CSE5001 · Advanced Artificial Intelligence Fall 2024

董骏博 / 12432995

Assignment 2

November 21, 2024

1 Problem 1

(1)

Step 1:
From the topic, we can get the probability $P(y^{(n)}|x^{(n)};\theta)$ by using $h_\theta(x)$:

$$P(y^{(n)}|x^{(n)};\theta)=\begin{cases} h_\theta(x^{(n)}) & \text{if } y^{(n)}=0, \\ 1-h_\theta(x^{(n)}) & \text{if } y^{(n)}=1. \end{cases}$$

This can be compactly written as:

$$P(y^{(n)}|x^{(n)};\theta)=[h_\theta(x^{(n)})]^{1-y^{(n)}}\cdot[1-h_\theta(x^{(n)})]^{y^{(n)}}$$

Step 2: Express the negative log-likelihood Taking the log of the likelihood:

$$L(\theta)=\prod_{n=1}^{N}P\left(y^{(n)}\mid \mathbf{x}^{(n)};\theta\right)=\prod_{n=1}^{N}h_\theta(x^{(n)})^{1-y^{(n)}}\cdot[1-h_\theta(x^{(n)})]^{y^{(n)}}$$

negative log-likelihood:

$$\begin{aligned}
-\ln L(\theta) &= -\sum_{n=1}^{N}P\left(y^{(n)}\mid \mathbf{x}^{(n)};\theta\right) \\
&= -\sum_{n=1}^{N}\ln\left(h_\theta(x^{(n)})^{1-y^{(n)}}\cdot(1-h_\theta(x^{(n)}))^{y^{(n)}}\right) \\
&= \sum_{n=1}^{N}\left[-\ln h_\theta(x^{(n)})^{1-y^{(n)}}-\ln\left((1-h_\theta(x^{(n)}))^{y^{(n)}}\right)\right] \\
&= \sum_{n=1}^{N}\left[-(1-y^{(n)})\ln h_\theta(x^{(n)})-(y^{(n)})\ln(1-h_\theta(x^{(n)}))\right] \\
&= \sum_{n=1}^{N}\left[-y^{(n)}\ln\left(1-h_\theta\left(x^{(n)}\right)\right)-\left(1-y^{(n)}\right)\ln\left(h_\theta\left(x^{(n)}\right)\right)\right]
\end{aligned}$$

Step 3:
Relate L_θ to the likelihood:

$$L_\theta(\text{cross entropy})=-\frac{1}{N}\ln L(\theta)(\text{likelihood}).$$

It can be seen that the cross entropy loss is the average of the negative log-likelihood.

(2)

$$L_\theta=\frac{1}{N}\sum_{n=1}^{N}\left[-y^{(n)}\ln\left(1-h_\theta\left(x^{(n)}\right)\right)-\left(1-y^{(n)}\right)\ln\left(h_\theta\left(x^{(n)}\right)\right)\right]$$

Step 1: Derivative of the first term $-y^{(n)} \ln \left(1 - h_\theta \left(x^{(n)}\right)\right)$:

$$\frac{\partial}{\partial \theta}\left(-y^{(n)} \ln \left(1 - h_\theta \left(x^{(n)}\right)\right)\right) = \frac{-y^{(n)} \frac{\partial}{\partial \theta} h_\theta \left(x^{(n)}\right)}{1 - h_\theta \left(x^{(n)}\right)}$$

where,

$$\frac{\partial}{\partial \theta} h_\theta \left(x^{(n)}\right) = h_\theta \left(x^{(n)}\right)\left(1 - h_\theta \left(x^{(n)}\right)\right) \cdot \frac{\partial f_\theta \left(x^{(n)}\right)}{\partial \theta}$$

Step 2: Derivative of the second term $(1 - y^{(n)}) \ln(h_\theta(x^{(n)}))$:

$$\frac{\partial}{\partial \theta}\left[-(1 - y^{(n)}) \ln(h_\theta(x^{(n)}))\right] = \frac{-(1 - y^{(n)}) \frac{\partial}{\partial \theta} h_\theta \left(x^{(n)}\right)}{h_\theta \left(x^{(n)}\right)}$$

Step 3: Combine the results:

$$\frac{\partial L_\theta}{\partial \theta} = \frac{-y^{(n)} \frac{\partial}{\partial \theta} h_\theta \left(x^{(n)}\right)}{1 - h_\theta \left(x^{(n)}\right)} + \frac{-(1 - y^{(n)}) \frac{\partial}{\partial \theta} h_\theta \left(x^{(n)}\right)}{h_\theta \left(x^{(n)}\right)}$$

where,

$$\frac{\partial}{\partial \theta} h_\theta \left(x^{(n)}\right) = h_\theta \left(x^{(n)}\right)\left(1 - h_\theta \left(x^{(n)}\right)\right) \cdot \frac{\partial f_\theta \left(x^{(n)}\right)}{\partial \theta}$$

Factorizing:

$$\frac{\partial L_\theta}{\partial \theta} = \frac{1}{N} \sum_{n=1}^{N} \left[\left(h_\theta \left(x^{(n)}\right) + y^{(n)} - 1\right) \cdot \frac{\partial f_\theta \left(x^{(n)}\right)}{\partial \theta}\right]$$

2 Problem 2

(1)

Step 1:

$$\begin{cases} h_1 &= f(w_1 i_1 + w_3 i_2 + b_1 \text{bias}_1), \\ h_2 &= f(w_2 i_1 + w_4 i_2 + b_1 \text{bias}_1) \end{cases} \implies \begin{cases} h_1 = f(0.2 \cdot 0.1 + 0.3 \cdot 0.15 + 0.2 \cdot 1) \\ h_2 = f(0.15 \cdot 0.1 + 0.25 \cdot 0.15 + 0.2 \cdot 1) \end{cases}$$

Step 2: Activation function

$$f(x) = \frac{1}{1 + e^{-x}}$$

$$\begin{cases} h_1 = f(0.2650) \\ h_2 = f(0.2525) \end{cases} \implies \begin{cases} h_1 = \frac{1}{1+e^{-0.2650}} \\ h_2 = \frac{1}{1+e^{-0.2525}} \end{cases} \implies \begin{cases} h_1 \approx 0.5659 \\ h_2 \approx 0.5628 \end{cases}$$

Step 3:

$$\begin{cases} o_1 = w_5 h_1 + w_7 h_2 + b_2 \text{bias}_2 \\ o_2 = w_6 h_1 + w_8 h_2 + b_2 \text{bias}_2 \end{cases} \implies \begin{cases} o_1 = 0.8 \cdot 0.5659 + 0.55 \cdot 0.5628 + 0.4 \cdot 1 \\ o_2 = 0.2 \cdot 0.5659 + 0.3 \cdot 0.5628 + 0.4 \cdot 1 \end{cases}$$

Therefore,

$$\begin{cases} o_1 \approx 1.1623 \\ o_2 \approx 0.6820 \end{cases}$$

(2)

Step 1:

$$\text{Squared error function} = (y - \hat{y}^2)$$

Step 2:

$$\begin{aligned} E_{\text{total}} &= E_{o_1} + E_{o_2} \\ &= (y_1 - o_1)^2 + (y_2 - o_2)^2 \\ &= (0.99 - 1.1623)^2 + (0.01 - 0.6820)^2 \\ &= 0.0297 + 0.4516 \\ &= 0.4813 \end{aligned}$$

So, we get the E_{total} = **0.4813**

(3)

Step 1: the gradient is

$$\frac{\delta E_{total}}{\delta w_5} = \frac{\delta E_{total}}{\delta \text{out}_{o1}} \cdot \frac{\delta \text{out}_{o1}}{\delta w_5}$$

where,

$$\frac{\delta E_{total}}{\delta \text{out}_{o1}} = \frac{(y_1 - o_1)^2 + (y_2 - o_2)^2}{\delta \text{out}_{o1}} = -2(y_1 - o_1)$$

$$\frac{\delta \text{out}_{o1}}{\delta w_5} = \frac{w_5 h_1 + w_7 h_2 + b_2 \text{bias}_2}{\delta w_5} = h_1$$

So, we can get:

$$\frac{\delta E_{total}}{\delta w_5} = -2(y - o_1)h_1 = -2 \cdot (0.99 - 1.1623) \cdot 0.5659 = 0.1950$$

Step 2:

$$\begin{aligned} w_5^{new} &= w_5 - \alpha \frac{\delta E_{total}}{\delta w_5} \\ &= 0.8 - 0.1 \cdot 0.1950 \\ &= 0.7805 \end{aligned}$$

3 Problem 3

Step 1:
Entropy is

$$\begin{aligned} H(Y) &= -\sum_{i=1}^{N} p_i \cdot \log_2(p_i) \\ &= -(\frac{3}{9}\log_2\frac{3}{9} + \frac{3}{9}\log_2\frac{3}{9} + \frac{3}{9}\log_2\frac{3}{9}) \\ &= \log_2 3 \end{aligned}$$

Step 2:
Conditional entropy are:

$$H(Y|X) = \sum_{i=1}^{N} p(X = x_i) \cdot H(Y|X = x_i)$$

Since attributes “height” and “weight” are numeric, we should select the split boundary according to

$$t^{(i)} = \frac{x^{(i)} + x^{(i+1)}}{2}, \quad i = 1, \ldots, n-1$$

$$H(Y \mid X = t) = p(X < t) \cdot H(Y \mid X < t) + p(X \geq t) \cdot H(Y \mid X \geq t)$$

For height:

$$x_1^{(n)} = \{165, 168, 171, 175, 177, 180, 182\}$$

$$t_1^{(n)} = \{166.5, 169.5, 173, 176, 178.5, 181\}$$

$$\begin{cases} H(Y \mid X = 166.5) = -\left(\frac{1}{9} \cdot \log_2 1 + \frac{8}{9}\left(\frac{3}{8}\log_2\frac{3}{8} + \frac{3}{8}\log_2\frac{3}{8} + \frac{2}{8}\log_2\frac{2}{8}\right)\right) = \frac{22}{9} - \frac{2}{3}\log_2 3 \\ H(Y \mid X = 169.5) = \frac{7}{9}\log_2 7 - \frac{1}{3}log_2 3 - \frac{2}{9} \\ H(Y \mid X = 173) = \frac{2}{3}\log_2 3 + \frac{2}{9} \\ H(Y \mid X = 176) = \frac{5}{9}\log_2 5 + \frac{2}{9} \\ H(Y \mid X = 178.5) = \frac{5}{9}\log_2 5 + \frac{2}{9} \\ H(Y \mid X = 181) = \log_2 3 \end{cases}$$

So, we can get Conditional entropy of the height:

$$t_1^{(*)} = \arg\min_{t_1^{(i)}} H(Y \mid \text{height} = t_1^{(i)}) = 173 \Rightarrow H(Y \mid \text{height} = t_1^{(*)}) = H(Y \mid \text{height} = 173) = \frac{2}{3}\log_2 3 + \frac{2}{9}$$

For weight:

$$x_2^{(n)} = \{61, 63, 67, 68, 72, 73, 75, 80\}$$

$$t_2^{(n)} = \{62, 65, 67.5, 70, 72.5, 74, 77.5\}$$

$$\begin{cases} H(Y \mid X = 62) = -\left(\frac{1}{9} \cdot \log_2 1 + \frac{8}{9}\left(\frac{3}{8}\log_2\frac{3}{8} + \frac{3}{8}\log_2\frac{3}{8} + \frac{2}{8}\log_2\frac{2}{8}\right)\right) = \frac{22}{9} - \frac{2}{3}\log_2 3 \\ H(Y \mid X = 65) = log_2 3 \\ H(Y \mid X = 67.5) = \frac{5}{9}\log_2 5 + \frac{2}{9} \\ H(Y \mid X = 70) = \frac{5}{9}\log_2 5 + \frac{2}{9} \\ H(Y \mid X = 72.5) = \log_2 3 \\ H(Y \mid X = 74) = \frac{7}{9}\log_2 7 - \frac{1}{3}\log_2 3 - \frac{2}{9} \\ H(Y \mid X = 77.5) = \frac{22}{9} - \frac{2}{3}\log_2 3 \end{cases}$$

So, we can get Conditional entropy of the weight:

$$t_2^{(*)} = \arg\min_{t_2^{(i)}} H(Y \mid weight = t_2^{(i)}) = 77.5; \Rightarrow H(Y \mid X_2 : t_2^{(*)}) = H(Y \mid weight = 77.5) = \frac{22}{9} - \frac{2}{3}\log_2 3$$

For eye-color:

$$\begin{cases} H(Y|eye - color = hazel) = -(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}) = 0.9183 \\ H(Y|eye - color = blue) = -(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}) = 0.9183 \\ H(Y|eye - color = brown) = -(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}) = 0.9183 \end{cases}$$

So, we can get Conditional entropy of the eye-color:

$$H(Y|eye - color) = 3 \cdot \frac{1}{3} * 0.9813 = 0.9813$$

For hair-color:

$$\begin{cases} H(Y|hair - color = black) = -(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}) = 0.9183 \\ H(Y|hair - color = brown) = -(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}) = 0.9183 \\ H(Y|hair - color = blond) = -(\frac{1}{3}\log_2\frac{1}{3} + \frac{1}{3}\log_2\frac{1}{3} + \frac{1}{3}\log_2\frac{1}{3}) = \log_2 3 \end{cases}$$

So, we can get Conditional entropy of the hair-color:

$$H(Y|hair - color) = \frac{1}{3} * 0.9813 + \frac{1}{3} * 0.9813 + \frac{1}{3} * log_2 3 = 1.1405$$

Step 3:
Information gain is

$$IG(X) = H(Y) - H(Y|X)$$

$$\begin{cases} IG(height) = H(region) - H(region|height) = \log_2 3 - \frac{2}{3}\log_2 3 - \frac{2}{9} = 0.3061 \\ IG(weight) = H(region) - H(region|weight) = \log_2 3 - 229 - +\frac{2}{3}\log_2 3 = 0.1972 \\ IG(eye-color) = H(region) - H(region|eye-color) = \log_2 3 - 0.9183 = 0.6667 \\ IG(hair-color) = H(region) - H(region|hair-color) = \log_2 3 - 1.1405 = 0.4444 \end{cases}$$

Step 4:
Since IG(eye-color) > IG(hair-color) > IG(height) > IG(weight), we choose **eye-color** as tree's root.