1 P1

Problem 1 *An AI system is trained to detect photos of lions. The system does not err when the photo actually has a lion, but 5% of photos in which a lion does not appear is detected as having a lion. The algorithm is executed on a set of photos in which 1 out of every 1000 photos actually has a lion. If the AI says a photo has a lion, what is the probability that it actually does have a lion?*

We will use Bayes' Theorem:

$$P(\textit{Actually lion} \mid \textit{AI finds lion}) = \frac{P(\textit{AI finds lion} \mid \textit{Actually lion})P(\textit{Actually lion})}{P(\textit{AI finds lion})}$$

Since the system is accurate whenever there is a lion, $P(\textit{AI finds lion} \mid \textit{Actually lion}) = 1$. The probability of there actually being a lion is given as 1 in 1000, or 0.001. Finally, we can rewrite

$$\begin{aligned} P(\textit{AI finds lion}) =& P(\textit{AI finds lion} \mid \textit{Actually lion})P(\textit{Actually lion}) \\ &+ P(\textit{AI finds lion} \mid \neg\textit{Actually lion})P(\neg\textit{Actually lion}) \end{aligned}$$

Since the system has a false positive rate of 5%, $P(\textit{AI finds lion} \mid \neg\textit{Actually lion}) = 0.05$. Substituting yields

$$\begin{aligned} P(\textit{AI finds lion}) =& 1 * 0.001 \\ &+ 0.05 * 0.999 \\ =& 0.05095 \end{aligned}$$

Now, plugging all of the numbers into our original equation, we have

$$P(\textit{Actually lion} \mid \textit{AI finds lion}) = \frac{1 * 0.001}{0.05095} \approx 1.96\%$$

2 P2

Problem 2 *An AI system trained to detect a leaf disease in photos of tomato bushes. The system does not err when the photo shows a diseased bush, but 11 out of 100 healthy bushes is also detected as diseased. The disease affects 4 out of every 1000 bushes. If the AI says a photo of a bush is diseased, what is the probability that it really is diseased?*

As an alternative approach, we may consider a sample of 25000 bushes. Since 4 of every 1000 bushes are diseased, we have 100 diseased bushes, and 24900 healthy bushes. Of the 100 diseased bushes, all of them are labeled as diseased by the system. Since 11 out of 100 healthy bushes is incorrectly labeled as diseased, of the 24900 healthy bushes, 2739 are marked as diseased. Thus, of the total 2839 bushes labeled as diseased, 100 are actually diseased. So the probability is $\frac{100}{2839} \approx 3.52\%$.