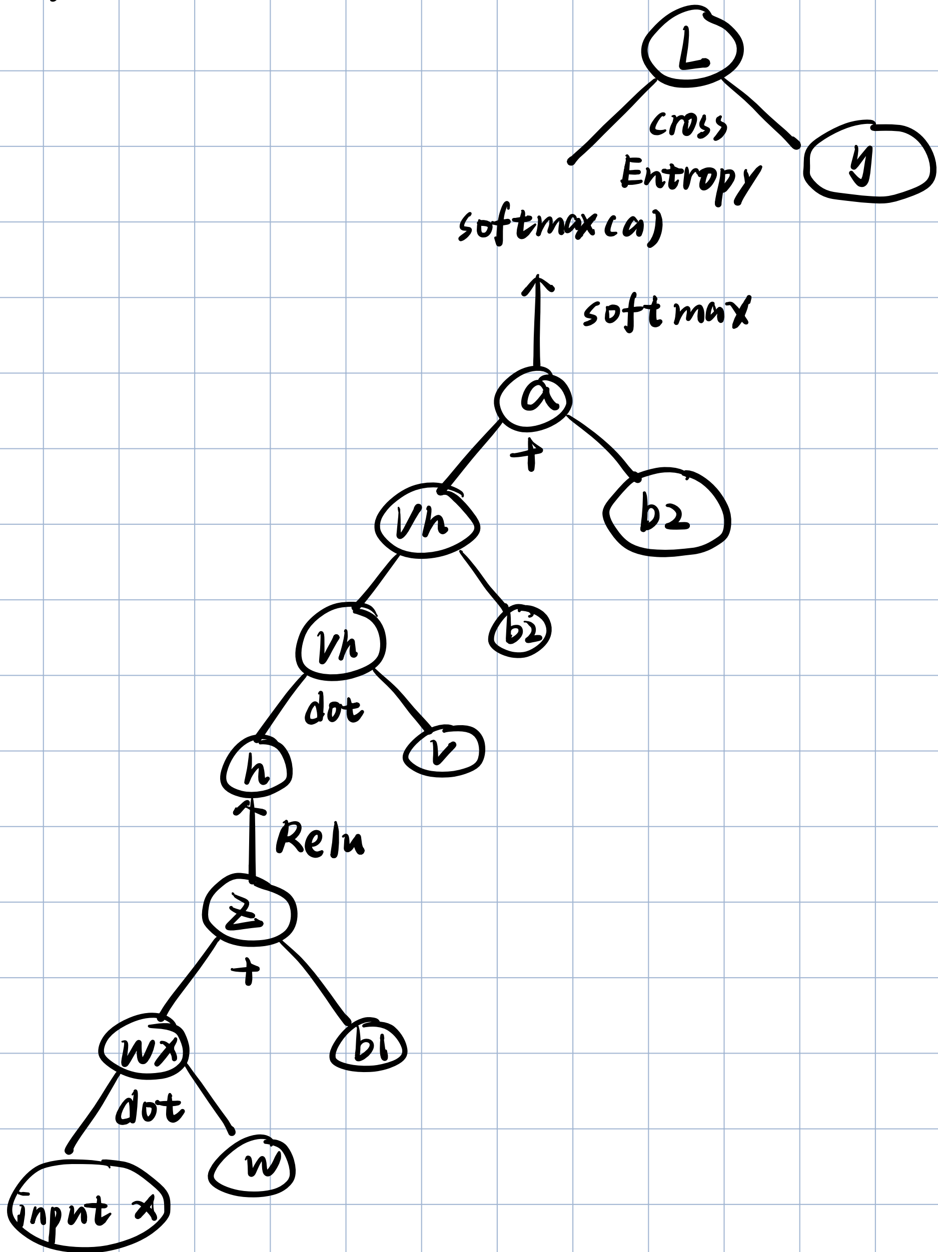
1.1 The computational graph for the loss function

(a)



(b) $L = \text{CrossEntropy}(y, \text{Softmax}(a)) \in R$

$$u_2 = \frac{\partial L}{\partial a} = \frac{\partial L}{\partial(\text{softmax}(a))} \cdot \frac{\partial(\text{softmax}(a))}{\partial(a)}$$

$$= (p - y) \in R^c \qquad (p = \text{softmax}(a))$$

$\because a = Vh + b_2$

$\therefore$ by chain rule

$$\psi_1(u_2, h) = \frac{\partial L}{\partial V} = \frac{\partial L}{\partial a} \cdot \frac{\partial a}{\partial V} = u_2 \cdot h^T \in R^{c \times k}$$

$\because b_2$ is a bias term $\Rightarrow \nabla = 1$

$\therefore \psi_2(u_2) = u_2 \in R^c$

For W, there is a hidden layer

Back propagation:

$$u_1 = \frac{\partial L}{\partial z} = V^T u_2 \odot H(z) \qquad (H_z = I(z>0))$$

$\because z = Wx + b_1$

$$\therefore \psi_3(u_1, x) = \frac{\partial L}{\partial W} = \frac{\partial L}{\partial z} \cdot \frac{\partial z}{\partial W} = u_1 \cdot x^T \quad \in R^{K \times D}$$

Similarly, b_1 is a bias term

$\therefore \psi_4(u_1) = u_1 \in R^K$

For gradients w.r.t. x

$\because z = Wx + b_1$

$$\therefore \psi_5(W, u_1) = \frac{\partial L}{\partial x} = \frac{\partial L}{\partial z} \cdot \frac{\partial z}{\partial W} = W^T u_1 \in R^{D \times 1}$$

1.1 CNN

Input with $32 \times 32 \times 3$

Output W:		Output shape
(1st) Conv5(10)	$\frac{32 + 2 \times \text{padding} - 5}{\text{stride}} + 1 = 32$	$32 \times 32 \times 10$
(1st) Maxpool2	$\frac{32 - 2}{\text{stride}} + 1 = 16$	$16 \times 16 \times 10$
(2nd) Conv3(20)	$\frac{16 + 2 \times \text{padding} - 3}{\text{stride}} + 1 = 16$	$16 \times 16 \times 20$
(2nd) Maxpool2	$\frac{16 - 2}{2} + 1 = 8$	$8 \times 8 \times 20$
FC10		10

二、

NO.	Layer	Activation Shape	# Parameters	Mark
1	Input Layer	$32\times32\times3$	0	-
2	Conv5(10)	$32\times32\times10$	$(5\times5\times3+1)\times10=760$	2 pts
3	Maxpool2	$16\times16\times10$	0	2 pts
4	Conv3(20)	$16\times16\times20$	$(3\times3\times10+1)\times20=1820$	2 pts
5	Maxpool2	$8\times8\times20$	0	2 pts
6	FC10	10	$(8\times8\times20+1)\times10=12810$	2 pts

1.2 i) Define the 3 classes as $\{1,2,3\} \Rightarrow \{$ blue, orange, grey$\}$

ii) Necessary formulas

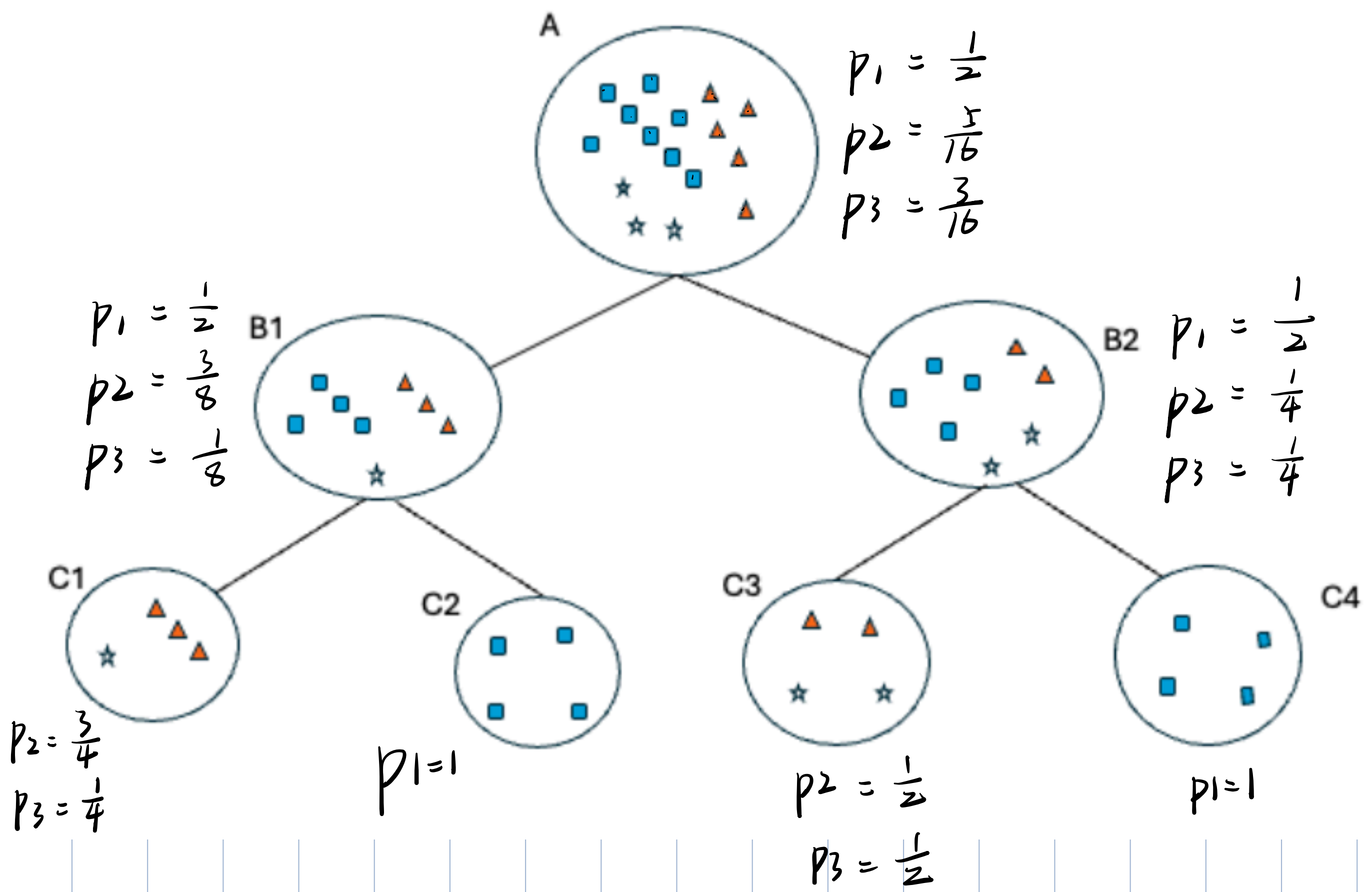
class proportion $p_i = \frac{n_i}{N}$

$(N = \#$ of points in the node, $i \in \{1,2,3\})$

Gini Index $G = 1 - \sum_{i=1}^{K} p_i^2$ $(K = \#$ of classes$)$

Entropy $H = -\sum_{i=1}^{K} p_i \cdot \log_2(p_i)$

Classification $E = 1 - \max_i(p_i)$



Node	G	H	E
A	$1-\left(\frac{1}{2}^2+(\frac{5}{16})^2+(\frac{3}{16})^2\right)$ $=\frac{79}{128}$	$=-\left(\frac{1}{2}\log_2\frac{1}{2}+\frac{5}{16}\log_2\frac{5}{16}+\frac{3}{16}\log_2 3\right)$ $=\frac{5}{2}-\frac{5}{16}\log_2 5-\frac{3}{16}\log_2 3$	$=1-\max(\frac{1}{2},\frac{5}{16},\frac{3}{16})$ $=\frac{1}{2}$
B1	$1-\left((\frac{1}{2})^2+(\frac{3}{8})^2+(\frac{1}{8})^2\right)$ $=\frac{19}{32}$	$=-\left(\frac{1}{2}\log_2\frac{1}{2}+\frac{3}{8}\log_2\frac{3}{8}+\frac{1}{8}\log_2\frac{1}{8}\right)$ $=2-\frac{3}{8}\log_2 3$	$=1-\max(\frac{1}{2},\frac{3}{8},\frac{1}{8})$ $=\frac{1}{2}$
B2	$1-\left((\frac{1}{2})^2+(\frac{1}{4})^2+(\frac{1}{4})^2\right)$ $=\frac{5}{8}$	$=-\left(\frac{1}{2}\log_2\frac{1}{2}+\frac{1}{4}\log_2\frac{1}{4}+\frac{1}{4}\log_2\frac{1}{4}\right)$ $=\frac{3}{2}$	$=1-\max(\frac{1}{2},\frac{1}{4},\frac{1}{4})$ $=\frac{1}{2}$
C1	$=1-\left(0^2+(\frac{3}{4})^2+(\frac{1}{4})^2\right)$ $=\frac{3}{8}$	$=-\left(0+\frac{3}{4}\log_2\frac{3}{4}+\frac{1}{4}\log_2\frac{1}{4}\right)$ $=2-\frac{3}{4}\log_2 3$	$=1-\max(0,\frac{3}{4},\frac{1}{4})$ $=\frac{1}{4}$
C2	$=1-(1^2+0+0)$ $=0$	$=-(\log_2 1+0+0)$ $=0$	$=1-1=0$
C3	$=1-\left(0+\frac{1}{2^2}+\frac{1}{2^2}\right)$ $=\frac{1}{2}$	$=-\left(0+\frac{1}{2}\log_2\frac{1}{2}+\frac{1}{2}\log_2\frac{1}{2}\right)$ $=1$	$=1-\max(0,\frac{1}{2},\frac{1}{2})$ $=\frac{1}{2}$
C4	$=1-(1^2+0+0)$ $=0$	$=-(\log_2 1+0+0)$ $=0$	$=1-1=0$

1.2 ii)

$$MSE(x,y) = E_D[(h_{D_i}(x) - y)^2]$$

$\because$ Given $E_{(x,y)D}[(h_{D_i}(x)-y)^2] = E_{(x,y)}[MSE(x,y)]$

$$\hat{MSE}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(h_D(x)-y)^2$$

$$= \frac{1}{10}\cdot[(7-8)^2+(8-8)^2+(9-8)^2+(6-8)^2+(9-8)^2$$
$$+(3-8)^2+(9-8)^2+(8-8)^2+(9-8)^2+(8-8)^2]$$

$$= \frac{1}{10}\cdot(1+0+1+4+1+25+1+0+1+0)$$

$$= \frac{34}{10} = 3.4$$

Average Prediction $\bar{\hat{y}} = \frac{1}{10}\cdot\sum_{i=1}^{10} h_{D_i}(x) = \frac{1}{10}(7+8+9+6+9+3+9+8+9+8) = 7.6$

$\therefore$ $Bias = \bar{\hat{y}} - y_{truth} = -0.4 \Rightarrow Bias^2 = 0.16$

For Variance $= \frac{1}{10}\cdot[(7-7.6)^2+(8-7.6)^2+(9-7.6)^2+(6-7.6)^2+(9-7.6)^2$
$+(3-7.6)^2+(9-7.6)^2+(8-7.8)^2+(9-7.6)^2]$

$$= \frac{1}{10}[0.6^2+0.4^2+1.4^2+1.6^2+1.4^2+4.6^2+1.4^2+0.4^2+1.4^2+0.4^2]$$

$$= \frac{1}{10}\times 32.4 = 3.24$$

$$\hat{MSE} = Variance + Bias$$