Spring 2024 CS4641/CS7641 Homework 1

Dr. Mahdi Roozbahani

Deadline: Friday, February 9th, 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.

- When submitting your assignment on Gradescope, **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 [15pts: 15pts + 2% Bonus for All]

Q4: Maximum Likelihood [25pts: 10pts + 15 pts Grad/6% Bonus for Undergrads]

- 4.1 Discrete Example [10pts]

- 4.2 Poisson Distribution [15pts Grad / 6% 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: Ethical Implications on Decision-Making [5 pts]

Q7: Programming [5pts]

Q8: Bonus for All [8%]

Points Totals:

- **Total Base:** 114 pts

- **Total Undergrad Bonus:** 6%

- **Total Bonus for All:** 10%

- **Total Possible Assignment Grade (Undergrad):** 116%

- **Total Possible Assignment Grade (Grad):** 110%

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

1.1 Determinant and Inverse of Matrix [15pts]

Given a matrix M:

$$M = \begin{bmatrix} 2 & 0 & 1 \\ -1 & r & -3 \\ 2 & 5 & -4 \end{bmatrix}$$

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

$$|\boldsymbol{M}| = 2 \times (-4r + 15) - 0 \times (4 + 6) + 1 \times (-5 - 2r) = -8r + 30 - 5 - 2r = \boxed{-10r + 25}$$

(b) For what value(s) of r does $\boldsymbol{M}^{-1}$ not exist? Why doesn't $\boldsymbol{M}^{-1}$ exist in this case? What does it mean in terms of rank and singularity for these values of r? *This question can be answered in less than 7 lines.* [3pts]

At $\boxed{r = \dfrac{5}{2}}$, the determinant of $\boldsymbol{M}$ is zero, indicating that $\boldsymbol{M}$ is singular. To have an inverse matrix, the determinant of $\boldsymbol{M}$ must be nonzero, which it is not in this question. This means that at $r = \frac{5}{2}$, $\boldsymbol{M}$ does not have an inverse and its rank is less than its dimension of 3, or simply said it is not full rank.

(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 second column $\boldsymbol{C_2}$ as a linear combination of the first column $\boldsymbol{C_1}$ and third column $\boldsymbol{C_3}$
- **or** the linear equation of the third row $\boldsymbol{R_3}$ as a linear combination of the first row $\boldsymbol{R_1}$ and second row $\boldsymbol{R_2}$.
- (calculation process is **NOT** required).

If no, explain why. [3pts]

$$\boxed{\textbf{Yes}, C_2 = \frac{1}{2}C_1 - 1C_3}$$

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

$$M^{-1} = \begin{bmatrix} \frac{3}{5} & \frac{1}{5} & 0 \\ -\frac{2}{5} & -\frac{2}{5} & \frac{1}{5} \\ -\frac{1}{5} & -\frac{2}{5} & 0 \end{bmatrix}$$

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

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

$$\boxed{\det(\boldsymbol{M}^{-1}) = \frac{1}{\det(\boldsymbol{M})}}$$

1.2 Characteristic Equation [8pts]

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

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.

Consider the eigenvalue problem: $\boldsymbol{Ax} = \lambda\boldsymbol{x}, \quad \boldsymbol{x} \neq 0$ where $\boldsymbol{x}$ is a non-zero eigenvector and λ is an eigenvalue of $\boldsymbol{A}$. To prove that $\det(\boldsymbol{A} - \lambda\boldsymbol{I}) = 0$:

1. Start by rearranging the eigenvalue equation:
$\boldsymbol{Ax} = \lambda\boldsymbol{x}$
can be rewritten as
$(\boldsymbol{A} - \lambda\boldsymbol{I})\boldsymbol{x} = \boldsymbol{0}$
Here, $\boldsymbol{I}$ is the identity matrix of the same dimensions as $\boldsymbol{A}$.

2. Observe that the equation $(\boldsymbol{A} - \lambda\boldsymbol{I})\boldsymbol{x} = \boldsymbol{0}$ implies that the vector $\boldsymbol{x}$ is in the null space of the matrix $\boldsymbol{A} - \lambda\boldsymbol{I}$. This means that the matrix $\boldsymbol{A} - \lambda\boldsymbol{I}$ does not have full rank and hence is not invertible.

3. From linear algebra, we know that a square matrix is invertible if and only if its determinant is non-zero. Consequently, for a matrix that is not invertible, its determinant must be zero.

4. Therefore, since $\boldsymbol{A} - \lambda\boldsymbol{I}$ is not invertible, we have $\det(\boldsymbol{A} - \lambda\boldsymbol{I}) = 0$ This completes the proof that the determinant of $\boldsymbol{A} - \lambda\boldsymbol{I}$ must be zero for λ to be an eigenvalue of $\boldsymbol{A}$.

1.3 Eigenvalues and Eigenvectors [5+10pts]

1.3.1 Eigenvalues [5pts]

Given the following matrix $\boldsymbol{A}$, find an expression for the eigenvalues λ of $\boldsymbol{A}$ in terms of a, b, and c. Express your answer in the form $\lambda = ...$ [5pts]

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

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

$$\boldsymbol{A} - \lambda\boldsymbol{I} = \begin{bmatrix} a - \lambda & b \\ b & c - \lambda \end{bmatrix}$$

$$\det(\boldsymbol{A} - \lambda\boldsymbol{I}) = (a - \lambda)(c - \lambda) - b^2$$

$$(a - \lambda)(c - \lambda) - b^2 = 0$$

$$\lambda^2 - (a + c)\lambda + ac - b^2 = 0$$

$$\lambda = \frac{(a+c) \pm \sqrt{(a+c)^2 - 4(ac - b^2)}}{2}$$

$$\lambda = \frac{(a+c) \pm \sqrt{a^2 + 2ac + c^2 - 4ac + 4b^2}}{2}$$

$$\lambda = \frac{(a+c) \pm \sqrt{a^2 - 2ac + c^2 + 4b^2}}{2}$$

$$\boxed{\lambda = \frac{(a + c) \pm \sqrt{(a - c)^2 + (2b)^2}}{2}}$$

1.3.2 Eigenvectors [10pts]

Given a matrix $\boldsymbol{A}$:

$$\boldsymbol{A} = \begin{bmatrix} x & 9 \\ 4 & x \end{bmatrix}$$

(a) Calculate the eigenvalues of $\boldsymbol{A}$ as a function of x (calculation process required). [3pts]

$$\det\left(\begin{bmatrix} x & 9 \\ 4 & x \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}\right) = 0$$

$$\det\left(\begin{bmatrix} x - \lambda & 9 \\ 4 & x - \lambda \end{bmatrix}\right) = 0$$

$$(x - \lambda)^2 - 36 = 0$$

$$(x - \lambda) = \pm 6$$

$$\boxed{\lambda = x - 6, x + 6}$$

(b) Find the normalized eigenvectors of matrix $\boldsymbol{A}$ (calculation process required). [7pts]

$$\text{For } \lambda = x - 6:$$

$$\begin{bmatrix} 6 & 9 \\ 4 & 6 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$6t + 9v_2 = 0 \implies v_2 = -\frac{2}{3}t$$

$$\boldsymbol{v} = t \begin{bmatrix} 1 \\ -\frac{2}{3} \end{bmatrix}$$

$$\|\boldsymbol{v}\| = \sqrt{t^2 + \left(-\frac{2}{3}t\right)^2}$$

$$\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|} = \frac{1}{\sqrt{1 + \left(-\frac{2}{3}\right)^2}} \begin{bmatrix} 1 \\ -\frac{2}{3} \end{bmatrix}$$

$$\boxed{\begin{bmatrix} \frac{3}{\sqrt{13}} \\ -\frac{2}{\sqrt{13}} \end{bmatrix}}$$

For $\lambda = x + 6$:

$$\begin{bmatrix} -6 & 9 \\ 4 & -6 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$-6s + 9v_2 = 0 \implies v_2 = \tfrac{2}{3}s$$

$$\boldsymbol{v} = s \begin{bmatrix} 1 \\ \tfrac{2}{3} \end{bmatrix}$$

$$\|\boldsymbol{v}\| = \sqrt{s^2 + \left(\tfrac{2}{3}s\right)^2}$$

$$\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|} = \frac{1}{\sqrt{1 + \left(\tfrac{2}{3}\right)^2}} \begin{bmatrix} 1 \\ \tfrac{2}{3} \end{bmatrix}$$

$$\boxed{\begin{bmatrix} \tfrac{3}{\sqrt{13}} \\ \tfrac{2}{\sqrt{13}} \end{bmatrix}}$$

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 two point Distribution. The probability mass function for X is:

$$p(x) = \begin{cases} 0.4 & x = c \\ 0.6 & x = -c \end{cases}$$

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 $Z = XY$.

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

For $Cov(Y,Z)$, where $Z = XY$, we have:

$$Cov(Y,Z) = E[Y \cdot XY] - E[Y]E[XY]$$

$$Cov(Y,Z) = E[Y^2 X] - E[Y]E[XY]$$

Since X and Y are independent, $E[XY] = E[X]E[Y]$. Given $Y \sim N(0,1)$, $E[Y] = 0$, so $E[XY] = E[X] \cdot 0 = 0$. Thus, the equation simplifies to:

$$Cov(Y,Z) = E[Y^2 X] - 0 \cdot 0$$

$$Cov(Y,Z) = E[Y^2 X]$$

Y^2 is independent of X, so $E[Y^2 X] = E[Y^2]E[X]$. For a standard normal variable, $E[Y^2] = Var(Y) = 1$. The expectation of X is calculated from its probability mass function:

$$E[X] = 0.4c + 0.6(-c) = -0.2c$$

Therefore, the covariance becomes:

$$Cov(Y,Z) = 1 \cdot (-0.2c) = -0.2c$$

Thus, the covariance between Yand Z is affected by the value of cand is proportional to -0.2c .

2.2 Correlation Coefficient [4pts]

Let X and Y be statistically independent random variables with $Var(X) = 3$ and $Var(Y) = 6$. We do not know $E[X]$ or $E[Y]$. Let $Z = 2X + 3Y$. 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]

$\text{Cov}(X, Z) = \text{Cov}(X, 2X + 3Y) = 2\text{Var}(X) + 3\text{Cov}(X, Y) = 2\text{Var}(X) = 2 \cdot 3 = 6$

$\text{Var}(Z) = \text{Var}(2X + 3Y) = 2^2\text{Var}(X) + 3^2\text{Var}(Y) = 4 \times 3 + 9 \times 6 = 12 + 54 = 66$

$\rho(X, Z) = \frac{\text{Cov}(X,Z)}{\sqrt{\text{Var}(X) \cdot \text{Var}(Z)}} = \frac{6}{\sqrt{3 \cdot 66}} = \frac{6}{\sqrt{198}}$

$\boxed{\rho(X, Z) \approx 0.426}$

3   Optimization [15pts + 2% 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:

$$\min_{x,y} \quad f(x,y) = xy - 8y$$
$$\text{s.t.} \quad g_1(x,y) = 2x^2 + y^2 \leq 20$$
$$g_2(x,y) = x \leq 1$$

(a) Write the Lagrange function for the minimization problem. Now change the minimum function to a maximum function (i.e. $\max_{x,y} \ f(x,y) = xy - 8y$) 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).

The Lagrange function for the given minimization problem is:
$L(x,y,\lambda_1,\lambda_2) = xy - 8y + \lambda_1(2x^2 + y^2 - 20) + \lambda_2(x - 1)$

The Lagrange function for the given maximization problem is:
$L(x,y,\lambda_1,\lambda_2) = xy - 8y - \lambda_1(2x^2 + y^2 - 20) - \lambda_2(x - 1)$

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

1. Stationarity: $\frac{\partial L}{\partial x} = y - 4\lambda_1 x - \lambda_2 = 0, \quad \frac{\partial L}{\partial y} = x - 8 - 2\lambda_1 y = 0$

2. Primal Feasibility: $2x^2 + y^2 - 20 \leq 0, \quad x - 1 \leq 0$

3. Dual Feasibility: $\lambda_1 > 0, \quad \lambda_2 > 0$

4. Complementary Slackness: $\lambda_1(2x^2 + y^2 - 20) = 0, \quad \lambda_2(x - 1) = 0$

(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]

Case 1: Both Constraints Active

$$2x^2 + y^2 = 20$$
$$x = 1$$
$$\lambda_1 > 0$$
$$\lambda_2 > 0$$

$$x = 1 \rightarrow y = 3\sqrt{2} \rightarrow \lambda_1 = \frac{-7\sqrt{2}}{12} \rightarrow \lambda_2 = \frac{16\sqrt{2}}{3}$$

$$x = 1 \rightarrow y = -3\sqrt{2} \rightarrow \lambda_1 = \frac{7\sqrt{2}}{12} \rightarrow \lambda_2 = -\frac{16\sqrt{2}}{3}$$

$$x = 1, y = 4.24264, \lambda_1 = -0.82495, \lambda_2 = 7.54247$$

$$x = 1, y = -4.24264, \lambda_1 = 0.82495, \lambda_2 = -7.54247$$

Case 2: g_1 Active, g_2 Inactive

$$2x^2 + y^2 = 20$$

$$x < 1$$

$$\lambda_1 > 0$$

$$\lambda_2 = 0$$

$$\lambda_2 = 0 \rightarrow y = \sqrt{20 - 2x^2} \rightarrow \lambda_1 = \frac{\sqrt{20 - 2x^2}}{4x} \rightarrow x = -1$$

$$\lambda_2 = 0 \rightarrow y = -\sqrt{20 - 2x^2} \rightarrow \lambda_1 = -\frac{\sqrt{20 - 2x^2}}{4x} \rightarrow x = -1$$

$$x = -1, y = 4.24264, \lambda_1 = -1.06066, \lambda_2 = 0$$

$$x = -1, y = -4.24264, \lambda_1 = 1.06066, \lambda_2 = 0$$

Case 3: g_1 Inactive, g_2 Active

$$2x^2 + y^2 < 20$$

$$x = 1$$

$$\lambda_1 = 0$$

$$\lambda_2 \geq 0$$

$$x = 1 \rightarrow -7 = 0 \qquad \textbf{Incorrect}$$

Case 4: Both Constraints Inactive

$$2x^2 + y^2 < 20$$

$$x < 1$$

$$\lambda_1 = 0$$

$$\lambda_2 = 0$$

$$\lambda_1 = 0, \lambda_2 = 0 \rightarrow y = 0 \rightarrow x = 8$$

$$x = 8, y = 0, \lambda_1 = 0, \lambda_2 = 0$$

(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]

In the form $(x, y, \lambda_1, \lambda_2)$:

Case 1: Both Constraints Active
$(1.000, 4.243, -0.825, 7.542)$
$(1.000, -4.243, 0.825, -7.542)$
Neither point is valid

Case 2: g_1 Active, g_2 Inactive
$(-1, 4.243, -1.061, 0.000)$
$(-1, -4.243, 1.061, 0.000)$
Only the second point is valid

Case 3: g_1 Inactive, g_2 Active
This case produces no points

Case 4: Both Constraints Inactive
$(8.000, 0.000, 0.000, 0.000)$
Invalid point

Candidate Point: $\boxed{(-1, -4.243)}$

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

Using the point $(-1, -4.243, 1.061, 0.000)$, we see $f(x, y) = 38.187$

$$H = \begin{bmatrix} 4\lambda_1 & 1 \\ 1 & 2\lambda_1 \end{bmatrix} = \begin{bmatrix} 4.244 & 1 \\ 1 & 2.122 \end{bmatrix}, \ |H| = 8.006 \text{ and } f_{xx}(x, y) = 4\lambda_1 = 4.244$$

This means that $L(x, y)$ is $\boxed{convex}$ at the candidate point $(-1, -4.243)$ given both $|H|$ and $f_{xx}(x, y)$ are positive

(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. [2%]
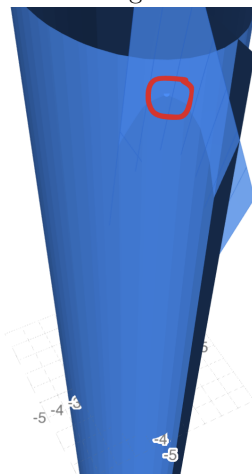
This point makes sense as we can see the curve reaches an apex at the candidate point.

NOTE: Use an explicit surface for the objective function, implicit surfaces for the constraints, and a point for the minimum candidate point.

HINT: Read the Example_optimization_problem.pdf in Canvas Files for HW1 to see an example with some explanations.

HINT: Click here for a video explaining the intuition behind KKT problems.

**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 Grads / 6% 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 12 sections, each numbered 1 through 12. 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 12 times such that it doesn't land on section 1 during the first 11 spins and lands on section 1 on the 12th 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

$\frac{Section's\ Area}{Total\ Spinner\ Area} = Probabilty\ of\ Landing\ on\ Section$
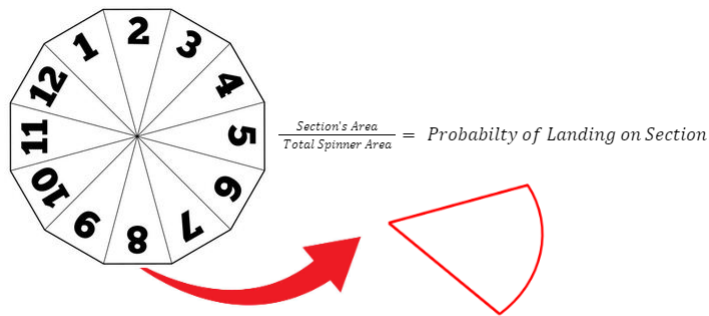
The likelihood function of this series of events can be portrayed as: $L(\theta) = (1 - \theta)^{11} \cdot \theta$

The log-likelihood function is: $l(\theta) = \ln((1 - \theta)^{11} \cdot \theta) = 11 \ln(1 - \theta) + \ln(\theta)$

MLE: $\frac{d}{d\theta}(l(\theta)) = \frac{d}{d\theta}(11 \ln(1 - \theta) + \ln(\theta)) = \frac{-11}{1-\theta} + \frac{1}{\theta}$

$\frac{-11}{1-\theta} + \frac{1}{\theta} = 0 \rightarrow \boxed{\theta = \frac{1}{12}}$

4.2 Poisson distribution [15 pts Grad / 6% 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 / 0.8%]

$$\boxed{L(\lambda) = P(X_1 = x_1) = \frac{\lambda^{x_1} e^{-\lambda}}{x_1!}}$$

(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 / 1.2%]

$$\boxed{L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}}$$

(c) What is the maximum likelihood estimator of λ? [10 pts / 4%]

$\ell(\lambda) = \sum_{i=1}^{n} \ln\left(\frac{\lambda^{x_i} e^{-\lambda}}{x_i!}\right)$

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

Taking the derivative of $\ell(\lambda)$ with respect to λ and setting it to zero will give us the MLE for λ. The derivative is:

$\frac{d}{d\lambda} \ell(\lambda) = \sum_{i=1}^{n} \left(\frac{x_i}{\lambda} - 1\right)$

Setting this equal to zero and solving for λ:

$\sum_{i=1}^{n} \left(\frac{x_i}{\lambda} - 1\right) = 0$

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

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

Thus, the MLE for λ is the sample mean of the observed values: $\boxed{\hat{\lambda} = \bar{x}}$

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.

X \ Y	0	1
0	$\frac{4}{12}$	$\frac{1}{12}$
1	$\frac{2}{12}$	$\frac{5}{12}$

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

For X:

$$P(X = 0) = \tfrac{4}{12} + \tfrac{1}{12} = \boxed{\frac{5}{12}}$$

$$P(X = 1) = \tfrac{2}{12} + \tfrac{5}{12} = \boxed{\frac{7}{12}}$$

For Y:

$$P(Y = 0) = \tfrac{4}{12} + \tfrac{2}{12} = \tfrac{6}{12} = \boxed{\frac{1}{2}}$$

$$P(Y = 1) = \tfrac{1}{12} + \tfrac{5}{12} = \tfrac{6}{12} = \boxed{\frac{1}{2}}$$

(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]

$$I(X, Y) = \sum_{x \in \{0,1\}} \sum_{y \in \{0,1\}} p(x,y) \log_2 \left(\frac{p(x,y)}{p(x)p(y)} \right)$$

$$p(X = 0) = \tfrac{5}{12}, \quad p(X = 1) = \tfrac{7}{12}, \quad p(Y = 0) = \tfrac{6}{12}, \quad p(Y = 1) = \tfrac{6}{12}$$

$$I(X, Y) = \tfrac{4}{12} \log_2 \left(\frac{\frac{4}{12}}{\frac{5}{12} \cdot \frac{6}{12}} \right) + \tfrac{1}{12} \log_2 \left(\frac{\frac{1}{12}}{\frac{5}{12} \cdot \frac{6}{12}} \right) + \tfrac{2}{12} \log_2 \left(\frac{\frac{2}{12}}{\frac{7}{12} \cdot \frac{6}{12}} \right) + \tfrac{5}{12} \log_2 \left(\frac{\frac{5}{12}}{\frac{7}{12} \cdot \frac{6}{12}} \right)$$

$$\boxed{I(X, Y) \approx 0.19571}$$

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	M	T	T	T
2	M	F	F	T
3	L	F	T	F
4	M	T	F	T
5	M	F	T	T
6	M	T	F	F
7	H	F	T	T
8	L	T	F	F
9	L	F	F	T
10	L	F	F	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]

$$\text{Symptomatic (T)} = 7$$
$$\text{Not Symptomatic (F)} = 3$$

$$P(T) = \frac{7}{10} = 0.7$$
$$P(F) = \frac{3}{10} = 0.3$$

$$H(Y) = -\sum P(y) \log_2 P(y)$$
$$H(Y) = -(0.7 \log_2 0.7 + 0.3 \log_2 0.3) \approx 0.881$$
$$\boxed{H(Y) \approx 0.881}$$

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

$$H(Y|X) = -\sum_{x \in X, y \in Y} P(x,y) \log_2 \frac{P(x,y)}{P(x)}$$

$$H(Y|X_1) = P(X_1 = H)H(Y|X_1 = H) + P(X_1 = M)H(Y|X_1 = M) + P(X_1 = L)H(Y|X_1 = L)$$
$$= \frac{1}{10}(-P(Y = T|X_1 = H)\log_2(P(Y = T|X_1 = H))) +$$
$$\frac{4}{10}(-P(Y = T|X_1 = M)\log_2(P(Y = T|X_1 = M)) - P(Y = F|X_1 = M)\log_2(P(Y = F|X_1 = M))) +$$
$$\frac{5}{10}(-P(Y = T|X_1 = L)\log_2(P(Y = T|X_1 = L)) - P(Y = F|X_1 = L)\log_2(P(Y = F|X_1 = L)))$$
$$= \frac{1}{10}(0) + \frac{4}{10}(-\frac{3}{4}\log_2(\frac{3}{4}) - \frac{1}{4}\log_2(\frac{1}{4})) + \frac{5}{10}(-\frac{2}{5}\log_2(\frac{2}{5}) - \frac{3}{5}\log_2(\frac{3}{5}))$$
$$\boxed{H(Y|X_1) \approx 0.761}$$

$$H(Y|X_3) = P(X_3 = T)H(Y|X_3 = T) + P(X_3 = F)H(Y|X_3 = F)$$

$$= \frac{4}{10}(-P(Y = T|X_3 = T)\log_2(P(Y = T|X_3 = T)) - P(Y = F|X_3 = T)\log_2(P(Y = F|X_3 = T))) +$$

$$\frac{6}{10}(-P(Y = T|X_3 = F)\log_2(P(Y = T|X_3 = F)) - P(Y = F|X_3 = F)\log_2(P(Y = F|X_3 = F)))$$

$$= \frac{4}{10}(-\frac{3}{4}\log_2(\frac{3}{4}) - \frac{1}{4}\log_2(\frac{1}{4})) + \frac{6}{10}(-\frac{4}{6}\log_2(\frac{4}{6}) - \frac{2}{6}\log_2(\frac{2}{6}))$$

$$\boxed{H(Y|X_3) \approx 0.875}$$

(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]

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

$$I(X_1, Y) = H(Y) - H(Y|X_1) \approx 0.881 - 0.761$$

$$\boxed{I(X_1, Y) = 0.120}$$

$$I(X_3, Y) = H(Y) - H(Y|X_3) \approx 0.881 - 0.875$$

$$\boxed{I(X_3, Y) = 0.006}$$

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

$$H(Y, X_2) = -\sum_{y \in Y, x_2 \in X_2} P(y, x_2)\log_2 P(y, x_2)$$

$$H(Y, X_2) = -(0.1\log_2 0.1 + 0.2\log_2 0.2 + 0.5\log_2 0.5 + 0.2\log_2 0.2)$$

$$\boxed{H(Y, X_2) \approx 1.761}$$

5.3 Entropy Proofs [10pts]

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

$$H(X|Y) = -\sum_{y \in Y} P(y) \sum_{x \in X} P(x|y) \log_2 P(x|y)$$

$$H(X) = -\sum_{x \in X} P(x) \log_2 P(x)$$

(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. You may use any theorem/ proof from the slides without having to re-prove it). [7pts]

Given $I(X,Y) = H(X) - H(X|Y)$, we can start by expressing mutual information as:

$$I(X,Y) = \sum_{x \in X, y \in Y} P(x,y) \log_2 \frac{P(x,y)}{P(x)P(y)}$$

For independent random variables X and Y, the joint probability $P(x,y)$ factorizes as:

$$P(x,y) = P(x)P(y)$$

Thus, the mutual information simplifies to:

$$I(X,Y) = \sum_{x \in X, y \in Y} P(x)P(y) \log_2(1) = \sum_{x \in X, y \in Y} P(x)P(y) \cdot 0 = 0$$

Now, considering the independence of X and Y and the definition of conditional entropy:

$$H(X|Y) = -\sum_{y \in Y} P(y) \sum_{x \in X} P(x|y) \log_2 P(x|y)$$

Since $P(x|y) = P(x)$ for independent variables:

$$H(X|Y) = -\sum_{y \in Y} P(y) \sum_{x \in X} P(x) \log_2 P(x)$$

By the law of total probability $\sum_{y \in Y} P(y) = 1$, it follows that:

$$H(X|Y) = H(X)$$

Therefore, by substitution into the initial expression:

$$I(X,Y) = H(X) - H(X|Y) = H(X) - H(X) = 0$$

This concludes the proof that mutual information is zero when the random variables are independent

6 Ethical Implications on Decision-Making [5 pts]

Real-world Implications

Loan eligibility determines who can receive a loan, typically based on financial history and demographics. It is a difficult problem, and often uses algorithms to make loan decisions. Often, this can result in reinforcing inequality and bias [1].

Suppose we're using a matrix to represent the attributes of individuals for loan approval. Each attribute (like income, credit score, years of employment, etc.) constitutes a column in our matrix. Here's a hypothetical toy example:

	Annual Income	Debt-to-Income Ratio	Employment History (years)	Credit Score
Candidate 1	50,000	0.2	5	700
Candidate 2	51,000	0.21	5.1	710
Candidate 3	45,000	0.19	4.9	690
Candidate 4	100,000	0.05	10	780

One algorithm used to predict credit score is linear regression, formulated as $\mathbf{y} = \mathbf{x}\mathbf{A}$. $\mathbf{y}$ are the target variables, $\mathbf{x}$ are the input features, and $\mathbf{A}$ is a matrix trained with an existing dataset. Training data $(\mathbf{x_D}, \mathbf{y_D})$ are taken from the training dataset D, $(\mathbf{x_D}, \mathbf{y_D}) \in D$. If $\mathbf{x_D}$ is linearly independent, $\mathbf{A}$ can be trained by simply inverting $\mathbf{x_D}$:

$$\mathbf{y_D} = \mathbf{x_D}\mathbf{A}$$

$$\mathbf{x_D}^{-1}\mathbf{y_D} = \mathbf{A}$$

The original equation can be rewritten as:

$$\mathbf{y} = \mathbf{x}\mathbf{A}$$

$$= \mathbf{x}\mathbf{x_D}^{-1}\mathbf{y_D}$$

Problems arise when the training data is close to linearly dependent. Recall that one way to invert a matrix is $\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})}\text{adj}(\mathbf{A})$. As $\mathbf{A}$ becomes more linearly dependent and $\det(\mathbf{A}) \to 0$, $||\mathbf{A}^{-1}||$ can become so large it causes numerical errors. Rewriting the original equation:

$$\mathbf{y} = \mathbf{x}\mathbf{x_D}^{-1}\mathbf{y_D}$$

$$= \frac{1}{\det(\mathbf{x_D})}\mathbf{x}\,\text{adj}(\mathbf{x_D})\mathbf{y_D}$$

The errors caused by $\det(\mathbf{x_D}) \to 0$ propagate to $\mathbf{y}$, causing predictions to be wildly inaccurate anywhere outside of the original training set.

Practical Implications

1. Instability: With a small determinant, minor variations in the attributes can lead to significant variations in the results. So, a small difference in income might result in a disproportionate change in loan eligibility.

2. Poor Generalization: If the matrix is based on data with limited variation (like our small community example), it's essentially trained on a very narrow subset of potential applicants. If someone from outside this narrow subset applies (e.g., a person with a 2-year employment but a $70,000 income), the system may not process their application fairly or accurately because it's unfamiliar with such profiles.

Given that a matrix used for determining loan approvals has a determinant close to zero due to limited variation in applicants' attributes:
Which of the following implications might this have on the decision-making process? Choose as all options that apply. Use "textbf{}" to select your answer.

A) It ensures a more uniform scoring system since most applicants have similar attributes.

B) It can lead to unpredictable scores, where tiny variations in attributes yield vastly different outcomes.

C) The system is more resilient to errors because of the limited attribute variation.

D) It might not generalize well to broader populations, potentially leading to biases when applied to more diverse applicant groups.

7 Programming [5 pts]

See the Programming subfolder in Canvas.

8 Bonus for All [8%]

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

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

Calculate the variance $Var(P+Q)$. *(This question may take substantial work to support, e.g. 25 to 30 lines)* [4%]

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 (Wolfram alpha calculator or other similar calculators can be used)
HINT: $Var(P+Q) = Var(P) + Var(Q) + 2Cov(P,Q)$ may be a good starting point.

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

$$\begin{aligned}
Var(P) &= Var(3X + 2XY^2) \\
&= 9Var(X) + 4Var(XY^2) + 2 \cdot 3 \cdot 2 \cdot Cov(X, XY^2) \\
&= 9Var(X) + 4(E[X^2Y^4] - [E(XY^2)]^2) \\
&= 9 \cdot 1 + 4(E[X^2]E[Y^4] - [E(X)E(Y^2)]^2) \\
&= 9 + 4(E[Y^4]) \quad \text{(since } E[X^2] = Var(X) + [E(X)]^2 = 1 \text{ and } E[X] = 0)
\end{aligned}$$

$E[Y^4] = 3$ according to https://www.integral-calculator.com/. Therefore:

$$\begin{aligned}
Var(P) &= 9 + 4 \cdot 3 \\
&= 21
\end{aligned}$$

Since $Q = X$, $Var(Q) = Var(X) = 1$

$$\begin{aligned}
Cov(P,Q) &= Cov(3X + 2XY^2, X) \\
&= 3Var(X) + 2Cov(XY^2, X) \\
&= 3 \cdot 1 + 2 \cdot 0 \quad \text{(since } X \text{ and } Y \text{ are independent, } Cov(XY^2, X) = 0) \\
&= 3
\end{aligned}$$

$$\begin{aligned}
Var(P+Q) &= Var(P) + Var(Q) + 2Cov(P,Q) \\
&= 21 + 1 + 2 \cdot 3 \\
&= \boxed{28}
\end{aligned}$$

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

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

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)* [2%]

$$f_X(x) = \int_0^\infty \frac{x}{2} e^{-y} dy$$

$$= \frac{x}{2} \int_0^\infty e^{-y} dy$$

$$= \frac{x}{2} \left[-e^{-y} \right]_0^\infty$$

$$= \frac{x}{2} \left[0 - (-1) \right]$$

$$= \frac{x}{2} \quad \text{for } 0 \le x \le 2$$

$$f_Y(y) = \int_0^2 \frac{x}{2} e^{-y} dx$$

$$= e^{-y} \int_0^2 \frac{x}{2} dx$$

$$= e^{-y} \left[\frac{1}{4} x^2 \right]_0^2$$

$$= e^{-y}$$

$$= e^{-y} \quad \text{for } y \ge 0$$

$$\boxed{f_X(x) = \frac{x}{2} \quad \text{for } 0 \le x \le 2}$$
$$\boxed{f_Y(y) = e^{-y} \quad \text{for } y \ge 0}$$

(c) A person decides to toss a biased coin with $P(heads) = 0.1$ 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)* [2%]

Probability distribution of Y:

$$P(Y = 0) = P(\text{no head in 5 tosses}) = 0.9^5$$
$$P(Y = 1) = 0.1$$
$$P(Y = k) = 0 \text{ for } k > 1 \text{ because we stop after the first head}$$

Variance of Y:

$$E[Y] = 1 \cdot P(Y = 1) + 0 \cdot P(Y = 0) = 0.1$$
$$E[Y^2] = 1^2 \cdot P(Y = 1) + 0^2 \cdot P(Y = 0) = 0.1$$
$$Var(Y) = E[Y^2] - (E[Y])^2 = 0.1 - (0.1)^2$$
$$= 0.1 - 0.01$$
$$= \boxed{0.09}$$

References

[1]   Cathy O'Neil. *Weapons of Math Destruction*. Penguin Books, 2017.