Before reading the solutions, we encourage you to go through the exam first. To help guide your studying, we went through and tagged each of the exam questions with relevant topics. If you are struggling with a particular question, we encourage you to brush up on the tagged topic. Note that these are not comprehensive are rather intended to serve as a starting point for your studying!

As always, please come to Office Hours or ask on Piazza for clarifications! Good luck studying! ☺

CS109 Final Spring 2017

1. Probability Axioms and Counting

2. a. Probability Axioms
2. b. Principle of Inclusion and Exclusion
2. c. Law of Total Probability
2. d. Conditional Probability

3. a. Identifying RVs (Poisson CDF)
3. b. Identifying RVs (Exponential RV)
3. c. Identifying RVs (Normal Approximation of Binomial)

4. a. Probability Computations
4. b. Sums and differences of Normals
4. c. Law of Total Probability

5. a. Joint distributions
5. b. Joint distributions

6. a. Pseudocode
6. b. Pseudocode

7. a. Expectation of RV
7. b. Variance of RV
7. c. Central limit theorem
7. d. Unbiased estimates
7. e. Central Limit Theorem

8. MLE

9. Naïve Bayes

10. MLE

Chris Piech
CS109

CS109 Final

1 Random Encounter Redux

You walk into a gathering at Stanford with R number of Stanford students. What is the probability that you know more than 5 people at the gathering?

Let P be the number of students at Stanford and let F be the number of Stanford students that you know. Assume that each Stanford student is equally likely to be at the gathering.

Solve this problem by recognizing that every combination of students is equally likely. Let E_k be the event that you know exactly k people.

$$P(E_k) = \frac{\binom{P-F}{R-k} \cdot \binom{F}{k}}{\binom{P}{R}}$$

Let X be the number of people that you know.

$$\begin{aligned} P(X > 5) &= 1 - P(X \leq 5) \\ &= 1 - \sum_{i=0}^{5} P(E_i) \\ &= 1 - \sum_{i=0}^{5} \frac{\binom{P-F}{R-i} \cdot \binom{F}{i}}{\binom{P}{R}} \end{aligned}$$

2 Letters of Recommendation

To get a job or internship next summer, you submit recommendation letters from two professors. Unfortunately, you can never be completely certain what your recommendation letters say.

You estimate that if you had two good letters, the probability that you would get the job is 0.75. If you only have one good letter, the probability is 0.2 and if you have no good letters the probability is 0.05.

You believe that the two letters are independent, the probability that the first letter is strong = 0.8 and the probability that the second letter is strong is 0.5.

Let J be the event that you get a job.

a. What is the probability of getting two good letters?

Let T be the event that you receive two good letters.
Let L_1 be the event that the first letter is good.
Let L_2 be the event that the second letter is good.

Since the two events are independent:

$$\begin{aligned} P(T) &= P(L_1) \cdot P(L_2) \\ &= 0.8 \cdot 0.5 = 0.4 \end{aligned}$$

b. What is the probability of getting exactly one good letter?

Let K be the event that you receive exactly one strong letter.

By the inclusion/exclusion principle:

$$\begin{aligned} P(K) &= P(L_1 \cap L_2^C) + P(L_2 \cap L_1^C) \\ &= P(L_1) \cdot P(L_2^C) + P(L_2) \cdot P(L_1^C) \\ &= 0.8 \cdot 0.5 + 0.5 \cdot 0.2 = 0.5 \end{aligned}$$

c. What is the probability of getting the job?

Let Z be the event that you receive exactly zero strong letters. Since the two letters are independent:

$$\begin{aligned} P(Z) &= P(L_1^C) \cdot P(L_2^C) \\ &= 0.2 \cdot 0.5 = 0.1 \end{aligned}$$

By the law of total probability:

$$\begin{aligned} P(J) &= P(J|T)P(T) + P(J|K)P(K) + P(J|Z)P(Z) \\ &= 0.75 \cdot P(T) + 0.2 \cdot P(K) + 0.05 \cdot P(Z) \\ &= 0.75 \cdot 0.4 + 0.2 \cdot 0.5 + 0.05 \cdot 0.1 \end{aligned}$$

d. You got the job. What is the probability that you had two good letters?

$$\begin{aligned} P(T|J) &= \frac{P(J|T)P(T)}{P(J)} \\ &= \frac{0.75 \cdot 0.4}{0.75 \cdot 0.4 + 0.2 \cdot 0.5 + 0.05 \cdot 0.1} \end{aligned}$$

3 Hindenbug

You are testing software and discover that your program has a non-deterministic bug that causes catastrophic failure. Your program was tested for 400 hours and the bug occurred **twice**.

a. Based on the rate of occurrence that you observed, what is the probability that the bug will occur fewer than five times if the program is used for another 400 hours?

Let X be the number of times the bug occurs in 400 hours. $X \sim \text{Poisson}(\lambda = 2)$.

$$P(X < 5) = \sum_{i=0}^{4} P(X = i) = \sum_{i=0}^{4} \frac{2^i}{i!} e^{-2}$$

b. Each user uses your program to complete a three hour long task. If the hindenbug manifests they will immediately stop their work. What is the probability that the bug manifests for a given user?

Let X be the amount of time, in hours, until the bug occurs. $X \sim \text{Exponential}(\lambda = 1/200)$. Let E be the even that a bug manifests for the user.

$$P(E) = P(X < 3) = 1 - e^{-\frac{3}{200}}$$

Alternatively, you can model the probability as

$$P(Y > 0)$$

with

$$Y$$

being the number of times the bug occurs in 3 hours.

c. Your program is used by one million users. Use a normal approximation to estimate the probability that more than 10000 users experience the bug. Let p be the solution to part (b).

Let X be the number of users who experience the bug. $X \sim \text{Bernoulli}(n = 10000, p)$
Let Y be a Normal approximation of X. $Y \sim N(\mu = 10^6 p, \sigma^2 = 10^6 p(1-p))$.

$$\begin{aligned} P(X > 10000) &= 1 - P(X \leq 10000) \\ &\approx 1 - P(Y < 10000.5) \\ &\approx 1 - \phi\left(\frac{10000.5 - 10^6 p}{\sqrt{10^6 p(1-p)}}\right) \end{aligned}$$

4 NBA Finals Week

Recall that a team's ability can be modeled by an *Elo score*, which predicts that if teams A and B have respective Elo scores E_A and E_B, then the probability that A wins a game against B, all else equal, is

$$P(A \text{ wins}) = \frac{1}{1 + 9^{\left(-\frac{E_A - E_B}{400}\right)}}$$

a. Suppose that team A has an Elo rating which is 200 less than the Elo rating for team B. What is the probability that team A wins a game?

$$E_A - E_B = -200$$

$$P(A \text{ wins}) = \frac{1}{1 + 9^{\left(\frac{200}{400}\right)}} = \frac{1}{4} = 0.25$$

b. Suppose the Elo scores of the two teams in the finals are drawn independently from a normal distribution with mean $\mu = 1600$ and variance $\sigma^2 = \frac{200^2}{2}$. What is the probability density function for the difference (D) between their Elo ratings? $D = E_A - E_B$.

Since E_A and E_B are independent Normals, $D \sim N(1600 - 1600, 2 \cdot \frac{200^2}{2}) = N(0, 200^2)$. The PDF is $P(D = d) = \frac{1}{200\sqrt{2\pi}} e^{-\frac{d^2}{2 \cdot 200^2}}$.

c. The difference between the elo scores of two teams is given by the probability density function from part (b). Write an expression for the probability that team A wins. It is ok to have an integral in your answer.

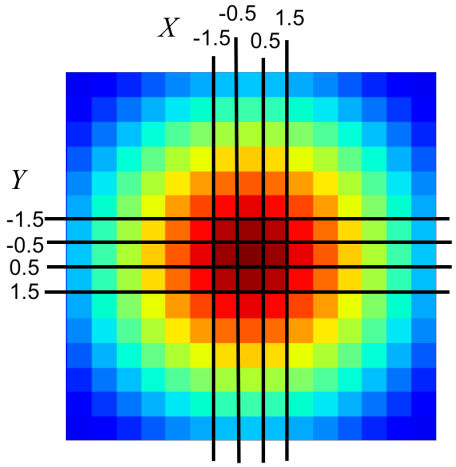
Let W be the event that team A wins. By the total law of probability:

$$\begin{aligned} P(W) &= \int_{-\infty}^{\infty} P(W | D = x) f(D = x) dx \\ &= \int_{-\infty}^{\infty} \frac{1}{1 + 9^{\left(-\frac{x}{400}\right)}} \frac{1}{200\sqrt{2\pi}} e^{\frac{-(x)^2}{2 \cdot 200^2}} dx \end{aligned}$$

This is equal to 1/2, which you could also argue intuitively based on the symmetry between the two teams. However, we did need an explanation for this intuition (in particular, it's incorrect to apply the ELO formula with $D = E[D] = 0$, or to simply find $P(D > 0)$, even though both give the same result).

5 Gaussian Blur

In image processing, a Gaussian blur is the result of blurring an image by a Gaussian function. It is a widely used effect in graphics software, typically to reduce image noise.



Gaussian blurring is based on a joint probability distribution of two **independent** random variables: $X \sim N(0,4)$ and $Y \sim N(0,4)$.

a. Write an expression for $P(X < x, Y < y)$. For full credit your expression should not have integrals.

Since X and Y are independent:

$$\begin{aligned} P(X < x, Y < y) &= P(X < x) \cdot P(Y < y) \\ &= \phi(\frac{x}{2}) \cdot \phi(\frac{y}{2}) \end{aligned}$$

b. Each pixel is given a weight equal to the probability that X and Y are both within the pixel bounds. The center pixel covers the area where $-0.5 \leq x \leq 0.5$ and $-0.5 \leq y \leq 0.5$. What is the weight of the center pixel?

$$\begin{aligned} &P(-0.5 < X < 0.5, -0.5 < Y < 0.5) \\ &= P(X < 0.5, Y < 0.5) - P(X < 0.5, Y < -0.5) - P(X < -0.5, Y < 0.5) + P(X < -0.5, Y < -0.5) \\ &= \phi\left(\frac{0.5}{2}\right) \cdot \phi\left(\frac{0.5}{2}\right) - 2\phi\left(\frac{0.5}{2}\right) \cdot \phi\left(\frac{-0.5}{2}\right) + \phi\left(\frac{-0.5}{2}\right) \cdot \phi\left(\frac{-0.5}{2}\right) \end{aligned}$$

Let $K = \phi(\frac{0.5}{2}) \approx 0.6$

$$\begin{aligned} &P(-0.5 < X < 0.5, -0.5 < Y < 0.5) \\ &= K^2 - 2K(1-K) + (1-K)^2 \\ &\approx (0.6)^2 - 2(0.6)(0.4) + (0.4)^2 \end{aligned}$$

6 Improved Exam Grading

You are experimenting with a new training course to prepare TAs for exam grading. You test your new course by giving the new training to 100 graders (group A) and giving the old, standard training to another set of 100 graders (group B). All 200 graders are asked to grade the same assignment.

The data collected by your experiment are the 100 grades given to the assignment by the graders in group A ($A_1 \ldots A_{100}$), and the 100 grades given by the graders in group B ($B_1 \ldots B_{100}$). You assume that each grade is independent given the grader's group.

You notice that the sample mean of the two groups is about the same. In expectation all graders are accurate. However the sample standard deviation of the grades given by group A was 5 percentage points, whereas the sample standard deviation of grades given by group B was 10 percentage points.

In this question we expect you to write pseudocode. You will be assessed on the quality of your algorithm, not on programming syntax. You may use any of the following methods: { **mean**, **sum** } on a list of numbers as well as any standard programming control flow. You may not refer to any statistics libraries.

a. Provide pseudo code for a method **sampleStandard** that can calculate the unbiased estimate of standard deviation for a list of IID samples $S = [S_1, S_2, \cdots, S_{100}]$.

```
def sampleStandard(S):
    sampleMean = mean(S)
    diffSum = 0
    for Si in S:
        diffSum += (Si - sampleMean)^2
    return sqrt(diffSum / (101))
```

b. Was the difference in standard deviations significant? Write pseudo-code for a method that could return the p-value of the claim. Under the assumption that all 200 grades are identically distributed, calculate the probability of observing a difference in sample standard deviation greater than or equal to five. You may use to the method **sampleStandard** from part (a).

```
def pValue(A, B):
    U = join(A, B)
    count = 0
    repeat(10000):
        sampleA = sampleWithReplace(U, 100)
        sampleB = sampleWithReplace(U, 100)
        stdA = sampleStandard(sampleA)
        stdB = sampleStandard(sampleB)
        # Either one tailed or two tailed is fine!
        if |stdA - stdB| > 5:
            count++
    return count / 10000
```

7 Differential Privacy

You have a dataset that consists of 100 IID values: $X_1 \ldots X_{100}$ where $X_i \sim \text{Bern}(p)$.

A researcher wants to calculate statistics on your data. Since you are mindful about privacy, you decide that you shouldn't give the raw data to the researcher. Instead for every sample X_i, you give the researcher a value Y_i using the following algorithm.

```
# Maximize accuracy, while preserving privacy.
def calculateYi(Xi):
    obfuscate = random()
    if obfuscate:
        return indicator(random())
    else:
        return Xi
```

Where random is a function that returns true or false with equal probability and indicator is a function that returns 1 if the input is true (and 0 otherwise).

a. What is $E[Y_i]$? Give your answer in terms of p. $Y_i \sim \text{Bernoulli}(p_y = 1/4 + p/2)$:

$$E[Y_i] = p_x = 1/4 + p/2$$

b. What is $\text{Var}(Y_i)$? Give your answer in terms of p.

$Y_i \sim \text{Bernoulli}(p_y = 1/4 + p/2)$:

$$\begin{aligned} Var(Y_i) &= p_y(1-p_y) \\ &= (1/4 + p/2)(3/4 - p/2) \\ &= 3/16 + p/4 - p^2/4 \end{aligned}$$

c. Write the distribution of the sample mean $\bar{Y}$ of the samples $Y_1 \ldots Y_{100}$. Explain why the sample mean follows that distribution. Your distribution parameters should be in terms of p.

By the Central Limit Theorem, we know that the sample mean must be normal.
$\bar{Y} \sim N(\mu = E[Y_i], \frac{\text{Var}(Y_i)}{100})$:

$$\begin{aligned} Var(Y_i) &= p_y(1-p_y) \\ &= (1/4 + p/2)(3/4 - p/2) \\ &= 3/16 + p/4 - p^2/4 \end{aligned}$$

d. Given the sample mean $\bar{Y}$, write an expression for an unbiased estimate of p. An unbiased estimate is one where the expectation of your estimate should be equal to the true value.

$$\begin{aligned} \hat{p} &= 2 \cdot \bar{Y} - 1/2 \\ E[\hat{p}] &= 2 \cdot E[\bar{Y}] - 1/2 \\ &= 2(1/4 + p/2) - 1/2 \\ &= 1/2 + p - 1/2 \\ &= p \end{aligned}$$

e. Write an expression for the probability that your estimate from the previous part is more than 0.1 greater than, or less than, the true probability p.

We use the central limit theorem and use the fact that $E[\hat{p}] = p_x$ and $Var(\hat{p}) = 4Var(\bar{Y}) = \frac{4}{100}\sigma_Y^2 = \frac{4}{100}(\frac{3}{16} + \frac{p}{4} - \frac{p^2}{4})$.

$$\begin{aligned} P(|\hat{p} - E[\hat{p}]| \geq 0.1) &= P(\hat{p} \geq p_x + 0.1) + P(\hat{p} \leq p_x - 0.1) \\ &= 2 * P(Z \leq \frac{(p_x - 0.1) - p_x}{\sqrt{\frac{4}{100}\sigma_Y^2}}) \\ &= 2 * P(Z \leq \frac{-0.1 \cdot \frac{10}{2}}{\sigma_Y}) \\ &= 2 * P(Z \leq \frac{-0.5}{\sigma_Y}) \\ &= 2 * \Phi(\frac{-0.5}{\sigma_Y}) = 2 * \Phi(\frac{-0.1}{2\sigma_{\bar{Y}}}) \end{aligned}$$

You could also use Chebyshev's inequality to find an upper bound, where $P(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2}$ for any random variable X with $E[X] = \mu$ and $Var(X) = \sigma^2$.

$$\begin{aligned} P(|\hat{p} - E[\hat{p}]| \geq 0.1) &= P(|\hat{p} - p_x| \geq 0.1) \\ &\leq \frac{4\sigma_Y^2}{100 \cdot 0.1^2} = 4\sigma_Y^2 \end{aligned}$$

8 Windfarm Modeling

In class we saw how climate sensitivity suggests that there is a fierce urgency to developing clean energy solutions. Wind energy presents many opportunities. However, wind is unpredictable and so using and expanding wind energy requires probability theory. The speed of the wind at a windfarm is a random variable that varies as a *Rayliegh Distribution*. A Rayliegh distribution is parameterized by a single scale parameter θ and has the following probability density function.

$$f_X(x) = \begin{cases} \frac{x}{\theta} e^{-x^2/2\theta} & x \geq 0 \\ 0 & else \end{cases}$$

We wish to model the wind speed on a wind farm. To this end we collect N independent measurements of wind speeds $w_1, w_2, \cdots, w_N$. Find a maximum likelihood estimate of θ if we are modeling the wind speed as coming from a Rayleigh distribution

Denote our wind speed as a random variable $W \sim \text{Rayleigh}(\sigma)$. The likelihood is given by

$$L(\sigma^2) = \prod_{i=1}^{N} f_W(w_i; \sigma^2)$$

We then take the logarithm and do some simplification

$$\begin{aligned}
\ell(\sigma^2) &= \log \prod_{i=1}^{N} f_W(w_i;\sigma^2) \\
&= \sum_{i=1}^{N} \log f_W(w_i;\sigma^2) \\
&= \sum_{i=1}^{N} \log \left(\frac{w_i}{\sigma^2} e^{-w_i^2/2\sigma^2} \right) \\
&= \sum_{i=1}^{N} \log w_i - \log \sigma^2 - w_i^2/(2\sigma^2)
\end{aligned}$$

Maximizing over σ^2 we have

$$\begin{aligned}
\arg\max_{\sigma^2} \ell(\sigma^2) &= \arg\max_{\sigma^2} \left[\sum_{i=1}^{N} \log w_i - \sum_{i=1}^{N} \log \sigma^2 - \sum_{i=1}^{N} w_i^2/(2\sigma^2) \right] \\
&= \arg\min_{\sigma^2} \left[N \log \sigma^2 + \frac{1}{2\sigma^2} \sum_{i=1}^{N} w_i^2 \right]
\end{aligned}$$

Taking the derivative with respect to σ^2 and equating to 0 we have

$$\begin{aligned}
0 &= N/\sigma^2 - \frac{1}{2\sigma^4} \sum_{i=1}^{N} w_i^2 \\
N/\sigma^2 &= \frac{1}{2\sigma^4} \sum_{i=1}^{N} w_i^2 \\
\sigma^2 &= \frac{1}{2N} \sum_{i=1}^{N} w_i^2
\end{aligned}$$

9 Multiclass Bayes

In this problem we are going to explore how to write Naive Bayes for multiple output classes.

We want to predict a single output variable Y which represents how a user feels about a book. Unlike in your homework the output variable Y can take on one of four values from the set $\{\text{Like}, \text{Love}, \text{Haha}, \text{Sad}\}$. We will base our predictions off of three binary feature variables $X_1, X_2,$ and X_3 which are indicators of the user's taste. All values $X_i \in \{0,1\}$.

We have access to a dataset with 10,000 users. Each user in the dataset has a value for X_1, X_2, X_3 and Y. You can use a special query method **count** that returns the number of users in the dataset with the given *equality* constraints.

Example usage of **count**:

count$(X_1 = 1, Y = \text{Haha})$	returns the number of users where $X_1 = 1$ and $Y = \text{Haha}$.
count$(Y = \text{Love})$	returns the number of users where $Y = \text{Love}$.
count$(X_1 = 0, X_3 = 0)$	returns the number of users where $X_1 = 0$, and $X_3 = 0$.

You are given a new user with $X_1 = 1$, $X_2 = 1$, $X_3 = 0$. What is the best prediction for how the user will feel about the book? You may leave your answer in terms of an argmax function. You should explain how you would calculate all probabilities used in your expression.

Use Laplace estimates for all probabilities.

$$Y|X_1 = 1, \hat{X_2} = 1, X_3 = 0 = \text{argmax}_y \frac{P(X_1 = 1, X_2 = 1, X_3 = 0|Y = y)P(Y = y)}{P(X_1 = 1, X_2 = 1, X_3 = 0)}$$
$$= P(X_1 = 1, X_2 = 1, X_3 = 0|Y = y)P(Y = y)$$
$$= P(X_1 = 1|Y = y)P(X_2 = 1|Y = y)P(X_3 = 0|Y = y)P(Y = y)$$

$$P(Y = y) = \textbf{count}(Y = y) + 1/10,000 + 4$$
$$P(X_1 = 1, Y = y) = [\textbf{count}(X_1 = 1, Y = y) + 1]/\textbf{count}(Y = y) + 2$$
$$P(X_2 = 1, Y = y) = [\textbf{count}(X_2 = 1, Y = y) + 1]/\textbf{count}(Y = y) + 2$$
$$'P(X_3 = 0, Y = y) = [\textbf{count}(X_3 = 0, Y = y) + 1]/\textbf{countcount}(X_1 = 1, Y = y) + 2$$

10 Logistic Vision Test

You decide that vision tests, given by eye doctors, could have more precise results if we employed some machine learning. In a vision test a user looks at a letter with a particular font size and either correctly guesses the letter, or incorrectly guesses the letter.

You assume that the probability that a particular patient is able to guess a letter correctly is:

$$p = \sigma(\theta - f)$$

Where θ is the user's vision and f is the font size of the letter.

Explain how you could estimate a user's vision θ based on their 20 responses $(f^{(1)}, y^{(1)}) \dots (f^{(20)}, y^{(20)})$, where $y^{(i)}$ is an indicator variable for whether the user correctly identified the ith letter and $f^{(i)}$ is the font size of the ith letter.

Derive any derivatives necessary.

We are going to solve this problem by finding the MLE estimate of θ. To find the MLE estimate, we are going to find the argmax of the log likelihood function. To calculate argmax we are going to use gradient ascent, which requires that we know the partial derivative of the log likelihood function with respect to theta.

First write the log likelihood

$$L(\theta) = \prod_{i=1}^{20} p^{y^{(i)}} (1-p)^{[1-y^{(i)}}$$
$$LL(\theta) = \sum_{i=1}^{20} y^{(i)} log(p) + (1 - (1-p)$$

Then, find the derivative of log likelihood with respect to θ. We first do this for one data point:

$$\frac{\partial LL}{\partial \theta} = \frac{\partial LL}{\partial p} \cdot \frac{\partial p}{\partial \theta}$$

We can calculate both the smaller partial derivatives independently:

$$\frac{\partial LL}{\partial p} = \frac{y^{(i)}}{p} - \frac{1 - y^{(i)}}{1 - p}$$
$$\frac{\partial p}{\partial \theta} = p[1 - p]$$

Putting it all together for one letter:

$$\begin{aligned}\frac{\partial LL}{\partial \theta} &= \frac{\partial LL}{\partial p} \cdot \frac{\partial p}{\partial \theta} \\ &= \Big[\frac{y^{(i)}}{p} - \frac{1-y^{(i)}}{1-p}\Big] p[1-p] \\ &= y^{(i)}(1-p) - p(1-y^{(i)}) \\ &= y^{(i)} - p \\ &= y^{(i)} - \sigma(\theta - f)\end{aligned}$$

For all twenty examples:

$$\frac{\partial LL}{\partial \theta} = \sum_{i=1}^{20} y^{(i)} - \sigma(\theta - f^{(i)})$$