Fall 2023 CS4641/CS7641 Homework 1

Dr. Mahdi Roozbahani

Deadline: Friday, September 22nd, 11:59 pm EST

- No unapproved extension of the deadline is allowed. For late submissions, please refer to the course website.

- Discussion is encouraged on Ed as part of the Q/A. However, all assignments should be done individually.

- Plagiarism is a **serious offense**. You are responsible for completing your own work. You are not allowed to copy and paste, or paraphrase, or submit materials created or published by others, as if you created the materials. All materials submitted must be your own. This also means you may not submit work created by generative models as your own.

- All incidents of suspected dishonesty, plagiarism, or violations of the Georgia Tech Honor Code will be subject to the institute's Academic Integrity procedures. If we observe any (even small) similarities/plagiarisms detected by Gradescope or our TAs, **WE WILL DIRECTLY REPORT ALL CASES TO OSI**, which may, unfortunately, lead to a very harsh outcome. **Consequences can be severe, e.g., academic probation or dismissal, grade penalties, a 0 grade for assignments concerned, and prohibition from withdrawing from the class**.

Instructions

- We will be using Gradescope for submission and grading of assignments.

- **Unless a question explicitly states that no work is required to be shown, you must provide an explanation, justification, or calculation for your answer.** Basic arithmetic can be combined (it does not need to each have its own step); your work should be at a level of detail that a TA can follow it.

- Your write-up must be submitted in PDF form, you may use either Latex, markdown, or any word processing software. We will **NOT** accept handwritten work. Make sure that your work is formatted correctly, for example submit $\sum_{i=0} x_i$ instead of sum_{i=0} x_i.

- **A useful video tutorial on LaTeX has been created by our TA team** and can be found here and an Overleaf document with the commands can be found here.

- Please answer each question on a new page. It makes it more organized to map your answers on GradeScope. When submitting your assignment, you are **required** to correctly map pages of your PDF to each question/subquestion to reflect where they appear. Improperly mapped questions will not be graded correctly.

- All assignments should be done individually, each student must write up and submit their own answers.

- **Graduate Students**: You are required to complete any sections marked as Bonus for Undergrads

Point Distribution

Q1: Linear Algebra [38pts]

- 1.1 Determinant and Inverse of a Matrix [15pts]

- 1.2 Characteristic Equation [8pts]

- 1.3 Eigenvalues and Eigenvectors [15pts]

Q2: Expectation, Co-variance and Statistical Independence [9pts]

- 2.1 Covariance [5pts]

- 2.2 Correlation [4pts]

Q3: Optimization [19pts: 15pts + 4pts Bonus for All]

Q4: Maximum Likelihood [25pts: 10pts + 15pts Bonus for Undergrads]

- 4.1 Discrete Example [10pts]

- 4.2 Poisson Distribution [15pts Bonus for Undergrads]

Q5: Information Theory [32pts]

- 5.1 Marginal Distribution [6pts]

- 5.2 Mutual Information and Entropy [16pts]

- 5.3 Entropy Proofs [10pts]

Q6: Programming [5pts]

Q7: Bonus for All [20pts]

1 Linear Algebra [15pts + 8pts + 15pts]

1.1 Determinant and Inverse of Matrix [15pts]

Given a matrix M:

$$M = \begin{bmatrix} 5 & -2 & 3 \\ -1 & r & 1 \\ 4 & -5 & 2 \end{bmatrix}$$

(a) Calculate the determinant of M in terms of r (calculation process is required). [4pts]

$$|M| = 5 * det\begin{bmatrix} r & 1 \\ -5 & 2 \end{bmatrix} - (-2) * det\begin{bmatrix} -1 & 1 \\ 4 & 2 \end{bmatrix} + 3 * det\begin{bmatrix} -1 & r \\ 4 & -5 \end{bmatrix}$$
$$= 5(2r+5)+2(-2-4)+3(5-4r)$$
$$= -2r + 28$$

(b) For what value(s) of r does M^{-1} not exist? Why doesn't M^{-1} exist in this case? What does it mean in terms of rank and singularity for these values of r? [3pts]

r can not be 14, since that would make the determinant 0. When r is 14, some of the column are linear dependent, which makes the matrix singlar, with a rank of 2, and prevents the matrix from being inverted.

(c) Will all values of r found in part (b) allow for a column (or row) to be expressed as a linear combination of the other columns (or rows)? **If yes**, provide

- **either** the linear equation of the third column C_3 as a linear combination of the first column C_1 and second column C_2

- **or** the linear equation of the second row R_2 as a linear combination of the first row R_1 and third row R_3.

If no, explain why. [3pts]

$C_3 = \frac{11}{17}C_1 + \frac{2}{17}C_2$.

(d) Write down M^{-1} for $r = 0$ (calculation process is **NOT** required). [2pts]

$$M^{-1} = \begin{bmatrix} 5/28 & -11/28 & -1/14 \\ 3/14 & -1/14 & -2/7 \\ 5/28 & 17/28 & -1/14 \end{bmatrix}$$

Change the values in the above matrix.

(e) Find the mathematical equation that describes the relationship between the determinant of M and the determinant of M^{-1}. [3pts]

NOTE: It may be helpful to find the determinant of M and M^{-1} for $r = 0$.

The determinant of the inverse is the inverse of the determinant of the original matrix.

1.2   Characteristic Equation [8pts]

Consider the eigenvalue problem:
$$\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}, \boldsymbol{x} \neq 0$$

where $\boldsymbol{x}$ is a non-zero eigenvector and λ is an eigenvalue of $\boldsymbol{A}$. Prove the determinant $|\boldsymbol{A} - \lambda\boldsymbol{I}| = 0$.

NOTE: There are many ways to solve this problem. You are allowed to use linear algebra properties as part of your solution.

we can write $(\boldsymbol{A} - \lambda\boldsymbol{I})\,\boldsymbol{x} = 0$ which are basic eigenvectors. Suppose the matrix $(\boldsymbol{A} - \lambda\boldsymbol{I})$ is invertible, that means that $(\boldsymbol{A} - \lambda\boldsymbol{I})^{-1}$ exists.

Therefore, $\boldsymbol{X} = \boldsymbol{I}\boldsymbol{X} = ((\boldsymbol{A} - \lambda\boldsymbol{I})^{-1}(\boldsymbol{A} - \lambda\boldsymbol{I}))\boldsymbol{X} = (\boldsymbol{A} - \lambda\boldsymbol{I})^{-1}((\boldsymbol{A} - \lambda\boldsymbol{I})\boldsymbol{X}) = (\boldsymbol{A} - \lambda\boldsymbol{I})^{-1}0 = 0$

This claims that $\boldsymbol{X}$ is 0, but $\boldsymbol{x} \neq 0$, therefore $(\boldsymbol{A} - \lambda\boldsymbol{I})$ can not have an inverse, which means that it's determinant is 0, therefore $|\boldsymbol{A} - \lambda\boldsymbol{I}| = 0$.

1.3 Eigenvalues and Eigenvectors [5+10pts]

1.3.1 Eigenvalues [5pts]

Given the following matrix A, find an expression for the eigenvalues λ of A in terms of a, b, and c. [5pts]

$$A = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$$

$$\begin{bmatrix} a & b \\ b & c \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} = \begin{bmatrix} a - \lambda & b \\ b & c - \lambda \end{bmatrix} = ac - a\lambda - c\lambda + \lambda^2 - b^2$$

$$\lambda_1 = 0.5(a + c - \sqrt{a^2 + 4b^2 - 2ac + c^2})$$
$$\lambda_2 = 0.5(a + c + \sqrt{a^2 + 4b^2 - 2ac + c^2})$$

1.3.2 Eigenvectors [10pts]

Given a matrix A:

$$A = \begin{bmatrix} x & 4 \\ 16 & x \end{bmatrix}$$

(a) Calculate the eigenvalues of A as a function of x (calculation process required). [3pts]

$$\begin{bmatrix} x & 4 \\ 16 & x \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} = \begin{bmatrix} x - \lambda & 4 \\ 16 & x - \lambda \end{bmatrix} = (x - \lambda)^2 = 64$$

$\lambda_1 = x + 8$
$\lambda_1 = x - 8$

(b) Find the normalized eigenvectors of matrix A (calculation process required). [7pts]

$Av = \lambda v$

$$\begin{bmatrix} x - (x + 8) & 4 \\ 16 & x - (x + 8) \end{bmatrix} \begin{bmatrix} v_{1,1} \\ v_{1,2} \end{bmatrix} = \begin{bmatrix} -8 & 4 \\ 16 & -8 \end{bmatrix} \begin{bmatrix} v_{1,1} \\ v_{1,2} \end{bmatrix}$$

-8$v_{1,1} + 4v_{1,2} = 0$,
$8v_{1,1} = 4v_{1,2}$,
$2v_{1,1} = v_{1,2}$
$v_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

$$\begin{bmatrix} x - (x - 8) & 4 \\ 16 & x - (x - 8) \end{bmatrix} \begin{bmatrix} v_{2,1} \\ v_{2,2} \end{bmatrix} = \begin{bmatrix} 8 & 4 \\ 16 & 8 \end{bmatrix} \begin{bmatrix} v_{2,1} \\ v_{2,2} \end{bmatrix}$$

$8v_{2,1} + 4v_{2,2} = 0$,
$8v_{2,1} = -4v_{2,2}$,
$2v_{2,1} = -v_{2,2}$
$v_2 = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$

$$v_1 = \begin{bmatrix} 1/5 \\ 2/5 \end{bmatrix}, v_2 = \begin{bmatrix} -1/5 \\ 2/5 \end{bmatrix}$$

2 Expectation, Co-variance and Statistical Independence [5pts + 4pts]

2.1 Covariance [5pts]

Suppose X, Y, and Z are three different random variables. Let X obey a Bernoulli Distribution. The probability mass function for X is:

$$p(x) = \begin{cases} 0.3 & x = c \\ 0.7 & x = -c \end{cases}$$

where c is a nonzero constant. Let Y obey the Standard Normal (Gaussian) Distribution, which can be written as $Y \sim N(0, 1)$. X and Y are statistically independent (i.e. $P(X|Y) = P(X)$). Meanwhile, let $Z = X - Y$.

Calculate the covariance of Y and Z (i.e. $Cov(Y, Z)$). Do values of c affect the covariance between Y and Z? [5pts]

$E[X] = p, E[Y] = 0$ (normal distribution), $E[XY] = E[X]E[Y]$ (x,y indep)
$Cov(Y, Z) = Cov(Y, X - Y) = E[Y(X - Y)] - E[Y]E[X - Y] = E[YX] - E[Y^2] - E[Y](E[X] - E[Y]) = E[XY] - E[Y^2] - E[X]E[Y] + E[Y]^2 = E[XY] - E[X]E[Y] - (E[Y^2] - E[Y]^2) = E[X]E[Y] - E[X]E[Y] - var(Y) = -1$, so the value of c does not affect the covariance of Y and Z

2.2 Correlation Coefficient [4pts]

Let X and Y be statistically independent random variables with $Var(X) = 4$ and $Var(Y) = 10$. We do not know $E[X]$ or $E[Y]$. Let $Z = 2X + 9Y$. Calculate the correlation coefficient defined as $\rho(X, Z) = \frac{Cov(X,Z)}{\sqrt{Var(X)Var(Z)}}$. If applicable, please round your answer to 3 decimal places. [4pts]

$E[XY] = E[X][Y]$ (x,y indep)

$Var(Z) = Var(2X + 9Y) = Cov(2X + 9Y, 2X + 9Y) = E[(2X + 9Y)(2X + 9Y)] - E[2X + 9Y]E[2X + 9Y] = E[4X^2] + E[81Y^2] + E[36XY] - ((E[2X] + E[9Y])(E[2X] + E[9Y])) = E[4X^2] + E[81Y^2] + E[36XY] - (E[2X]^2 + E[9Y]^2 + 36E[XY]) = 4Var(X) + 81Var(Y)$

$Cov(X, 2X + 9Y) = E[X(2X + 9Y)] - E[X]E[2X + 9Y] = 2E[X^2] + 9E[XY] - 2E[X]^2 - 9E[X]E[Y] = 2E[X^2] - 2E[X]^2 + 9E[X]E[Y] - 9E[X]E[Y] = 9Var(X)$

$\rho(X, Z) = \frac{Cov(X,Z)}{\sqrt{Var(X)Var(Z)}} = \frac{9Var(X)}{\sqrt{Var(x)(4Var(X) + 81Var(Y))}} = \frac{9*4}{\sqrt{4(4*4 + 81*10)}} = 0.626$

3 Optimization [15pts + 4pts Bonus for All]

Optimization problems are related to minimizing a function (usually termed loss, cost or error function) or maximizing a function (such as the likelihood) with respect to some variable x. The Karush-Kuhn-Tucker (KKT) conditions are first-order conditions for a solution in nonlinear programming to be optimal, provided that some regularity conditions are satisfied. In this question, you will be solving the following optimization problem:

$$\min_{x,y} \quad f(x,y) = xy - 4x$$
$$\text{s.t.} \quad g_1(x,y) = x^2 + 3y^2 \leq 18$$
$$g_2(x,y) = y \leq 1$$

(a) Write the Lagrange function for the minimization problem. Now change the minimum function to a maximum function (i.e. $\max_{x,y} \ f(x,y) = xy - 4x$) and provide the Lagrange function for the maximization problem with the same constraints g_1 and g_2. [2pts]

NOTE: The maximization problem is only for part (a).

$\min_{x,y}$	$L(x,y,\lambda_1,\lambda_2) = xy - 4x + \lambda_1(x^2 + 3y^2 - 18) + \lambda_2(y - 1)$
$\max_{x,y}$	$L(x,y,\lambda_1,\lambda_2) = xy - 4x - \lambda_1(x^2 + 3y^2 - 18) - \lambda_2(y - 1)$

(b) List the names of all 4 groups of KKT conditions and their corresponding mathematical equations or inequalities for this specific minimization problem. [2pts]

(a) **stationary condition**
$\frac{\partial L}{\partial x} = y - 4 + 2\lambda_1 x = 0$
$\frac{\partial L}{\partial y} = x + 6 * \lambda_1 * y + \lambda_2 = 0$

(b) **Primal feasibility**
$g_1(x,y) = x^2 + 3y^2 - 18 \leq 0$
$g_2(x,y) = y - 1 \leq 0$

(c) **Dual Feasibility**
$\lambda_1 \geq 0$
$\lambda_2 \geq 0$

(d) **complementary slackness**
$\lambda_1(x^2 + 3y^2 - 18) = 0$
$\lambda_2(y - 1) = 0$

(c) Solve for 4 possibilities formed by each constraint being active or inactive. Do not forget to check the inactive constraints for each point when applicable. Candidate points must satisfy all the conditions mentioned in part b). [7pts]

(a) $\lambda_1 = \lambda_2 = 0$
$y - 4 + 2x(0) = 0 \rightarrow y = 4$
$x + 6y(0) + 0 = 0 \rightarrow x = 0$

(b) $\lambda_1 = 0, y - 1 = 0$
$y = 1$
$1 - 4 - 0 = 0 \rightarrow$ not possible
$x + 6(1)(0) + \lambda_2 = 0 \rightarrow x = \lambda_2$

(c) $x^2 + 3y^2 - 18 = 0, \lambda_2 = 0$
$x = \sqrt{15}, y = -1, \lambda_1 = \frac{\sqrt{3}}{2\sqrt{5}}$ (candidate point)
$x = -\sqrt{15}, y = -1, \lambda_1 = -\frac{\sqrt{3}}{2\sqrt{5}}$ (not a candidate point)

(d) $x^2 + 3y^2 - 18 = 0, y - 1 = 0$
$y = 1 \rightarrow x^2 + 3(1)^2 - 18 = 0 \rightarrow x = \pm\sqrt{15}$
$y = 1, x = \sqrt{15} \rightarrow 1 - 4 + 2\sqrt{15}\lambda_1 = 0 \rightarrow \lambda_1 = \frac{\sqrt{3}}{2\sqrt{5}}$
$y = 1, x = -\sqrt{15} \rightarrow 1 - 4 - 2\sqrt{15}\lambda_1 = 0 \rightarrow \lambda_1 = -\frac{\sqrt{3}}{2\sqrt{5}}$
$y = 1, x = \sqrt{15}, \lambda_1 = \frac{\sqrt{3}}{2\sqrt{5}} \rightarrow \sqrt{15} + 6(\frac{\sqrt{3}}{2\sqrt{5}})(1) + \lambda_2 = 0 \rightarrow \lambda_2 = -\sqrt{15} - \frac{3\sqrt{3}}{\sqrt{5}}$ (not a candidate point)
$y = 1, x = -\sqrt{15}, \lambda_1 = -\frac{\sqrt{3}}{2\sqrt{5}} \rightarrow -\sqrt{15} - 6(\frac{\sqrt{3}}{2\sqrt{5}})(1) + \lambda_2 = 0 \rightarrow \lambda_2 = \sqrt{15} + \frac{3\sqrt{3}}{\sqrt{5}}$ (not a candidate point)

(d) List the candidate point(s) (there is at least 1) obtained from part c). Please round answers to 3 decimal points and use that answer for calculations in further parts. This part can be completed in one line per candidate point. [2pts]

(a) x = 0, y = 4
f(x,y) = (0)(4) - 4(0) = 0

(b) $x = \sqrt{15}, y = -1$
$f(x, y) = (\sqrt{15})(-1) - 4(\sqrt{15}) = -5\sqrt{15} \approx -19.365$

(e) Find the **one** candidate point for which $f(x, y)$ is smallest. Check if $L(x, y)$ is concave or convex at this point by using the Hessian in the second partial derivative test. [2pts]

$\frac{\partial^2 L}{\partial^2 x} = 2\lambda_1 > 0$
convex since second derivative is greater than 0

(f) **BONUS FOR ALL:** Make a 3D plot of the objective function $f(x, y)$ and constraints g_1 and g_2 using Math3d. Mark the minimum candidate point and include a screenshot of your plot. Briefly explain why your plot makes sense in one sentence. Although this is bonus, this is **VERY HELPFUL** in understanding what was accomplished in this problem. [4pts]

NOTE: Use an explicit surface for the objective function, implicit surfaces for the constraints, and a point for the minimum candidate point.
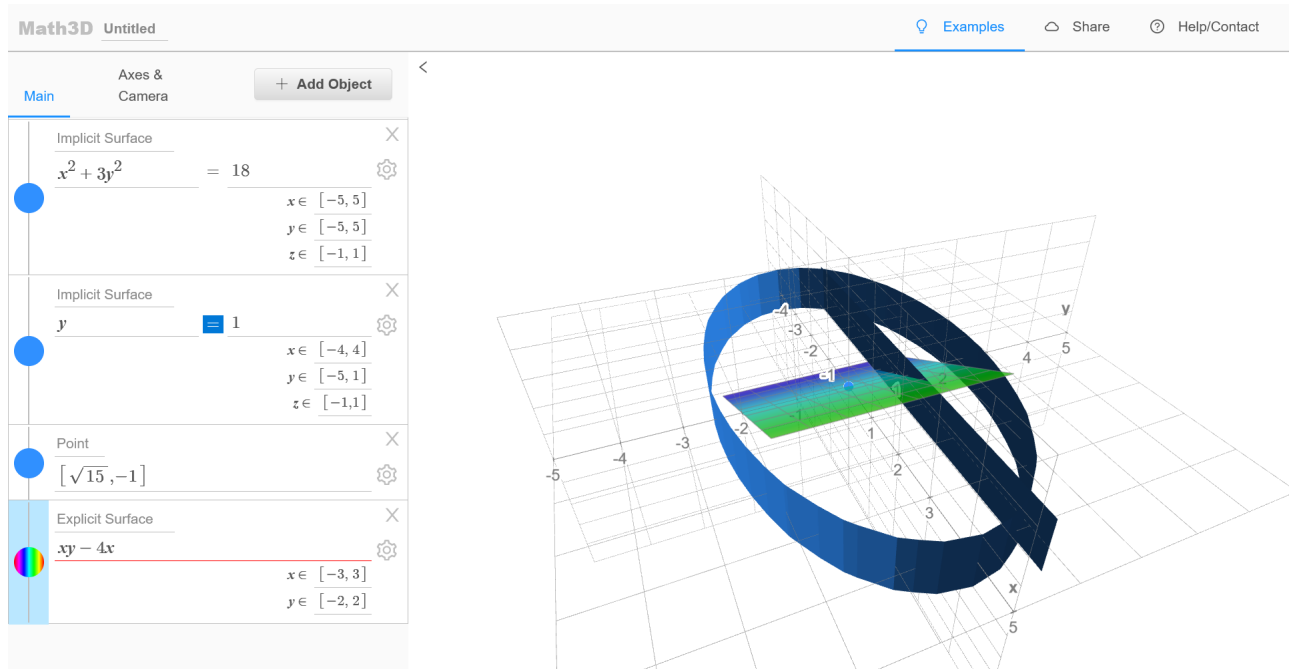
The point we found is on the function and is contained in our 2 constraints which we can see, and while it can not be seen very well we expect that to be the lowest point in our function.

HINT: Read the Example_optimization_problem.pdf in Canvas Files for HW1 to see an example with some explanations.
**HINT:** Click here for an example maximization problem. It's recommended to only watch up until 23:14.
HINT: Click here to determine how to set up the problem for minimization in part (a) and for KKT conditions in part (b).

4 Maximum Likelihood [10pts + 15pts Bonus for Undergrads]

4.1 Discrete Example [10pts]

Mastermind Mahdi decides to give a challenge to his students for their MLE Final. He provides a spinner with 8 sections, each numbered 1 through 8. The students can change the sizes of each section, meaning that they can select the probability the spinner lands on a certain section. Mahdi then proposes that the students will get a 100 on their final if they can spin the spinner 8 times such that it doesn't land on section 1 during the first 7 spins and lands on section 1 on the 8th spin. If the probability of the spinner landing on section 1 is θ, what value of θ should the students select to most likely ensure they get a 100 on their final? Use your knowledge of Maximum Likelihood Estimation to get a 100 on the final.

NOTE: You must specify the log-likelihood function and use MLE to solve this problem for full credit. You may assume that the log-likelihood function is concave for this question
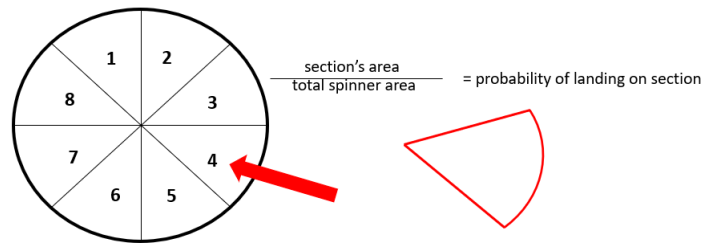
$$L = \log P(X = 8|\theta) + 7 \log P(X \neq 1|\theta)$$
θ is the possibility of landing on section 1

derivative of log-likelihood function
$$\frac{\partial L}{\partial \theta} = \frac{1}{P(X=8|\theta)} * \frac{\partial P(X=8|\theta)}{\partial p} + 7\frac{1}{P(X \neq 1 = 8|\theta)} * \frac{\partial P(X \neq 1|\theta)}{\partial \theta}$$

$\frac{\partial P(X=8|\theta)}{\partial \theta} = \frac{\partial (\theta)}{\partial \theta} = 1$
possibility of spinner landing on 1 on the 8th try, with respect to θ

$\frac{\partial P(X \neq 1|\theta)}{\partial \theta} = \frac{-7(\theta)(1-\theta)}{(1-\theta)^2} = \frac{-7\theta}{1-\theta}$
possibility of spinner not landing on section 1 on the first 7 tries, with respect to θ

$\frac{\partial L}{\partial \theta} = \frac{1}{P(X=8|\theta)} * \frac{\partial P(X=8|\theta)}{\partial \theta} + 7\frac{1}{P(X \neq 1 = 8|\theta)} * \frac{\partial P(X \neq 1|\theta)}{\partial \theta} \rightarrow (\frac{1}{\theta})(1) + 7(\frac{1}{1-\theta})(\frac{-7\theta}{1-\theta}) \rightarrow \frac{1}{\theta} - \frac{49\theta}{(1-\theta)^2} = 0$
setting the derivative to 0 and finding θ gives us the maximum likelihood estimate of θ, which is -1/6 and 1/8, -1/6 is not possible therefore $\theta = 1/8$

If we set section 1 portion so that $\theta = 1/8$, we will get 100 on the final.

4.2   Poisson distribution [15 pts]: Bonus for undergrads

The Poisson distribution is defined as:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} (k = 0, 1, 2, ...).$$

(a) Let $X_1 \sim Poisson(\lambda)$, and X_1 be an observed value of X_1. What is the likelihood given λ? [2 pts]

$\boxed{\lambda = X_1}$

(b) Now, assume we are given n such values. Let $(X_1, ..., X_n) \sim Poisson(\lambda)$ where $X_1, ..., X_n$ are i.i.d. random variables, and $X_1, ..., x_n$ be observed values of $X_1, ..., X_n$. What is the likelihood of this data given λ? You may leave your answer in product form. [3 pts]

$\boxed{L(\lambda) = \Pi_{i=1}^n P(X = k) = \Pi_{i=1}^n \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}}$

(c) What is the maximum likelihood estimator of λ? [10 pts]

$$\frac{\lambda^{\sum_{i=1}^n x_i} e^{-n\lambda}}{\Pi_{i=1}^n (x_i!)}$$

$$\log L(\lambda) = (\sum_{i=1}^n x_i) \log \lambda - n\lambda - \sum_{i=1}^n \log(x_i!)$$

$$\frac{\partial}{\partial \lambda} \log L(\lambda) = \frac{\sum_{i=1}^n x_i}{\lambda} - n = 0$$

$$\lambda = \frac{1}{n} \sum_{i=1}^n x_i$$

5   Information Theory [6pts + 16pts + 10pts]

5.1   Marginal Distribution [6pts]

Suppose the joint probability distribution of two binary random variables X and Y are given as follows. X are the rows, and Y are the columns.

Y X	0	1
0	$\frac{6}{16}$	$\frac{3}{16}$
1	$\frac{1}{4}$	$\frac{3}{16}$

(a) Show the marginal distribution of X and Y, respectively. [3pts]

Y X	0	1	Marginal distribution of Y
0	$\frac{6}{16}$	$\frac{3}{16}$	$\frac{9}{16}$
1	$\frac{1}{4}$	$\frac{3}{16}$	$\frac{7}{16}$
Marginal distribution of X	$\frac{5}{8}$	$\frac{3}{8}$	

(b) Find mutual information $I(X, Y)$ for the joint probability distribution in the previous question to at least 5 decimal places (please use base 2 to compute logarithm) [3pts]

$I(X, Y)$
$= P(0,0) * \log_2 \frac{P(0,0)}{P(X=0)P(Y=0)} + P(0,1) * \log_2 \frac{P(0,1)}{P(X=0)P(Y=1)} + P(1,0) * \log_2 \frac{P(1,0)}{P(X=1)P(Y=0)} + P(1,1) * \log_2 \frac{P(1,1)}{P(X=1)P(Y=1)}$

$= \frac{6}{16} * \log_2 \frac{\frac{6}{16}}{\frac{5}{8}\frac{9}{16}} + \frac{3}{16} * \log_2 \frac{\frac{3}{16}}{\frac{5}{8}\frac{7}{16}} + \frac{1}{4} * \log_2 \frac{\frac{1}{4}}{\frac{3}{8}\frac{9}{16}} + \frac{3}{16} * \log_2 \frac{\frac{3}{16}}{\frac{3}{8}\frac{7}{16}}$

≈ 0.03025

5.2 Mutual Information and Entropy [16pts]

A recent study has shown symptomatic infections are responsible for higher transmission rates. Using the data collected from positively tested patients, we wish to determine which feature(s) have the greatest impact on whether or not some will present with symptoms. To do this, we will compute the entropies, conditional entropies, and mutual information of select features. Please use base 2 when computing logarithms.

ID	Vaccine Doses (X_1)	Wears Mask? (X_2)	Underlying Conditions (X_3)	Symptomatic (Y)
1	H	F	T	T
2	H	F	F	F
3	H	F	T	F
4	M	F	T	F
5	L	T	T	T
6	L	T	F	F
7	L	T	F	T
8	L	T	T	T
9	L	T	T	T
10	M	T	T	T

Table 1: Vaccine Doses: {(H) booster, (M) 2 doses, (L) 1 dose, (T) True, (F) False}

(a) Find entropy $H(Y)$ to at least 3 decimal places. [3pts]

True	6
False	4

$H(Y) = -P(T)\log_2(P(T)) - P(F)\log_2(P(F)) = -(6/10)\log_2(6/10) - (4/10)\log_2(4/10) \approx 0.971$

(b) Find conditional entropy $H(Y|X_1)$ and $H(Y|X_3)$ to at least 3 decimal places. [7pts]

Y \ X_1	H	M	L
T	1	1	4
F	2	1	1
M_{X_1}	3	2	5

$H(Y|X_1)$
$= P(T)H(Y|X_1 = H) + P(T)H(Y|X_1 = M) + P(T)H(Y|X_1 = L) + P(F)H(Y|X_1 = H) + P(F)H(Y|X_1 = M) + P(F)H(Y|X_1 = L)$
$= P(T, H) \log_2 \frac{P(x_1=H)}{P(T, H)} + P(F, H) \log_2 \frac{P(x_1=H)}{P(F, H)} + P(T, M) \log_2 \frac{P(x_1=M)}{P(T, M)} + P(F, M) \log_2 \frac{P(x_1=M)}{P(F, M)} + P(T, L) \log_2 \frac{P(x_1=L)}{P(T, L)} + P(F, L) \log_2 \frac{P(x_1=L)}{P(F, L)}$
$= 0.1 \log_2 \frac{0.3}{0.1} + 0.2 \log_2 \frac{0.3}{0.2} + 0.1 \log_2 \frac{0.2}{0.1} + 0.1 \log_2 \frac{0.2}{0.1} + 0.4 \log_2 \frac{0.5}{0.4} + 0.1 \log_2 \frac{0.5}{0.1}$
≈ 0.836

Y \ X_3	T	F
T	5	1
F	2	2
M_{X_3}	7	3

$H(Y|X_3)$
$= P(T)H(Y|X_3 = T) + P(T)H(Y|X_3 = F) + P(F)H(Y|X_3 = T) + P(F)H(Y|X_3 = F)$
$= P(T,T) \log_2 \frac{P(x_3=T)}{P(T,T)} + P(T, F) \log_2 \frac{P(x_3=F)}{P(T, F)} + P(F,T) \log_2 \frac{P(x_3=T)}{P(F,T)} + P(1,1) \log_2 \frac{P(x_3=F)}{P(F,F)}$
$= 0.5 \log_2 \frac{0.7}{0.5} + 0.1 \log_2 \frac{0.3}{0.1} + 0.2 \log_2 \frac{0.7}{0.2} + 0.2 \log_2 \frac{0.3}{0.2}$
≈ 0.880

(c) Find mutual information $I(X_1, Y)$ and $I(X_3, Y)$ to at least 3 decimal places and determine which one (X_1 or X_3) is more informative. [3pts]

$I(X_1, Y) = H(Y) - H(Y|X_1) \approx 0.971 - 0.836 \approx 0.135$
$I(X_3, Y) = H(Y) - H(Y|X_3) \approx 0.971 - 0.880 \approx 0.091$
X_3 is more informative since it have a larger reduction in entropy

(d) Find joint entropy $H(Y, X_2)$ to at least 3 decimal places. [3pts]

Y \ X_2	T	F
T	5	1
F	1	3

$H(Y, X_2) = \Sigma_{Y, X_2} P(Y, X_2) * \log_2 \frac{1}{P(Y, X_2)} = H(5/10, 1/10, 1/10, 3/10) \approx 1.685$

5.3 Entropy Proofs [10pts]

(a) Write the discrete case mathematical definition for $H(X|Y)$ and $H(X)$. [3pts]

$H(X|Y) =$

$$\sum_{y \in Y} = p(y)H(X|Y = y) = \sum_{x \in X, y \in Y} P(x,y) \log \frac{p(y)}{p(x,y)}$$

$H(X) =$

$$\sum_{x} P(x) * \log \frac{1}{P(x)}$$

(b) **Using the mathematical definition of $H(X)$ and $H(X|Y)$ from part (a)**, prove that $I(X,Y) = 0$ if X and Y are statistically independent. (Note: you must provide a mathematical proof and cannot use the visualization shown in class found here) [7pts]

Start from: $I(X,Y) = H(X) - H(X|Y)$

$$I(X,Y) = H(X) - H(X|Y)$$

$$= \sum_{x} P(x) * \log \frac{1}{P(x)} - \sum_{x \in X, y \in Y} P(x,y) \log \frac{P(y)}{P(x,y)}$$

$$= \sum_{x \in X, y \in Y} P(x,y) * \log \frac{P(x,y)}{P(x)P(y)}$$

So I(X,Y)=0 when, at all points, either:

- $P(x,y) = 0$
- $\log \frac{P(x,y)}{P(x)P(y)} = 0$

the first case, states that the joint possibility is 0, which is that x and y can never happen together, while the second case needs $\frac{P(x,y)}{P(x)P(y)} = 1$, which implies $P(x,y) = P(x)P(y)$, which is independence, where both events can happen together. Therefore, I(X,Y) = 0, if X and Y are independent

6 Programming [5 pts]

See the Programming subfolder in Canvas.

7   Bonus for All [20 pts]

(a) Let X, Y be two statistically independent $N(0,1)$ random variables, and P, Q be random variables defined as:

$$P = 5X + 3XY^2$$
$$Q = X$$

Calculate the variance $Var(P+Q)$. *(This question may take substantial work to support, e.g. 25 to 30 lines)* [10pts]

HINT: The following equality may be useful: $Var(XY) = E[X^2Y^2] - [E(XY)]^2$
HINT: $E[Y^4] = \int_{-\infty}^{\infty} y^4 f_Y(y) dy$ where $f_Y(y)$ is the probability density function of Y.
HINT: $Var(P+Q) = Var(P) + Var(Q) + 2Cov(P,Q)$ may be a good starting point.

$\mu = 0, E[X] = \mu$ (normal distribution)
$E[Y] = 0, E[X] = 0$ (normal distribution)
$Var[X] = 1, Var[Y] = 1$ (normal distribution)
$E[XY] = E[X]E[Y]$ (X,Y indep)

$Var(P + Q) = Var(P) + Var(Q) + 2Cov(P,Q)$
$Var(P + Q) = Var(5X + 3XY^2) + Var(X) + 2Cov(5X + 3XY^2, X)$

$Var(5X + 3XY^2) = Var(5X) + Var(3XY^2) + 2Cov(5X, 3XY^2)$
$Var(5X) = 25Var(X) = 25(E[X^2] - E[X]^2) = 25$
$Var(3XY^2) = 9Var(XY^2) = 9(E[X^2Y^4] - [E(XY^2)]^2 = 9(E[X^2]E[Y^4] - (E[X]E[Y^2])^2) = 9E[X^2]E[Y^4]$
$E[Y^4] = \int_{-\infty}^{\infty} \frac{y^4}{\sqrt{2\pi}} exp(\frac{-y^2}{2}) dy = 3$
$E[X^2] = \int_{-\infty}^{\infty} \frac{x^2}{\sqrt{2\pi}} exp(\frac{-x^2}{2}) dy = 1$
$Var(3XY^2) = 9Var(XY^2) = 9E[X^2]E[Y^4] = 9 * 3 * 1 = 27$
$2Cov(5X, 3XY^2) = 2(E[5X * 3XY^2] - E[5X]E[3XY^2]) = 2(15E[X^2]E[Y^2] - 15E[X]^2E[Y^2]) = 2(15 * 1 * 1 - 15 * 0^2 * 1) = 30$
$2Cov(5X + 3XY^2, X) = 2(E[X(5X + 3XY^2)] + E[5X + 3XY^2]E[X]) = 2(E[5X^2 + 3X^2Y^2] + E[5X + 3XY^2] * 0) = 2(5E[X^2] + 3E[X^2][Y^2]) = 2(5 * 1 + 3 * 1 * 1) = 16$

$Var(P + Q) = Var(5X) + Var(3XY^2) + 2Cov(5X, 3XY^2) + Var(X) + 2Cov(5X + 3XY^2, X) \rightarrow$
$25 + 27 + 30 + 1 + 16 = 99$

(b) Suppose that X and Y have joint pdf given by:

$$f_{X,Y}(x,y) = \begin{cases} \dfrac{x}{2}e^{-y} & 0 \le x \le 2, y \ge 0 \\ 0 & otherwise \end{cases}$$

What are the marginal probability density functions for X and Y? *(It is possible to thoroughly support your answer to this question in 8 to 10 lines)* [5 pts]

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_0^{\infty} \frac{x}{2}e^{-y}dx = -\frac{x}{2}e^{-y}\,|_0^{\infty} = \frac{x}{2}$$

$$f_X(x) = \begin{cases} \dfrac{x}{2} & 0 \le x \le 2 \\ 0 & otherwise \end{cases}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx = \int_0^2 \frac{x}{2}e^{-y}dx = \frac{x^2}{4}e^{-y}\,|_0^2 = e^{-y}$$

$$f_Y(y) = \begin{cases} e^{-y} & y \ge 2 \\ 0 & otherwise \end{cases}$$

(c) A person decides to toss a biased coin with $P(heads) = 0.2$ repeatedly until he gets a head. He will make at most 5 tosses. Let the random variable Y denote the number of heads. Find the probability distribution of Y. Then, find the variance of Y. *(It is possible to thoroughly support your answer to this question in 5 to 10 lines)* [5 pts]

$P(Y = 0) = \binom{5}{0} * 0.2^0 * (1 - 0.2)^5 = 0.32768$
$P(Y = 1) = \binom{5}{1} * 0.2^1 * (1 - 0.2)^4 = 0.4096$
$P(Y = 2) = \binom{5}{2} * 0.2^2 * (1 - 0.2)^3 = 0.2048$
$P(Y = 3) = \binom{5}{3} * 0.2^3 * (1 - 0.2)^2 = 0.0512$
$P(Y = 4) = \binom{5}{4} * 0.2^4 * (1 - 0.2)^1 = 0.0064$
$P(Y = 5) = \binom{5}{5} * 0.2^5 * (1 - 0.2)^0 = 0.00032$
$\mu = 0.32768 * 0 + 0.4096 * 1 + 0.2048 * 2 + 0.0512 * 3 + 0.0064 * 4 + 0.00032 * 5 = 1$
$var(Y) = \sum(Y - \mu)^2 P(Y) = 0.32768 * (0 - 1)^2 + 0.4096 * (1 - 1)^2 + 0.2048 * (2 - 1)^2 + 0.0512 * (3 - 1)^2 +$
$0.0064 * (4 - 1)^2 + 0.00032 * (5 - 1)^2 = 0.8$