3.21

Intro

In this question we will derive the expression 3.76 for the mutual information between the binary feature j and the output y. We will use the notation from that equation (e.g. $\pi_c = p(y = c)$).

Solution

The general expression for the mutual information between a feature j and the output is

$$I_j = \sum_{x_j} \sum_{y} p(x_j, y) log \frac{p(x_j, y)}{p(x_j)p(y)} \tag{1}$$

For binary features, this expression turns into:

$$I_j = \sum_{y} \left(p(x_j = 0, y) log \frac{p(x_j = 0, y)}{p(x_j = 0)p(y)} + p(x_j = 1, y) log \frac{p(x_j = 1, y)}{p(x_j = 1)p(y)} \right) \tag{2}$$

We can use the chain rule, to rewrite (2) as follow:

$$I_j = \sum_{y} \left(p(y)p(x_j = 0|y) log \frac{p(x_j = 0|y)}{p(x_j = 0)} + p(y)p(x_j = 1|y) log \frac{p(x_j = 1|y)}{p(x_j = 1)} \right) \tag{3}$$

Now, from equation 3.76 we know that:

$$\begin{aligned} &\pi_c = p(y = c) \\ &\theta_{jc} = p(x_j = 1|y = c) \\ &\theta_j = p(x_j = 1) = \sum_c \pi_c \theta_{jc} \end{aligned} \tag{4}$$

Substituting (4) in (3) we get the desired result:

$$I_j = \sum_{y} \left(\pi_c (1 - \theta_{jc}) log \frac{1 - \theta_{jc}}{1 - \theta_j} + \pi_c \theta_{jc} log \frac{\theta_{jc}}{\theta_j} \right) \tag{5}$$

Conclusion

In this question we derived Equation 3.76 for the mutual information between a binary feature and the output.