Chris Piech
CS109

Fall 2022
Nov 2$^{\text{nd}}$

CS109 Midterm Exam (Wednesday)

This is a closed calculator/computer exam. You are, however, allowed to use notes in the exam. You have 2 hours (120 minutes) to take the exam. The exam is 120 points, meant to roughly correspond to one point per minute of the exam. You may want to use the point allocation for each problem as an indicator for pacing yourself on the exam.

In the event of an incorrect answer, any explanation you provide of how you obtained your answer can potentially allow us to give you partial credit for a problem. For example, describe the distributions and parameter values you used, where appropriate. It is fine for your answers to include summations, products, factorials, exponentials, and combinations.

You can leave your answer in terms of Φ (the CDF of the standard normal). For example $\Phi(3/4)$ is an acceptable final answer..

LELAND STANFORD JUNIOR UNIVERSITY
DIE LUFT DER FREIHEIT WEHT
1891

I acknowledge and accept the letter and spirit of the honor code. I pledge to write more neatly than I have in my entire life:

Signature: ____

Family Name (print): ____

Given Name (print): ____

Stanford Email (@stanford.edu): ______________________________________

1 Weather Prediction [20 points]

We are building a weather prediction service for Cloud City. We have access to the full history of weather stored in a list called `history`. `history[i]` is `True` if it rained on day `i`. For example the history list for the last week at Stanford would look like: `history = [False, False, False, False, False, False, True]` if the only time it rained was the last day of the week. The `history` list for Cloud City has over 10,000 elements and represents the daily weather for the last 30 years.

a. (7 points) Write pseudo-code to estimate the probability of rain on a random day.

(python code **not** required)

```
rainy_days = 0
days = len(history)
for i in range(days):
  if history[i] == True:
    rainy_days += 1
return rainy_days / days
```

b. (7 points) Write pseudo-code to estimate the probability of rain tomorrow given that it was sunny today.

(python code **not** required)

```
rain_tomorrow_sun_today = 0
sun_today = 0
days = len(history)
for i in range(days-1):
  if history[i] == False:
      sun_today += 1
      if history[i+1] == True:
          rain_tomorrow_sun_today += 1
return rain_tomorrow_sun_today / sun_today
```

c. (6 points) How would you test if rain on a given day is independent of rain on the previous day, using the values calculated in part a and b?

In part a, we calculate $a = P(\text{rain})$, and in part b, we calculate $b = P(\text{rain}|\text{sun on the previous day})$. To check for independence between rain and whether it rained on the previous day, we can equivalently check for independence between rain and it *not* raining on the previous day. So, we can test

$$\begin{aligned} P(\text{rain}) &\overset{?}{\approx} P(\text{rain}|(\text{rain on the previous day})^C) \\ &= P(\text{rain}|\text{sun on the previous day}). \end{aligned}$$

In other words, we test $a \overset{?}{\approx} b$.

2 Down to Lunch [20 points]

You are writing an app called Down to Lunch which invites people to share a meal. You want to minimize spam while maximizing the chance of creating a good sized group.

a. (8 points) Assume that each person checks their phones an average of 262 times a day (or 0.182 times a minute), this rate is uniform over the day and events are independent. What is the probability that a person checks their phone in the next 5 minutes?

Using an exponential distribution:
Let X be the time in minutes until a person checks their phone. $X \sim \text{Exp}(\lambda = 0.182)$.

$$\begin{aligned} P(X < 5) &= 1 - e^{(-0.182*5)} \\ &= 0.597 \end{aligned}$$

Using a Poisson distribution: Let Y be the number of times a person checks their phone in the next 5 minutes. $Y \sim \text{Poi}(\lambda = 0.182 * 5)$.

$$\begin{aligned} P(X \geq 1) &= 1 - P(X = 0) \\ &= 1 - [\frac{(0.182*5)^0 e^{(-0.182*5)}}{0!}] \\ &= 1 - e^{(-0.182*5)} \\ &= 0.597 \end{aligned}$$

b. (12 points) Let p_a be your answer to part (a). Each person you invite to lunch will come if they check their phone in the next five minutes and they accept the invite. The probability that someone accepts an invitation, given that they check their phone in the next 5 minutes is $^1/_2$. If you invite 7 people, what is the probability of having a good sized lunch? A good sized lunch group has 1 to 4 guests accept an invite.

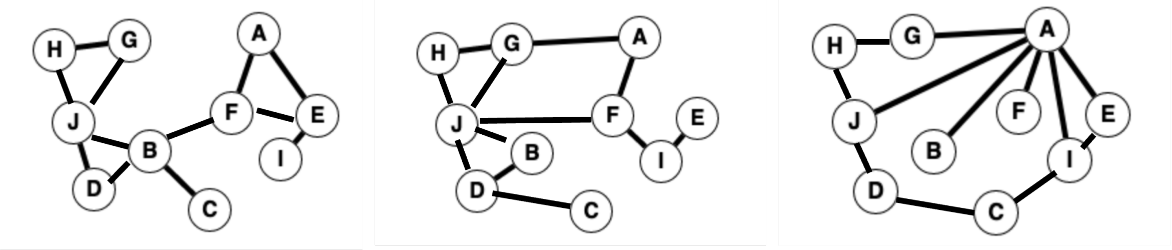
We are given that:

$$\begin{aligned} P(accept|check) &= 0.5 \\ P(accept \cap check) &= P(accept|check) * P(check) \\ &= 0.5 * p_a \end{aligned}$$

There are n = 7 people, and a good sized lunch occurs when k = 1,2,3,4 (either of these values). Let Y $\sim$ Binomial(n = 7, p = 0.5 * P$_a$) :
$P(lunch) = \sum_{i=1}^{4} \binom{7}{i} (0.5p_a)^i (1 - 0.5p_a)^{7-i}$

3 Random Edges [25 points]

Here are three different networks each with 10 nodes and 12 random edges:



a. (4 points) Consider a network with 10 nodes. Count the number of possible locations for undirected edges. An undirected edge from node A to node B is not distinct from an edge from node B to node A. You can assume that an edge does not connect a node to itself.

Each edge connects 2 nodes, and there are 10 possible options for each node, so the answer is

$$\binom{10}{2} = \boxed{45}.$$

b. (6 points) Assume the same network (with 10 nodes) has 12 random edges. If each pair of nodes is equally likely to have an edge, how many unique ways could we chose the location of the 12 edges? Let k be your answer to part a.

There are k possible pairs of nodes for an edge to connect, and we have 12 distinct (undirected) edges, so there are $\boxed{\binom{k}{12}}$ ways to do this.

c. (10 points) In the same network with 10 nodes and 12 edges, select a node uniformly at random. Let X be the number of edges connected to the node. What is the probability that $X = i$ for $0 \leq i \leq 9$. Recall that there are only 9 nodes to connect to since there are 10 nodes.

Let the random note selected be N, and fix $N = j$ for some $j = 0, 1, 2, \dots, 9$.

We will first compute the distribution of $P(X = i|N = j)$ using $|E|/|S|$. The sample space is the set of ways to choose 12 edges, and the event space is the set of ways to do so such that we've chosen exactly i of the edges incident to node j (which has 9 edges incident to it), so the answer is

$$P(X = i|N = j) = \frac{\binom{9}{i}\binom{k-9}{12-i}}{\binom{k}{12}}.$$

Since the answer did not depend on j, $P(X = i)$ must be the same value. Explicitly,

$$
\begin{aligned}
P(X=i) &= \sum_{j=0}^{9} P(X=i|N=j)P(N=j) \\
&= \sum_{j=0}^{9} \frac{\binom{9}{i}\binom{k-9}{12-i}}{\binom{k}{12}} \cdot \frac{1}{10} \\
&= \boxed{\frac{\binom{9}{i}\binom{k-9}{12-i}}{\binom{k}{12}}}
\end{aligned}
$$

Remark 1. This is precisely a hypergeometric distribution.

Remark 2. The binomial distribution gets very close to this solution, however, notice each of the 9 “trials” (corresponding to an edge incident to a node) are not independent of each other.

d. (5 points) Let $\mathrm{P}(X=i)$ be the answer to the previous part. What is $\mathrm{Var}(X)$?

Using the law of the unconscious statistician,

$$
\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\
&= \left(\sum_{i=0}^{9} i^2 \cdot P(X=i)\right) - \left(\sum_{i=0}^{9} i \cdot P(X=i)\right)^2
\end{aligned}
$$

Remark. How might we actually compute this for the distribution in part (c)? We could break X up into indicator random variables $X = X_1 + \ldots + X_9$, then using linearity of expectation we just have to compute $\mathbb{E}[X_i]$ and $\mathbb{E}[X_i X_j]$.

4 Expressed Trait [15 points]

We would like to predict the probability that a trait is “on”. This particular trait is turned on for a cell if either gene$_1$ or gene$_2$ are “active”. A gene is considered active if at least one RNA copy of the gene was created in the last 10 seconds.

For gene$_1$, on average 30 RNA copies are created every 10 seconds,
For gene$_2$ on average 20 RNA copies are created every 10 seconds.

You may assume that the process of creating an RNA copy for a gene occurs independently of the last time an RNA copy was created, that RNA copies are created at a constant average rate, and that genes create RNA copies independently of each other.

What is the probability that the trait is on?

Let $X \sim \text{Poi}(\lambda = 30)$ be the number of RNA copies gene$_1$ creates in 10 seconds.
Let $Y \sim \text{Poi}(\lambda = 20)$ be the number of RNA copies gene$_2$ creates in 10 seconds.

$$\begin{aligned} P(\text{gene}_1 \text{ is active}) &= 1 - P(\text{gene}_1 \text{ is NOT active}) \\ &= 1 - P(X = 0) \\ &= 1 - \frac{30^0 e^{-30}}{0!} \\ &= 1 - e^{-30} \end{aligned}$$

Similarly, $P(\text{gene}_2 \text{ is active}) = 1 - e^{-20}$.

Solution 1 (Inclusion-Exclusion):

$$\begin{aligned} P(\text{trait is on}) &= P(\text{gene}_1 \text{ is active OR gene}_2 \text{ is active}) \\ &= P(\text{gene}_1 \text{ is active}) + P(\text{gene}_2 \text{ is active}) - P(\text{gene}_1 \text{ is active AND gene}_2 \text{ is active}) \\ &= P(\text{gene}_1 \text{ is active}) + P(\text{gene}_2 \text{ is active}) - P(\text{gene}_1 \text{ is active}) \cdot P(\text{gene}_2 \text{ is active}) \\ &= 1 - e^{-30} + 1 - e^{-20} + (1 - e^{-30})(1 - e^{-20}) \\ &= 1 - e^{-50} \end{aligned}$$

Solution 2 (De Morgan’s Law):

$$\begin{aligned} P(\text{trait is on}) &= 1 - P(\text{trait is NOT on}) \\ &= 1 - P(\text{gene}_1 \text{ is NOT active AND gene}_2 \text{ is NOT active}) \\ &= 1 - P(\text{gene}_1 \text{ is NOT active}) \cdot P(\text{gene}_2 \text{ is NOT active}) \\ &= 1 - e^{-30} * e^{-20} \\ &= 1 - e^{-50} \end{aligned}$$

Solution 3 (Mutually Exclusive Outcomes):

$$\begin{aligned} P(\text{trait is on}) &= P(\text{gene}_1 \text{ active, gene}_2 \text{ active}) + P(\text{gene}_1 \text{ active, gene}_2 \text{ NOT active}) + P(\text{gene}_1 \text{ NOT active, gene}_2 \text{ active}) \\ &= (1 - e^{-30})(1 - e^{-20}) + (1 - e^{-30})(e^{-20}) + (e^{-30})(1 - e^{-20}) \\ &= 1 - e^{-50} \end{aligned}$$

5 Mixing up Gaussians [20 points]

Consider the following function:

```
def sample():
    X = rand_bern(0.3)                     # bernoulli sample
    if X == 0:
       return rand_gauss(mu = -1, std = 2) # gaussian sample
    return rand_gauss(mu = 1, std = 1)     # gaussian sample
```

a. (10 points) What is the probability that the function returns a value less than 0?

Let Y be the output of the function. Then:

$$\begin{aligned} P(Y<0) &= P(Y<0|X=0)P(X=0)+P(Y<0|X=1)P(X=1) \\ &= \Phi(\frac{0-(-1)}{2})*(1-0.3)+\Phi(\frac{0-1}{1})*(0.3) \\ &= 0.7*\Phi(0.5)+0.3*\Phi(-1) \end{aligned}$$

b. (10 points) The function returns back the value 0.1. What is the probability that X was 1?

a. Define target: Let X be the outcome of Bernoulli random variable. Let S be the output of the function.

$$P(X=1|S=0.1)$$

b. Bayes

$$P(X=1|S=0.1)=\frac{P(S=0.1|X=1)P(X=1)}{P(S=0.1)}$$

c. Use law of total probability to expand denominator

$$P(X=1|S=0.1)=\frac{P(S=0.1|X=1)P(X=1)}{P(S=0.1|X=1)P(X=1)+P(S=0.1|X=0)P(X=0)}$$

d. Cancel out epsilons in Gaussian probabilities to be able to use PDF of Gaussian Distribution. Let $f_{X=0}$ be the PDF of the Gaussian Distribution when $X=0$ and $f_{X=1}$ the PDF of the Gaussian Distribution when $X=1$.

$$P(X=1|S=0.1)=\frac{f_{X=1}(0.1)P(X=1)}{f_{X=1}(0.1)P(X=1)+f_{X=0}(0.1)P(X=0)}$$

e. Substitute values using PDF of Gaussian Distribution

$$P(X=1|S=0.1)=\frac{\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{0.1-1}{1})^2}P(X=1)}{\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{0.1-1}{1})^2}P(X=1)+\frac{1}{2\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{0.1+1}{2})^2}P(X=0)}$$

f. Simplify expression

$$\begin{aligned} P(X=1|S=0.1) &= \frac{e^{-\frac{1}{2}(\frac{0.1-1}{1})^2}\cdot 0.3}{e^{-\frac{1}{2}(\frac{0.1-1}{1})^2}\cdot 0.3+\frac{1}{2}e^{-\frac{1}{2}(\frac{0.1+1}{2})^2}\cdot 0.7} \\ &= \frac{e^{-\frac{1}{2}(-0.9)^2}\cdot 0.3}{e^{-\frac{1}{2}(-0.9)^2}\cdot 0.3+\frac{1}{2}e^{-\frac{1}{2}(0.55)^2}\cdot 0.7} \end{aligned}$$

6 Bayesian Viral Load Test [20 points]

We are going to build a Bayesian Viral Load Test which updates a belief distribution regarding a patient's viral load. Though viral load is continuous, in our test we represent it by discretizing the quantity into whole numbers between 0 and 99, inclusive. The units of viral load are the number of viral instances per million samples.

a. (3 points) If a person has a viral load of 9 (in other words, 9 viruses out of every 1 million samples) what is the probability that a random sample from the person is a virus?

$$\frac{9}{1,000,000}$$

b. (7 points) We test 100,000 samples from one person for the virus. If the person's true viral load is 9, what is the probability that exactly 1 of our 100,000 samples is a virus? Use a computationally efficient approximation to compute your answer. Your approximation should respect that there is 0 probability of getting negative virus samples.

Let's define a random variable X, the number of samples that are viral given the true viral load is 9. The question is asking for $P(X = 1)$.

We can think about this as a binomial process, where the number of trials n is the number of samples and the probability p is the probability that a sample is viral.

$$n = 100,000, p = \frac{9}{1,000,000}$$

Notice that n is very small and p is very large, so we can use the Poisson approximation to approximate our answer. We find $\lambda = np = 100,000 \cdot 9/1,000,000 = 0.9$, so $X \sim Poi(\lambda = 0.9)$. The last step is to use the PMF of the Poisson distribution.

$$P(X = 1) = \frac{(0.9)^1 e^{-0.9}}{1!}$$

c. (10 points) Based on what we know about a patient (their symptoms and personal history) we have encoded a prior belief in a list `prior` where `prior[i]` is the probability that the viral load equals i. `prior` is of length 100 and has keys 0 through 99.

Write an equation for the updated probability that the true viral load is i given that we observe a count of 1 virus sample out of 100,000 tested. Recall that $0 \leq i \leq 99$. You may use approximations.

We want to find

$$P(\text{viral load} = i|\text{observed count of} \frac{1}{100000}) \tag{1}$$

We can apply Bayes Rule to get

$$= \frac{P(\text{observed count of} \frac{1}{100000}|\text{viral load} = i)P(\text{viral load} = i)}{P(\text{observed count of} \frac{1}{100000})} \tag{2}$$

From part (b), we know that we can define a random variable $X \sim \text{observed count out of 100,000}|\text{viral load} = i$, and we can model X as a Poisson approximation to a binomial with $n = 100000$ and $p = \frac{i}{1000000}$, with

$$\lambda = np = 100000 \cdot \frac{i}{1000000} = \frac{i}{10}$$

So X can be written as

$$X \sim Poi(\lambda = \frac{i}{10})$$

Now we can rewrite our Bayes Rule equation (equation 2) as

$$= \frac{P(X = 1)P(\text{viral load} = i)}{P(\text{observed count of} \frac{1}{100000})}$$

We can now use the Poisson PMF and our given `prior` to get:

$$= \frac{\frac{\frac{i}{10} e^{\frac{-i}{10}}}{1!} \cdot \texttt{prior[i]}}{P(\text{observed count of} \frac{1}{100000})} \tag{3}$$

We now need to expand our denominator. We can use the General Law of Total Probability to expand

$$P(\text{observed count of} \frac{1}{100000}) = \sum_{j=0}^{99} P(\text{observed count of} \frac{1}{100000} | \text{viral load} = i) P(\text{viral load} = i)$$

We can rewrite this as

$$= \sum_{j=0}^{99} \frac{\frac{j}{10} e^{\frac{-j}{10}}}{1!} \cdot \texttt{prior[j]}$$

$$= \sum_{j=0}^{99} \frac{j}{10} e^{\frac{-j}{10}} \cdot \texttt{prior[j]}$$

And finally, we can plug this into Equation 3 to get

$$\boxed{\frac{\frac{i}{10} e^{\frac{-i}{10}} \cdot \texttt{prior[i]}}{\sum_{j=0}^{99} \frac{j}{10} e^{\frac{-j}{10}} \cdot \texttt{prior[j]}}}.$$

7 Ranked Evaluation [1 point]

Bonus point: please rank the questions from this exam in order of how confident you feel in the correctness of your answer. For example: writing $3 < 1 < 5$ would mean you are most confident in your answer for question 5, second most confident in your answer for question 1, and least confident in your answer for question 3. Please rank whole questions, not subparts.

$$\_\_\_\_ < \_\_\_\_ < \_\_\_\_ < \_\_\_\_ < \_\_\_\_ < \_\_\_\_$$