Homework 3

伴泽为 122090814

Due: Apr. 16 23:59 Tue

Q1 (a) Given $|D| = 8$,

There are 2 classes: Yes: $|C_1| = 4$; No: $|C_2| = 4$

$$H(D) = -\sum_{k=1}^{2} \frac{|C_k|}{|D|} \log_2 \frac{|C_k|}{|D|} = -\frac{4}{8}\log_2\frac{4}{8} - \frac{4}{8}\log_2\frac{4}{8} = 1$$

"Animated"

For attribute "Animated": $n=2$, a_1: Yes, $|D_1|=3$; a_2: No, $|D_2|=5$

$|D_{11}| = 2$, $|D_{12}| = 1$, $|D_{21}| = 2$, $|D_{22}| = 3$

$D_{ik} = D_i \cap C_k$

$$H(D|\text{Animated}) = \sum_{i=1}^{2} \frac{|D_i|}{|D|} H(D_i) = -\sum_{i=1}^{2} \frac{|D_i|}{|D|} \sum_{k=1}^{2} \frac{|D_{ik}|}{|D_i|} \log_2 \frac{|D_{ik}|}{|D_i|}$$

$$= \frac{3}{8}\left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right) + \frac{5}{8}\left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right)$$

$$= \left(\frac{3}{8}\log_2 3 - \frac{1}{4}\right) - \left(\frac{1}{4}\log_2\frac{2}{5} - \frac{3}{8}\log_2\frac{3}{5}\right)$$

$$= \frac{5}{8}\log_2 5 - \frac{1}{2}$$

$$g(D, \text{Animated}) = H(D) - H(D|\text{Animated}) = \frac{3}{2} - \frac{5}{8}\log_2 5$$

$$H_{\text{Animated}}(D) = -\sum_{k=1}^{2} \frac{|D_i|}{|D|} \log_2 \frac{|D_i|}{|D|} = -\frac{3}{8}\log_2\frac{3}{8} - \frac{5}{8}\log_2\frac{5}{8} = 3 - \frac{3}{8}\log_2 3 - \frac{5}{8}\log_2 5$$

$$g_R(D, \text{Animated}) = \frac{g(D,\text{Animated})}{H_{\text{Animated}}(D)} = \frac{\frac{3}{2} - \frac{5}{8}\log_2 5}{3 - \frac{3}{8}\log_2 3 - \frac{5}{8}\log_2 5}$$

"Popup"

For attribute "Popup": $n=2$, a_1: Yes, $|D_1| = 3$; a_2: No, $|D_2| = 5$

$|D_{11}| = 0$, $|D_{12}| = 3$, $|D_{21}| = 4$, $|D_{22}| = 1$

$$H(D|\text{Popup}) = \sum_{i=1}^{2} \frac{|D_i|}{|D|} H(D_i) = \frac{3}{8} \times 0 + \frac{5}{8}\left(-\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5}\right)$$

$$= -\frac{1}{2}\log_2\frac{4}{5} - \frac{1}{8}\log_2\frac{1}{5} = \frac{1}{2}\log_2 5 - 1 + \frac{1}{8}\log_2 5$$

$$= \frac{5}{8}\log_2 5 - 1$$

Q1 $g(D, \text{Popup}) = H(D) - H(D|\text{Popup}) = 2 - \frac{5}{8}\log_2 5$

$$H_{Popup}(D) = -\frac{3}{8}\log_2\frac{3}{8} - \frac{5}{8}\log_2\frac{5}{8} = 3 - \frac{3}{8}\log_2 3 - \frac{5}{8}\log_2 5$$

$$g_R(D, \text{Popup}) = \frac{g(D, \text{Popup})}{H_{Popup}(D)} = \frac{2 - \frac{5}{8}\log_2 5}{3 - \frac{3}{8}\log_2 3 - \frac{5}{8}\log_2 5}$$

"Colorful" For attribute "Colorful", $n=2$, a_1: Yes, $|D_1| = 4$; a_2: No, $|D_2| = 4$
$|D_{11}| = 3$, $|D_{12}| = 1$, $|D_{21}| = 1$, $|D_{22}| = 3$

$$H(D|\text{Colorful}) = \frac{4}{8}\left(-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}\right) + \frac{4}{8}\left(-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}\right)$$
$$= -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = -\frac{3}{4}\log_2 3 + \frac{3}{4}\log_2 4 + \frac{1}{4}\log_2 4$$
$$= \frac{3}{2} + \frac{1}{2} - \frac{3}{4}\log_2 3 = 2 - \frac{3}{4}\log_2 3$$

$$g(D, \text{Colorful}) = H(D) - H(D|\text{Colorful}) = \frac{3}{4}\log_2 3 - 1$$

$$H_{Colorful}(D) = -\frac{4}{8}\log_2\frac{4}{8} - \frac{4}{8}\log_2\frac{4}{8} = -\log_2\frac{1}{2} = 1$$

$$g_R(D, \text{Colorful}) = \frac{g(D, \text{Colorful})}{H_{Colorful}(D)} = \frac{3}{4}\log_2 3 - 1$$

It's obvious that $\frac{2 - \frac{5}{8}\log_2 5}{3 - \frac{3}{8}\log_2 3 - \frac{5}{8}\log_2 5} > \frac{\frac{3}{2} - \frac{5}{8}\log_2 5}{3 - \frac{3}{8}\log_2 3 - \frac{5}{8}\log_2 5}$

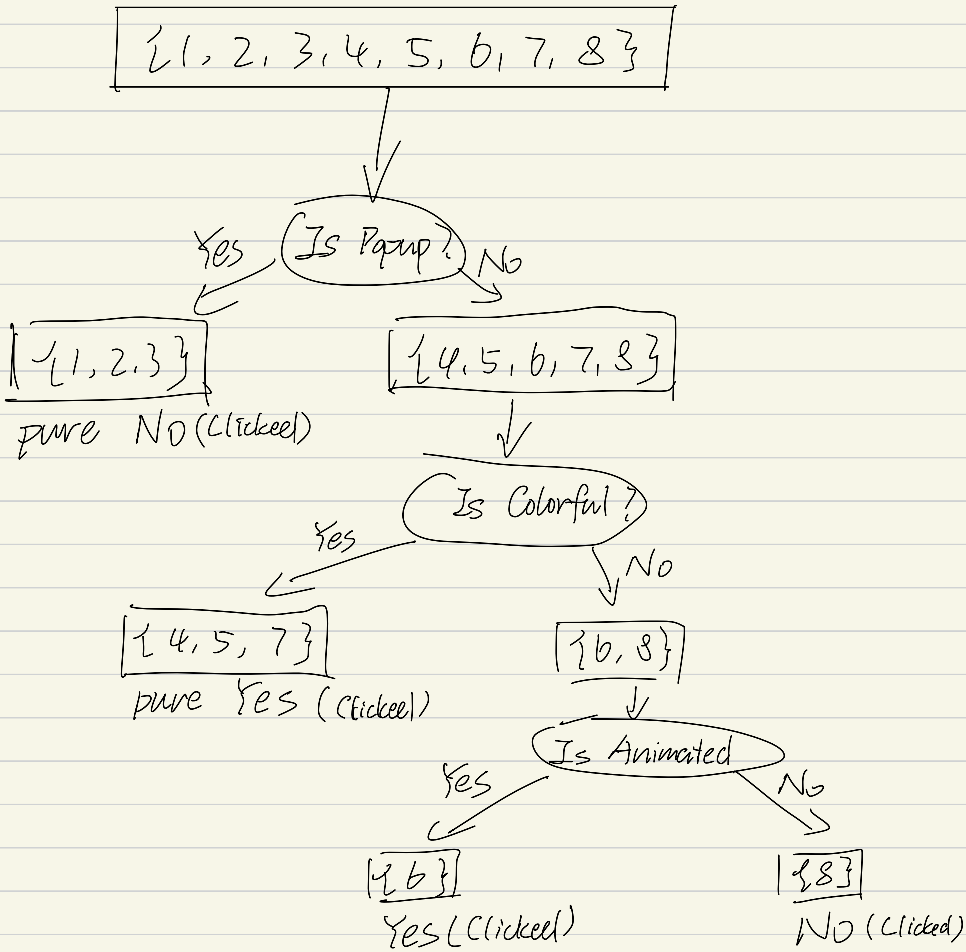
that $g_R(D, \text{Popup}) > g_R(D, \text{Animated})$

Then, use calculator, we have $g_R(D, \text{Popup}) > g_R(D, \text{Colorful})$

Thus, the information gain for the best split at the root is

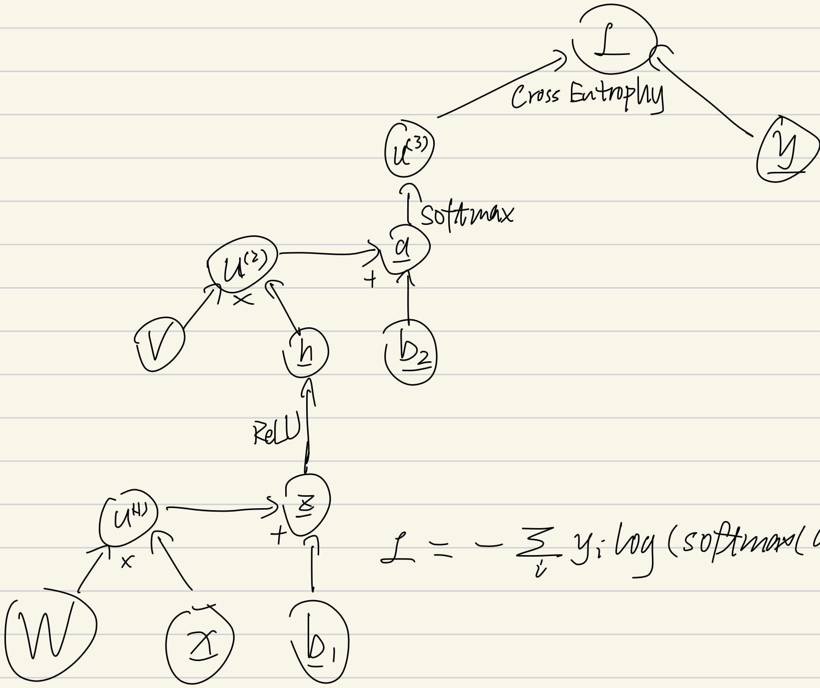
$$\frac{2 - \frac{5}{8}\log_2 5}{3 - \frac{3}{8}\log_2 3 - \frac{5}{8}\log_2 5}$$

Q1 (b)


{1, 2, 3, 4, 5, 6, 7, 8}
Is Popup?
Yes
No
{1, 2, 3}
pure No (Clicked)
{4, 5, 6, 7, 8}
Is Colorful?
Yes
No
{4, 5, 7}
pure Yes (Clicked)
{6, 8}
Is Animated
Yes
No
{6}
Yes (Clicked)
{8}
No (Clicked)

$$\mathcal{L} = -\sum_i u_i^{(3)} \log_2 y_i$$

Q2 (1)



$$\mathcal{L} = -\sum_i y_i \log(\mathrm{softmax}(a_i))$$

(2) In the classification problem, $\underline{y}$ is one-hot matrix

$y_i \in \{0, 1\}$

$$\frac{\partial \mathcal{L}}{\partial u_i^{(3)}} = -\sum_{i=1}^{C} y_i \frac{1}{u_i^{(3)}} \Rightarrow \frac{\partial \mathcal{L}_i}{\partial u_i^{(3)}} = -y_i \frac{1}{u_i^{(3)}}$$

$$\frac{\partial u_i^{(3)}}{\partial a_j} = \begin{cases} u_i^{(3)}(1-u_i^{(3)}) = \mathrm{softmax}(a_i)(1-\mathrm{softmax}(a_i)), & i = j \\ -u_i^{(3)} u_j^{(3)} = -\mathrm{softmax}(a_i)\,\mathrm{softmax}(a_j), & i \neq j \end{cases}$$

$$\frac{\partial \mathcal{L}}{\partial a_i} = \sum_{j=1}^{C}\left(\frac{\partial \mathcal{L}_j}{\partial u_j^{(3)}} \cdot \frac{\partial u_j^{(3)}}{\partial a_i}\right) = \sum_{j=1, j\neq i}^{C}\left(\frac{\partial \mathcal{L}_j}{\partial u_j^{(3)}} \cdot \frac{\partial u_j^{(3)}}{\partial a_i}\right) + \frac{\partial \mathcal{L}_i}{\partial u_i^{(3)}} \cdot \frac{\partial u_i^{(3)}}{\partial a_i} = \sum_{j\neq i}\left[-y_j \frac{1}{u_j^{(3)}} \cdot (-u_j^{(3)} u_i^{(3)})\right] - y_i \frac{1}{u_i^{(3)}} \cdot (u_i^{(3)}(1-u_i^{(3)}))$$

$$= \sum_{j\neq i} y_j u_i^{(3)} - y_i + y_i u_i^{(3)} = \sum_{j=1}^{C} y_j u_i^{(3)} - y_i = u_i^{(3)} \sum_{j=1}^{C} y_j - y_i \quad (\underline{y} \text{ is one-hot})$$

$$= u_i^{(3)} - y_i = \mathrm{softmax}(a_i) - y_i$$

$$\nabla_{\underline{a}} \mathcal{L} = \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial a_1} \\ \frac{\partial \mathcal{L}}{\partial a_2} \\ \vdots \\ \frac{\partial \mathcal{L}}{\partial a_C} \end{bmatrix} = \begin{bmatrix} \mathrm{softmax}(a_1) - y_1 \\ \mathrm{softmax}(a_2) - y_2 \\ \vdots \\ \mathrm{softmax}(a_3) - y_3 \end{bmatrix} = \mathrm{softmax}(V\underline{h} + \underline{b}_2) - \underline{y}$$

$$\frac{\partial a_i}{\partial \vec{v}_j} = \begin{cases} 0, & i \neq j \\ \underline{h}^T, & i = j \end{cases}$$

$$\nabla_V \mathcal{L} = \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial a_1} \cdot \sum_{j=1}^{C} \frac{\partial a_i}{\partial \vec{v}_j} \\ \frac{\partial \mathcal{L}}{\partial a_2} \cdot \sum_{j=1}^{C} \frac{\partial a_i}{\partial \vec{v}_j} \\ \vdots \\ \frac{\partial \mathcal{L}}{\partial a_C} \cdot \sum_{j=1}^{C} \frac{\partial a_i}{\partial \vec{v}_j} \end{bmatrix} = \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial a_1} \cdot \underline{h}^T \\ \frac{\partial \mathcal{L}}{\partial a_2} \cdot \underline{h}^T \\ \vdots \\ \frac{\partial \mathcal{L}}{\partial a_3} \cdot \underline{h}^T \end{bmatrix} = (\mathrm{softmax}(V\underline{h} + \underline{b}_2) - \underline{y}) \cdot \underline{h}^T$$

✓

Q2 (2) $\frac{\partial a_i}{\partial b_{2j}} = \begin{cases} 1, & i=j \\ 0, & i \neq j \end{cases}$

$$\nabla_{\underline{b_2}} \mathcal{L} = \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial a_1} \cdot \sum_{j=1}^{c} \frac{a_i}{b_{2j}} \\ \vdots \\ \frac{\partial \mathcal{L}}{\partial a_c} \cdot \sum_{j=1}^{c} \frac{a_c}{b_{2c}} \end{bmatrix} = \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial a_1} \\ \vdots \\ \frac{\partial \mathcal{L}}{\partial a_c} \end{bmatrix} = \text{softmax}(V\underline{h} + \underline{b_2}) - \underline{y}$$

$V \in \mathbb{R}^{c \times k}$ $\quad \underline{h} \in \mathbb{R}^{c}$

$$\frac{\partial a}{\partial \underline{h}} = V \Rightarrow \frac{\partial a_i}{\partial \underline{h}} = (\underline{v}_i^T)^T \text{ (i-th row of } V\text{)}$$

$$\nabla_{\underline{h}} \mathcal{L} = V^T \cdot \nabla_{\underline{a}} \mathcal{L} = V^T (\text{softmax}(V\underline{h} + \underline{b_2}) - \underline{y}) \in \mathbb{R}^k$$

$$\frac{\partial h_i}{\partial z_i} = \mathbb{I}(z_i > 0) = \begin{cases} 1, & z_i > 0 \\ 0, & z_i \le 0 \end{cases} \quad \forall i = 1, \dots, K$$

$$\nabla_{\underline{z}} \underline{h} = \begin{bmatrix} \mathbb{I}(z_1 > 0) \\ \vdots \\ \mathbb{I}(z_K > 0) \end{bmatrix}$$

$$\nabla_{\underline{z}} \mathcal{L} = \nabla_{\underline{h}} \mathcal{L} \cdot \nabla_{\underline{z}} \underline{h} = \left[V^T(\text{softmax}(V\underline{h} + \underline{b_2}) - \underline{y})\right]^T \begin{bmatrix} \mathbb{I}(z_1 > 0) \\ \vdots \\ \mathbb{I}(z_K > 0) \end{bmatrix}$$

$$\frac{\partial \underline{z}}{\partial W} = \underline{x}^T \quad , \quad \frac{\partial \underline{z}}{\partial \underline{b_1}} = 1$$

$$\nabla_W \mathcal{L} = \left[V^T(\text{softmax}(V \,\text{ReLU}(W\underline{x} + \underline{b_1}) + \underline{b_2}) - \underline{y})\right] \cdot \begin{bmatrix} H(\underline{w}_1^T \underline{x} + b_{1,1}) \\ \vdots \\ H(\underline{w}_K^T \underline{x} + b_{1K}) \end{bmatrix} \cdot \underline{x}^T$$

$$\nabla_{\underline{b_1}} \mathcal{L} = \left[V^T(\text{softmax}(V \,\text{ReLU}(W\underline{x} + \underline{b_1}) + \underline{b_2}) - \underline{y})\right] \cdot \begin{bmatrix} H(\underline{w}_1^T \underline{x} + b_{1,1}) \\ \vdots \\ H(\underline{w}_K^T \underline{x} + b_{1K}) \end{bmatrix}$$

Q3 (a) The width of input layer is $n_{in} = 32$, the depth is 1
The width of output n_{out} is 28
Denote the width of the convolution filters by k
We have,

$$n_{in} - k + 1 = n_{out} \Rightarrow k = 32 - 28 + 1 = 5$$

(b) $K_1 = 6$, $F_1 = 5$, $S_1 = 1$, $P_1 = 0$, $C_1 = 1$

Numbers of parameters in each filter (with a bias term)

$$F_1^2 \times C_1 + 1 = 26$$

Total number of neurons needed, also the total number of parameters:

$$26 \times K_1 = 156$$

156 neurons are needed in the convolutional layer C1

(c) $W_1 = H_1 = 28$; $F_2 = 2$, $P_2 = 0$; $W_2 = H_2 = 14$, $K_2 = 6$

$$W = (W_1 - F_2 + 2P_2)/S_2 + 1$$

$$\Rightarrow S_2 = \frac{W_1 - F_2 + 2P_2}{W_2 - 1} = \frac{28 - 2}{13} = 2$$

The stride distance required for this filter is 2

Total neurons needed

$$K_2 \times (F^2 \times 1 + 1) = 30$$

We need 30 neurons in total

Q3 (d) The purpose in using the hidden layers for convolution and pooling is that we want to convert the input data into the output form with the most likely values. During the converting process, we filter the data in each layers is we hope we filter according to some "features" such that the result in the next layer is more "featured" and can be used to distinguish the output. However, there may be "features", which will never be understood by human.

Besides, pooling layers does better in extracting and highlighting features. It also do well in reducing the size of data with a lower cost. Besides, buy simplify the model with pooling layer, it might help us avoid overfitting.

Since the convolution layers and pooling layers are supposed to learn only the local features, the fully connected layers, intuitively, mean to combine the local features into the output.

Q4 (a) In convolution layers, only a few parameters for filters are used, which while reduce cost of calculation in training models. If use FC layers, the loads on storing and calculating parameters between many layers is extremely heavy.

And the same time, fewer parameters of convolutional layers means the model is simpler such that it could help avoiding overfitting.

(b) $[1, 4, 0, -2, 3]$

Given the length-3 filter, we assume it to be linear:

$$f(x_1, x_2, x_3) = w_1 x_1 + w_2 x_2 + w_3 x_3$$

$$\begin{cases} w_1 + 4w_2 = -2 \\ 4w_1 - 2w_3 = 2 \\ -2w_2 + 3w_3 = 11 \end{cases} \Rightarrow \underline{w} = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} = \begin{bmatrix} 2 \\ -1 \\ 3 \end{bmatrix}$$

(c)

$$[a, b, c] \xrightarrow{[x, y, z]} [ax, ay, az+bx, by, bz+cx, cy, cz]$$

Since the input is a 2×2 matrix, the size of the filter is:

$$H_2 = (H_1 - F + 2P)/S + 1 \Rightarrow H_1 = S(H_2 - 1) + F - 2P$$

$$H_1 = 1 \times (2-1) + 2 - 0 = 3$$

The output of transpose convolution is

$$-1\begin{bmatrix} +1 & -1 & 0 \\ 0 & +1 & 0 \\ 0 & 0 & 0 \end{bmatrix} + 2\begin{bmatrix} 0 & +1 & -1 \\ 0 & 0 & +1 \\ 0 & 0 & 0 \end{bmatrix} + 3\begin{bmatrix} 0 & 0 & 0 \\ +1 & -1 & 0 \\ 0 & +1 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & +1 & -1 \\ 0 & 0 & +1 \end{bmatrix} = \begin{bmatrix} -1 & 3 & -2 \\ 3 & -3 & 1 \\ 0 & 3 & 1 \end{bmatrix}$$