CS2402 Assignment 2

Lecturer: Kede Ma & Shuaicheng Li **Due on Apr 19, 2024 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 (10pts)

Independent random variables X and Y follow distributions $\mathcal{N}(4,11)$ and $\mathcal{N}(6,14)$, respectively. Show that the probability that $(X-Y)^2$ has a value between 4 and 16 is approximately 0.25.

SOLUTIONS:

Let $Z = X - Y$, $4 \le (X-Y)^2 \le 16$, then $2 \le Z \le 4$ or $-4 \le Z \le -2$

$$E(Z) = E(X-Y) = E(X) - E(Y) = -2$$

$$Var(Z) = Var(X-Y) = Var(X) + Var(Y) = 25$$

$$P(2 \le Z \le 4) = P\left(\frac{2+2}{5} \le \frac{Z-(-2)}{5} \le \frac{4+2}{5}\right) = F(1.2) - F(0.8) = 0.0968$$

$$P(-4 \le Z \le -2) = P\left(\frac{-4+2}{5} \le \frac{Z-(-2)}{5} \le \frac{-2+2}{5}\right) = F(0) - F(-0.4) = F(0.4) - F(0)$$
$$= 0.1554$$

$$P(2 \le Z \le 4) + P(-4 \le Z \le -2) \approx 0.25$$

Question 2 (10pts)

In a sequence of Bernoulli trials with probability p of success, find the probability that a successes will occur before the b^{th} failures. (Suppose $a = 2$, $b = 3$, s represents success and f represents failures, then the events of 2 successes will occur before the 3^{rd} failures include: $ssfff$, $sfsff$, $sffsf$, $fssff$, $fsfsf$, $ffssf$.)

SOLUTIONS:

Supposes the event is A. There are a successes with $b-1$ failures and obtain an additional failure (the b^{th})

$$P(A) = \binom{a+b-1}{a} p^a (1-p)^{b-1} \cdot (1-p) = \binom{a+b-1}{a} p^a (1-p)^b$$

Question 3 (10pts)

(a) The random variable X is defined as the larger value obtained in two throws of an unbiased six-sided die. Show that the probability massive function is

$$p(x) = \frac{2x-1}{36}, x = 1,2,\ldots,6.$$

(b) The random variable Y represents the highest value obtained in k throws of an unbiased six-sided die. Find an expression for the probability massive function of Y.

[Hint: find the distribution first.]

SOLUTIONS:

a) **(5pts)**

$$X = \max(X_1, X_2)$$

$$P(X \le 1) = P(X_1 \le 1, X_2 \le 1) = (\frac{1}{6})^2$$

$$P(X \le 2) = P(X_1 \le 2, X_2 \le 2) = (\frac{2}{6})^2$$

$$P(X \le 3) = P(X_1 \le 3, X_2 \le 3) = (\frac{3}{6})^2$$

$$P(X \le 4) = P(X_1 \le 4, X_2 \le 4) = (\frac{4}{6})^2$$

$$P(X \le 5) = P(X_1 \le 5, X_2 \le 5) = (\frac{5}{6})^2$$

$$P(X \le 6) = P(X_1 \le 6, X_2 \le 6) = (\frac{6}{6})^2$$

$$P(X = 1) = P(X \le 1) = \frac{1}{36}$$

$$P(X = 2) = P(X \le 2) - P(X \le 1) = \frac{3}{36}$$

$$P(X = 3) = P(X \le 3) - P(X \le 2) = \frac{5}{36}$$

$$P(X = 4) = P(X \le 4) - P(X \le 3) = \frac{7}{36}$$

$$P(X = 5) = P(X \le 5) - P(X \le 4) = \frac{9}{36}$$

$$P(X = 6) = P(X \le 6) - P(X \le 5) = \frac{11}{36}$$

Thus

$$p(x) = \frac{2x-1}{36}, x = 1,2,\dots,6.$$

b) **(5pts)**

$$Y = \max\,(X_1, \dots, X_k)$$

$$P(Y \le 1) = P(X_1 \le 1, \dots, X_k \le 1) = (\frac{1}{6})^k$$

$$P(Y \le 2) = P(X_1 \le 2, \dots, X_k \le 2) = (\frac{2}{6})^k$$

$$P(Y \le 3) = P(X_1 \le 3, \dots, X_k \le 3) = (\frac{3}{6})^k$$

$$P(Y \le 4) = P(X_1 \le 4, \dots, X_k \le 4) = (\frac{4}{6})^k$$

$$P(Y \le 5) = P(X_1 \le 5, \dots, X_k \le 5) = (\frac{5}{6})^k$$

$$P(Y \le 6) = P(X_1 \le 6, \dots, X_k \le 6) = (\frac{6}{6})^k$$

$$P(Y = 1) = P(Y \le 1) = (\frac{1}{6})^k$$

$$P(Y = 2) = P(Y \le 2) - P(Y \le 1) = (\frac{2}{6})^k - (\frac{1}{6})^k$$

$$P(Y = 3) = P(Y \le 3) - P(Y \le 2) = (\frac{3}{6})^k - (\frac{2}{6})^k$$

$$P(Y = 4) = P(Y \le 4) - P(Y \le 3) = (\frac{4}{6})^k - (\frac{3}{6})^k$$

$$P(Y = 5) = P(Y \le 5) - P(Y \le 4) = (\frac{5}{6})^k - (\frac{4}{6})^k$$

$$P(Y = 6) = P(Y \le 6) - P(Y \le 5) = (\frac{6}{6})^k - (\frac{5}{6})^k$$

Question 4(10pts)

Consider a small town where each family can have 0, 1 or 2 children. The probability of gender is equal and independent. Let the probability of a family of zero or one child be r, and the probability of having a second child be $1 - r$. Calculate the following probabilities:

(a) P(a family has two girls).

(b) P(the elder child is a boy and the younger child is a girl).

(c) P(a family has at least one boy).

Please show the problem solving process and answers.

SOLUTIONS:

a) (3pts)

Since the genders of the two children are independent, we can multiply the probabilities of each event. The probability that each child is a girl is $\frac{1}{2}$. Therefore, the probability that both children are girls is

$$\mathrm{P(has\ two\ girls)} = \mathrm{P(first\ child\ is\ a\ girl)} \times \mathrm{P(second\ child\ is\ a\ girl\ |\ second\ child\ exists)} \times \mathrm{P(second\ child\ exists)}$$

$$= \frac{1}{2} \times \frac{1}{2} \times (1 - r) = \frac{1}{4} \times (1 - r)$$

b) (3pts)

$$\mathrm{P(the\ elder\ child\ is\ a\ boy\ and\ the\ younger\ child\ is\ a\ girl)} = \mathrm{P(elder\ child\ is\ a\ boy)} \times \mathrm{P(younger\ child\ is\ a\ girl\ |\ second\ child\ exists)} \times \mathrm{P(second\ child\ exists)} = \frac{1}{2} \times \frac{1}{2} \times (1 - r) = \frac{1}{4} \times (1 - r)$$

c) **(4pts)**

We use **s** to denote the probability of a family with one child, (r-s) to denote the probability of a family with no children, and 1-r to denote the probability of a family with two children. Having at least one boy means that we are looking for a complementary set of a set that with no boys at all.

$$\mathrm{P(no\ boys)} = \mathrm{P(no\ child)} + \mathrm{P(only\ one\ child, the\ child\ is\ a\ girl)} + \mathrm{P(both\ children\ are\ girls)}$$

$$= (\mathrm{r} - \mathrm{s}) + \mathrm{s} \times \frac{1}{2} + \frac{1}{4} \times (1 - \mathrm{r})$$

$$= (\mathrm{r} - \mathrm{s}) + \frac{\mathrm{s}}{2} + \frac{1 - \mathrm{r}}{4}$$

$$= \frac{3\mathrm{r} - 2\mathrm{s} + 1}{4}$$

$$\mathrm{P(the\ family\ has\ at\ least\ one\ boy)} = 1 - \mathrm{P(no\ boys)}$$

$$= 1 - \frac{3r - 2s + 1}{4}$$

$$= \frac{3 - 3r + 2s}{4}$$

The variable `s` can be assumed to take any reasonable value, such as $\frac{1}{2}r, \frac{1}{3}r$, or r. For instance, if we do not consider families with no children, and let s = r, then the answer is $\frac{3-r}{4}$. If we assume that families with no children and families with one child each comprise $\frac{1}{2}r$, then the answer is $\frac{r}{4} + \frac{3}{4}(1 - r)$.

Question 5 (10pts)

Suppose a diagnostic test for a rare disease has the following characteristics:

- The probability that a person has the disease is 1% (prior probability).
- The probability that the test correctly identifies a sick person as having the disease is 99%.
- The probability that the test correctly identifies a healthy person as not having the disease is 95%.

(a) Fill in the table below with the probability (expressed as a percentage) of each situation:

	Disease Present (D+)	**Disease Absent (D-)**
Test Positive (T+)	TP: ____	FP: ____
Test Negative (T-)	FN: ____	TN: ____

(b) Calculate the posterior probability that a patient actually has the disease when given the positive test result.

SOLUTION:

a) (4pts)

	Disease Present (D+)	**Disease Absent (D-)**
Test Positive (T+)	TP: 99% (Sensitivity)	FP: 5%
Test Negative (T-)	FN: 1%	TN: 95% (Specificity)

b) **(6pts)**

$$\mathrm{P(Disease|TestPositive)} = \frac{\mathrm{P(TestPositive|Disease)} \times \mathrm{P(Disease)}}{\mathrm{P(TestPositive)}}$$

The total probability of a positive test is:

$$\begin{aligned}\mathrm{P(TestPositive)} &= \mathrm{P(TestPositive|Disease)} \times \mathrm{P(Disease)} \\ &\quad + \mathrm{P(TestPositive|NoDisease)} \times \mathrm{P(NoDisease)}\end{aligned}$$

$$= (0.99 \times 0.01) + (0.05 \times 0.99)$$

Now calculate the posterior probability:

$$\mathrm{P(Disease|TestPositive)} = \frac{(0.99 \times 0.01)}{(0.99 \times 0.01) + (0.05 \times 0.99)}$$

$$= \frac{0.0099}{0.0099 + 0.0495}$$

$$= \frac{0.0099}{0.0594} = \frac{1}{6}$$

$$= \ 0.1666 \text{ or } 16.66\%$$

Question 6 (10pts)

Suppose you have two fair coins and you flip them both. Let X be the number of heads that show. Then, you grab a standard deck of 52 cards and draw X cards with replacement. Let Y be the number of aces you draw.

(a) Find $P(Y = 1|X = 2)$.

(b) Find $P(Y = 0)$, $P(Y = 1)$, and $P(Y = 2)$.

(c) Find $P(X = 2|Y = 1)$.

SOLUTION:

(a) (3pts)

Each card drawn has a probability $1/13$ of being an ace because there are 13 possible ranks $A, 2, 3, \ldots, 10, J, Q, K$, and each draw is independent of the others because of the replacement.

Therefore, $Y \sim Binomial(X, 1/13)$, and $P(Y = 1|X = 2) = \binom{2}{1}\left(\frac{1}{13}\right)^{1}\left(1 - \frac{1}{13}\right)^{2-1} = \frac{24}{169}$.

(b) (4pts)

Each coin has a probability $1/2$ of being heads, independently of each other, so $X \sim Binomial(2, 1/2)$. Using the fact, alone with the law of total probability and the fact that $Y \sim Binomial(X, 1/13)$, we obtain for $i = 0, 1, 2$:

$$P(Y = i) = P(Y = i|X = 0)P(X = 0) + P(Y = i|X = 1)P(X = 1) + P(Y = i|X = 2)P(X = 2)$$

$$P(Y = 0) = 1 \times \left[\binom{2}{0}\left(\frac{1}{2}\right)^{0}\left(1 - \frac{1}{2}\right)^{2-0}\right] + \left[\binom{1}{0}\left(\frac{1}{13}\right)^{0}\left(1 - \frac{1}{13}\right)^{1-0}\right] \times \left[\binom{2}{1}\left(\frac{1}{2}\right)^{1}\left(1 - \frac{1}{2}\right)^{2-1}\right]$$
$$+ \left[\binom{2}{0}\left(\frac{1}{13}\right)^{0}\left(1 - \frac{1}{13}\right)^{2-0}\right] \times \left[\binom{2}{2}\left(\frac{1}{2}\right)^{2}\left(1 - \frac{1}{2}\right)^{2-2}\right] = \frac{625}{676}$$

$$P(Y = 1) = 0 \times \left[\binom{2}{0}\left(\frac{1}{2}\right)^{0}\left(1 - \frac{1}{2}\right)^{2-0}\right] + \left[\binom{1}{1}\left(\frac{1}{13}\right)^{1}\left(1 - \frac{1}{13}\right)^{1-1}\right] \times \left[\binom{2}{1}\left(\frac{1}{2}\right)^{1}\left(1 - \frac{1}{2}\right)^{2-1}\right]$$
$$+ \left[\binom{2}{1}\left(\frac{1}{13}\right)^{1}\left(1 - \frac{1}{13}\right)^{2-1}\right] \times \left[\binom{2}{2}\left(\frac{1}{2}\right)^{2}\left(1 - \frac{1}{2}\right)^{2-2}\right] = \frac{50}{676}$$

$$P(Y = 2) = 0 \times \left[\binom{2}{0}\left(\frac{1}{2}\right)^{0}\left(1 - \frac{1}{2}\right)^{2-0}\right] + 0 \times \left[\binom{2}{1}\left(\frac{1}{2}\right)^{1}\left(1 - \frac{1}{2}\right)^{2-1}\right]$$
$$+ \left[\binom{2}{2}\left(\frac{1}{13}\right)^{2}\left(1 - \frac{1}{13}\right)^{2-2}\right] \times \left[\binom{2}{2}\left(\frac{1}{2}\right)^{2}\left(1 - \frac{1}{2}\right)^{2-2}\right] = \frac{1}{676}$$

(c) (3pts)

Using Bayes' theorem,

$$P(X=2|Y=1)=\frac{P(Y=1|X=2)P(X=2)}{P(Y=1)}=\frac{\left[\binom{2}{1}\left(\frac{1}{13}\right)^{1}\left(1-\frac{1}{13}\right)^{2-1}\right]\times\left[\binom{2}{2}\left(\frac{1}{2}\right)^{2}\left(1-\frac{1}{2}\right)^{2-2}\right]}{\frac{50}{676}}$$

$$=\frac{16224}{33800}=\frac{12}{25}$$

Question 7 (10pts)

A producer of a certain type of electronic component ships to suppliers in lots of twenty. 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 defective exists in the lot?

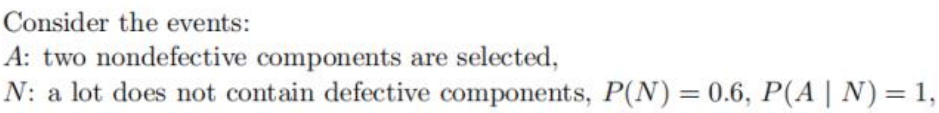
(c) What is the probability that two defectives exist in the lot?

SOLUTIONS:

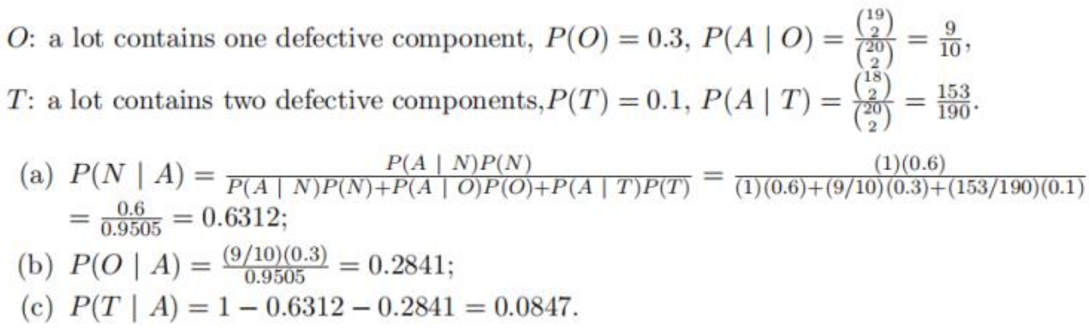
a) (4pts)

b) (3pts)

c) (3pts)



Consider the events:
A: two nondefective components are selected,
N: a lot does not contain defective components, $P(N)=0.6$, $P(A \mid N)=1$,



O: a lot contains one defective component, $P(O)=0.3$, $P(A \mid O)=\frac{\binom{19}{2}}{\binom{20}{2}}=\frac{9}{10}$,

T: a lot contains two defective components,$P(T)=0.1$, $P(A \mid T)=\frac{\binom{18}{2}}{\binom{20}{2}}=\frac{153}{190}$.

(a) $P(N \mid A)=\frac{P(A \mid N)P(N)}{P(A \mid N)P(N)+P(A \mid O)P(O)+P(A \mid T)P(T)}=\frac{(1)(0.6)}{(1)(0.6)+(9/10)(0.3)+(153/190)(0.1)}$
$=\frac{0.6}{0.9505}=0.6312$;

(b) $P(O \mid A)=\frac{(9/10)(0.3)}{0.9505}=0.2841$;

(c) $P(T \mid A)=1-0.6312-0.2841=0.0847$.

Question 8 (10pts)

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

SOLUTIONS:

(a) **(4pts)** Use linear regression to calculate the relationship between the advertising and sales (slop, and intercept).

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 36.2 \quad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = 1036$$

x	y	x-x̄	y-ȳ	(x-x̄)*(y-ȳ)
22	650	-14.2	-386	5481.2
31	855	-5.2	-181	941.2
33	1064	-3.2	28	-89.6
44	1191	7.8	155	1209
51	1420	14.8	384	5683.2

$S_{xx} = 518.8,\ S_{xy} = 13225,$

$\beta = \frac{S_{xy}}{S_{xx}} = 25.49$ and $\alpha = \bar{y} - \beta\bar{x} = 113.2.$

Then the linear regression equation is y = 25.49x + 113.2, which describes the relationship between advertising (x) and sales (y).

(b) **(3pts)** Calculate the correlation coefficient between the amounts of advertising and sales.

$S_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = 11.389$ and $S_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2} = 297.499$.

The correlation coefficient $r = \frac{1}{n-1}\sum_{i=1}^{n}(\frac{x_i-\bar{x}}{S_x})(\frac{y_i-\bar{y}}{S_y})$ = 0.976.

(c) **(3pts)** 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.

If there is no advertising, x=0, the expected sales y= 25.49*0 + 113.2 = 113.2,

if the amount of advertising x=58, then the sales will be y =25.49*58 + 113.2 = 1591.74.

Question 9 (10pts)

Let $\theta > 0$, and consider the p.d.f. $p(x) = \theta e^{-\theta x}, x \geq 0$. Based on a random sample $X_1, ..., X_n$ from this distribution, find the maximum likelihood estimator for θ.

Hint: $a \ln \theta - b\theta$ achieves its maximum when $\theta = \frac{a}{b}$.

SOLUTIONS:

Log likelihood:

$$\sum_{i=1}^{n}(\ln\theta - \theta X_i) = n\ln\theta - \theta\sum_{i=1}^{n}X_i$$

It achieves its maximum when $\hat{\theta} = \frac{n}{\sum_{i=1}^{n}X_i}$.

Question 10 (10pts) Implementation

You need to implement a python program detector.py (included in assignment2 fold) 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 (29, March, 2024) 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 is provided in "detector.py". You need to implement Bayes rule to compute the probability of spam email for the given two emails.

(a) **(6pts)** Please fill in the blanks and execute "detector.py" with the two exemplified emails, respectively.

(b) **(4pts)** Please complete the tables.

SOLUTIONS:

$$W = \{word1, word2, word3, word4, word5\}$$

$$P(W \mid spam) = \prod_{w \in W} P(w \mid spam), P(W \mid \neg spam) = \prod_{w \in W} P(w \mid \neg spam)$$

$$\begin{aligned} P(W) &= P(W, spam) + P(W, \neg spam) \\ &= P(W \mid spam)P(spam) + P(W \mid \neg spam)P(\neg spam) \end{aligned}$$

$$P(spam \mid W) = \frac{P(W, spam)}{P(W)} = \frac{P(W \mid spam)P(spam)}{P(W)}$$

Email I:

Hypothesis	**Prior odds**	**Likelihood ratio**	**Posterior odds**	**P(spam | email)**
spam	0.8	0.000163	0.000130	0.101751
¬spam	0.2	0.005753	0.001151	

Email II:

Hypothesis	**Prior odds**	**Likelihood ratio**	**Posterior odds**	**P(spam | email)**
spam	0.8	0.296208	0.236966	0.980170
¬spam	0.2	0.023970	0.004794	

```
# 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 = input('Please enter the second key word: ')                          # gets the second
keyword
word2 = word2.lower()                                  # convert the second keyword to lowercase
```

```
word3 = input('Please enter the third key word: ')                    # gets the third keyword
word3 = word3.lower()                           # convert the third keyword to lowercase

word4 = input('Please enter the fourth key word: ')                   # gets the fourth
keyword
word4 = word4.lower()                           # convert the fourth keyword to lowercase

word5 = input('Please enter the fifth key word: ')                    # 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 = 0.8                      # 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_spam * 0.66
    p_no_spam = p_no_spam * 0.10
if n_word3 != 0:                                # the third keyword occurred in the email
    p_spam = p_spam * 0.11
    p_no_spam = p_no_spam * 0.60
if n_word4 != 0:                                # the fourth keyword occurred in the email
    p_spam = p_spam * 0.51
    p_no_spam = p_no_spam * 0.51
if n_word5 != 0:                                # the fifth keyword occurred in the email
    p_spam = p_spam * 0.005
    p_no_spam = p_no_spam * 0.40

# Calculate the Posterior odds #
posterior_odds_spam = p_spam * prior_odds_spam              # posterior odds of spam
posterior_odds_no_spam = p_no_spam * (1-prior_odds_spam)              # posterior odds of no
spam

p_isSpam = posterior_odds_spam / (posterior_odds_spam+posterior_odds_no_spam)
# 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))
```