Mini-Project 1 – Part 2

ECE 471
Fall 2024

Karthik Vasu (kvasu2), Wilmer Smilde (wsmilde2), Xinran Yu (xinran4)

Task 3

Q1: Based on the above assumptions, answer the following questions on basic probability.

a) The assumption of at most one disengagement per mile allows us to treat the occurrence of a disengagement in a mile as a random variable with a ___Bernoulli_____ distribution

b) Based on the above assumptions, calculate the probability of disengagement per mile on a cloudy day.

```
Probability of disengagement per mile on a cloudy day is:
0.005902556775527249
```

c) Based on the above assumptions, calculate the probability of disengagement per mile on a clear day.

```
Probability of disengagement per mile on a clear day is:
0.0005195663748517998
```

Task 3

Q1: Based on the above assumptions, answer the following questions on basic probability

d) Similarly, calculate the probability of an automatic disengagement per mile on a cloudy day, and the probability of an automatic disengagement per mile on a clear day.

```
Probability of an automatic disengagement per mile on a clear day is:
0.00026390673008345393
Probability of disengagement per mile on a cloudy day is:
0.0028063653172267283
```

e) How likely is it that there are 100 or more disengagements in 10,000 miles under cloudy conditions? (**hint**: use Central Limit Theorem)

We have our boolean probability parameter p_{cloudy} for the binomial distribution for the engamement per mile under cloudy conditions. Let X_i be the boolean variable that is 1 if there is a disengagement in mile i and 0 if there is no disengagment. Let

$$Y = \sum_{i=1}^{10000} X_i$$

We have to calculate

$$P(Y > 100)$$

By the central limit theorem, the normalized variable

$$Y^* = \frac{Y - 10000 \cdot p_{\text{cloudy}}}{\sqrt{10000 p_{\text{cloudy}}(1 - p_{\text{cloudy}})}}$$

is (very accurately) described by a normal distribution with mean 0 and standard deviation 1. We have:

$$P(Y > 100) = P\left(Y^* > \frac{100 - 10000 p_{\text{cloudy}}}{\sqrt{10000 p_{\text{cloudy}}(1 - p_{\text{cloudy}})}}\right).$$

```
The probability that there are 100 disengagements per 10,000 miles under cloudy conditions is:
4.420337695876242e-08
Sanity check: the probablity that there is 1 disengagment should be rather high:
0.999999999999982
```

Task 3

Q2: Assuming that the disengagement per mile is a random variable with the distribution you answered in Task 3.1.a, and the weather condition is *cloudy* .

a) What is the distribution of "the number of miles until the next disengagement"? Explain your reasoning. Calculate and state the values of the parameters of the distribution

Let X_k be the random variable of disengagement in mile k ($X_k = 0$ if no disengagement in mile k, and $X_k = 1$ if there is a disengagement in mile k, with probability p_{cloudy}), and M the number of miles until next engagement. We calculate:

$$P(M = n) = \left(\prod_{k=1}^{n-1} P(X_k = 0)\right) \cdot P(X_n = 1) = (1 - p_{\text{cloudy}})^{n-1} \cdot p_{\text{cloudy}}$$

We see that M follows a geometric distribution.

b) What is the distribution of "the number of disengagements in 10,000 miles"? (**hint:** this is equivalent to drawing n=10,000 independent trials from the distribution of disengagement per mile you calculated from Task 3.1.a). Calculate and state the values of the parameters of the distribution.

As stated in our answer to Task 3.1.a., the number of disengagements Y_n per n miles under cloudy conditions follows a binomial distribution, so

$$P(Y_n = k) = \binom{k}{n} p_{\text{cloudy}}^k (1 - p_{\text{cloudy}})^{n-k}$$

It is of course understood that $n = 10,000$ and $p_{\text{cloudy}} \cong 0.006$ as calculated above, but it's way more transparant to write n and p_{cloudy} instead of their actual values.

c) Notice that the number of disengagements "$\boldsymbol{n}$" in Task 3.2.b is large while the probability of disengagement per mile "$\boldsymbol{p}$" is very small, what distribution does your answer in Task 3.2.b approximate? Calculate and state the values of the parameters of the distribution.

Under the assumption that n is very large and p is very small (but keeping the average $\mu = n \cdot p$ constant), the binomial distribution can be approximated by a Poisson distribution. There is only one parameter, which is the average. The value of the Poisson that approximates Y_{10000} is $\mu = 10000 \cdot p_{\text{cloudy}}$. See the value below:

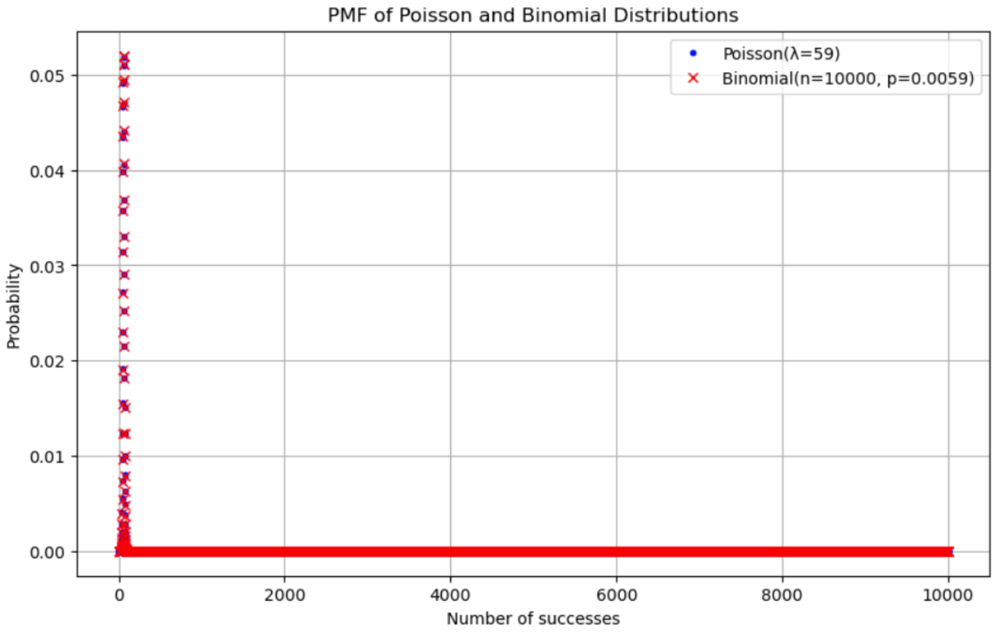
```
59.02556775527249
```

Task 3

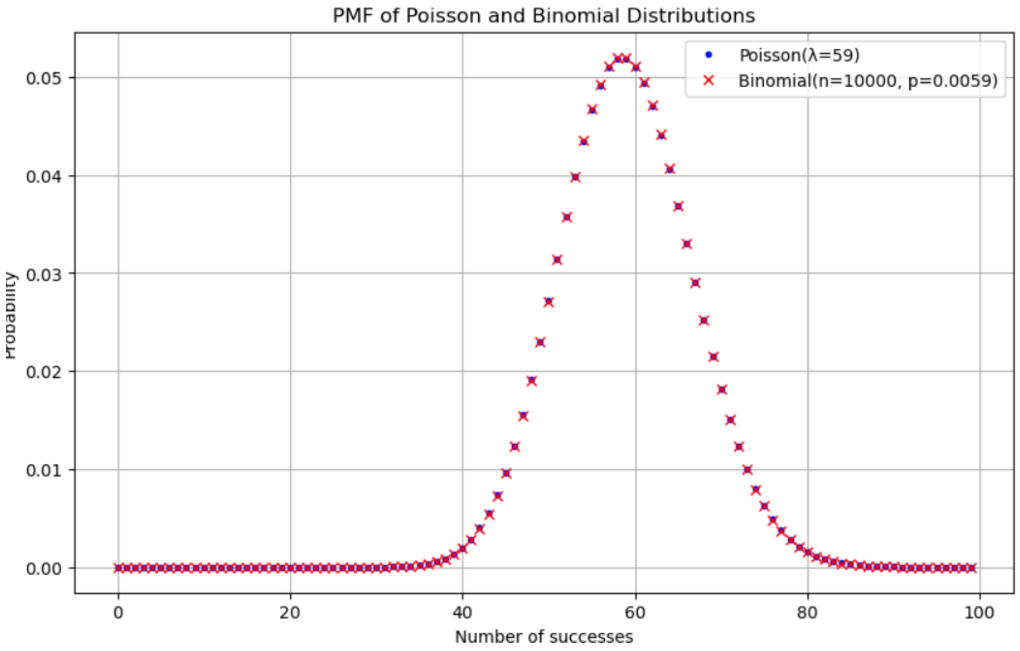
Q2: Assuming that the disengagement per mile is a random variable with the distribution you answered in Task 3.1.a, and the weather condition is *cloudy* .

d) Plot the probability mass function (PMF) of the distribution in Task 3.2.b and Task 3.2.c for:

1. x-axis ranging between 0 and 10000.



2. The x-axis ranging between 0 and 100



We see that the distributions are indistinguishable, and their absolute difference is negligeble. Their relative difference tends to grow as x increases, but after $x \sim 550$, the probabilities become essentially zero for the python memory anyway, so the relative difference is not a useful metric.

e) Solve Task 3.1.e by using the cumulative distribution function (CDF) of the distribution you computed in Task 3.2.c and compare the results. Discuss your findings

If we approximate the binomial distribution with a Poisson distribution, we find:

```
probability = 1 - sp.stats.poisson.cdf(100, 60)
print(probability)
```

✓ 0.0s

```
8.685389315532532e-07
```

This probability is one order of magnitude larger than our answer in 3.1.e., but still very small.

Task 3

Q3: What’s the conditional probability that the reaction time is:

a) Greater than 0.4s given that the weather was cloudy? Reaction time is measured only in cases where there was an **automatic disengagement**.

```
The probability of reaction time being greater than 0.4 seconds on a cloudy day is: 0.6120906801007556
```

b) Greater than 0.7s given that the weather was clear? Reaction time is measured only in cases where there was an **automatic disengagement**.

```
The probability of reaction time being greater than 0.7 seconds on a clear day is: 0.3854166666666667
```

Task 3

Q4: A study found that an **automatic AV disengagement** will result in an accident if the human driver is *slow* in reacting. Following reactions are considered slow: (i) a reaction time greater than 0.4s under cloudy conditions and, (ii) a reaction time greater than 0.7s under clear conditions. Find the probability of an accident per mile due to automatic AV disengagement and slow reaction.

```
The probability of an accident is: 0.5679513184584178
```

Task 3

Q5: You will investigate how to diagnose the cause of an AV disengagement based on new observations:

a) An AV had a disengagement with a reaction time greater than 0.4s on a cloudy day. What is the posterior probability that the root cause of the disengagement was “Software Froze”?

We are given that a disengagement occurs with a reaction time greater than 0.4s and on a cloudy day. We have to find $P(C = Software | R > 0.4, W = cloudy)$

```
The probability that the accident was caused by software failure on a cloudy day is: 0.053497942386831275
```

b) What is the posterior probability that the root cause of the disengagement was “Software Froze if the disengagement happened on a clear day with reaction time greater than 0.7s. Based on the probabilities calculated in Tasks 3.5.a and 3.5.b, discuss your findings.

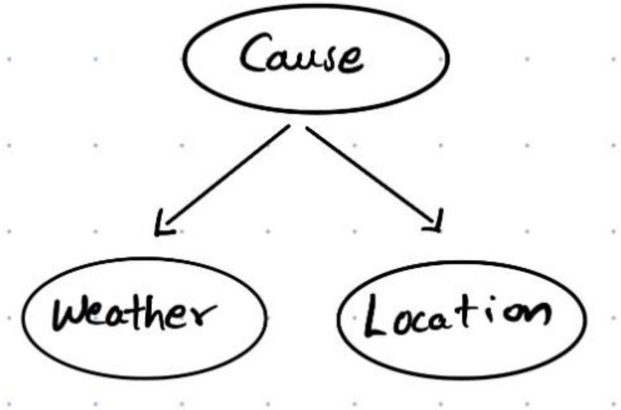
We are given that a disengagement occurs with a reaction time greater than 0.7s and on a clear day. We have to find $P(C = Software | R > 0.7, W = clear)$

```
The probability that the accident was caused by software failure on a clear day is: 0.24324324324324326
```

Task 3

Q6: In this question, you will construct a Naive Bayes model to infer the root cause of disengagement scenarios of AVs. Naive Bayes assumes that the factors are class conditionally independent. We assume that both Location (urban-street or highway) and Weather (cloudy or clear) are factors related to the Cause (consider the Cause has 3 different values, "Software Froze", "Hardware Fault" or "Other"), and Location and Weather are independent given the Cause. Answer the following questions

a) Draw a graph for the Naive Bayes model described in the question.



b) Count the number of parameters needed to define the Naive Bayes model (including the prior and the conditional probability distributions

Using Naive Bayes, the conditional probabilities are given by

$$P(C_k|x) \propto P(W, L, C_k)$$

$$= P(L|C_k)P(W|C_k)P(C_k)$$

There are 3 values of k, so we have to compute 3 values for $P(L|C_k)$ and 3 for $P(W|C_k)$. We will assume that $P(C_k) = 1/3$. In total we need to compute 6 parameters

Task 3

Q6 **(continued):**

c) Based on the number of parameters needed, derive, and show the conditional probability tables and prior probability from the given dataset to infer the Cause .

	Software Froze	Hardware Fault	Other
urban-street	0.938776	0.913462	0.993917
highway	0.061224	0.086538	0.006083

	Software Froze	Hardware Fault	Other
cloudy	0.387755	0.442308	0.913625
clear	0.612245	0.557692	0.086375

$$P(C_k) = 1/3$$

d) b) According to the conditional probability tables you derived, what is the most probable root cause of disengagement given the Weather was **cloudy** and the Location was **urban-street**

	Software Froze	Hardware Fault	Other
Probability	0.121338	0.134677	0.302689

From the probability calculations we can see that Other has the largest value, therefore the root cause of disengagement on a cloudy day on an urban-street is most likely Other causes