3.22

Intro

In this exercise, we will fit a naive Bayes spam filter by hand. In order to do that, we need to calculate the number of samples of each class $\{spam, non-spam\}$ and the number of times each word of interest occurs on each class.

Solution

The first step is to count the number of samples for each class

$$\begin{aligned}
N_{non-spam} &= 4 \\
N_{spam} &= 3
\end{aligned} \tag{1}$$

The second step is to calculate the number of relevant word occurences in each sentence, and group them up by word and class:

$$\begin{aligned}
N_{spam,secret} &= 2 \\
N_{spam,offer} &= 2 \\
N_{spam,million} &= 1 \\
N_{spam,dollar} &= 1 \\
N_{spam,today} &= 1 \\
N_{spam,is} &= 1 \\
N_{spam,the\_rest} &= 0
\end{aligned} \tag{2}$$

$$\begin{aligned}
N_{non-spam,low} &= 2 \\
N_{non-spam,price} &= 2 \\
N_{non-spam,for} &= 1 \\
N_{non-spam,valued} &= 1 \\
N_{non-spam,custom} &= 1 \\
N_{non-spam,play} &= 1 \\
N_{non-spam,secret} &= 1 \\
N_{non-spam,sports} &= 2 \\
N_{non-spam,today} &= 1 \\
N_{non-spam,is} &= 1 \\
N_{non-spam,healthy} &= 1 \\
N_{non-spam,pizza} &= 1 \\
N_{non-spam,the\_rest} &= 0
\end{aligned} \tag{3}$$

With the values of (1), (2) and (3) we can calculate all the MLE's asked in the question:

$$\begin{aligned}
\pi_{spam} &= \frac{3}{7} \\
\pi_{non-spam} &= \frac{4}{7} \\
\theta_{spam,secret} &= \frac{2}{3} \\
\theta_{non-spam,secret} &= \frac{1}{4} \\
\theta_{sports,non-spam} &= \frac{2}{4} = \frac{1}{2} \\
\theta_{dollar,spam} &= \frac{1}{3}
\end{aligned} \tag{4}$$

Conclusion

In this exercise, we fitted a naive Bayes spam filter by hand. We saw that the procedure to calculate the MLE parameters was straighfoward. In practice, we would write a code to fit the model rather than do it by hand. The code that perfom the fitting would be as straighfoward as the procedure above.