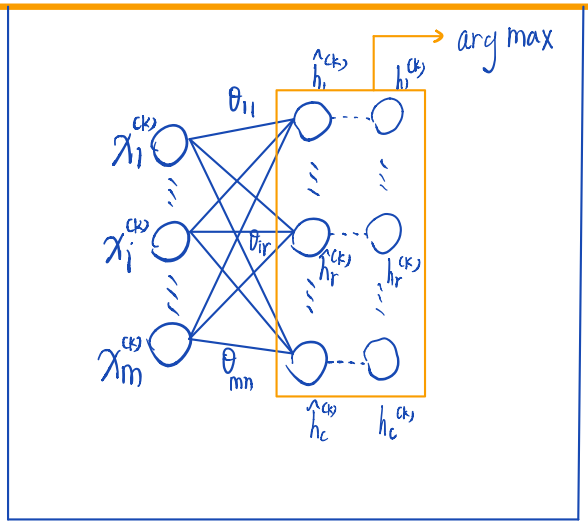
$$\frac{\partial L^{(k)}}{\partial \theta_i} = (h^{(k)} - y^{(k)}) \cdot x_i$$

$$= x^T \cdot (h - y)$$

多分类感知机 (Multi-class Perceptron)

假设： $h(x) = \arg\max_{r=1,\dots,c} \sum_{i=1}^{m} x_i \theta_{ir}$

损失： $L = \max_{r=1,\dots,c} \sum_{i=1}^{m} x_i \theta_{ir} - \sum_{i=1}^{m} x_i \theta_{ic^*}$



$$\hat{h}_r^{(k)} = \sum_{i=1}^{m} x_i^{(k)} \theta_{ir}$$

$$h_r^{(k)} = \begin{cases} \hat{h}_r^{(k)} & r = \arg\max_{d=1,\dots,c} \hat{h}_c^{(k)} \\ 0 & \text{其它} \end{cases}$$

$$L^{(k)} = h_r^{(k)} - \hat{h}_{c^*}^{(k)} \quad (\text{注：} y_{c^*}^{(k)} = 1 \text{、} y_{\text{非}c^*\text{部}} = 0)$$

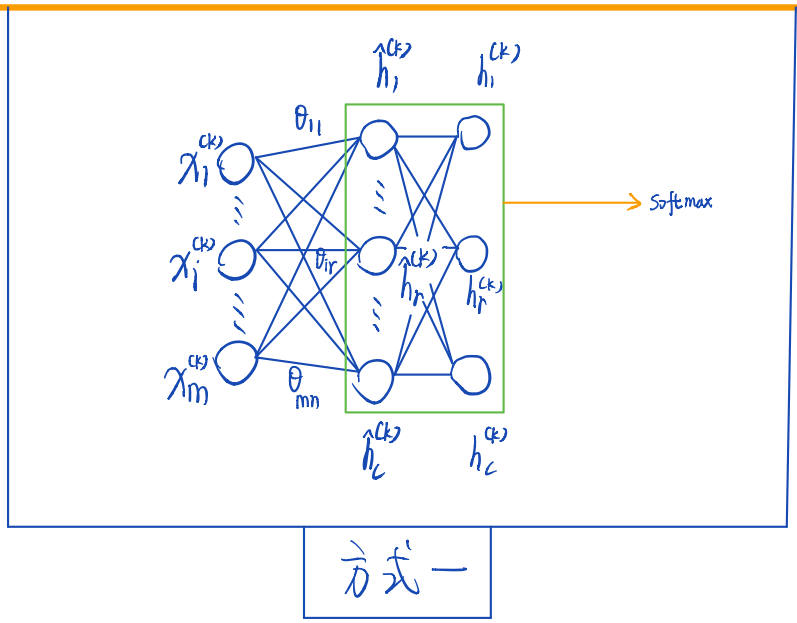
$$\frac{\partial L^{(k)}}{\partial \theta_{ir}} = (1\{r = r^*\} - 1\{r = c^*\}) \cdot x_i^{(k)}$$

$$= x^T \cdot (h - y)$$

Softmax（多分类）

假设：$h_\theta(x) = \text{softmax}(x\theta)$

损失：交叉熵（CE）



方式一

$$\hat{h}_r^{(k)} = \sum_{i=1}^{m} x_i^{(k)} \theta_{ir}$$

$$h_r^{(k)} = \text{Softmax}(\hat{h}_r^{(k)}) = \frac{e^{\hat{h}_r^{(k)}}}{\sum_{t=1}^{c} e^{\hat{h}_t^{(k)}}}$$

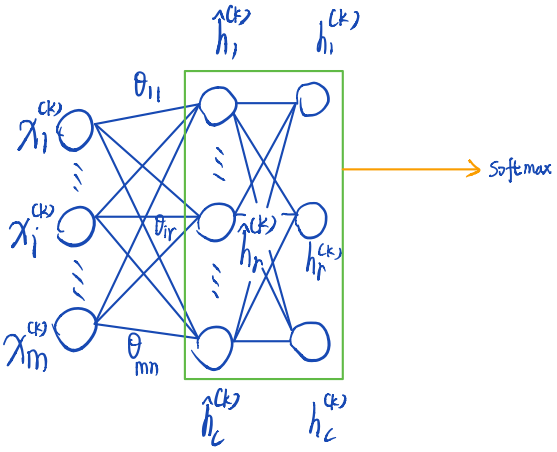
$$L^{(k)} = -\log\left(\prod_{r=1}^{c} h_r^{y_r^{(k)}}\right) = -\sum_{r=1}^{c} y_r^{(k)} \log h_r^{(k)} = -y_{c^*}^{(k)} \log h_{c^*}^{(k)}$$

$$\frac{\partial L^{(k)}}{\partial \theta_{ir}} = \frac{\partial L^{(k)}}{\partial h_{c^*}^{(k)}} \cdot \frac{\partial h_{c^*}^{(k)}}{\partial \hat{h}_r^{(k)}} \cdot \frac{\partial \hat{h}_r^{(k)}}{\partial \theta_{ir}}$$

$$= \begin{cases} r = c^* \text{ 时} & \left(-\frac{\sum_{t=1}^{c} e^{\hat{h}_t^{(k)}}}{e^{h_r^{(k)}}}\right)\left(\frac{e^{h_r^{(k)}}\left(\sum_{t=1}^{c} e^{\hat{h}_t^{(k)}}\right) - e^{h_r^{(k)}} \cdot e^{h_r^{(k)}}}{\left(\sum_{t=1}^{c} e^{\hat{h}_t^{(k)}}\right)^2}\right) \cdot x_i^{(k)} = (h_r^{(k)} - 1) \cdot x_i^{(k)} \\ r \neq c^* \text{ 时} & \left(-\frac{\sum_{t=1}^{c} e^{\hat{h}_t^{(k)}}}{e^{h_{c^*}^{(k)}}}\right)\left(\frac{0 \quad - e^{h_{c^*}^{(k)}} \cdot e^{h_r^{(k)}}}{\left(\sum_{t=1}^{c} e^{\hat{h}_t^{(k)}}\right)^2}\right) \cdot x_i^{(k)} = (h_r^{(k)} - 0) \cdot x_i^{(k)} \end{cases}$$

$$= (h_r^{(k)} - y_r^{(k)}) \cdot x_i^{(k)}$$

$$= x^T \cdot (h - y)$$



方式二

$$\hat{h}_r^{(k)} = \sum_{i=1}^{m} x_i^{(k)} \theta_{ir}$$

$$h_r^{(k)} = \text{Softmax}(\hat{h}_r^{(k)}) = \frac{e^{\hat{h}_r^{(k)}}}{\sum_{f=1}^{c} e^{\hat{h}_f^{(k)}}}$$

$$L^{(k)} = -\log\left(\prod_{f=1}^{c} h_f^{y_f^{(k)}}\right) = -\sum_{f=1}^{c} y_f^{(k)} \log h_f^{(k)}$$

$$\frac{\partial L^{(k)}}{\partial \theta_{ir}} = \sum_{d=1}^{c} \frac{\partial L^{(k)}}{\partial h_d^{(k)}} \cdot \frac{\partial h_d^{(k)}}{\partial \hat{h}_r^{(k)}} \cdot \frac{\partial \hat{h}_r^{(k)}}{\partial \theta_{ir}}$$

$$= \left(-\frac{y_r^{(k)}\left(\sum_{f=1}^{c} e^{\hat{h}_f^{(k)}}\right)}{e^{\hat{h}_r^{(k)}}}\right)\left(\frac{e^{\hat{h}_r^{(k)}}\left(\sum_{f=1}^{c} e^{\hat{h}_f^{(k)}}\right) - e^{\hat{h}_r^{(k)}} \cdot e^{\hat{h}_r^{(k)}}}{\left(\sum_{f=1}^{c} e^{\hat{h}_f^{(k)}}\right)^2}\right) x_j^{(k)} + \sum_{d \neq r}^{c}\left(-\frac{y_r^{(k)}\left(\sum_{f=1}^{c} e^{\hat{h}_f^{(k)}}\right)}{e^{\hat{h}_d^{(k)}}}\right)\left(\frac{0 - e^{\hat{h}_d^{(k)}} \cdot e^{\hat{h}_r^{(k)}}}{\left(\sum_{f=1}^{c} e^{\hat{h}_f^{(k)}}\right)^2}\right) x_i^{(k)}$$

$$= \left(-y_r^{(k)}\left(1 - h_r^{(k)}\right) \cdot z_s^{(k)}\right) + \sum_{d \neq r}^{c}\left(-y_d^{(k)}\left(0 - h_r^{(k)}\right) \cdot z_s^{(k)}\right)$$

$$= \left(h_r^{(k)} - y_r^{(k)}\right) \cdot z_s^{(k)}$$

注：① 如果 $y_r^{(k)} = 1$，则 $y_{d \neq r}^{(k)}$ 全 $= 0$，第二项化掉

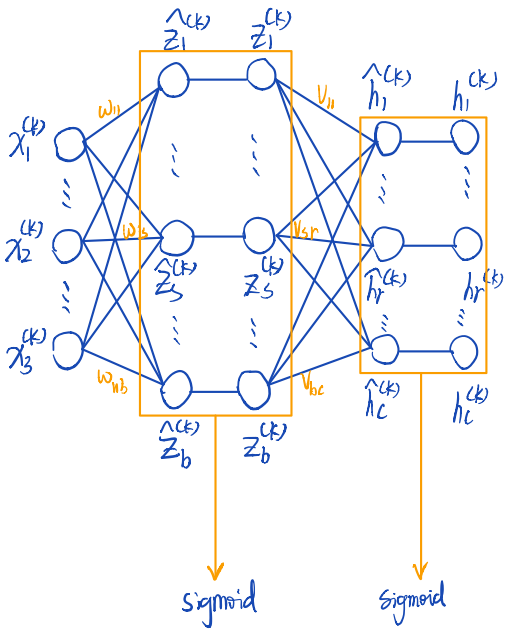
② 如果 $y_r^{(k)} = 0$，则 $y_{d \neq r}^{(k)}$ 中会有一个 $= 1$，第一项化掉，第二项累加中只会保留一项

$$= x^T \cdot (h - y)$$

人工神经网络(ANN)、多层感知机(MLP)

① 假设：xw → Sigmoid → zV → sigmoid

损失：mean square error (mse)



$$\hat{z}_s^{(k)} = \sum_{i=1}^{m} x_i^{(k)} \cdot w_{is}$$

$$z_s^{(k)} = sigmoid(\hat{z}_s^{(k)}) = \frac{1}{1+e^{-\hat{z}_s^{(k)}}}$$

$$\hat{h}_r^{(k)} = \sum_{s=1}^{b} z_s^{(k)} \cdot V_{sr}$$

$$h_r^{(k)} = sigmoid(\hat{h}_r^{(k)}) = \frac{1}{1+e^{-\hat{h}_r^{(k)}}}$$

$$L^{(k)} = \frac{1}{2}\sum_{r=1}^{c}(h_r^{(k)} - y_r^{(k)})^2$$

(假有一个为1，其余全为0)

$$\frac{\partial L^{(k)}}{\partial V_{sr}} = \frac{\partial L^{(k)}}{\partial h_r^{(k)}} \cdot \frac{\partial h_r^{(k)}}{\partial \hat{h}_r^{(k)}} \cdot \frac{\partial \hat{h}_r^{(k)}}{\partial V_{sr}}$$

$$= (h_r^{(k)} - y_r^{(k)}) \cdot (h_r^{(k)} \cdot (1 - h_r^{(k)})) \cdot z_s^{(k)}$$

$(h_r^{(k)} - y_r^{(k)}) \cdot (h_r^{(k)} \cdot (1 - h_r^{(k)}))$ → $OutputLayer\,L_r^{(k)}$

$$= z^T \cdot [(h - y) \times (h \times (1 - h))]$$

$$\frac{\partial L^{(k)}}{\partial w_{is}} = \sum_{r=1}^{c} \frac{\partial L^{(k)}}{\partial h_r^{(k)}} \cdot \frac{\partial h_r^{(k)}}{\partial \hat{h}_r^{(k)}} \cdot \frac{\partial \hat{h}_r^{(k)}}{\partial z_s^{(k)}} \cdot \frac{\partial z_s^{(k)}}{\partial \hat{z}_s^{(k)}} \cdot \frac{\partial \hat{z}_s^{(k)}}{\partial w_{is}}$$

$$= \sum_{r=1}^{c} OutputLayer\,L_r^{(k)} \cdot \frac{\partial \hat{h}_r^{(k)}}{\partial z_s^{(k)}} \cdot \frac{\partial z_s^{(k)}}{\partial \hat{z}_s^{(k)}} \cdot \frac{\partial \hat{z}_s^{(k)}}{\partial w_{is}}$$

$$= \sum_{r=1}^{c} OutputLayer\,L_r^{(k)} \cdot V_{sr} \cdot (z_s^{(k)} \cdot (1 - z_s^{(k)})) \cdot x_i^{(k)}$$

(n,c) (h,c) (n,h) (n,m)

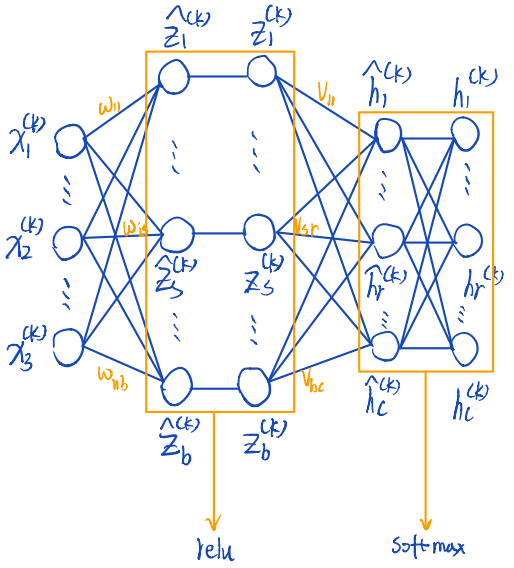
$OutputLayer\,L_r^{(k)}$ → $HiddenLayer\,L_s^{(k)}$

$$= X^T \cdot \left(\left((h - y) \times h \times (1 - h) \right) \cdot V^T \right) \times z \times (1 - z)$$

2.1

假设： $xw \rightarrow relu \rightarrow zv \rightarrow softmax$

损失： cross entropy



$$\hat{z}_s^{(k)} = \sum_{i=1}^{m} x_i^{(k)} \cdot w_{is}$$

$$z_s^{(k)} = relu(\hat{z}_s^{(k)}) = \hat{z}_s^{(k)} \times mask_s$$

$$\hat{h}_r^{(k)} = \sum_{s=1}^{b} z_s^{(k)} \cdot v_{sr}$$

$$h_r^{(k)} = softmax(\hat{h}_r^{(k)}) = \frac{e^{\hat{h}_r^{(k)}}}{\sum_{f=1}^{c} e^{\hat{h}_f^{(k)}}}$$

$$L^{(k)} = -\log\left(\prod_{r=1}^{c} h_r^{y_r^{(k)}}\right) = -\sum_{r=1}^{c} y_r^{(k)} \log h_r^{(k)} = -y_{c^*}^{(k)} \log h_{c^*}^{(k)} \quad (y_{c^*}^{(k)}=1,\ y_{r\neq c^*}^{(k)}=0)$$

$$\frac{\partial L^{(k)}}{\partial v_{sr}} = \frac{\partial L^{(k)}}{\partial h_{c^*}^{(k)}} \cdot \frac{\partial h_{c^*}^{(k)}}{\partial \hat{h}_r^{(k)}} \cdot \frac{\partial \hat{h}_r^{(k)}}{\partial v_{sr}}$$

$r = c^*$

$$= -\left[\left(\frac{\sum_{f=1}^{c} e^{\hat{h}_f^{(k)}}}{e^{\hat{h}_r^{(k)}}}\right) \cdot \left(\frac{e^{\hat{h}_r^{(k)}} \cdot \left(\sum_{f=1}^{c} e^{\hat{h}_f^{(k)}}\right) - e^{\hat{h}_r^{(k)}} \cdot e^{\hat{h}_r^{(k)}}}{\left(\sum_{f=1}^{c} e^{\hat{h}_f^{(k)}}\right)^2}\right) \cdot z_s^{(k)}\right]$$

$$= -(1 - h_r^{(k)}) \cdot z_s^{(k)}$$

$r \neq c^*$

$$= -\left[\left(\frac{\sum_{f=1}^{c} e^{\hat{h}_f^{(k)}}}{e^{\hat{h}_{c^*}^{(k)}}}\right) \cdot \left(\frac{0 - e^{\hat{h}_{c^*}^{(k)}} \cdot e^{\hat{h}_r^{(k)}}}{\left(\sum_{f=1}^{c} e^{\hat{h}_f^{(k)}}\right)^2}\right) \cdot z_s^{(k)}\right]$$

$$= -(0 - h_r^{(k)}) \cdot z_s^{(k)}$$

$$= (h_r^{(k)} - y_r^{(k)}) \cdot z_s^{(k)}$$

$(h_r^{(k)} - y_r^{(k)})$ → $OutputLayer\,L_r^{(k)}$

$$= z^T \cdot (h - y)$$

$$\frac{\partial L^{(k)}}{\partial w_{is}} = \sum_{r=1}^{c} \frac{\partial L^{(k)}}{\partial h_r^{(k)}} \cdot \frac{\partial h_r^{(k)}}{\partial \hat{h}_r^{(k)}} \cdot \frac{\partial \hat{h}_r^{(k)}}{\partial z_s^{(k)}} \cdot \frac{\partial z_s^{(k)}}{\partial \hat{z}_s^{(k)}} \cdot \frac{\partial \hat{z}_s^{(k)}}{\partial w_{is}}$$

$$= \sum_{r=1}^{c} OutputLayer\,L_r^{(k)} \cdot \frac{\partial \hat{h}_r^{(k)}}{\partial z_s^{(k)}} \cdot \frac{\partial z_s^{(k)}}{\partial \hat{z}_s^{(k)}} \cdot \frac{\partial \hat{z}_s^{(k)}}{\partial w_{is}}$$

$$= \sum_{r=1}^{c} OutputLayer\,L_r^{(k)} \cdot v_{sr} \cdot mask_s \cdot x_i^{(k)}$$

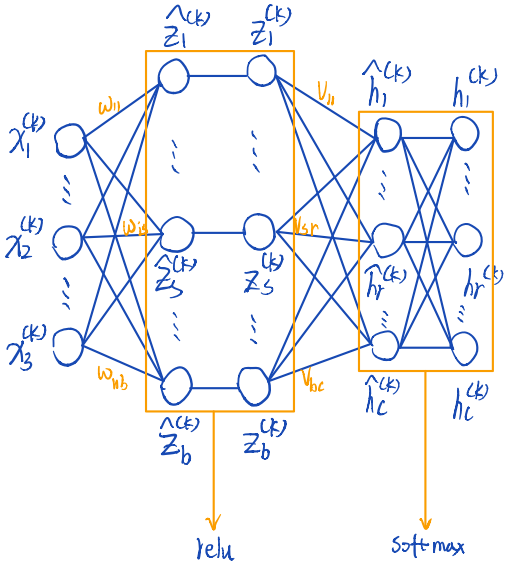
$\sum_{r=1}^{c} OutputLayer\,L_r^{(k)} \cdot v_{sr}$ → $HiddenLayer\,L_s^{(k)}$

$$= x^T \cdot \left((h - y) \cdot V^T\right) \times mask$$

2.2

假设：$xw \rightarrow relu \rightarrow zv \rightarrow softmax$

损失：cross entropy



$$\hat{z}_s^{(k)} = \sum_{i=1}^{m} x_i^{(k)} \cdot w_{is}$$

$$z_s^{(k)} = relu(\hat{z}_s^{(k)}) = \hat{z}_s^{(k)} \times mask_s$$

$$\hat{h}_r^{(k)} = \sum_{s=1}^{b} z_s^{(k)} \cdot v_{sr}$$

$$h_r^{(k)} = softmax(\hat{h}_r^{(k)}) = \frac{e^{\hat{h}_r^{(k)}}}{\sum_{f=1}^{c} e^{\hat{h}_f^{(k)}}}$$

$$L^{(k)} = -\log\left(\prod_{r=1}^{c} h_r^{(k)\,y_r^{(k)}}\right) = -\sum_{r=1}^{c} y_r^{(k)} \log h_r^{(k)}$$

仅有一个为1，其余全为0（one-hot）

$$\frac{\partial L^{(k)}}{\partial v_{sr}} = \sum_{d=1}^{c} \frac{\partial L^{(k)}}{\partial h_d^{(k)}} \cdot \frac{\partial h_d^{(k)}}{\partial \hat{h}_r^{(k)}} \cdot \frac{\partial \hat{h}_r^{(k)}}{\partial v_{sr}}$$

$$= -\left[\left(\frac{y_r^{(k)} \cdot \left(\sum_{f=1}^{c} e^{\hat{h}_f^{(k)}}\right)}{e^{\hat{h}_r^{(k)}}}\right) \cdot \left(\frac{e^{\hat{h}_r^{(k)}} \cdot \left(\sum_{f=1}^{c} e^{\hat{h}_f^{(k)}}\right) - e^{\hat{h}_r^{(k)}} \cdot e^{\hat{h}_r^{(k)}}}{\left(\sum_{f=1}^{c} e^{\hat{h}_f^{(k)}}\right)^2}\right) \cdot z_s^{(k)}\right] - \sum_{d \neq r}^{c}\left[\left(\frac{y_d^{(k)} \cdot \left(\sum_{f=1}^{c} e^{\hat{h}_f^{(k)}}\right)}{e^{\hat{h}_d^{(k)}}}\right) \cdot \left(\frac{0 - e^{\hat{h}_d^{(k)}} \cdot e^{\hat{h}_r^{(k)}}}{\left(\sum_{f=1}^{c} e^{\hat{h}_f^{(k)}}\right)^2}\right) \cdot z_s^{(k)}\right]$$

$$= -\left[y_r^{(k)}\left(1 - h_r^{(k)}\right) \cdot z_s^{(k)}\right] - \sum_{d \neq r}^{c}\left[y_d^{(k)}\left(0 - h_r^{(k)}\right) \cdot z_s^{(k)}\right]$$

注：①如果 $y_r^{(k)} = 1$，则 $y_{d \neq r}^{(k)}$ 全=0，第二项化掉

②如果 $y_r^{(k)} = 0$，则 $y_{d \neq r}^{(k)}$ 中含有一个=1，第一项化掉，第二项累加中又会保留一项

$$= \left(h_r^{(k)} - y_r^{(k)}\right) \cdot z_s^{(k)}$$

$\rightarrow$ OutputLayer $L_r^{(k)}$

$$= z^T \cdot (h - y)$$

$$\frac{\partial L^{(k)}}{\partial w_{is}} = \sum_{r=1}^{c} \frac{\partial L^{(k)}}{\partial h_r^{(k)}} \cdot \frac{\partial h_r^{(k)}}{\partial \hat{h}_r^{(k)}} \cdot \frac{\partial \hat{h}_r^{(k)}}{\partial z_s^{(k)}} \cdot \frac{\partial z_s^{(k)}}{\partial \hat{z}_s^{(k)}} \cdot \frac{\partial \hat{z}_s^{(k)}}{\partial w_{is}}$$

$$= \sum_{r=1}^{c} OutputLayer\,L_r^{(k)} \cdot \frac{\partial \hat{h}_r^{(k)}}{\partial z_s^{(k)}} \cdot \frac{\partial z_s^{(k)}}{\partial \hat{z}_s^{(k)}} \cdot \frac{\partial \hat{z}_s^{(k)}}{\partial w_{is}}$$

$$= \sum_{r=1}^{c} OutputLayer\,L_r^{(k)} \cdot v_{sr} \cdot mask_s \cdot x_i^{(k)}$$

$\rightarrow$ HiddenLayer $L_s^{(k)}$

$$= x^T \cdot \left((h - y) \cdot v^T\right) \times mask$$

① 参数的初始值非常重要。一开始直接 np.random.randn(*,*)，太大了，此时 L2 正则化影响很小，但 loss 大部分都是正则项

当将参数初始化为 0.01 * np.random.randn(*,*)，网络此时的效果真的很好了，
L2 影响不明显，且此时损失(Loss)可能会上升，因此又通过让权值上升大于 L2 损失

② drop out 实现步骤

(1) 每个神经元的输出：$p\vec{x} + (1-p)\cdot 0$

- 方法一：
  - 训练时补偿：乘以 $\frac{1}{p}$
  - 测试时不用管
- 方法二：
  - 训练时不用管
  - 测试时补偿：乘以 p

(2)
- 随机删除神经元：`drop_index = np.random.rand(x.shape[1])` `drop_index = (drop_index > ratio)` `dropout.mask[:, drop_index] = 0`
- 随机矩阵删除 随机指定：`dropout.mask = (np.random.rand(*x.shape) < ratio)`

(经测试，更优)