DDA3020 Assignment 3 Student ID: 121090429 Ou Ziyi 欧子裔

Written Problems

1.1 Tree

① Root Node

$$H(\text{class}) = H(\tfrac{4}{6},\tfrac{2}{6}) = -[\tfrac{4}{6}\cdot\log_2(\tfrac{4}{6}) + \tfrac{2}{6}\cdot\log_2(\tfrac{2}{6})]$$
$$= -[\tfrac{2}{3} + \log_2(\tfrac{1}{3})]$$
$$= 0.9182958$$

$$H(\text{class}\mid\text{gender}) = \tfrac{4}{6}H(\tfrac{1}{2},\tfrac{1}{2}) + \tfrac{2}{6}H(\tfrac{2}{2},\tfrac{0}{2})$$
$$= \tfrac{4}{6}\times(-1)\times[\tfrac{1}{2}\cdot\log_2(\tfrac{1}{2}) + \tfrac{1}{2}\cdot\log_2(\tfrac{1}{2})] + \tfrac{2}{6}\times(-1)\times[\tfrac{2}{2}\cdot\log_2(\tfrac{2}{2}) + 0]$$
$$= -\tfrac{2}{3}\cdot(\log_2(\tfrac{1}{2}))$$
$$= \tfrac{2}{3} = 0.6666667$$

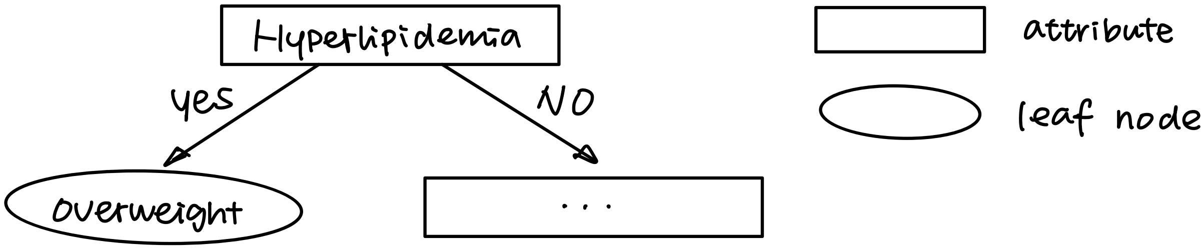
$$H(\text{class}\mid\text{hyperlipidemia}) = \tfrac{3}{6}H(\tfrac{3}{3},0) + \tfrac{3}{6}H(\tfrac{1}{3},\tfrac{2}{3})$$
$$= \tfrac{1}{2}\times(-1)\times[\tfrac{1}{3}\cdot\log_2(\tfrac{1}{3}) + \tfrac{2}{3}\cdot\log_2(\tfrac{2}{3})]$$
$$= -\tfrac{1}{2}\times[\tfrac{2}{3} + \log_2(\tfrac{1}{3})]$$
$$= -\tfrac{1}{3} + \tfrac{1}{2}\log_2 3$$
$$= 0.4591479$$

$$H(\text{class}\mid\text{unhealthy diet}) = \tfrac{4}{6}H(\tfrac{3}{4},\tfrac{1}{4}) + \tfrac{2}{6}H(\tfrac{1}{2},\tfrac{1}{2})$$
$$= \tfrac{2}{3}\times(-1)\times[\tfrac{3}{4}\log_2(\tfrac{3}{4}) + \tfrac{1}{4}\log_2(\tfrac{1}{4})] + \tfrac{1}{3}\times(-1)\times[\tfrac{1}{2}\log_2(\tfrac{1}{2}) + \tfrac{1}{2}\log_2(\tfrac{1}{2})]$$
$$= \tfrac{5}{3} - \tfrac{1}{2}\log_2 3$$
$$= 0.8741854$$

$$H(\text{class}\mid\text{exercise}) = \tfrac{4}{6}H(\tfrac{1}{2},\tfrac{1}{2}) + \tfrac{2}{6}H(\tfrac{2}{2},\tfrac{0}{2})$$
$$= \tfrac{2}{3} = 0.6666667$$

$$I(\text{Class};\text{attribute}) = H(\text{class}) - H(\text{Class}\mid\text{attribute})$$

Since the information gained from the hyperlipidemia is the largest, we choose it as root node.



② All students with Hyperlipidemia are overweight ⇒ pure node

For students who don't have Hyperlipidemia:

denote "NO hyperlipidemia" as "NO H". $P(\text{NO H}) = \frac{1}{2}$

Since $H(X|Y) = -\sum_{(x,y)\in(X,Y)} p(x,y)\log_2 p(x|y)$,

$$\Rightarrow H(\text{class}\,|\,\text{NO H}) = -\left[\frac{1}{6}\times\log_2\left(\frac{1}{3}\right)+\frac{2}{6}\times\log_2\left(\frac{2}{3}\right)\right]$$

$$= -\left[\frac{1}{3}+\frac{1}{2}\log_2\left(\frac{1}{3}\right)\right] = 0.4591479$$

$H(\text{class}\,|\,\text{gender}, \text{NO H}) = 0$

since $p(\text{overweight}\,|\,\text{girl}, \text{NO H}) = p(\text{not overweight}\,|\,\text{boy}, \text{NO H}) = 1$

$$H(\text{class}\,|\,\text{unhealthy diet}, \text{NO H}) = -\left[\frac{2}{3}\cdot H\left(\frac{1}{2},\frac{1}{2}\right)+\frac{1}{3}\cdot H(0,1)\right]\cdot P(\text{NO H})$$

$$= -\frac{2}{3}\left[\frac{1}{2}\log_2\left(\frac{1}{2}\right)+\frac{1}{2}\log_2\left(\frac{1}{2}\right)\right]\cdot\frac{1}{2}$$

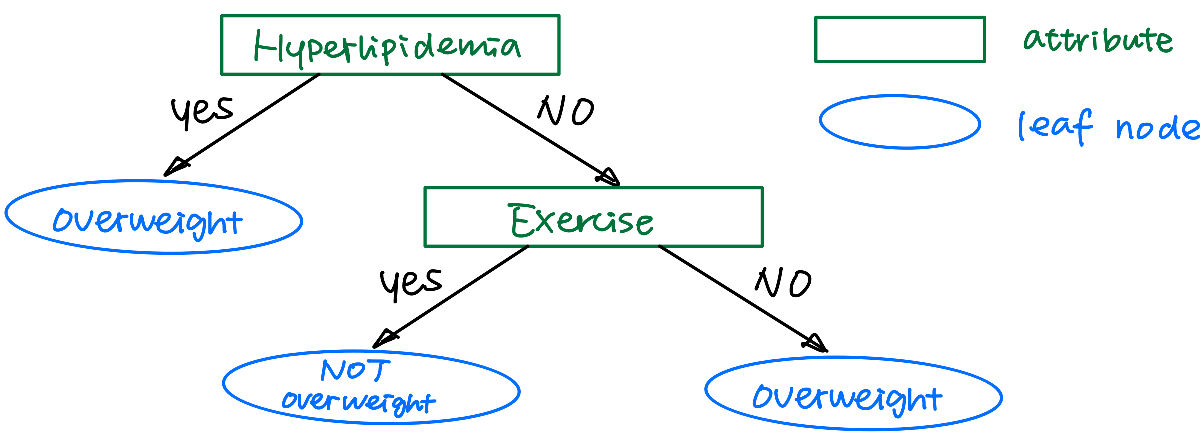
$$= \frac{1}{3} = 0.333333$$

$H(\text{class}\,|\,\text{exercise}, \text{NO H}) = 0$

since $P(\text{overweight}\,|\,\text{no exercise}, \text{NO H}) = P(\text{not overweight}\,|\,\text{exercise}, \text{NO H}) = 1$

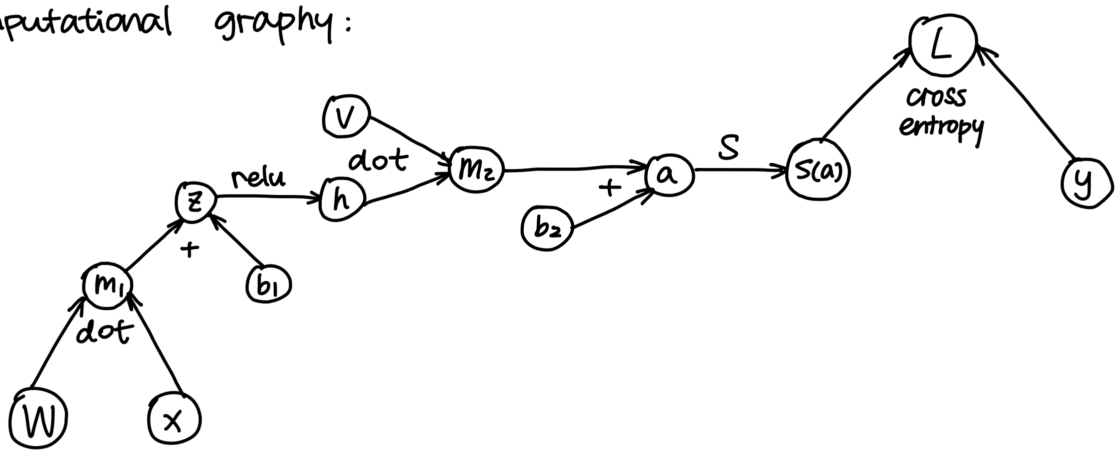
So we can either choose gender or exercise as attributes.

Choose "exercise", we update the decision tree:



1.2 Backpropagation for MLP

computational graphy:



Let $Wx = m_1 \in \mathbb{R}^K$, $Vh = m_2 \in \mathbb{R}^C$

$$u_2 = \nabla_a L = \left(\frac{\partial L}{\partial a}\right)^T = (p-y)$$

$$\Rightarrow L = \text{CrossEntropy}(y, S(a)) = (p-y)^T a = (p-y)^T(Vh+b_2) = u_2^T(Vh+b_2)$$

$$\nabla_V L = \left(\frac{\partial}{\partial V}(u_2^T Vh + u_2^T b_2)\right)^T = \psi_1(u_2, h) = u_2 h^T \in \mathbb{R}^{C\times K}$$

$$\nabla_{b_2} L = \left(\frac{\partial}{\partial b_2}(u_2^T Vh + u_2^T b_2)\right)^T = \psi_2(u_2) = u_2 \in \mathbb{R}^C$$

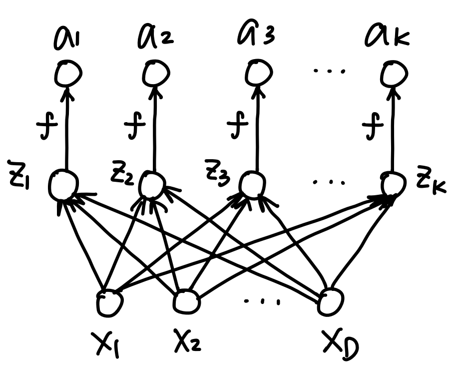
$$\nabla_W L = \left[\frac{\partial L}{\partial W}\right] = \left(\frac{\partial z}{\partial W}\right)^T \cdot \left(\frac{\partial L}{\partial z}\right)^T = \psi_3(u_1, x) = u_1 x^T \in \mathbb{R}^{K\times D}$$

$$\nabla_{b_1} L = \left(\frac{\partial L}{\partial b_1}\right)^T = \left(\frac{\partial z}{\partial b_1}\right)^T \cdot \left(\frac{\partial L}{\partial z}\right)^T = \psi_4(u_1) = u_1 \in \mathbb{R}^K$$

$$\nabla_x L = \left(\frac{\partial L}{\partial x}\right)^T = \left(\frac{\partial z}{\partial x}\right)^T \cdot \left(\frac{\partial L}{\partial z}\right)^T = \psi_5(W, u_1) = W^T u_1 \in \mathbb{R}^D$$

1.3 Connection of neural network to logistic regression

Proof: For a neural network for a K class outcome that uses cross entropy loss, let a_i denote the output of neuron i. Suppose the input layer has D features. (i.e., $X \in \mathbb{R}^D$)



$$z_i = \sum_{j=1}^{D} W_{ij} \cdot x_j + b_i = W_i X + b_i, \ i = 1, 2, \cdots, K$$

$$a = \text{softmax}(WX+b) \Rightarrow a_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

Suppose there are N data points ($y \in \mathbb{R}^N$)

y is the ground-truth label.

$$L = \text{CrossEntropy}(y, a)$$

$$= -\sum_{j=1}^{N}\sum_{i=1}^{K} p(y_j = i) \cdot \log(a_i)$$

Consider a multinomial logistic model with K class,

for class i:

$$P(\hat{y}=i|x) = \frac{e^{W_i x+b_i}}{\sum_{j=1}^{K} e^{W_j x+b_j}}$$

The loss should be $L' = -\sum_{j=1}^{N}\sum_{i=1}^{K} P(y_j=i)\cdot \log(\hat{y})$,

which is equivalent to the neural network for a K class outcome that uses cross entropy loss

1.4 CNN

layers: Conv5(10) + Maxpool$_2$ + Conv5(10) + Maxpool$_2$ + FC$_{10}$
(① ② ③ ④ ⑤)

As shown above, we denote each layer with ①. ②. ③. ④. ⑤ respectively.

(a) ① input: 32×32×3 filter: 5×5×3 stride: 1 padding: 2

$$\frac{N-F+2P}{S}+1 = \frac{32-5+2\times2}{1}+1 = 32$$

$\boxed{\text{shape: } 32\times32\times10}$

② input: 32×32×10 filter: 2×2 stride: 2

$$\frac{N-F}{S}+1 = \frac{32-2}{2}+1 = 16$$

$\boxed{\text{shape: } 16\times16\times10}$

③ input: 16×16×10 filter: 5×5×3 stride: 1 padding: 2

$$\frac{N-F+2P}{S}+1 = \frac{16-5+2\times2}{1}+1 = 16$$

$\boxed{\text{shape: } 16\times16\times10}$

④ input: 16×16×10 filter: 2×2 stride: 2

$$\frac{N-F}{S}+1 = \frac{16-2}{2}+1 = 8$$

$\boxed{\text{shape: } 8\times8\times10}$

⑤ input: 8×8×10 → stretch to 640×1

$640\times1 \xrightarrow{} 10\times640 \text{ weights} \xrightarrow{} 10\times1$ (WX)

$\boxed{\text{shape: } 10\times1}$

(b) ① each filter has $5 \times 5 \times 3 + 1 = 76$ parameters

10 filters, $76 \times 10 = 760$ parameters in total in this layer.

② 0 (zero) parameters

③ each filter has $5 \times 5 \times 10 + 1 = 251$ parameters

10 filters, $251 \times 10 = 2510$ parameters in total in this layer.

④ 0 (zero) parameters

⑤ $10 \times 640 + 10 = 6410$ parameters

1.5 Dropout

(a) Dropout prevents complex co-adaptation because it forces the network to learn more robust representation that are less dependent on the presence of specific neurons, which reduces the risk of overfitting.

(b) we have proportional coefficient $\frac{1}{1-p}$ to ensure $E(h') = h$ (to preserve the mean value of activation function)

$$E(h') = p \cdot 0 + (1-p) \cdot \frac{h}{1-p} = h$$

(c) Since $h^{(l+1)} = h^{(l)} \odot \text{mask}$, $\frac{\partial h^{(l+1)}}{\partial h^{(l)}} = \text{mask}$

By Chain Rule,

$$\frac{\partial L}{\partial h^{(l)}} = \frac{\partial L}{\partial h^{(l+1)}} \odot \frac{\partial h^{(l+1)}}{\partial h^{(l)}}$$

$$= \frac{\partial L}{\partial h^{(l+1)}} \odot \text{mask}$$

1.6 Assessing AUC and Performance of a Binary Classifier

(a) sigmoid function $\sigma(z) = \frac{1}{1+e^{-z}}$

① positive samples:

$$w_1x_1 + w_2x_2 = \begin{bmatrix} 2 \\ 0.7 \\ -1.3 \\ 2.1 \\ -0.3 \end{bmatrix} \Rightarrow f(x) = \sigma(w_1x_1 + w_2x_2) = \begin{bmatrix} 0.8808 \\ 0.6682 \\ 0.2142 \\ 0.8909 \\ 0.4256 \end{bmatrix}$$

② negative samples:

$$w_1x_1+w_2x_2 = \begin{bmatrix} -3 \\ 1.2 \\ -0.6 \\ 3.5 \\ -4.1 \end{bmatrix} \Rightarrow f(x) = \sigma(w_1x_1+w_2x_2) = \begin{bmatrix} 0.0474 \\ 0.7685 \\ 0.3543 \\ 0.9707 \\ 0.0163 \end{bmatrix}$$

(b) threshold $= 0.5 \Rightarrow \begin{cases} f(x) \geq 0.5, & \hat{y} \text{ is positive} \\ f(x) < 0.5, & \hat{y} \text{ is negative.} \end{cases}$

① Accuracy

$TP = 3,\ TN = 3,\ FP = 2,\ FN = 2$

$TP + FP + FN + TN = 5+5$

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} = \frac{3+3}{5+5} \times 100\% = 60\%$$

② Precision

$$\text{precision} = \frac{TP}{TP+FP} = \frac{3}{3+2} = \frac{3}{5}$$

③ Recall

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{3}{3+2} = \frac{3}{5}$$

④ F1 score

$$f_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \times \frac{3}{5} \times \frac{3}{5}}{\frac{3}{5} + \frac{3}{5}} = \frac{3}{5}$$

⑤ Confusion matrix

$$\begin{array}{c|cc} & \text{positive (predicted)} & \text{negative (predicted)} \\ \hline \text{positive (actual)} & 3 & 2 \\ \text{negative (actual)} & 2 & 3 \end{array}$$

(c) ① threshold $= 0$

$TP = 5,\ TN = 0,\ FP = 5,\ FN = 0$

$$\text{FPR (false positive rate)} = \frac{FP}{FP+TN} = \frac{5}{5+0} = 1$$

$$\text{TPR (true positive rate)} = \frac{TP}{TP+FN} = \frac{5}{5+0} = 1$$

② threshold $= 0.2$

$TP = 5,\ TN = 2,\ FP = 3,\ FN = 0$

$$FPR = \frac{FP}{FP+TN} = \frac{3}{3+2} = \frac{3}{5} = 0.6$$

$$TPR = \frac{TP}{TP+FN} = \frac{5}{5+0} = 1$$

③ threshold = 0.4

$TP = 4$, $TN = 3$, $FP = 2$, $FN = 1$

$$FPR = \frac{FP}{FP+TN} = \frac{2}{2+3} = \frac{2}{5} = 0.4$$

$$TPR = \frac{TP}{TP+FN} = \frac{4}{4+1} = \frac{4}{5} = 0.8$$

④ threshold = 0.6

$TP = 3$, $TN = 3$, $FP = 2$, $FN = 2$

$$FPR = \frac{FP}{FP+TN} = \frac{2}{2+3} = \frac{2}{5} = 0.4$$

$$TPR = \frac{TP}{TP+FN} = \frac{3}{3+2} = \frac{3}{5} = 0.6$$

⑤ threshold = 0.8

$TP = 2$, $TN = 4$, $FP = 1$, $FN = 3$

$$FPR = \frac{FP}{FP+TN} = \frac{1}{1+4} = \frac{1}{5} = 0.2$$

$$TPR = \frac{TP}{TP+FN} = \frac{2}{2+3} = \frac{2}{5} = 0.4$$

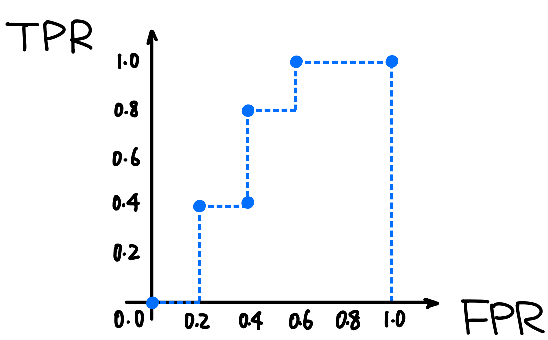
⑥ threshold = 1

$TP = 0$, $TN = 5$, $FP = 0$, $FN = 5$

$$FPR = \frac{FP}{FP+TN} = \frac{0}{0+5} = 0$$

$$TPR = \frac{TP}{TP+FN} = \frac{0}{0+5} = 0$$

The ROC curve is as follows:



(d) AUC = 0.64