Question:	1	2	3	4	Total
Points:	20	40	20	20	100
Score:					

SKKU SWE3026_41 Probability and Random Process
2021 Spring Final Exam

Student ID: ______________________________

Name: ______________________________

- Please check if you received all 13 pages of the test material.

  시작하기 전에 반드시 페이지가 총 13쪽인지 확인 하십시오.

- We do this exam online. You should write your answer in the 'Final Exam' page in i-Campus. We will grade your answers in the **i-Campus** only.

  우리는 온라인으로 기말고사를 진행합니다. 문제에 대한 답은 i-Campus 내 'Final Exam' 페이지에 제출하세요. 채점은 **i-Campus**에 적힌 답으로만 진행됩니다.

- **Lecturer and TAs will not answer your questions about the exam.** If you think that there is anything ambiguous, unclear or wrong about a problem, please write the reasons and make necessary assumptions to solve the problem. We will take your explanation into consideration while grading. If you want to get partial points, please upload the explanations (pdf format) to i-Campus.

  시험시간동안 질문을 받지 않습니다. 만일 문제에 오류나 이상이 있을 경우, 왜 문제가 이상이 있다고 생각하는지에 대해 기술하시면 됩니다. 문제가 애매하다고 생각되는 경우 문제를 풀 때 본인이 생각하는 가정을 함께 작성하면 됩니다. 채점 시 가정 및 설명을 고려하겠습니다. 만약 부분 점수를 받고 싶으시다면 pdf 포멧인 설명을 i-Campus에 올려주세요.

- **Write your answer in Python 3.** We will grade your answers based only on Python 3.

  Python 3를 사용해 문제를 해결하세요. 채점은 Python 3 기준으로만 진행됩니다.

1. (20 points) (No partial points) Answer the following questions.

   1-1. (15 points) Answer the questions with "T" if the answer is true or "F" otherwise. Other characters are not accepted.

   1-1-1. (1 point) The instructor name of this class is JinYeong Bak.

   [Answer box]

   Solution: T

   1-1-2. (1 point) Negative rational number is countable infinite set.

   [Answer box]

   Solution: T

   1-1-3. (1 point) Disjoint and independent of events are the same.

   [Answer box]

   Solution: F

   1-1-4. (1 point) Law of the unconscious statistician (LOTUS) for discrete random variables is $E\left[g\left(X\right)\right] = \sum_{x_k \in R_X} g\left(x_k\right) P_X\left(g\left(x_k\right)\right)$.

   [Answer box]

   Solution: F

   1-1-5. (1 point) If X and Y are random variables, then $E[E[X|Y]] = E[Y]$.

   [Answer box]

   Solution: F

1-1-6. (1 point) When we find upper bounds on $P(X \geq \alpha n)$ for $X \sim Binomial(n, p)$, the upper bound from Chebyshev's inequalities is tighter than the upper bound from Chernoff bounds.

[Answer box]

Solution: F

1-1-7. (1 point) The delta function of zero $\delta(0)$ is 0.

[Answer box]

Solution: F

1-1-8. (1 point) If X and Y are independent random variables, then $E[g(X)h(Y)] = g(E[X])h(E[Y])$.

[Answer box]

Solution: F

1-1-9. (1 point) If X and Y are independent random variables and then the moment generating function of a random variables $Z = X + Y$ is $M_Z(s) = M_X(s)M_Y(s)$.

[Answer box]

Solution: T

1-1-10. (1 point) Bayesian statistician think that the unknown quantity is assumed to be a fixed one.

[Answer box]

Solution: F

1-1-11. (1 point) The normalized random variable is as follows: $Z = \frac{\{X-\mu\}}{\sigma}$ when $E[X] = \mu,\ Var[X] = \sigma^2$.

[Answer box]

Solution: T

1-1-12. (1 point) A systematic way of parameter estimation that finds the parameter value that maximizes the posterior distribution is MLE.

[Answer box]

Solution: F

1-1-13. (1 point) We should minimize the cost function to find the best parameters of linear regression.

[Answer box]

Solution: T

1-1-14. (1 point) A random sequence is a random process.

[Answer box]

Solution: T

1-1-15. (1 point) In a Markov chain, X_m random variable depends on X_{m-1} and X_{m-2} that are previous values.

[Answer box]

Solution: F

1-2. (5 points) Choose the best answer to fill in the blank from the options given.

1-2-1. (1 point) We want to model the arrival of customers at a service facility. The most useful random variable is $\boxed{}$.

① Pascal
② Hypergeometric
③ Geometric
④ Poisson

Answer: Poisson

1-2-2. (1 point) Bayes‘ rule is

$$P(B|A) = \frac{P(A|B)P(B)}{\boxed{}}$$

① $P(A)$
② $P(A, B)$
③ $P(B)$
④ $P(A|B)$
⑤ $P(B|A)$
⑥ $P(B, A)$

Answer: $P(A)$

1-2-3. (1 point) We have a set with n elements, and we want to draw k samples from the set such that ordering does not matter and repetition is not allowed. The number of k-element subsets of the set is $\boxed{}$.

① P_k^n
② n^k
③ k^n
④ $\binom{n}{k}$
⑤ $\binom{k}{n}$
⑥ $\binom{n+k-1}{k}$

Answer: n^k

1-2-4. (1 point) If the correlation coefficient of two random variables X and Y is zero, we say that X and Y are $\boxed{}$.

① correlated
② uncorrelated
③ positively correlated
④ negatively correlated
⑤ independent
⑥ dependent

Answer: uncorrelated

1-2-5. (1 point) For a discrete random variable X and event A, the ________ of X given A is defined as

$$\frac{P(X = x_i \text{ and } A)}{P(A)}$$

① marginal PDF
② marginal PMF
③ marginal CDF
④ conditional PDF
⑤ conditional PMF
⑥ conditional CDF

Answer: conditional PMF

2. (40 points) Choose the best answer for the following questions. (Caution) Select the last option 'I do not know' when you do not solve the problem. We will give zero scores if you select the option. If not, we will check your answer and give **negative points if it is incorrect**. For example, when you choose the wrong option except 'I do not know', you will get −6 points for each wrong answer. Please select your answer carefully.

2-1. (6 points) There are 15 people in a party, including Hannah and Sarah. We divide the 15 people into 3 groups, where each group has 5 people. What is the probability that Hannah and Sarah are in the same group?

① 0
② $\frac{1}{3}$
③ $\frac{1}{7}$
④ $\frac{2}{7}$
⑤ $\frac{2}{9}$
⑥ $\frac{5}{9}$
⑦ I do not know

Answer: $\frac{2}{7}$

2-2. (6 points) Let N be the number of phone calls made by the customers of a phone company in a given hour. Suppose that $N \sim Poisson(\beta)$, where $\beta > 0$ is known. Let X_i be the length of the i'th phone call, for $i = 1, 2, \ldots, N$. We assume X_i's are independent of each other and also independent of N. We further assume $X_i \sim Exponential(\lambda)$, where $\lambda > 0$ is known. Let Y be the sum of the lengths of the phone calls, i.e.,

$$Y = \sum_{i=1}^{N} X_i$$

What is $E[Y]$?

Hint: $E[N] = \beta, E[X] = \frac{1}{\lambda}$

① $\beta + \frac{1}{\lambda}$
② $\beta^2 + \frac{1}{\lambda}$
③ $\beta\frac{1}{\lambda}$
④ $\beta\lambda$
⑤ $N\beta\lambda$
⑥ $N\beta\lambda^2$
⑦ I do not know

Answer: $\beta\frac{1}{\lambda}$

2-3. (6 points) In a communication system, each codeword consists of 1000 bits. Due to the noise, each bit may be received in error with probability 0.1. It is assumed bit errors occur independently. Since error correcting codes are used in this system, each codeword can be decoded reliably if there are less than or equal to 125 errors in the received codeword, otherwise the decoding fails. Using the central limit theorem, find the probability of decoding failure.

Hint: $E[X] = p, Var[X] = p(1-p)$ where $X \sim Bernoulli(p)$

$1000 \times \frac{1}{10} \times \frac{9}{10} = 90 \quad \sigma = \sqrt{90}$

$m \quad \frac{x - \mu}{\sqrt{6}}$

① $\Phi(\frac{10}{9})$
② $\Phi(\frac{25}{90})$
③ $\Phi(\frac{25}{\sqrt{90}})$
④ $1-\Phi(\frac{10}{9})$
⑤ $1-\Phi(\frac{25}{90})$
⑥ $1-\Phi(\frac{25}{\sqrt{90}})$
⑦ I do not know

Answer: $1-\Phi(\frac{25}{\sqrt{90}})$

10↑

2-4. (6 points) Suppose we would like to test the hypothesis that at least 10% of students suffer from allergies. We collect a random sample of 225 students and 21 of them suffer from allergies. ~~Compute~~ the P-value of the hypothesis.

① $\Phi(-\frac{1}{2})$
② $\Phi(-\frac{1}{3})$
③ $\Phi(-\frac{1}{4})$
④ $\Phi(\frac{1}{4})$
⑤ $\Phi(\frac{1}{3})$
⑥ $\Phi(\frac{1}{2})$
⑦ I do not know

$225 \times \frac{1}{10} \times \frac{9}{10}$

$\frac{21}{225}$

$21 - 22.5$

1.5

Answer: $\Phi(-\frac{1}{3})$

2-5. (6 points) Consider the Markov chain with three states $S = \{1, 2, 3\}$, that has the state transition matrix as below:

$$P = \begin{bmatrix} 1/4 & 0 & 3/4 \\ 1/2 & 0 & 1/2 \\ 1/2 & 1/4 & 1/4 \end{bmatrix}$$

Suppose $P(X_1 = 1) = 1/2$ and $P(X_1 = 2) = 1/4$. Find $P(X_1 = 3, X_3 = 1)$

① 1/32
② 1/16
③ 3/32
④ 3/16
⑤ 5/32
⑥ 5/16
⑦ I do not know

$\frac{1}{2} \times \frac{1}{4} + \frac{1}{4} \times \frac{1}{2} + \frac{1}{4} \times \frac{1}{2} = \frac{3}{8}$

Answer: 3/32

2-6. (10 points) There are 1000 households in a town. Specifically, there are 100 households with one member, 200 households with 2 members, 300 households with 3 members, 200 households with 4 members, 100 households with 5 members, and 100 households with 6 members.

2-6-1. ~~(5 points) We pick~~ a household at random, and define the random variable X as the number of people in the chosen household. Find the expected value of X.

① 2.2

② 3.3
③ 4.4
④ 5.5
⑤ 6.6
⑥ 7.7
⑦ I do not know

Answer: 3.3

10 × 1 +

1 + 4 + 9 + 8 + 5 + 6 $= \frac{33}{10}$

2-6-2. (5 points) We pick a person in the town at random, and define the random variable Y as the number of people in the household where the chosen person lives. Find the expected value of Y. Please raise decimals to the next whole number. Ex) $9.63 \to 10, 7.21 \to 7$

① 1
② 2
③ 3
④ 4
⑤ 5
⑥ 6
⑦ I do not know

Answer: 4

1 2 3 4 5 6

$\frac{1}{33}$ $\frac{4}{33}$ $\frac{9}{33}$ $\frac{8}{33}$

$\frac{121}{33}$

100 1 4 9 8 5 6

3300

1 8 27 32 25 36

3. (20 points) Jun wants to find a lower bound of $\log P(X)$ where X is a random variable. Jin suggests that one of the lower bound is $\sum_Z P(Z|X) \log \frac{P(Z,X)}{P(Z|X)}$ where Z is another random variable. However, Jun is hard to know the reason. He asks Jin the reason and Jin writes the steps of finding the lower bound of $\log P(X)$ as below. Write the in-depth reason for each step for Jun.

$$\log P(X) = \log \sum_Z P(Z,X) \tag{1}$$

$$= \log \sum_Z P(Z,X) \frac{P(Z|X)}{P(Z|X)} \tag{2}$$

$$= \log \sum_Z P(Z|X) \frac{P(Z,X)}{P(Z|X)} \tag{3}$$

$$\geq \sum_Z P(Z|X) \log \frac{P(Z,X)}{P(Z|X)} \tag{4}$$

[Answer box for (1)]

Solution: Law of total probability

[Answer box for (2)]

Solution: $\frac{P(Z|X)}{P(Z|X)} = 1$

[Answer box for (3)]

Solution: Commutative property: The order of the factors does not change the product.

[Answer box for (4)]

Solution: Jensen's Inequality where $\sum_Z P(Z|X)$ is the form of expectation

4. (20 points) This question is related to your homework assignments - Naive Bayes Classifier. Read the instructions carefully and answer each question according to the instruction.

[Program]

```
import re
import math

def main():
    training1_sentence = "John likes to watch movies. Mary likes movies too."
    training2_sentence = "In the machine learning, " \
                         "naive Bayes classifiers are a family of simple probabilistic classifiers."
    testing_sentence = "John also likes to watch football games."

    alpha = 0.1
    prob1 = 0.4
    prob2 = 0.6

    print(naive_bayes(training1_sentence, training2_sentence, testing_sentence, alpha, prob1, prob2))

def naive_bayes(training1_sentence, training2_sentence, testing_sentence, alpha, prob1, prob2):
    bow_train1 = create_BOW(training1_sentence)
    bow_train2 = create_BOW(training2_sentence)
    bow_test = create_BOW(testing_sentence)

    classify1 = math.log( (4-1-1) ) + (4-1-2)
    classify2 = math.log( (4-1-3) ) + (4-1-4)

    return normalize_log_prob(classify1, classify2)


def calculate_doc_prob(bow_train, bow_test, alpha):
    total_dict = list(bow_train[0].keys())
    for token in bow_test[0].keys():
        if token not in total_dict:
            total_dict.append(token)
    N = 0
    for n in bow_train[1]:
        N = N + n
    prob_dic = {}
    for word in bow_test[0].keys():
        if word not in bow_train[0].keys():
            (4-1-5) = alpha / (N + alpha * len(total_dict))
        else:
            n = bow_train[0][word]
            (4-1-5) = (bow_train[1][n] + alpha) / (N + alpha * len(total_dict))
    logprob = 0
    for word in bow_test[0].keys():
        index = bow_test[0][word]
        logprob += math.log( (4-1-5) ) * bow_test[1][index]

    return logprob

```

```python
def normalize_log_prob(prob1, prob2):
    maxprob = max(prob1, prob2)
    prob1 -= maxprob
    prob2 -= maxprob
    prob1 = math.exp(prob1)
    prob2 = math.exp(prob2)
    normalize_constant = 1.0 / float(prob1 + prob2)
    prob1 *= normalize_constant
    prob2 *= normalize_constant
    return prob1, prob2

def replace_non_alphabetic_chars_to_space(sentence):
    return re.sub(r'[^a-z]+', ' ', sentence)

def create_BOW(sentences):
    bow_dict = {}
    bow = []
    sentences = sentences.lower()
    sentences = replace_non_alphabetic_chars_to_space(sentences)
    words = sentences.split()
    for token in words:
        if len(token) < 1:
            continue
        if token not in bow_dict:
            new_idx = len(bow)
            bow.append(0)
            bow_dict[token] = new_idx
        bow[bow_dict[token]] += 1
    return bow_dict, bow

if __name__ == "__main__":
    main()

```

4-1. (15 points) Fill in the blanks (4-1-1), (4-1-2), (4-1-3), and (4-1-4) to complete the program code of Naive Bayes Classifier.

[Answer box for (4-1-1)]

Solution:

```
prob1

```

[Answer box for (4-1-2)]

Solution:

```
calculate_doc_prob(bow_train1, bow_test, alpha)

```

[Answer box for (4-1-3)]

Solution:

```
prob2
```

[Answer box for (4-1-4)]

Solution:

```
calculate_doc_prob(bow_train2, bow_test, alpha)
```

[Answer box for (4-1-5)]

Solution:

```
prob_dic[word]
```

4-2. (5 points) Explain the role of the ‘alpha’ variable in line 10. Hint: What if the value of ‘alpha’ is zero?

[Answer box]

Solution: Simple answer: The divide by zero error or exception occurs at line 45 when the ‘alpha’ is zero.

Better answer: Describe the situation when the test sentence has words that do not appear in training data.

Best answer: This is a prior belief of a word in the test data.