CS2402 Assignment 2

Lecturer: Kede Ma & Shuaicheng Li **Due on Apr 15, 2025 23:59**

You should submit your solution to canvas before the deadline. Late submission halves the score.

For all the problems in this* assignment, *you SHOULD present your solution step by step instead of giving the answers only.

Question 1

A certain disease affects 1% of the population. A medical test for this disease has 99% sensitivity (probability of testing positive if you have the disease) and 98% specificity (probability of testing negative if you don't have the disease).

(a) What is the probability that a randomly selected person tests positive?

(b) If a person tests positive, what is the probability they actually have the disease?

(c) If a person tests negative, what is the probability they do not have the disease?

Question 2

In a sequence of Bernoulli trials with probability p of success, find the probability that there are exactly d-consecutive successes and $N - d$ fails in a sequence of N independent trials. For example, if N=3, d=2, the sets of sequences with exactly d-consecutive successes are {ssf, fss}.

Question 3

A standard deck of 52 cards is shuffled. Consider the following events:

A: The card is a heart.

B: The card is a face card (Jack, Queen, King).

C: The card is red (hearts or diamonds).

(a) Are events A and B independent?

(b) Given that event C (red card) has occurred, are A and B now conditionally independent (P(A|C) and P(B|C))?

Question 4

In a diverse community, 85% of left-handed individuals are athletes, and 80% of right-handed individuals are athletes.

(a) Suppose the fraction of left-handed individuals in the community is q, i.e., $P(\text{left\_handed}) = q$. Show that the overall proportion of athletes is between 0.8 and 0.85.

(b) If the proportion of athletes is 83%, determine the ratio of left-handed to right-handed individuals.

(c) Suppose there are twice as many right-handed individuals as left-handed individuals in the community. Show that the overall proportion of athletes is at least 81%.

Question 5

Let $\theta > 0$ and $0 < x < 1$, consider the probability density function $f(x|\theta) = \theta x^{\theta-1}$. Based on a random sample X_1, X_2,..., X_n drawn from this distribution, find the maximum likelihood estimator for θ. (*Hint: Start with the likelihood function and simplify the log-likelihood.*)

Question 6

Suppose that the birthday of each of three people is equally likely to be any one of the 365 days of the year, independently of the others. Let B_{ij} denote the event that person *i* has the same birthday as person *j*.

(a) Are the events B_{12} and B_{23} independent? Why?

(b) Are the events B_{12} and B_{13} independent? Why?

(c) Are the events B_{12}, B_{13} and B_{23} independent? Why?

Question 7

Consider rolling two fair six-sided dice. Let *X* be the random variable representing the **minimum of the two dice**. For example, when the two dice are (3,4), then *X=3*.

(a) Derive P(*X*=*x*) (probability mass function of *X*) in **analytic form** (a small fraction of the marks will be deducted if you only list all possible values of *X* and their corresponding probabilities).

(b) Calculate E[*X*] and Var[*X*].

Question 8

The table below shows some data from the early days of a clothing company. Each row shows the sales for a year, and the amount spent on advertising in that year.

Year	Advertising (Million Dollars)	Sales (Million Dollars)
1	22	650
2	31	855
3	33	1064

4	44	1191
5	51	1420

(a) Use linear regression to calculate the relationship between the advertising and sales (slop, and intercept).

(b) Calculate the correlation coefficient between the amounts of advertising and sales.

(c) If there is no advertising, what is the expected sales? If the amount of advertising is 58 million dollars in the next year, please predict the sales.

Question 9

A producer of a certain type of electronic component ships to suppliers in lots, each lot has twenty components. Suppose that 60% of all such lots contain no defective components, 30% contain one defective component, and 10% contain two defective components. A lot is picked, two components from the lot are randomly selected and tested, and neither is defective.

(a) What is the probability that zero defective components exist in the lot?

(b) What is the probability that one defect exists in the lot?

(c) What is the probability that two defects exist in the lot?

Question 10

You need to implement a python program detector.py (included in assignment2 folder) to solve the problem. However, you only need to include the solutions with blanks filled out in your submitted PDF file.

Use Bayes' rule to detect the spam email. The prior probability of spam emails are 80%.

The key words are listed in the following table

Word	**P(word | spam)**	**P(word | ¬spam)**
$	88%	47%
donate	66%	10%
research	11%	60%
contact	51%	51%
CS2402	0.5%	40%

The two emails are:

Email I.

Dear research student, we have just uploaded the competition questions for CS2402. Please sign up and contact me before this Friday (31, March, 2023) if you have any questions. The award for the winner is $10,000,000! Also, you can donate your award to the whole class.

Email II.

I have decided to donate what I have to you. I was diagnosed with cancer of the lungs few years ago. I have been inspired by God to donate my inheritance to you for the good work of God and charity purpose. I am doing this because my family are unbelievers and I will not allow them inherit this money for their own selfishness. I decided to bequeath the sum of $10,000,000.00 to you. If you are much more interested, Contact Thomas with this specified email: thmasbfd@gmail.com).

Python sketch codes for detecting spam email are provided in "detector.py". You need to implement Bayes rule to compute the probability of spam email for the given two emails.

1) **Please fill in the blanks and execute "detector.py" with the two exemplified emails, respectively.**
2) **Please complete the tables.**

Email I:

Hypothesis	**Prior odds**	**Likelihood ratio**	**Posterior odds**	**P(spam | email)**
spam	0.8			
¬spam	0.2			

Email II:

Hypothesis	**Prior odds**	**Likelihood ratio**	**Posterior odds**	**P(spam | email)**
spam	0.8			
¬spam	0.2			

```
# Input the email with your keyboard #
email = input('Please enter the test Email: ')     # gets the test email
email = email.lower()                              # convert string to lowercase
print(email)

# Input the spam detection keyword #
```

```
word1 = input('Please enter the first key word: ') # gets the first keyword
word1 = word1.lower()                  # convert the first key word to lowercase

word2 = ____________                    # gets the second keyword
word2 = word2.lower()                  # convert the second keyword to lowercase

word3 = ____________                    # gets the third keyword
word3 = word3.lower()                  # convert the third keyword to lowercase

word4 = ____________                    # gets the fourth keyword
word4 = word4.lower()                  # convert the fourth keyword to lowercase

word5 = ____________                    # gets the fifth keyword
word5 = word5.lower()                  # convert the fifth keyword to lowercase

# Detect whether the key word occured in your email #
n_word1 = email.count(word1)  # count the number of the first keyword in the email
n_word2 = email.count(word2)  # count the number of the second keyword in the email
n_word3 = email.count(word3)  # count the number of the third keyword in the email
n_word4 = email.count(word4)  # count the number of the fourth keyword in the email
n_word5 = email.count(word5)  # count the number of the fifth keyword in the email

# The prior odds of spam is 80% #
p_spam = 1.0                          # initialize the likelihood ratio of the spam email
p_no_spam = 1.0                         # initialize the likelihood ratio of the no spam email
prior_odds_spam = ____________                 # prior odds of spam

# Calculate the Likelihood ratio #
if n_word1 != 0:                      # the first keyword occurred in the email
   p_spam = p_spam * 0.88
   p_no_spam = p_no_spam * 0.47
if n_word2 != 0:                      # the second keyword occurred in the email
   p_spam = ____________
   p_no_spam = ____________
if n_word3 != 0:                      # the third keyword occurred in the email
   p_spam = ____________
   p_no_spam = ____________
if n_word4 != 0:                      # the fourth keyword occurred in the email
   p_spam = ____________
   p_no_spam = ____________
if n_word5 != 0:                      # the fifth keyword occurred in the email
   p_spam = ____________
   p_no_spam = ____________

# Calculate the Posterior odds #
posterior_odds_spam = ____________               # posterior odds of spam
posterior_odds_no_spam = ____________              # posterior odds of no spam
```

```
p_isSpam = ____________                    # probability of spam P(spam | email)

print("p_spam: %.6f, p_no_spam: %.6f, posterior_odds_spam: %.6f, posterior_odds_no_spam: %.6f,
p_isSpam: %.6f"%(p_spam, p_no_spam, posterior_odds_spam, posterior_odds_no_spam, p_isSpam))
```