Chris Piech May 9, 2017

CS109 Midterm Examination

This is a closed calculator/computer exam. You are, however, allowed to use notes in the exam. The last page of the exam is a Standard Normal Table, in case you need it.

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, unless the question specifically asks for a numeric quantity or closed form. Where numeric answers are required, the use of fractions is fine.

Problem	Score
1 (17 pts)	
2 (16 pts)	
3 (26 pts)	
4 (20 pts)	
5 (16 pts)	
6 (25 pts)	
Total (120 pts)	

THE STANFORD UNIVERSITY HONOR CODE

A. The Honor Code is an undertaking of the students, individually and collectively:

(1) that they will not give or receive aid in examinations; that they will not give or receive unpermitted aid in class work, in the preparation of reports, or in any other work that is to be used by the instructor as the basis of grading;

(2) that they will do their share and take an active part in seeing to it that others as well as themselves uphold the spirit and letter of the Honor Code.

B. The faculty on its part manifests its confidence in the honor of its students by refraining from proctoring examinations and from taking unusual and unreasonable precautions to prevent the forms of dishonesty mentioned above. The faculty will also avoid as far as practicable, academic procedures that create temptations to violate the Honor Code.

C. While the faculty alone has the right and obligation to set academic requirements, the students and faculty will work together to create optimal conditions for honorable academic work.

I acknowledge and accept the letter and spirit of the honor code:

Signature: Xiaoqi Zhou.

Name (print): XIAOQI ZHOU

1. **Spotify** (17 points)

You ask Spotify (a music streaming service) to shuffle songs from a playlist with 10 songs by 5 different artists (A, B, C, D and E). There are 2 songs by each artist.

a. (4 points) What is the total number of orderings of the 10 songs (each song is distinct)?

permutation of 10 songs is

$10!$

b. (4 points) Spotify realizes that if it plays two songs from the same artist in a row, users don't trust that the shuffle was truly random. What is the probability that a random ordering of the songs has two songs from artist A in a row?

we can combine 2 songs from artist A and the total number of ordering is

$9!2!$

$P\{2 \text{ songs from artist A in a row}\}$

$= \frac{9!2!}{10!} = \frac{2}{10} = \frac{1}{5}$

c. (5 points) What is the probability that a random ordering of the songs has two songs from artist A in a row **or** two songs from artist B in a row?

A = {2 songs from Artist A in a row}

B = {2 songs from Artist B in a row}

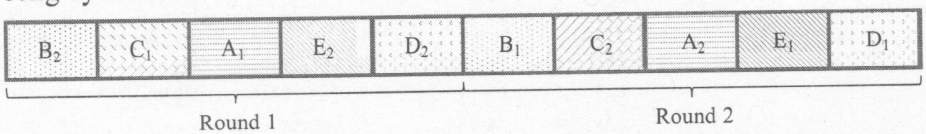
$P(A \cup B) = P(A) + P(B) - P(AB)$

$= \frac{1}{5} + \frac{1}{5} - \frac{8!2!2!}{10!}$

$= \frac{2}{5} - \frac{4}{90} = \frac{16}{45}$

d. (4 points) Spotify tries an algorithm where they play the 10 songs in two rounds:
   i. In the first round, they play one song from each artist.
   ii. In the second round, they play the other song from each artist, preserving the **same artist ordering** as in the first round.

Here is an example of one ordering produced by the algorithm. A_2 is the second song by artist A:



How many ways are there or ordering songs under this new algorithm?

Spotify uses a shuffling algorithm like the one in part (d) for the reason described in (b)

When the first round songs are selected, there are only one ordering available for the second round. We can decide the order for the 5 artists in the first round and there are 2 choise for each artist

ordering number for the new algorithm is

$5! 2^5$

2. **Secure Passwords** (16 points)

You are writing a password manager app that generates random strings of digits 0-9 to use as passcodes.

a) (4 points) How many distinct passcodes of length 8 are possible?

For each number in the passcode there are 10 different choice. number of distinct passcodes is:

10^8

b) (4 points) Suppose your app generates all length-8 passcodes with equal probability. A hacker tries to crack a passcode from this app by brute force, starting with 00000000 and guessing passcodes in sequential order (00000001, 00000002, etc.). If the hacker can try 1 million passcodes in a second, what is the probability that a generated 8-digit passcode is guessed in one second?

The probability is the random passcode matches the first 10^6 trial

$$P = \frac{10^6}{10^8} = \frac{1}{100}$$

c) (8 points) To slow down the hacker you make your website wait 2^n seconds before responding to a login request, where n is the number of times a visitor has incorrectly guessed their password. If the hacker has a 1/1000 chance of correctly guessing a password on each attempt, what is the expected amount of time it will take the hacker to crack a user's password.

X = {the failed trail before the hacker get the password}

$P\{X=n\} = (1-p)^n p$

$$E[X] = \sum_{i=0}^{\infty} 2^n (1-p)^n p$$

$$= \sum_{i=0}^{\infty} (2-2p)^n p$$

Since $p = \frac{1}{1000}$

$E[X] = \infty$

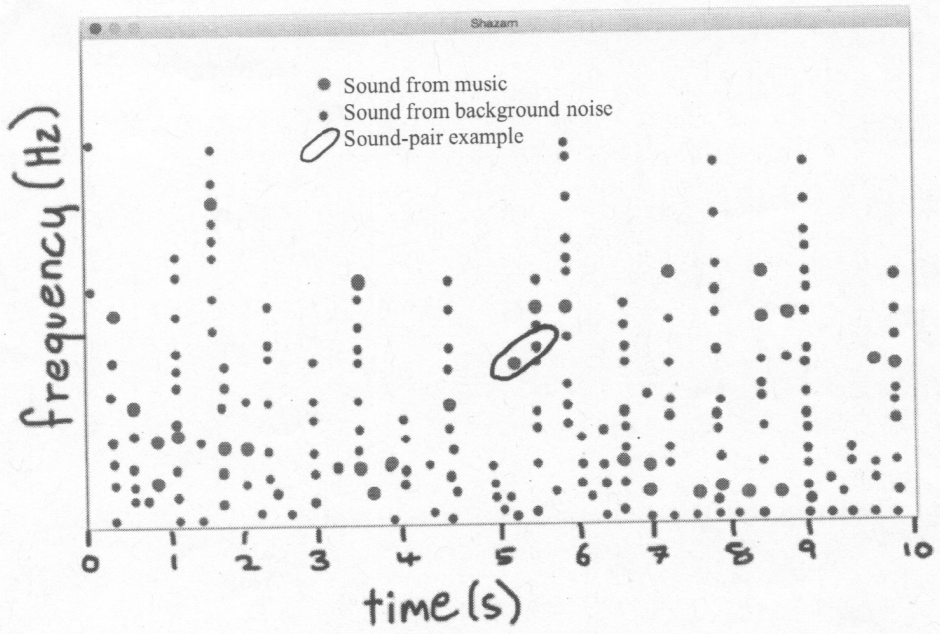
The strategy used in part (d) is a great password protection algorithm.

3. **Shazam** (26 points)

Shazam is an app that listens to a 10 second sample of background music playing (for example in a restaurant) and guesses the song. A user has just sent in a sample which we will use to explore how Shazam works. Within the 10 second sample there are:

50 sounds heard from the background ***music*** and
2000 sounds heard from background ***noise***

A "sound" is a frequency heard at a particular time. Here is a visualization of our 10 second sample (a few sounds are omitted for visual clarity):



a. (3 points) There are 2050 total sounds in the sample. How many distinct ***pairs*** of two sounds are there? Sounds can not be paired with themselves. Sound-pairs with the same two sounds are ***not*** distinct.

The number of distinct sound-pairs is

$$\binom{2050}{2} = \frac{2050 \times 2049}{2} = 1025 \times 2049$$

b. (4 points) How many distinct sound pairs exist such that ***both*** sounds in the sound-pair are from the music (as opposed to being from background noise)?

$$\binom{50}{2} = \frac{49 \times 50}{2} = 25 \times 49$$

c. (5 points) Every distinct pair of sounds from the 10 second sample casts a vote as to what song the pair thinks is playing. If both sounds in a sound-pair are from the music, the pair always casts a vote for the correct song. Otherwise, since at least one sound is from noise, the pair casts a vote uniformly at random from a set of 5 songs (always the same five songs, including the correct song).

We want **more than 1/5 of the total number of votes** to go to the correct song. How many of the pairs containing background noise must vote for the correct song in order for the correct song to get 1/5 of the votes?

total number = 1025×2049

sound-pair with both songs = 25×49

required votes = $(1025 \times 2049 - 25 \times 49) \times \frac{1}{5}$

$= (2000 \times 1000 + 1000 \times 49 + 2000 \times 25) \times \frac{1}{5}$

$= (2099000) \times \frac{1}{5}$

$= 419800$

d. (6 points) Let d be the number of sound-pairs that you calculated in part (c). What is the probability that the correct song receives more than 1/5 of all the votes?

Since the volume of sound pairs are very high, we can use poisson distribution to represent the probability

$X \sim Poi(\lambda)$

$\lambda = pn$ where $n = 2049 \times 1050$ $p = \frac{1}{5}$

$P(X \geq d) = \sum_{i=d}^{\infty} e^{-\lambda} \frac{\lambda^i}{i!}$

e. (8 points) You are at the Tree House and hear music playing. You believe that there is a:

80% chance the song is: Hold Up by Beyonce (event X_1)
20% chance the song is: Can't Get Used by Andy Williams (event X_2)

You run "Shazam" and the app returns that it predicts the Andy Williams song is playing. Let q be the probability that the correct song is returned by Shazam. Let $(1 - q)$ be the probability that an incorrect song is returned. What is your new probability for X_1 and X_2? Express your answer in terms of q.

$A = \{$ Andy Williams song is playing $\}$

$P(A|X_1) = (1-q)$

$P(A|X_2) = q$

$P(AX_1) = P(A|X_1)P(X_1) = (1-q)0.8$

$P(AX_2) = P(A|X_2)P(X_2) = q \times 0.2$

$P(A) = P(AX_1) + P(AX_2) = 0.8 - 0.6q$

$P(X_1|A) = \frac{P(X_1A)}{P(A)} = \frac{(1-q)0.8}{0.8-0.6q}$ $P(X_2|A) = \frac{P(X_2A)}{P(A)} = \frac{0.2q}{0.8-0.6q}$

The new probability is

~~$P(X_1') = P(X_1|A)$~~

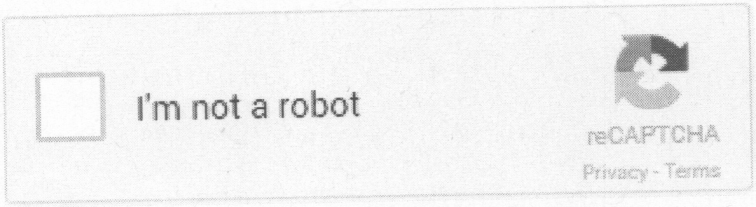
$P(X_1') = P(X_1|A) = \frac{0.8-0.8q}{0.8-0.6q}$

$P(X_2') = P(X_2|A) = \frac{0.2q}{0.8-0.6q}$

This is the basic version of Shazam. Wisdom of the crowds is a phenomenon where a crowd (in this case a crowd of note-pairs) is correct consistently and wrong randomly.

4 **Recaptcha** (20 points)

Based on browser history, Google believes that **there is a 0.2 probability that a particular visitor to a website is a robot.** They decide to give the visitor a recaptcha:



Google presents the visitor with a box, 10 pixels wide by 10 pixels tall. The visitor must click inside the box to show that they are not a robot.

Google has observed that robots click very close to the center of a recaptcha. The distance D of a robot click from the center of the box, in pixels, is normally distributed with mean 0 and variance 2. Humans, on the other hand, click uniformly in the box (all locations are equally likely).

a. (6 points) What is the probability that a robot clicks on a pixel that has a **distance from the center** of the box which is greater than or equal to 1.2 pixels?

$P(x \geq 1.2) = 1 - P(x < 1.2) = 1 - \Phi\left(\frac{1.2-\mu}{\sigma}\right) = 1 - \Phi\left(\frac{1.2-0}{\sqrt{2}}\right)$

$= 1 - \Phi(0.85)$

$= 1 - 0.8023$

$= 0.1977$

b. (6 points) What the the Probability Density Function (PDF) of a human clicking X pixels from the left of the box and Y pixels from the top of the box?

$f_{XY}(x,y) = C$

$\int_{-5}^{5}\int_{-5}^{5} C\,dx\,dy = 1$

$\int_{-5}^{5} 10C\,dy = 1$

$100C = 1$

$C = \frac{1}{100}$

$f_{XY}(x,y) = \frac{1}{100}$

c. (8 points) The visitor clicks in the box at pixel (3,3) which has a distance of 2 pixels from the center of the box. What is Google's **new** belief that the visitor is a robot?

Human $f_H(x,y) = \frac{1}{100}$

Robot $f_R(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$

$P_H(X=3, Y=3) = \frac{1}{100} \cdot \varepsilon = 0.01\varepsilon$

$P_R(d=2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(2-\mu)^2/2\sigma^2} \cdot \varepsilon$

$= \frac{1}{2\sqrt{\pi}} e^{-1} \cdot \varepsilon$

$= 0.1038\varepsilon$

Assume Probability is equally likely on the circle

$P_R(X=3, Y=3) = \frac{P_R(d=2)}{2\pi \cdot d} = \frac{0.1038\varepsilon}{4\pi} = 0.0083$

$P(A|H) = 0.01\varepsilon$

$P(A|R) = 0.0083\varepsilon$

$P(AH) = 0.01\varepsilon \cdot 0.8$

$P(AR) = 0.0083\varepsilon \cdot 0.2$

$P(A) = 0.0097\varepsilon$

$P(H|A) = \frac{P(AH)}{P(A)} = 0.825$

$P(R|A) = \frac{P(RA)}{P(A)} = 0.171$

$P(H') = 0.825$

$P(R') = 0.171$

Recaptcha uses more sophisticated statistics of natural human house gestures and clicks, but this problem covers the central idea behind the new click based recaptchas.

5. Exponentials (16 points)

a. (8 points) What is the probability that an exponential random variable $X \sim \text{Exp}(\lambda)$ takes on a value that is within one standard deviation of its mean?

$\sigma = \frac{1}{\lambda}$

$P(X \leq \frac{1}{\lambda}) = \int_0^{\frac{1}{\lambda}} \lambda e^{-\lambda x} dx$

$= -e^{\lambda x} \Big|_0^{\frac{1}{\lambda}}$

$= -e + 1$

$= 1 - e$

b. (8 points) Let $X_1 \sim \text{Exp}(\lambda_1)$ and $X_2 \sim \text{Exp}(\lambda_2)$ be independent exponential random variables. Let $M = \max(X_1, X_2)$ where max is a function which returns the larger of the two values. Give an expression for the cumulative density function of M.

$P_m(X_m \leq a) = P_1(X_1 \leq a) \cap P_2(X_2 \leq a)$

since X_1 and X_2 are independent

$P_m(X_m \leq a) = P_1(X_1 \leq a) \cdot P_2(X \leq a)$

$F_m(a) = (1 - e^{-\lambda_1 a})(1 - e^{-\lambda_2 a})$

6. Eleven-eleven (25 points)

You are running Alibaba, an online retailer. On the busiest hour of the busiest day, 6pm, November 11th (光棍节), you expect to receive 15 million requests per minute.

Each server you own can handle 10,000 requests in a minute. If the number of requests to your site during any minute is greater than 10,000 times the number of servers, you will "drop" requests. You may assume that the number of requests in any minute is **independent** of the number of requests in any other minute.

a. (5 points) Let p be the probability that you drop a request in any given minute of the busiest hour. What is the maximum value of p such that the probability that you don't drop a single request in the busiest hour is > 0.99?

$(1-p)^{60} > 0.99$

$1-p > 0.99^{\frac{1}{60}}$

$p < 1 - 0.99^{\frac{1}{60}}$

b. (7 points) Write an expression for the **exact** number of machines that you need (K) such that the probability that you don't drop a single request in the busiest hour is > 0.99. Use p from part a. The expression does *not* have to be closed form.

The distribution of request in each minute is

$X \sim poi(\lambda)$ where $\lambda = 15 \times 10^6$

To guarantee that the chance of request over the capablity of server lower than p

$P(X > 1 \times 10^5 K) < p$

$\sum_{i=10^5 K}^{\infty} e^{-\lambda} \frac{\lambda^i}{i!} < p$

c. (7 points) A Poisson random variable with a parameter of $\lambda \geq 1000$, a can be approximated by a normal. The approximating normal should match the mean and the variance of the Poisson. Let $X \sim \text{Poi}(1000)$. What is an approximation for $P(990 < X < 1000)$. Give your answer to two decimal places.

$\mu = E[X] = \lambda$

$\sigma^2 = \text{Var}(X) = \lambda$

$X \sim N(\lambda, \lambda)$

$P(990 < x < 1000) = P(X < 1000) - P(X < 990)$

$= \Phi\left(\frac{1000-\lambda}{\sqrt{\lambda}}\right) - \Phi\left(\frac{990-\lambda}{\sqrt{\lambda}}\right)$

$= 0.5 - \left(1 - \Phi\left(\frac{10}{\sqrt{\lambda}}\right)\right)$

$= 0.5 - \left(1 - \Phi\left(\frac{10}{\sqrt{1000}}\right)\right)$

$= 0.5 - (1 - \Phi(0.32))$

$= 0.5 + 0.6217 - 1 \approx 0.12$

d. (6 points) The probit function, $\Phi^{-1}(x)$ is the inverse of the CDF of a standard normal. It maps from probabilities to the standard normal CDF input that would produce said probability. For example, you can confirm using the Standard Normal Table that $\Phi(0.1) = 0.5398$, so $\Phi^{-1}(0.5398) = 0.1$. Give a closed form expression, using the probit function, for an approximation of K.

$X \sim N(\lambda, \lambda) \quad \lambda = 1.5 \times 10^6$

$P(x > 10^4 k) = 1 - P(X < 10^4 k) =$

~~$= 1 - \Phi\left(\frac{\lambda - 10^4 k}{\sqrt{\lambda}}\right)$~~

~~$= \Phi\left(\frac{10^4 k - \lambda}{\sqrt{\lambda}}\right)$~~

$= 1 - \Phi\left(\frac{10^4 k - \lambda}{\sqrt{\lambda}}\right)$

$1 - \Phi\left(\frac{10^4 k - \lambda}{\sqrt{\lambda}}\right) < P$

$\Phi\left(\frac{10^4 k - \lambda}{\sqrt{\lambda}}\right) > 1 - P$

$\frac{10^4 k - \lambda}{\sqrt{\lambda}} > \Phi^{-1}(1-P)$

$10^4 k - \lambda > \sqrt{\lambda}\,\Phi^{-1}(1-P)$

$k > \frac{\sqrt{\lambda}\,\Phi^{-1}(1-P) + \lambda}{10^4}$