2023Fall Probability & Statistics for EECS 2023/10/12

Homework 2

Professor: Ziyu Shao & Dingzhu Wen *Due: 2022/10/22 10:59pm*

1. Alice is trying to communicate with Bob, by sending a message (encoded in binary) across a channel.

   (a) Suppose for this part that she sends only one bit (a 0 or 1), with equal probabilities. 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. Given that Bob receives a 1, what is the probability that Alice actually sent a 1?

   (b) 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 that the error probabilities are as in (a), with error events for different bits independent of each other. Given that Bob receives 110, what is the probability that Alice intended to convey a 1?

2. 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.

3. 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

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

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 $A_1 \cap A_2 ... \cap A_{100}$ 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).

4. In Monty Hall problem, now suppose the car is not placed randomly with equal probability behind the three doors. Instead, the car is behind door one with probability p_1, behind door two with probability p_2, and behind door three with probability p_3. Here $p_1 + p_2 + p_3 = 1$ and $p_1 \geq p_2 \geq p_3 > 0$. You are to choose one of the three doors, after which Monty will open a door he knows to conceal a goat. Monty always chooses randomly with equal probability among his options in those cases where your initial choice is correct. What strategy should you follow?

5. Consider the Monty Hall problem, except that Monty enjoys opening door 2 more than he enjoys opening door 3, and if he has a choice between opening these two doors, he opens door 2 with probability p, where $\frac{1}{2} \leq p \leq 1$. To recap: there are three doors, behind one of which there is a car (which you want), and behind the other two of which there are goats (which you dont want). Initially, all possibilities are equally likely for where the car is. You choose a door, which for concreteness we assume is door 1.

(a) Find the unconditional probability that the strategy of always switching succeeds (unconditional in the sense that we do not condition on which of doors 2 or 3 Monty opens).

(b) Find the probability that the strategy of always switching succeeds, given that Monty opens door 2.

(c) Find the probability that the strategy of always switching succeeds, given that Monty opens door 3.

6. *A/B testing* is a form of randomized experiment that is used by many companies to learn about how customers will react to different treatments. For example, a company may want to see how users will respond to a new feature on their website (compared with how users respond to the current version of the website) or compare two different advertisements.
As the name suggests, two different treatments, Treatment A and Treatment B, are being studied. Users arrive one by one, and upon arrival are randomly assigned to one of the two treatments. The trial for each user is classified as "success" (e.g., the user made a purchase) or "failure". The probability that the nth user receives Treatment A is allowed to depend on the outcomes for the previous users. This set-up is known as a *two-armed bandit*.
Many algorithms for how to randomize the treatment assignments have been studied. Here is an especially simple (but fickle) algorithm, called a "stay-with-a-winner" procedure:

(i) Randomly assign the first user to Treatment A or Treatment B, with equal probabilities.

(ii) If the trial for the nth user is a success, stay with the same treatment for the $(n+1)$st user; otherwise, switch to the other treatment for the $(n+1)$st user.

Let a be the probability of success for Treatment A, and b be the probability of success for Treatment B. Assume that $a \neq b$, but that a and b are unknown (which is why the test is needed). Let p_n be the probability of success on the nth trial and a_n be the probability that Treatment A is assigned on the nth trial (using the above algorithm).

(a) Show that
$$p_n = (a-b)a_n + b, a_{n+1} = (a+b-1)a_n + 1 - b$$

(b) Use the results from (a) to show that p_{n+1} satisfies the following recursive equation:
$$p_{n+1} = (a+b-1)p_n + a + b - 2ab$$

(c) Use the result from (b) to find the long-run probability of success for this algorithm, $\lim_{n\to\infty} p_n$, assuming that this limit exists.

7. (**Optional: Challenging Problem**)

(a) An event E_{n+1} is mutually independent of the set of events $E_1, \ldots, E_n$ if for any subset $I \subseteq [1, n]$

$$P\left(E_{n+1} | \bigcap_{j \in I} E_j\right) = P(E_{n+1}).$$

(b) A dependence graph for the set of events $E_1, \ldots, E_n$ is a graph $G = (V, E)$ such that $V = \{1, \ldots, n\}$, and for $i = 1, \ldots, n$, event E_i is mutually independent of the events $\{E_j | (i, j) \notin E\}$.

(c) Assume there exist real numbers $x_1, \ldots, x_n \in [0, 1]$ such that, for any i $(1 \leq i \leq n)$,

$$P(E_i) \leq x_i \prod_{j:(i,j) \in E} (1 - x_j).$$

Then show the following inequality hold:

$$P\left(\bigcap_{i=1}^{n} E_i^c\right) \geq \prod_{i=1}^{n} (1 - x_i).$$

(d) Find the possible applications of the above inequality in the field of EECS.