SI 140A-02 Probability & Statistics for EECS, Fall 2024
Homework 2

Name: **Wenye Xiong**
Student ID: 2023533141

Due on Oct. 15, 2024, 11:59 UTC+8

Read all the instructions below carefully before you start working on the assignment, and before you make a submission.

- You are required to write down all the major steps towards making your conclusions; otherwise you may obtain limited points of the problem.
- Write your homework in English; otherwise you will get no points of this homework.
- Any form of plagiarism will lead to 0 point of this homework.

Problem 1

Alice is trying to communicate with Bob, by sending a message (encoded in binary) across a channel. If she sends a 0, there is a 5% chance of an error occurring, resulting in Bob receiving a 1; if she sends a 1, there is a 10% chance of an error occurring, resulting in Bob receiving a 0. To reduce the chance of miscommunication, Alice and Bob decide to use a repetition code. Again Alice wants to convey a 0 or a 1, but this time she repeats it two more times, so that she sends 000 to convey 0 and 111 to convey 1. Bob will decode the message by going with what the majority of the bits were. Assume error events for different bits are independent of each other. Given that Bob receives 110, what is the probability that Alice intended to convey a 1?

Solution

Before we use the Bayes' theorem to solve this problem, we need to have a prior probability of Alice's intention. Since Alice has an equal chance of sending 0 or 1, we have $P(A_0) = P(A_1) = 0.5$. Let A_i be the event that Alice sends i for $i \in \{0, 1\}$, and B_j be the event that Bob receives j for $j \in \{0, 1\}$, $B_{j,k,l}$ be the event that Bob receives jkl for $j, k, l \in \{0, 1\}$. According to Bayes' theorem, we have the following formula:

$$P(A_1|B_{110}) = \frac{P(A_1)P(B_{110}|A_1)}{P(A_1)P(B_{110}|A_1)+P(A_0)P(B_{110}|A_0)} = \frac{0.5*0.9^2*0.1}{0.5*0.9^2*0.1+0.5*0.05^2*0.95} = 0.97$$

Problem 2

To battle against spam, Bob installs two anti-spam programs. An email arrives, which is either legitimate (event L) or spam (event L^c), and which program j marks as legitimate (event M_j) or marks as spam (event M_j^c) for $j \in \{1, 2\}$. Assume that 10% of Bobs email is legitimate and that the two programs are each "90% accurate" in the sense that $P(M_j|L) = P(M_j^c|L^c) = 9/10$. Also assume that given whether an email is spam, the two programs' outputs are conditionally independent.

(a) Find the probability that the email is legitimate, given that the 1st program marks it as legitimate (simplify).

(b) Find the probability that the email is legitimate, given that both programs mark it as legitimate (simplify).

(c) Bob runs the 1st program and M_1 occurs. He updates his probabilities and then runs the 2nd program. Let $\tilde{P}(A) = P(A|M_1)$ be the updated probability function after running the 1st program. Explain briefly in words whether or not $\tilde{P}(L|M_2) = P(L|M_1 \cap M_2)$: is conditioning on $M_1 \cap M_2$ in one step equivalent to first conditioning on M_1, then updating probabilities, and then conditioning on M_2?

Solution

(a):
According to the Bayes' theorem, we have the following formula:

$$P(L|M_1) = \frac{P(L)P(M_1|L)}{P(L)P(M_1|L)+P(L^c)P(M_1|L^c)} = \frac{0.1*0.9}{0.1*0.9+0.9*0.1} = 0.5$$

(b):
According to the Bayes' theorem, we have the following formula:

$$P(L|M_1 \cap M_2) = \frac{P(L)P(M_1|L)P(M_2|L)}{P(L)P(M_1|L)P(M_2|L)+P(L^c)P(M_1|L^c)P(M_2|L^c)} = \frac{0.1*0.9^2}{0.1*0.9^2+0.9*0.1^2} = 0.9$$

(c):
Yes, $\tilde{P}(L|M_2) = P(L|M_1 \cap M_2)$, because the **Coherency of Bayes' rule** and the two programs' outputs are conditionally independent given whether an email is spam. So conditioning on $M_1 \cap M_2$ in one step is equivalent to first conditioning on M_1, then updating probabilities, and then conditioning on M_2. According to the Bayes' theorem, the order of various conditions does not affect the final result.

Problem 3

Fred decides to take a series of n tests, to diagnose whether he has a certain disease (any individual test is not perfectly reliable, so he hopes to reduce his uncertainty by taking multiple tests). Let D be the event that he has the disease, $p = P(D)$ be the prior probability that he has the disease, and $q = 1 - p$. Let T_j be the event that he tests positive on the jth test.

(a) Assume for this part that the test results are conditionally independent given Fred's disease status. Let $a = P(T_j \mid D)$ and $b = P(T_j \mid D^c)$, where a and b don't depend on the jth test. Find the posterior probability that Fred has the disease, given that he tests positive on all n of the n tests.

(b) Suppose that Fred tests positive on all n tests. However, some people have a certain gene that makes them always test positive. Let G be the event that Fred has the gene. Assume that $P(G) = 1/2$ and that D and G are independent. If Fred does not have the gene, then the test results are conditionally independent given his disease status. Let $a_0 = P(T_j \mid D, G^c)$ and $b_0 = P(T_j \mid D^c, G^c)$, where a_0 and b_0 don't depend on j. Find the posterior probability that Fred has the disease, given that he tests positive on all n of the tests.

Solution

(a):
Let $T = T_1 \cap T_2 \cap \cdots \cap T_n$ be the event that Fred tests positive on all n tests. According to the Bayes' theorem, we have the following formula:

$$P(D|T) = \frac{P(D)P(T|D)}{P(D)P(T|D)+P(D^c)P(T|D^c)} = \frac{pa^n}{pa^n+qb^n}$$

(b):
This time we have to consider whether Fred has the gene, so we add the gene event G to the formula and update the value of $P(T|D)$ and $P(T|D^c)$.

$$P(T|D) = P(T|D,G)P(G|D) + P(T|D,G^c)P(G^c|D) = \frac{1}{2} + \frac{1}{2}a_0^n$$
$$P(T|D^c) = P(T|D^c,G)P(G|D^c) + P(T|D^c,G^c)P(G^c|D^c) = \frac{1}{2} + \frac{1}{2}b_0^n$$

And according to the Bayes' theorem, we have the following formula:

$$P(D|T) = \frac{P(D)P(T|D)}{P(D)P(T|D)+P(D^c)P(T|D^c)} = \frac{p\left(\frac{1}{2}+\frac{1}{2}a_0^n\right)}{p\left(\frac{1}{2}+\frac{1}{2}a_0^n\right)+q\left(\frac{1}{2}+\frac{1}{2}b_0^n\right)} = \frac{p+pa_0^n}{p+pa_0^n+q+qb_0^n} = \frac{p+pa_0^n}{1+pa_0^n+qb_0^n}$$

Problem 4

We want to design a spam filter for email. A major strategy is to find phrases that are much more likely to appear in a spam email than in a no spam email. In that exercise,we only consider one such phrase:"free money". More realistically, suppose that we have created a list of 100 words or phrases that are much more likely to be used in spam than in non-spam. Let W_j be the event that an email contains the jth word or phrase on the list. Let

$$p = P(spam), p_j = P(W_j|spam), r_j = P(W_j|not \quad spam)$$

where "spam" is shorthand for the event that the email is spam.
Assume that $W_1, ..., W_{100}$ are conditionally independent given that the email is spam, and also conditionally independent given that it is not spam. A method for classifying emails (or other objects) based on this kind of assumption is called a naive Bayes classifier. (Here "naive" refers to the fact that the conditional independence is a strong assumption, not to Bayes being naive. The assumption may or may not be realistic, but naive Bayes classifiers sometimes work well in practice even if the assumption is not realistic.)
Under this assumption we know, for example, that

$$P(W_1, W_2, W_3^c, W_4^c, ..., W_{100}^c|spam) = p_1p_2(1-p_3)(1-p_4)...(1-p_{100}).$$

Without the naive Bayes assumption, there would be vastly more statistical and computational difficulties since we would need to consider $2^{100} \approx 1.3 \times 10^{30}$ events of the form $A1 \cap A2... \cap A100$ with each A_j equal to either W_j or W_j^c . A new email has just arrived, and it include the 23rd, 64th, and 65th words or phrases on the list (but not the other 97). So we want to compute

$$P(spam|W_1^c, ..., W_{22}^c, W_{23}, W_{24}^c, ..., W_{63}^c, W_{64}, W_{65}, W_{66}^c, ..., W_{100}^c).$$

Note that we need to condition on all the evidence, not just the fact that $W_{23} \cap W_{64} \cap W_{65}$ occurred. Find the condition probability that the new email is spam (in terms of p and the p_j and r_j).

Solution

Let $W = W_1^c \cap W_2^c \cap \cdots \cap W_{22}^c \cap W_{23} \cap W_{24}^c \cap \cdots \cap W_{63}^c \cap W_{64} \cap W_{65} \cap W_{66}^c \cap \cdots \cap W_{100}^c$ be the event that the new email includes the 23rd, 64th, and 65th words or phrases on the list. According to the Bayes' theorem and LOTP, we have the following formula:

$$P(spam|W) = \frac{P(spam)P(W|spam)}{P(spam)P(W|spam)+P(not \quad spam)P(W|not \quad spam)} = \frac{p(1-p_1)(1-p_2)...(1-p_{22})p_{23}(1-p_{24})...(1-p_{63})p_{64}p_{65}(1-p_{66})...(1-p_{100})}{p(1-p_1)(1-p_2)...(1-p_{22})p_{23}(1-p_{24})...(1-p_{63})p_{64}p_{65}(1-p_{66})...(1-p_{100})+(1-p)(1-r_1)(1-r_2)...(1-r_{22})r_{23}(1-r_{24})...(1-r_{63})r_{64}r_{65}(1-r_{66})...(1-r_{100})}$$

Problem 5

Consider a family that has n children, and $n \geq 2$. We are interested in the children's genders, where each child can be a Boy or a girl with equal probability.

(a) What is the probability that all children are girls given that the first (elder) child is a girl?

(b) We ask the father: "Do you have at least one daughter?" He responds "Yes!" Given this extra information, what is the probability that all children are girls? In other words, what is the probability that all children are girls given that we know at least one of them is a girl?

(c) If we randomly ran into one child in the shopping mall, and see that she is a girl. Given this extra information, what is the probability that all children are girls?

(d) We ask the father, "Do you have at least one daughter named Catherine?" He replies, "Yes!" What is the probability that all children are girls? In other words, we want to find the probability that all children are girls, given that the family has at least one daughter named Catherine. Here we assume that if a child is a girl, her name will be Catherine with probability α independently from other children's names. If the child is a boy, his name will not be Catherine.

Solution

(a):
Consider the event that all children are girls and the first child is a girl. This event is equivalent to the event that all children are girls, thus the probability is $\frac{1}{2^n}$. Then consider the probability that the first child is a girl, which is $\frac{1}{2}$. According to the conditional probability formula, we have the following formula:

$$P(\text{all girls}|\text{first girl}) = \frac{P(\text{all girls, first girl})}{P(\text{first girl})} = \frac{1/2^n}{1/2} = \frac{1}{2^{n-1}}$$

(b):
Consider the event that all children are girls and the family has at least one daughter. This event is equivalent to the event that all children are girls, thus the probability is $\frac{1}{2^n}$. Then consider the probability that the family has at least one daughter, which is the complement of the event that all children are boys, thus the probability is $1 - \frac{1}{2^n}$. According to the conditional probability formula, we have the following formula:

$$P(\text{all girls}|\text{at least one girl}) = \frac{P(\text{all girls, at least one girl})}{P(\text{at least one girl})} = \frac{1/2^n}{1-1/2^n} = \frac{1}{2^n-1}$$

(c):
We now reconsider the problem a, we can easily find that if the extra information is a random child is a girl, the probability stays the same. Now we consider the case that we choose a random child and know she is girl, this is exactly the same as the case that we ran into a random child in the shopping mall and see that she is a girl. Thus the probability is $\frac{1}{2^{n-1}}$.
(d):
Consider the event that all children are girls and we has a child called Catherine. Having all girls has a probability of $\frac{1}{2^n}$. The probability that we have a child called Catherine when all children are girls is $1 - (1 - \alpha)^n$, which is the complement of the event that all girls are not named Catherine. Thus the probability of these two events happening together is $(1 - (1 - \alpha)^n)/2^n$.
Next, we consider the probability that we have a child called Catherine, which is the complement of the event that all children are not named Catherine. The probability of a single child being named Catherine is the result of the probability that the child is girl times the probability that the girl is named Catherine, which is $\frac{1}{2} * \alpha$. Thus the probability is $1 - (1 - \frac{1}{2}\alpha)^n$. According to the conditional probability formula, we have the following formula:

$$P(\text{all girls}|\text{one Catherine}) = \frac{P(\text{all girls, one Catherine})}{P(\text{one Catherine})} = \frac{(1-(1-\alpha)^n)/2^n}{1-(1-\frac{1}{2}\alpha)^n} = \frac{1-(1-\alpha)^n}{2^n-(2-\alpha)^n}$$