Question:	1	2	3	4	Total
Points:	20	40	10	30	100
Score:					

SKKU SWE3026_41 Probability and Random Process
2020 Fall Final Exam

Student ID: _____

Name: _____

- Please check if you received all 10 pages of the test material.
  시작하기 전에 반드시 페이지가 총 10쪽인지 확인 하십시오.
- Due to the COVID-19 pandemic, 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.
  코로나19 범유행으로 인하여 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.

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

   1-1. (10 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]

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

      [Answer box]

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

      [Answer box]

      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]

      1-1-5. (1 point) The delta function of zero $\delta\left(0\right)$ is 0.

      [Answer box]

      1-1-6. (1 point) If X and Y are independent random variables, then $E\left[g\left(X\right)h\left(Y\right)\right] = g\left(E\left[X\right]\right)h\left(E\left[Y\right]\right)$.

      [Answer box]

1-1-7. (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]

1-1-8. (1 point) The normalized random variable is as follows: $Z = \frac{\{X-\mu\}}{\sigma}$ when $E[X] = \mu$, $Var[X] = \sigma^2$.

[Answer box]

1-1-9. (1 point) A random sequence is a random process.

[Answer box]

1-1-10. (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]

1-2. (10 points) Choose the best answer to fill in the blank from the options given.

1-2-1. (2 points) We want to model the arrival of customers at a service facility. The most useful random variable is ☐.

① Pascal
② Hypergeometric
③ Geometric
④ Poisson

1-2-2. (2 points) Bayes' rule is

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

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

1-2-3. (2 points) 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 ☐.

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

1-2-4. (2 points) If the correlation coefficient of two random variables X and Y is zero, we say that X and Y are ☐.

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

1-2-5. (2 points) 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

2. (40 points) Choose the best answer for the following questions.

2-1. (5 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 $\frac{5}{9}$

2-2. (5 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 $N\beta\lambda^2$

2-3. (10 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 $X \sim Bernoulli(p)$

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

2-4. (10 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 $\Phi(\frac{1}{2})$

2-5. (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.

2-5-1. (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 1.1
② 2.2
③ 3.3
④ 4.4
⑤ 5.5
⑥ 6.6

2-5-2. (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) $9.63 \rightarrow 10, 7.21 \rightarrow 7$

① 1
② 2
③ 3
④ 4
⑤ 5
⑥ 6

3. (10 points) Often when working with maximum likelihood functions, out of ease we maximize the log-likelihood rather than the likelihood to find the maximum likelihood estimator. Why is maximizing $L(x;\Theta)$ as a function of Θ equivalent to maximizing $\log L(x;\Theta)$?

[Answer box]

4. (30 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]

```
1  import re
2  import math
3
4  def main():
5      training1_sentence = "John likes to watch movies. Mary likes movies
       too."
6      training2_sentence = "In the machine learning, " \
7                           "naive Bayes classifiers are a family of
       simple probabilistic classifiers."
8      testing_sentence = "John also likes to watch football games."
9
10     alpha = 0.1
11     prob1 = 0.4
12     prob2 = 0.6
13
14     print(naive_bayes(training1_sentence, training2_sentence,
       testing_sentence, alpha, prob1, prob2))
15
16 def naive_bayes(training1_sentence, training2_sentence, testing_sentence
       , alpha, prob1, prob2):
17     bow_train1 = create_BOW(training1_sentence)
18     bow_train2 = create_BOW(training2_sentence)
19     bow_test = create_BOW(testing_sentence)
20
21     classify1 = math.log(    (4-1-1)    ) + |          (4-1-2)          |
22     classify2 = math.log(    (4-1-3)    ) + |          (4-1-4)          |
23
24     return normalize_log_prob(classify1, classify2)
25
26
27 def calculate_doc_prob(bow_train, bow_test, alpha):
28     total_dict = list(bow_train[0].keys())
29     for token in bow_test[0].keys():
30         if token not in total_dict:
31             total_dict.append(token)
32     N = 0
33     for n in bow_train[1]:
34         N = N + n
35     prob_dic = {}
36     for word in bow_test[0].keys():
37         if word not in bow_train[0].keys():
38             prob_dic[word] = alpha / (N + alpha * len(total_dict))
39         else:
40             n = bow_train[0][word]
41             prob_dic[word] = (bow_train[1][n] + alpha) / (N + alpha *
       len(total_dict))
42     logprob = 0
43     for word in bow_test[0].keys():
44         index = bow_test[0][word]
45         logprob += math.log(prob_dic[word]) * bow_test[1][index]
46
47     return logprob
48
```

```
49 def normalize_log_prob(prob1, prob2):
50     maxprob = max(prob1, prob2)
51     prob1 -= maxprob
52     prob2 -= maxprob
53     prob1 = math.exp(prob1)
54     prob2 = math.exp(prob2)
55     normalize_constant = 1.0 / float(prob1 + prob2)
56     prob1 *= normalize_constant
57     prob2 *= normalize_constant
58     return prob1, prob2
59
60 def replace_non_alphabetic_chars_to_space(sentence):
61     return re.sub(r'[^a-z]+', ' ', sentence)
62
63 def create_BOW(sentences):
64     bow_dict = {}
65     bow = []
66     sentences = sentences.lower()
67     sentences = replace_non_alphabetic_chars_to_space(sentences)
68     words = sentences.split()
69     for token in words:
70         if len(token) < 1:
71             continue
72         if token not in bow_dict:
73             new_idx = len(bow)
74             bow.append(0)
75             bow_dict[token] = new_idx
76         bow[bow_dict[token]] += 1
77     return bow_dict, bow
78
79 if __name__ == "__main__":
80     main()
81
```

4-1. (20 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)]

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

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

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

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

[Answer box]