BÀI TẬP LỚN

Môn: Xác suất thống kê.

Hạn nộp: Đầu tiết học thứ 5 ngày 23/11/2023

Bài 1 2,5 điểm Trên 1 bảng quảng cáo người ta mắc 2 hệ thống bóng đèn. Hệ thống I gồm 3 bóng đèn mắc nối tiếp, hệ thống II gồm 3 bóng đèn mắc song song. Khả năng bị hỏng của mỗi bóng đèn sau 60 giờ thắp sáng liên tục là 10%, việc hỏng bóng coi như độc lập.

a) Tìm xác suất để có ít nhất một hệ thống bị hỏng.

b) Biết rằng có đúng một hệ thống bị hỏng, tính xác suất để đó là hệ thống I.

Bài 2 2,5 điểm Một khu dân cư có tỉ lệ người mắc bệnh A là 20%.

a) Chọn ngẫu nhiên 15 người trong khu dân cư nói trên. Giả sử nhiều khả năng nhất có m người mắc bệnh A. Tìm m và tính xác suất của biến cố "có đúng m người mắc bệnh A".

b) Được biết 90% dân số trong khu dân cư đó đã tiêm phòng bệnh A. Tỷ lệ không mắc bệnh A đối với người đã tiêm phòng là 95% và đối với người chưa tiêm phòng là 30%. Chọn ngẫu nhiên 2 người trong khu dân cư đó. Tính xác suất để có ít nhất 1 người mắc bệnh A.

Bài 3 3,0 điểm Cho X, Y là các biến ngẫu nhiên liên tục có hàm mật độ xác suất đồng thời

$$f_{X,Y}(x,y) = \begin{cases} ke^{-2x-y} & \text{nếu } 0 < y < x \\ 0 & \text{nếu ngược lại} \end{cases}$$

a) Tìm k và các hàm mật độ xác suất của X, Y.

b) Tính $E(X|Y = y)$.

c) Tính $P(X < 1, Y < 1)$ và $P(Y > 2)$.

Bài 4 2,0 điểm Đọc bài viết bên dưới rồi làm bài tập sau (có thể viết bằng Tiếng Việt hoặc Tiếng Anh).

Suppose that a Bayesian spam filter is trained on a set of 10,000 spam messages and 5000 messages that are not spam. The word "enhancement" appears in 1500 spam messages and 20 messages that are not spam, while the word "herbal" appears in 800 spam messages and 200 messages that are not spam. Estimate the probability that a received message containing both the words "enhancement" and "herbal" is spam. Will the message be rejected as spam if the threshold for rejecting spam is 0.9?

Bayesian Spam Filters Most electronic mailboxes receive a flood of unwanted and unsolicited messages, known as spam. Because spam threatens to overwhelm electronic mail systems, a tremendous amount of work has been devoted to filtering it out. Some of the first tools developed for eliminating spam were based on Bayes' theorem, such as Bayesian spam filters. A Bayesian spam filter uses information about previously seen e-mail messages to guess whether an incoming e-mail message is spam. Bayesian spam filters look for occurrences of particular words in messages. For

a particular word w, the probability that w appears in a spam e-mail message is estimated by determining the number of times w appears in a message from a large set of messages known to be spam and the number of times it appears in a large set of messages known not to be spam. When we examine e-mail messages to determine whether they might be spam, we look at words that might be indicators of spam, such as "offer," "special," or "opportunity," as well as words that might indicate that a message is not spam, such as "mom," "lunch," or "Jan" (where Jan is one of your friends). Unfortunately, spam filters sometimes fail to identify a spam message as spam; this is called a false negative. And they sometimes identify a message that is not spam as spam; this is called a false positive. When testing for spam, it is important to minimize false positives, because filtering out wanted e-mail is much worse than letting some spam through.

We will develop some basic Bayesian spam filters. First, suppose we have a set B of messages known to be spam and a set G of messages known not to be spam. (For example, users could classify messages as spam when they examine them in their inboxes.) We next identify the words that occur in B and in G. We count the number of messages in the set containing each word to find $n_B(w)$ and $n_G(w)$, the number of messages containing the word w in the sets B and G, respectively. Then, the empirical probability that a spam message contains the word w is $p(w) = \dfrac{n_B(w)}{|B|}$, and the empirical probability that a message that is not spam contains the word w is $q(w) = \dfrac{n_G(w)}{|G|}$. We note that $p(w)$ and $q(w)$ estimate the probabilities that an incoming spam message, and an incoming message that is not spam, contain the word w, respectively.

Now suppose we receive a new e-mail message containing the word w. Let S be the event that the message is spam. Let E be the event that the message contains the word w. The events S , that the message is spam, and $\overline{S}$, that the message is not spam, partition the set of all messages. Hence, by Bayes' theorem, the probability that the message is spam, given that it contains the word w, is

$$P(S|E) = \frac{P(S)P(E|S)}{P(S)P(E|S) + P(\overline{S})P(E|\overline{S})}$$

To apply this formula, we first estimate $P(S)$, the probability that an incoming message is spam, as well as $p(\overline{S})$, the probability that the incoming message is not spam. Without prior knowledge about the likelihood that an incoming message is spam, for simplicity we assume that the message is equally likely to be spam as it is not to be spam. That is, we assume that $P(S) = P(\overline{S}) = 0.5$. Using this assumption, we find that the probability that a message is spam, given that it contains the word w, is

$$P(S|E) = \frac{P(E|S)}{P(E|S) + P(E|\overline{S})}$$

(Note that if we have some empirical data about the ratio of spam messages to messages that are not spam, we can change this assumption to produce a better estimate for $P(S)$ and for $P(\overline{S})$).

Next, we estimate $P(E|S)$, the conditional probability that the message contains the word w given that the message is spam, by $p(w)$. Similarly, we estimate $P(E|\overline{S})$, the conditional probability that the message contains the word w, given that the message is not spam, by $q(w)$. Inserting these estimates for $P(E|S)$ and $P(E|\overline{S})$ tells us that $P(S|E)$ can be estimated by

$$r(w) = \frac{p(w)}{p(w) + q(w)}$$

that is, $r(w)$ estimates the probability that the message is spam, given that it contains the word

w. If $r(w)$ is greater than a threshold that we set, such as 0.9, then we classify the message as spam.

Detecting spam based on the presence of a single word can lead to excessive false positives and false negatives. Consequently, spam filters look at the presence of multiple words. For example, suppose that the message contains the words w_1 and w_2. Let E_1 and E_2 denote the events that the message contains the words w_1 and w_2, respectively. To make our computations simpler, we assume that E_1 and E_2 are independent events and that $E_1|S$ and $E_2|S$ are independent events and that we have no prior knowledge regarding whether or not the message is spam. (The assumptions that E_1 and E_2 are independent and that $E_1|S$ and $E_2|S$ are independent may introduce some error into our computations; we assume that this error is small.) Using Bayes'theorem and our assumptions, we can show that $P(S|E_1 \cap E_2)$, the probability that the message is spam given that it contains both w_1 and w_2, is

$$P(S|E_1 \cap E_2) = \frac{P(E_1|S)P(E_2|S)}{P(E_1|S)P(E_2|S) + P(E_1|\overline{S})P(E_2|\overline{S})}$$

We estimate the probability $P(S|E_1 \cap E_2)$ by

$$r(w_1, w_2) = \frac{p(w_1)p(w_2)}{p(w_1)p(w_2) + q(w_1)q(w_2)}$$

That is, $r(w_1, w_2)$ estimates the probability that the message is spam, given that it contains the words w_1 and w_2. When $r(w_1, w_2)$ is greater than a preset threshold, such as 0.9, we determine that the message is likely spam.

The more words we use to estimate the probability that an incoming mail message is spam, the better is our chance that we correctly determine whether it is spam.