ECE/CS/ME 539 – Fall 2024 — Activity 4

1.

Consider the egg classification example in the lecture note.

Label\Weight	**1g**	**2g**	**3g**	**4g**
Large	15	40	10	5
Jumbo	0	8	17	5

(a) Compute the following empirical probabilities using the values in the table. Note that there are 100 eggs in total.

Answer:

$$P\{\text{Label} = \text{Large}\} = \frac{(15+40+10+5)}{100} = 0.7$$

$$P\{\text{Label} = \text{Jumbo and Weight} = 4\text{g}\} = \frac{5}{100} = \frac{1}{20} = 0.05$$

$$P\{\text{Weight} = 2\text{g} \mid \text{Label} = \text{Large}\} = \frac{P\{2\text{g, Large}\}}{P\{\text{Large}\}} = \frac{(40/100)}{0.7} = \frac{4}{7}$$

$$P\{\text{Weight} = 2\text{g}\} = \frac{(40+8)}{100} = 0.48$$

$$P\{\text{Weight} = 2\text{g} \mid \text{Label} = \text{Jumbo}\} = \frac{P\{2\text{g, Jumbo}\}}{P\{\text{Jumbo}\}} = \frac{(8/100)}{(30/100)} = \frac{4}{15}$$

$$P\{\text{Label} = \text{Jumbo} \mid \text{Weight} = 2\text{g}\} = \frac{P\{\text{Jumbo, } 2\text{g}\}}{P\{2\text{g}\}} = \frac{(8/100)}{(48/100)} = \frac{1}{6}$$

(b) Compute the posterior probability by filling the following table. Each entry should be the value of $P\{\text{row heading (Large, Jumbo)} \mid \text{column heading (1g, 2g, 3g, 4g)}\}$. Then, based on your answer, deduce an optimal decision rule for each weight measurement. For example, if $P\{\text{Large} \mid 1g\} > P\{\text{Jumbo} \mid 1g\}$, then decide if the egg weighs 1g, then determine it is a Large egg. If the probabilities are equal, choose one decision to make the overall decision rule easier.

Label\Weight	**1g**	**2g**	**3g**	**4g**
Large	$\frac{15}{15}$	$\frac{5}{6}$	$\frac{10}{27}$	$\frac{1}{2}$
Jumbo	0	$\frac{1}{6}$	$\frac{17}{27}$	$\frac{1}{2}$
Decision (L/J)	Large	Large	Jumbo	Jumbo

Answer: See table. With Weight = 4g, we decide the egg is Jumbo. This way, the decision rule is summarized as following: If weight $\leq$ 2g, decide Large; else decide Jumbo.

(c) In this part of the problem, we will verify that even though the decision rule is "optimal", it does not imply no mistakes will be made. Using the Bayesian decision rule established in part (b), fill in the "confusion matrix" below. The row headings are "ground truth" labels and the column headings are labels according to the decision rule. Therefore, the diagonal elements represent the number of correctly classified eggs, and the off-diagonal elements represent the number of eggs that are misclassified by the decision rule. Define

$$P\{\text{Misclassification}\} = \frac{\sum \text{values of off-diagonal elements in the confusion matrix}}{\sum \text{values of all elements in the confusion matrix}} \times 100\%$$

Compute the probability of misclassification of the decision rule in part (b).

Ground truth \decision rule predicted	**Large**	**Jumbo**
Large		
Jumbo		

Answer: Use the decision rule: if weight $\leq$ 2g, then Large, else Jumbo. We have

Ground truth \decision rule predicted	**Large**	**Jumbo**
Large	55	15
Jumbo	8	22

$P\{\text{Mis-classification}\} = \frac{(8+15)}{(55+15+8+22)} = 23\%$

2.

A rare disease affects 1% of the population. Three medical tests have been developed to detect this disease. If a person has the disease, the tests will correctly identify them as having the disease with a probability of 98%, 80% and 70%, respectively (these are called the test sensitivity). If a person does not have the disease, the tests will correctly identify them as not having the disease with a probability of 95%, 60% and 99%, respectively (these are called the test specificity).

(a) If a person is tested using only the first test and the result is positive, what is the probability they actually have the disease? (This is called the positive predictive value.)

Answer: Let $D(D^C)$ be the event that a person has the disease (not have, respectively) and $T_i^+(T_i^-)$ the event that the ith test is positive (negative). Using Bayes' Theorem

$$P(D \mid T_1^+) = \frac{P(T_1^+ \mid D)P(D)}{P(T_1^+)} = \frac{P(T_1^+ \mid D)P(D)}{P(T_1^+ \mid D)P(D) + P(T_1^+ \mid D^C)P(D^C)}$$

where $P(T_1^+ \mid D) = 0.98$ (sensitivity), $P(D) = 0.01$ (prevalence), $P(T_1^+ \mid D^C) = 1 - P(T_1^- \mid D^C) = 1 - 0.95$, and $P(D^C) = 1 - 0.01$.

Substituting in the equation above, we get $P(D \mid T_1^+) = \frac{0.98 \times 0.01}{0.0593} = 0.165$.

(b) If a person is tested using only the first test and the result is negative, what is the probability they actually don't have the disease? (This is called the negative predictive value.)

Answer:

$$P(D^C \mid T_1^-) = \frac{P(T_1^- \mid D^C)P(D^C)}{P(T_1^-)} = \frac{P(T_1^- \mid D^C)P(D^C)}{P(T_1^- \mid D)P(D) + P(T_1^- \mid D^C)P(D^C)}$$

Where $P(T_1^- \mid D^C) = 0.95$ (specificity) and $P(T_1^- \mid D) = 1 - P(T_1^+ \mid D) = 1 - 0.98$.
Substituting in the equation above, we get $P(T_1^-) = \frac{0.95 \times 0.99}{(1-0.98)\times 0.01 + 0.95 \times 0.99} = 0.9998$.

(c) Are the events "test result is positive" and "having the disease" statistically independent? Justify your answer.

Answer:

$$\begin{aligned} P(T_1^+, D) &= P(T_1^+ \mid D)P(D) = 0.98 \times 0.01 = 0.0098 \\ P(T_1^+) &= P(T_1^+ \mid D)P(D) + P(T_1^+ \mid D^C)P(D^C) = 0.0593 \\ P(D) &= 0.01 \end{aligned}$$

Since $P(T_1^+, D) \neq P(T_1^+) \times P(D)$, the two events are not independent.

(d) Now suppose a person is tested using all three tests and the results are positive, negative, positive, respectively. Assuming that the tests are statistically independent when conditioned on random variable D, what is the probability that the person actually has the disease.

Answer:

$$P(D \mid T_1^+, T_2^-, T_3^+) = \frac{P(T_1^+, T_2^-, T_3^+ \mid D)P(D)}{P(T_1^+, T_2^-, T_3^+)} = \frac{P(T_1^+ \mid D)P(T_2^- \mid D)P(T_3^+ \mid D)P(D)}{P(T_1^+, T_2^-, T_3^+)}$$

where

$$\begin{aligned} P(T_1^+, T_2^-, T_3^+) &= P(T_1^+, T_2^-, T_3^+ \mid D)P(D) + P(T_1^+, T_2^-, T_3^+ \mid D^C)P(D^C) \\ &= P(T_1^+ \mid D)P(T_2^- \mid D)P(T_3^+ \mid D)P(D) + P(T_1^+ \mid D^C)P(T_2^- \mid D^C)P(T_3^+ \mid D^C)P(D^C) \end{aligned}$$

Substituting all quantities above, we get

$$\begin{aligned} P(D \mid T_1^+, T_2^-, T_3^+) &= \frac{0.98 \times 0.2 \times 0.7 \times 0.01}{0.98 \times 0.2 \times 0.7 \times 0.01 + 0.05 \times 0.6 \times 0.01 \times 0.99} \\ &= \frac{0.001372}{0.001372 + 0.000297} \\ &= 0.822 \end{aligned}$$