Chris Piech
CS109

Section #4
Feb 6th, 2025

Section #4 Solutions

Problems by Chris

1 Approximating Normal

Your website has 100 users. Each day, each user independently has a 20% chance of logging into your website. Use a normal approximation to estimate the probability that more than 21 users log in on the same day.

Let B be the number of users that log in. $B \sim \text{Bin}(n = 100, p = 0.2)$. Since n is big enough and p is relatively moderate, B can be approximated with a normal with a matching the mean and variance.

Let C be the normal that approximates B. We have $E[B] = np = 20$ and $Var(B) = np(1-p) = 16$ (these are the formulas for any Binomial). Then, $C \sim N(\mu = 20, \sigma^2 = 16)$.

To find the probability that B is more than 21, we'll use the CDF of the Normal, ϕ, for C. Note that because we are approximating a discrete value with a continuous random variable, we need to continuity correct:

$$\begin{aligned} P(B > 21) \approx P(C > 21.5) &= P\Big(\frac{C - 20}{\sqrt{16}} > \frac{21.5 - 20}{\sqrt{16}}\Big) \\ &= P(Z > 0.375) \\ &= 1 - P(Z < 0.375) \\ &= 1 - \phi(0.375) = 1 - 0.6462 = 0.3538 \end{aligned}$$

2 Conditional Flu

If a person has the flu, the distribution of their temperature is Gaussian with mean 101 and variance 1. If a person does not have the flu, the distribution of their temperature is Gaussian with mean 98 and variance 1. All you know about a person is that they have a temperature of 100. What is the probability they have the flu? Historically, 20% of people you analyze have had the flu.

This is an inference problem, as it involves doing Bayes' Theorem with random variables. We are going to define two random variables:

F is an indicator variable which is 1 if the person has the flu.
X is the distribution of the person's temperature.

The question asks: what is $P(F = 1|X = 100)$?

The problem tells us that $F \sim \text{Bern}(p = 0.2)$ and that X is distributed as following, conditioned on F:

$$X|F = 1 \sim N(\mu = 101, \sigma^2 = 1)$$
$$X|F = 0 \sim N(\mu = 98, \sigma^2 = 1)$$

We can solve this using the inference version of Bayes, which allows for a mixture of discrete and continuous random variables.

$$P(F = 1|X = 100) = \frac{f(X = 100|F = 1)P(F = 1)}{f(X = 100|F = 1)P(F = 1) + f(X = 100|F = 0)P(F = 0)}$$

The next step is to substitute the PDF of the Normal distribution:

$$\begin{aligned} P(F = 1|X = 100) &= \frac{\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(100-101)^2} \cdot 0.2}{\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(100-101)^2} \cdot 0.2 + \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(100-98)^2} \cdot 0.8} \\ &= \frac{e^{-\frac{1}{2}} \cdot 0.2}{e^{-\frac{1}{2}} \cdot 0.2 + e^{-2} \cdot 0.8} \\ &\approx .528 \end{aligned}$$

3 Algorithmic Fairness

An AI model makes a binary prediction (G for guess) for whether a person will repay a loan. We want to know: is the model “fair” with respect to a binary demographic (D for demographic)? To answer this question, let’s analyze the historical predictions of the model and compare the predictions to the true outcome (T for truth). Consider the following joint probability table from the model’s history:

	D = 0		**D = 1**	
	G = 0	G = 1	G = 0	G = 1
T = 0	0.21	0.32	0.01	0.01
T = 1	0.07	0.28	0.02	0.08

D: is the demographic of an individual (binary).
G: is the “repay” prediction made by the algorithm. 1 means predicted repay.
T: is the true “repay” result. 1 means did repay.

Recall that cell (D = i,G = j,T = k) is the probability P(D = i,G = j,T = k).

a. (4 points) What is $P(D = 1)$?

$P(D = 1) = 0.01 + 0.01 + 0.02 + 0.08 = 0.12$

b. (4 points) What is $P(G = 1|D = 1)$?

$P(G = 1|D = 1)$ = (0.01 + 0.08) / 0.12 = 0.75

c. (6 points) Fairness definition 1: Parity
An algorithm satisfies "parity" if the probability that the algorithm makes a positive prediction ($G = 1$) is the same regardless of the demographic variable. Does this algorithm satisfy parity?

We want to see if $P(G = 1|D = 1) = P(G = 1|D = 0)$.
$P(G = 1|D = 0)$ = (0.32 + 0.28) / (0.21 + 0.07 + 0.32 + 0.28) = 0.60/0.88 = 0.68.
Thus, we see that $P(G = 1|D = 1) > P(G = 1|D = 0)$ and the algorithm does not satisfy parity.

d. (6 points) Fairness definition 2: Calibration
An algorithm satisfies "calibration" if the probability that the algorithm is correct ($G = T$) is the same regardless of demographics. Does this algorithm satisfy calibration?

We essentially want to see if $P(G = 0, T = 0|D = 0) = P(G = 0, T = 0|D = 1)$ and if $P(G = 1, T = 1|D = 0) = P(G = 1, T = 1|D = 1)$.

First we check if $P(G = 0, T = 0|D = 0) = P(G = 0, T = 0|D = 1)$.
$P(G = 0, T = 0|D = 0)$ = 0.21 / (0.21 + 0.07 + 0.32 + 0.28) = 0.239
$P(G = 0, T = 0|D = 1)$ = 0.01 / (0.01 + 0.02 + 0.01 + 0.08) = 0.083

So we can see that the algorithm does not satisfy calibration.

e. (6 points) Fairness definition 3: Equality of odds
An algorithm satisfies "equality of odds" if the probability that the algorithm predicts a positive outcome given given that the true outcome is positive ($G = 1|T = 1$) is the same regardless of demographics. Does this algorithm satisfy equality of odds?

$P(G = 1|T = 1, D = 0)$ = 0.28 / (0.28 + 0.07) = 0.8
$P(G = 1|T = 1, D = 1)$ = 0.08 / (0.08 + 0.02) = 0.8

We see that $P(G = 1|T = 1, D = 0) = P(G = 1|T = 1, D = 1)$ and thus, the algorithm does satisfy equality of the odds!

4 (Optional) Daycare.ai

Providing affordable (or better, free) daycare would have a tremendously positive effect on society. California mandates that the ratio of babies to staff must be ≤ 4. We have a challenge: just because a baby is **enrolled**, doesn't mean they will **show up**. At a particular location, 6 babies are enrolled. We estimate that the probability an enrolled child actually shows up on a given day is $\frac{5}{6}$. Assume that babies show up independent of one another.

a. (4 points) What is the probability that either 5 or 6 babies show up?

Let X be the number of babies that show up. $X \sim \text{Bin}(n = 6, p = \frac{5}{6})$.

$$\begin{aligned}
P(X = 5 \text{ or } X = 6) &= P(X = 5) + P(X = 6) \\
&= \binom{n}{5} p^5 (1 - p)^{n-5} + \binom{n}{6} p^6 (1 - p)^{n-6} \\
&= \binom{6}{5} \frac{5}{6}^5 (1 - \frac{5}{6}) + \frac{5}{6}^6
\end{aligned}$$

b. (8 points) If 0 to 4 babies show up our costs are \$200. If 5 or 6 babies show up our costs are \$500. If the daycare charges \$50 per child, what is their expected profit? Recall that Profit = Revenue - Cost. You may leave your answer as an expression with summations.

Let M be our profit. We will use the general formula for expectation:

$$E[M] = \sum_{m \in M} mP(M = m)$$

To plug in, we'll need the probability of each possible outcome. We can find the probability of any specific number of babies showing up based on part a ($P(X = x)$), so we just need to relate these probabilities to profits.
If 4 or fewer babies show up ($X \leq 4$), our profit is $50X - 200$. If 5 or 6 babies show up, our profit is $50X - 500$.

$$\mathrm{E}[M] = \sum_{x=0}^{4} (50x - 200) P(X = x) + \sum_{x=5}^{6} (50x - 500) P(X = x)$$

c. (8 points) Each family is unique. With our advanced analytics we were able to estimate a show-up probability for each of the six enrolled babies: $p_1, p_2, \ldots, p_6$ where p_i is the probability that baby i shows up. Write a new expression for the probability that 5 or 6 babies show up. You may still assume that babies show up independent of one another.

We can't use the binomial distribution here because not all of the trials have the same probability! We must "loop" over them and compute them manually rather than just computing the probability of a particular $X = x$ and multiplying it by some quantity.
Let X be the number of babies who show up and B_i be a Bernoulli variable representing if baby i shows up.

$$\begin{aligned} P(X = 5 \text{ or } X = 6) &= P(X = 5) + P(X = 6) \\ &= \left(\sum_{i=1}^{6} P(B_i = 0) \left[\prod_{j \in \{1,2,\ldots 6\}; i \neq j} P(B_j = 1) \right] \right) + \prod_{i=1}^{6} (P(B_i = 1)) \end{aligned}$$

The first term is for 5 of the 6 showing up, and the second term is for all babies showing up.

5 Midterm Prep Guiding Questions

The midterm exam is coming up. Below are a few broad, guiding questions you might use to help solidify your thinking, prepare a study guide, etc.

1. **Counting** What are event and sample spaces? What's the significance of equally likely events in probability problem-solving? How do we reason differently about distinct vs. indistinct outcomes? What's the difference between combinations and permutations? What are the sum/or rule, step/product rule, inclusion-exclusion, and when do we use them?

2. **Probability Rules** When do we use the definition of conditional probability, the chain rule, the law of total probability, Bayes' theorem, the Complement Rule, DeMorgan's law etc.? What are independence and mutual exclusion?

3. **Random Variables** What is the difference between a random variable and a standard variable? What are expectation and variance, generally? What's the difference between continuous and discrete random variables? We've seen lots of random variables - in which situations would each of them be appropriate? Which ones can be used to approximate others, and in which cases? What's the difference between PMF, PDF, and CDF?

4. **Inference** You want to mix Bayes' theorem and discrete random variables to answer a question of the form: What is the probability that $X = 4$ given that $Y = 2$. How could you solve this problem? What would change if Y or X were continuous?