Naïve Bayes Classification

Stu Field

1 Bayes' Rule

$$P(\theta|Data) = \frac{P(Data|\theta) \times P(\theta)}{P(Data)} \tag{1}$$

2 Bayes Classifier

$$\begin{aligned}
\text{odds ratio} &= \frac{P(x=c_1|Data)}{P(x=c_2|Data)} = \frac{P(Data|x=c_1)\cdot P(x=c_1)}{P(Data|x=c_2)\cdot P(x=c_2)} \\
log(\text{odds}) &= log\left(\frac{P(Data|x=c_1)}{P(Data|x=c_2)}\cdot\frac{P(x=c_1)}{P(x=c_2)}\right) \\
log(\text{odds}) &= log\left(\frac{P(Data|x=c_1)}{P(Data|x=c_2)}\right) + log\left(\frac{P(x=c_1)}{P(x=c_2)}\right)
\end{aligned}$$

3 Bayes In Practice

$$\begin{aligned}
\frac{P(Data|x=c_1)}{P(Data|x=c_2)} &= \frac{P(A_1=a_1, A_2=a_2, \ldots, A_n=a_n|x=c_1)}{P(A_1=a_1, A_2=a_2, \ldots, A_n=a_n|x=c_2)} \\
&= \frac{P(A_1=a_1|x=c_1)P(A_1=a_1|x=c_1), \ldots, P(A_n=a_n|x=c_1)}{P(A_1=a_1|x=c_2)P(A_1=a_1|x=c_2), \ldots, P(A_n=a_n|x=c_2)}
\end{aligned}$$

$$\begin{aligned}
log\left(\frac{P(Data|x=c_1)}{P(Data|x=c_2)}\right) &= log\left(\frac{P(A_1=a_1|x=c_1)}{P(A_1=a_1|x=c_2)}\right) + \\
&\quad log\left(\frac{P(A_2=a_2|x=c_1)}{P(A_2=a_2|x=c_2)}\right) + \cdots + log\left(\frac{P(A_n=a_n|x=c_1)}{P(A_n=a_n|x=c_2)}\right)
\end{aligned}$$

Therefore, putting together gives,

$$\begin{aligned} log(\text{odds}) &= log\left(\frac{P(A_1 = a_1|x = c_1)}{P(A_1 = a_1|x = c_2)}\right) + log\left(\frac{P(A_2 = a_2|x = c_1)}{P(A_2 = a_2|x = c_2)}\right) + \ldots \\ &\quad + log\left(\frac{P(A_n = a_n|x = c_1)}{P(A_n = a_n|x = c_2)}\right) + log\left(\frac{P(x = c_1)}{P(x = c_2)}\right) \end{aligned}$$

where,

$$log\left(\frac{P(x = c_1)}{P(x = c_2)}\right)$$

is the Bayesian prior and is typically set to zero (unless known previously).

3.1 Calculation of the Terms

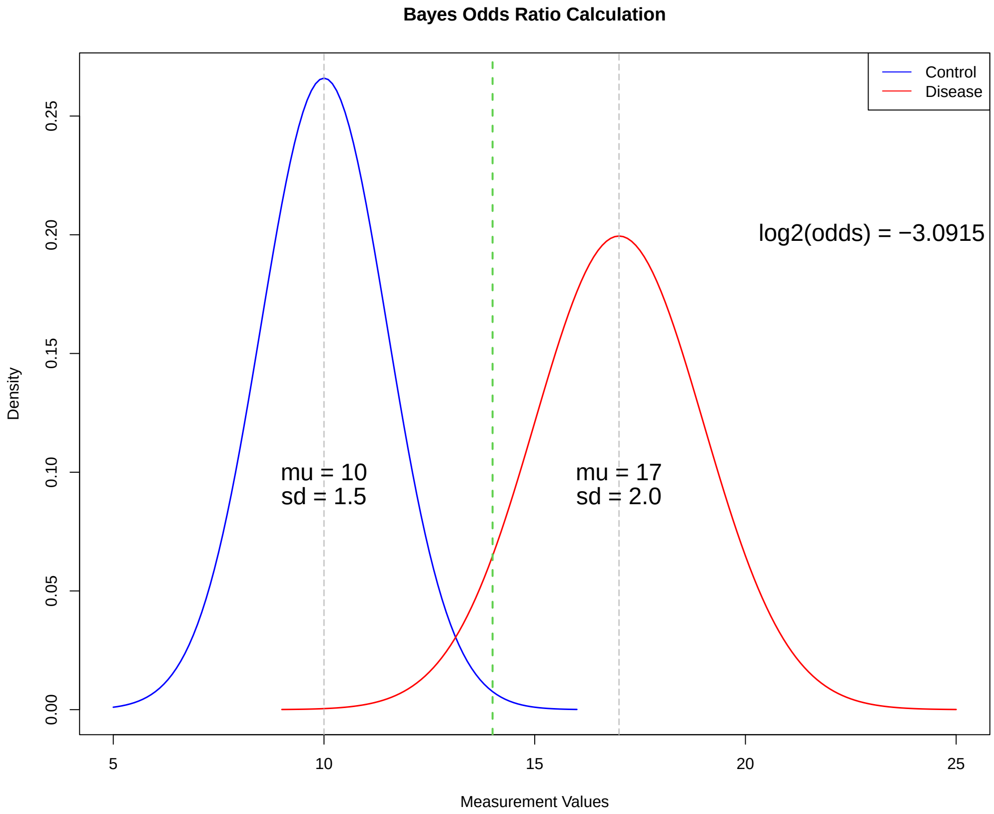

Figure 1: PDFs of the a theoretical control population (blue) and disease population (red) for a single analyte. The population estimates are shown and the green line represents a sample value of 14. The odds ratio is the probability of control given the sample value over the probability of disease given the sample (`dnorm(14, 10, 1.5) / dnorm(14, 17, 2)`).