(Fall 2024) 515512 Probability (**Early Bird: 2024/10/25, 9pm; Normal: 2024/10/26, 9pm**)

Homework 2: Random Variables, Independence, and Expectation

Submission Guidelines: You will submit two files via E3: (1) Please compress all your technical report and write-ups (photos/scanned copies are acceptable; please make sure that the electronic files are of good quality and reader-friendly) into one .pdf file (2) Please also submit your Jupyter Notebook file.

Problem 1 (Communication Over a Noisy Channel With Erasure) (6+6+6=18 points)

At each time, the transmitter sends a bit (either 0 or 1), and there are three possible outcomes: (1) The bit is delivered correctly; (2) The receiver gets the wrong bit ("flipped"); (3) The receiver receives a message that the bit was not received ("erased").

- If a "0" is sent, the erasure probability is ε_0 and the flipping probability is α_0.
- Similarly, if a "1" is sent, the erasure probability is ε_1 and the flipping probability is α_1.
- Moreover, at each transmission, the transmitter chooses to send a "0" with probability p and send a "1" with probability $1-p$ ($p \in [0, 1]$).
- The choices of the sent bits are all independent.
- Given the knowledge about the sent bit ("0" or "1"), the result of the corresponding reception is independent from another transmission or reception.

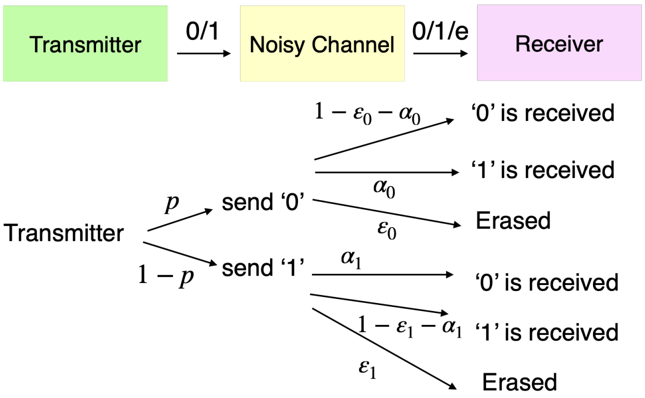

Figure 1: Communication over a noisy channel with erasure.

(a) Define the following two events: $A :=$ {The first sent bit is "1"} and $B :=$ {The first bit is erased}. Then, are the two events A, B independent? (Hint: You may need to specify the values of $p, \varepsilon_0, \varepsilon_1, \alpha_0, \alpha_1$ under which independence would hold)

(b) Based on (a), let us further define an event $C :=$ {The first bit is NOT delivered correctly}. Then, are the two events A, B (defined in the subproblem (a) above) *conditionally independent* given the event C?

(c) What is the probability that a transmitted bit is delivered correctly on the receiver side? Please express your answer using $\varepsilon_0, \varepsilon_1, \alpha_0, \alpha_1, p$. (Hint: Total probability theorem)

Problem 2 (Special Discrete Random Variables) (8+8+8=24 points)

(a) A wireless transmitter sends out either a "1" with probability p, or a "0" with probability $1-p$, independent of earlier transmissions. If the number of transmissions within a given time interval has a Poisson PMF with parameter λ, T, show that the number of 1s transmitted in that same time interval has a Poisson PMF with parameter $p\lambda$. (Hint: This is usually called *splitting a Poisson random variable.* You may define two random variables X, Y to be the numbers of 1s and 0s transmitted, respectively. Let $Z = X + Y$. Then, analyze $P(X = n, Y = m)$ by using multiplication rule with the help of Z).

(b) Let $X_1, X_2, \cdots, X_n$ be n independent Geometric random variables with the same success probability $p \in (0, 1)$. Define $X = \max(X_1, \cdots, X_n)$ and $Y = \min(X_1, \cdots, X_n)$. What is the PMF of X? Moreover, what is the PMF of Y? What kind of random variable is Y?

(c) DNA was first discovered in 1869 by Swiss physician Friedrich Miescher, and later on scientists found that a DNA sequence can be represented as a sequence of letters, where the alphabets has 4 letters: A, C, T, G. Suppose such a sequence is generated randomly in a letter-by-letter manner, and the letters are independent and probabilities of A, C, T, G are p_A, p_C, p_T, and p_G, respectively. Given a DNA sequence (denoted by S_2) of length 123. Define a random variable X_{S_2} to be the number of letter T in the sequence S_2. What is the PMF of the random variable X_{S_2}? (Hint: Binomial distributions)

Problem 3 (PMF and Shannon Entropy) (6+6+6+6=24 points)

Quantifying the amount of information carried by a random variable has been a central issue in various domains, including machine learning and wireless communication. In this problem, let us take a quick look at one fundamental and super useful concept in information theory, namely the *Shannon entropy* of the distribution of a discrete random variable.

Consider a discrete random variable X with the set of possible values $\{1, 2, \cdots, n\}$. Define $p_i := P(X = i)$, for all $i = 1, \cdots, n$. Next, we define a metric called *entropy*:

$$H(X) := -\sum_{i=1}^{n} p_i \ln p_i.$$

(Note: “ln” means the base of the logarithm is e).

(a) Recall that one of the property of entropy is that “if a choice is broken down into two successive choices, the original entropy shall be the weighted sum of the individual entropy.” Based on the example provided in Lecture 9, please verify the following:

$$H\big(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}\big) = H\big(\frac{1}{2}, \frac{1}{2}\big) + \frac{1}{2} \cdot H\big(\frac{2}{3}, \frac{1}{3}\big).$$

(b) Consider the basic case with $n = 2$. In this case, we could simply write $H(X) = -\big(p_1 \ln p_1 + (1 - p_1)\ln(1 - p_1)\big)$. Please plot $H(X)$ as a function of p. What are the maximum and minimum of $H(X)$ and under what p_1?

(c) Now let us consider the general case of $n \geq 2$. What is the maximum possible value of $H(X)$? Find the PMF $\{p_i\}_{i=1}^n$ that achieves this maximum. (Hint: You may use the weighted inequality of arithmetic and geometric means, i.e., $\frac{w_1x_1+w_2x_2+\cdots+w_nx_n}{w} \geq (x_1^{w_1} \cdot x_2^{w_2} \cdots x_n^{w_n})^{\frac{1}{w}}$, where $\{w_i\}$ and $\{x_i\}$ are non-negative real numbers and $w = w_1 + \cdots + w_n$)

(d) Given the same setting of (c), what is the minimum possible value of $H(X)$? Please find out all the PMFs $\{p_i\}_{i=1}^n$ that achieve this minimum.

Problem 4 (Expected Value and Moments) (8+10=20 points)

(a) Suppose $X \sim$ Geometric(p). Show that (i) $E[X] = 1/p$; (ii) $E[e^{tX}] = \frac{pe^t}{1-(1-p)e^t}$, for $t < -\ln(1 - p)$. (Hint: Regarding $E[X]$, write down the PMF and try to reuse the fact that the total probability is 1.)

(b) Let $z_n = (-1)^n\sqrt{n}$, for $n = 1, 2, 3 \cdots$. Let Z be a discrete random variable with the set of possible values $\{z_n : n = 1, 2, 3 \cdots\}$. The PMF of Z is

$$p_Z(z_n) = P(Z = z_n) = \frac{6}{(\pi n)^2}, \ \forall n \in \mathbb{N}.$$

What is Var$[Z]$? How about the value of $\sum_{n=1}^{\infty} z_n^3 \cdot p_Z(z_n)$? Does $E[Z^3]$ exist? Please carefully justify each step of your answer.

Problem 5 (Programming: Naive Bayes Classifier for Spam Filtering) (16+10=26 points)

As described in Lecture 6, Naive Bayes classification is a simple and classic tool for machine learning problems, especially for spam filtering and text classification. In this problem, we will implement a naive Bayes classifier in python with the help of the *scikit-learn* package. You will do this Python programming task on Jupyter Notebook (See `https://docs.jupyter.org/en/latest/`). The resulting classifier will be trained and tested on a SMS spam collection dataset provided on E3 (Source: `https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset`).

Here is a brief summary of naive Bayes classification in the context of spam filtering:

- Our goal is to learn a classifier that determines whether a piece of SMS is spam or not (i.e. there are only two types of labels, "spam" or "ham") given its raw text.
- Specifically, we would leverage maximum a posteriori (MAP) estimation to determine the label of a SMS: Given a piece of SMS with n words $\{x_1, x_2, \cdots, x_n\}$, we would like to determine whether it is a spam or not. In naive Bayes classification, we typically assume *conditional independence* for the likelihood, i.e.

  $$p(x_1, x_2, \cdots, x_n|\text{label}) = \prod_{i=1}^{n} p(x_i|\text{label}).$$

  Then, we could apply MAP by calculating the following posterior distribution:

  $$p(\text{spam}|x_1, x_2, \cdots, x_n) \propto p(\text{spam}) \cdot p(x_1, x_2, \cdots, x_n|\text{spam})$$
  $$p(\text{ham}|x_1, x_2, \cdots, x_n) \propto p(\text{ham}) \cdot p(x_1, x_2, \cdots, x_n|\text{ham})$$

- To find $p(x_i|\text{label})$, we shall leverage a training dataset to find the empirical frequency of each word for each type of SMS.
- The prior distribution (i.e. $p(\text{spam})$ and $p(\text{ham})$) is to be configured by the designer.

Based on the above summary, in this implementation there are mainly 3 mini-tasks (normally you need no more than 20 lines of code in total):

- **Read and split the dataset**: The SMS spam dataset is provided in "spam.csv", which contains 5572 English text messages. Each message has two fields, namely the label (either "ham" or "spam") and the corresponding raw text. Moroever, we need to divide the dataset into a training set and a testing set (Note: This part has already been done in "naive_bayes.ipynb").
- **Train the classifier**: First, find the empirical frequency of each word for both spam and ham SMS using the training dataset (possibly with the help of some existing parsing/counting functions). Next, implement the MAP estimation (Hint: You may use MultinomialNB in the scikit-learn package)
- **Test the classifier**: Predict the labels of the messages in the testing dataset and find the accuracy of your classifier.

(a) Please finish the remaining parts of "naive_bayes.ipynb". What is the accuracy of your classifier with a 70%/30% partition of the dataset and a uniform prior?

(b) What if we use different prior distributions (please try at least 2 priors other than the uniform prior)? Also, if we change the partition of the dataset, will there be any significant change in the accuracy?

Please briefly summarize your observation in a technical report (no more than 1 page) and turn in your code and the report via E3.