CSE 312: Foundations of Computing II
Quiz Section #4: Discrete Random Variables, Naive Bayes

Review: Main Theorems and Concepts

Random Variable (rv): A numeric function $X : \Omega \to \mathbb{R}$ of the outcome.

Range/Support: The support/range of a random variable X, denoted Ω_X, is the set of all possible values that X can take on.

Discrete Random Variable (drv): A random variable taking on a ___________ (either finite or countably infinite) number of possible values.

Probability Mass Function (pmf) for a discrete random variable X: a function $p_X : \Omega_X \to [0, 1]$ with $p_X(x) = P(X = x)$ that maps possible values of a discrete random variable to the probability of that value happening, such that $\sum_x p_X(x) =$ ____________.

Expectation (expected value, mean, or average): The expectation of a discrete random variable is defined to be $\mathbb{E}[X] =$ ______________________________. The expectation of a function of a discrete random variable $g(X)$ is $\mathbb{E}[g(X)] =$ ______________________________.

Linearity of Expectation: Let X and Y be random variables, and $a, b, c \in \mathbb{R}$. Then, $\mathbb{E}[aX + bY + c] =$ ______________________________.

Exercises

1. Suppose we have N items in a bag, K of which are successes. Suppose we draw (without replacement) until we have k successes, $k \le K \le N$. Let X be the number of draws until the k^{th} success. What is Ω_X? What is $p_X(n) = P(X = n)$? (We say X is a "negative hypergeometric" random variable).

2. A frog starts on a 1-dimensional number line at 0. At each second, independently, the frog takes a unit step right with probability p_1, to the left with probability p_2, and doesn't move with probability p_3, where $p_1 + p_2 + p_3 = 1$. After 2 seconds, let X be the location of the frog. Find $p_X(k)$, the probability mass function for X. Find $\mathbb{E}[X]$. Find $p_Y(k)$, the probability mass function for $Y = |X|$, and $\mathbb{E}[Y]$.

3. Suppose we have r independent random variables $X_1, \ldots, X_r$ that each represent the number of coins flipped up to and including the first head, where $\mathbb{P}(\text{head}) = p$. Recall that each X_i has probability mass function,

$$p_{X_i}(k) = \mathbb{P}(X_i = k) = (1 - p)^{k-1} p$$

   (a) What do you think $\mathbb{E}[X_i]$ should be (without calculations) if $p = \frac{1}{2}$? If $p = \frac{1}{3}$? In the general case? (Proof in lecture soon.)

   (b) Suppose we define $X = X_1 + \ldots + X_r$. What does X represent, in English words? (Hint: think of performing each "trial" one after the other.)

   (c) What is Ω_X? Find $p_X(k)$, the probability mass function for X.

(d) Find $\mathbb{E}[X]$ using linearity of expectation.

4. Questions about the Naive Bayes Classifier:

   (a) Naive Bayes assumes conditional independence of words in an email, given that we know the label (ham or spam) of the email. Why is that assumption necessary to make Naive Bayes work?

   (b) Is the conditional independence assumption actually true in the real world? That is, are the occurrences of words in an email independent of each other, if we know the label of the email? Explain.

   (c) Do you expect the Naive Bayes Classifier to correctly classify all emails in a test set? Explain why or why not.

   (d) If you were a spammer and you knew we used Naive Bayes to filter spam, how would you change your emails to try to get past the filter?

5. Let the random variable X be the sum of two independent rolls of a fair 3-sided die. (If you are having trouble imagining what that looks like, you can use a 6-sided die and change the numbers on 3 of its faces.)

   (a) What is the probability mass function of X?

   (b) Find $\mathbb{E}[X]$ directly by applying the definition of expectation to the result from part (a).

   (c) Find $\mathbb{E}[X]$ again using linearity of expectation.

   (d) Check that your answers to parts (b) and (c) are the same. Which way of computing the expectation was simpler, (a)+(b), or (c)?

6. Let the random variable X be the number of heads in n independent flips of a fair coin.

   (a) What is the probability mass function of X?

   (b) Find $\mathbb{E}[X]$ directly by applying the definition of expectation to the result from part (a).

   Hint: prove and use the identity $i\binom{n}{i} = n\binom{n-1}{i-1}$.

7. You have 10 pairs of socks (so 20 socks in total), with each pair being a different color. You put them in the washing machine, but the washing machine eats 4 of the socks chosen at random. Every subset of 4 socks is equally probable to be the subset that gets eaten. Let X be the number of complete pairs of socks that you have left.

   (a) What is the probability mass function of X?

   (b) Find $\mathbb{E}[X]$ directly by applying the definition of expectation to the result from part (a). Give your answer exactly as a simplified fraction.

   (c) Find $\mathbb{E}[X]$ again using linearity of expectation. Give your answer exactly as a simplified fraction.

   (d) Check that your answers to parts (b) and (c) are the same. Which way of computing the expectation was simpler, (a)+(b), or (c)?

8. Find the expected number of bins that remain empty when m balls are distributed into n bins randomly and independently. For each ball, each bin has an equal probability of being chosen. (Notice that two bins being

empty are not independent events: if one bin is empty, that decreases the probability that the second bin will also be empty. This is particularly obvious when $n = 2$ and $m > 0$.)

9. At a reception, n people give their hats to a hat-check person. When they leave, the hat-check person gives each of them a hat chosen at random from the hats that remain. What is the expected number of people who get their own hats back? (This is closely related to, but much simpler than, the challenge problem from the worksheet from quiz section #2. Notice that the hats returned to two people are not independent events: if a certain hat is returned to one person, it cannot also be returned to the other person.)

10. (This exercise is the same as Exercise 5, but with an ordinary 6-sided die rather than a 3-sided die.) Let the random variable X be the sum of two independent rolls of a fair 6-sided die.

    (a) What is the probability mass function of X?

    (b) Find $\mathbb{E}[X]$ directly by applying the definition of expectation to the result from part (a).

    (c) Find $\mathbb{E}[X]$ again using linearity of expectation.

    (d) Check that your answers to parts (b) and (c) are the same. Which way of computing the expectation was simpler, (a)+(b), or (c)?