

Spring 2022 CS4641/CS7641 A Homework 1

Dr. Mahdi Roozbahani

Deadline: Feb 10, Thursday, 11:59 pm AOE

YOUR NAME

- No unapproved extension of the deadline is allowed. Late submission will lead to 0 credit.
- Discussion is encouraged, but each student must write their own answers and explicitly mention any collaborators.
- **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.
- 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

- This assignment has no programming, only written questions.
- 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.
- 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 guide on LaTeX is provided in [this Edstem post](#).
- Please answer each question on a new page. It makes it more organized to map your answers on GradeScope. When submitting your assignment, you must correctly map pages of your PDF to each question/subquestion to reflect where they appear. Improperly mapped questions may not be graded correctly.
- Discussion is encouraged on Edstem as part of the Q/A. You may discuss high-level ideas with other students at the "whiteboard" level (e.g. how cross validation works, using `matmul` instead of `dot`) and review any relevant materials online. However, 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 [43pts]

- 1.1 Determinant and Inverse of a Matrix [15pts]
- 1.2 Characteristic Equation [8pts]
- 1.3 Eigenvalues and Eigenvectors [20pts]

Q2: Covariance, Correlation, and Independence [9pts]

- 2.1 Covariance [5pts]
- 2.2 Correlation [4pts]

Q3: Optimization [15pts]

Q4: Maximum Likelihood [25pts]

- 4.1 Discrete Example [10pts]
- 4.2 Poisson Distribution [15pts: Bonus for Undergrads]

Q5: Information Theory [35pts]

- 5.1 Marginal Distribution [6pts]
- 5.2 Mutual Information and Entropy [19pts]
- 5.3 Entropy Proofs [10pts]

Q6: Bonus for All [15pts]

1 Linear Algebra [15pts + 8pts + 20pts]

1.1 Determinant and Inverse of Matrix [15pts]

Given a matrix M :

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

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

Solution:

$$|M| = 2 * (-2r - 0) - r(r + 15) + 2(0 + 10) = -4r - r^2 - 15r + 20 = -r^2 - 19r + 20$$

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

Solution:

M^{-1} does not exist whenever $|M| = 0$. Since we previously calculated $|M| = -r^2 - 19r + 20 = (r - 1)(r + 20)$, this occurs when either when $r = 1$ or $r = -20$. For these values of r , the matrix M is singular and it is not full rank.

- (c) Will all values of r found in part b allow for a row (or a column) to be expressed as a linear combination of the other rows (or columns) respectively? If yes, provide one example, if no, explain why. [3pts]

Solution:

Yes. For both values of r from part b, the matrix is of rank 2. Therefore the columns (and rows) are linearly dependent and there is at least one column (row) that can be expressed as a linear combination of the other two columns/rows.

When $r = 1$:

$$\begin{bmatrix} 2 & 1 & 2 \\ 1 & -2 & -3 \\ 5 & 0 & 1 \end{bmatrix}$$

R_3 can be expressed as $2R_1 + R_2$

When $r = -20$:

$$\begin{bmatrix} 2 & -20 & 2 \\ 1 & -2 & -3 \\ 5 & 0 & -20 \end{bmatrix}$$

C_3 can be expressed as $-4C_1 - \frac{1}{2}C_2$

$$-4 \begin{bmatrix} 2 \\ 1 \\ 5 \end{bmatrix} - \frac{1}{2} \begin{bmatrix} -20 \\ -2 \\ 0 \end{bmatrix} = \begin{bmatrix} -8 \\ -4 \\ -20 \end{bmatrix} + \begin{bmatrix} 10 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ -1 \end{bmatrix}$$

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

Solution:

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

- (e) Find the determinant of M^{-1} for $r = 0$. What is the relationship between the determinant of M and the determinant of M^{-1} ? [3pts]

Solution:

$$|M^{-1}| = 0 * \det \begin{bmatrix} -1/2 & 2/5 \\ 0 & -1/5 \end{bmatrix} - 0 * \det \begin{bmatrix} -3/4 & 2/5 \\ 1/2 & -1/5 \end{bmatrix} + \frac{1}{5} \det \begin{bmatrix} -3/4 & -1/2 \\ 1/2 & 0 \end{bmatrix}$$

$$|M^{-1}| = \frac{1}{5} * \frac{1}{4} = \frac{1}{20}$$

The determinant of M^{-1} is the reciprocal of the determinant of M .

1.2 Characteristic Equation [8pts]

Consider the eigenvalue problem:

$$Ax = \lambda x, x \neq 0$$

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

Solution:

Consider the equation $Ax = 0$. There exists a solution other than $x = \vec{0}$ iff there is linear combination of the columns of A that results in the zero vector. If this is the case, then it implies that the columns of A form a linearly dependent set. It follows that the determinant of A must be 0, since it is not linearly independent; that is, if $Ax = 0, x \neq 0$ has a solution, then $|A| = 0$.

Now we can rewrite the given equation $Ax = \lambda x$ as $(A - \lambda I)x = 0$. The logic shown above can be applied to the equation to show that the determinant $|A - \lambda I|$ must equal 0.

1.3 Eigenvalues and Eigenvectors [5pts + 5pts + 10pts]

1.3.1 Eigenvalues [5pts]

Given a matrix \mathbf{A} :

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

- (a) Find an expression for the eigenvalues (λ) of \mathbf{A} in the terms given. [4pts]

Solution:

$$\begin{aligned} \det \begin{bmatrix} a - \lambda & b \\ b & c - \lambda \end{bmatrix} &= 0 \\ (a - \lambda)(c - \lambda) - b^2 &= 0 \\ ac - a\lambda - c\lambda + \lambda^2 - b^2 &= 0 \\ \lambda &= \frac{a + c - \sqrt{a^2 + 4b^2 + c^2 - 2ac}}{2} \end{aligned}$$

- (b) Find a simple expression for the eigenvalues if $c = a$. Under what conditions are all $\lambda > 0$? [1pt]

Solution:

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

$$\lambda = \frac{2a \pm \sqrt{4b^2}}{2}$$

$$\lambda = \frac{2a \pm 2b}{2}$$

$$\lambda = a + b, a - b$$

All $\lambda > 0$ when $a > b$.

1.3.2 Eigenvectors [5pts]

A symmetric matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$ can be decomposed as

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T = \sum_{n=1}^N \lambda_n \mathbf{v}_n \mathbf{v}_n^T$$

Where \mathbf{V} is a matrix whose columns are the eigenvectors of \mathbf{A} , \mathbf{v}_n are the columns of \mathbf{V} and $\mathbf{\Lambda}$ is a diagonal matrix whose elements are the eigenvalues of \mathbf{A} . The eigenvectors are orthonormal to each other, i.e., $\mathbf{v}_i^T \mathbf{v}_j = \begin{cases} 1, i = j \\ 0, i \neq j \end{cases}$.

- (a) Show that $\text{trace}(\mathbf{A}) = \sum_{n=1}^N \lambda_n$ [3pts]

NOTE: $\mathbf{v}_i^T \mathbf{v}_j \neq \mathbf{v}_i \mathbf{v}_j^T$

Solution:

By definition, the columns of \mathbf{V} are orthogonal, so $\mathbf{V}^T = \mathbf{V}^{-1}$. So we can multiply by the inverse and rewrite the decomposition as:

$$\begin{aligned} \mathbf{\Lambda} &= \mathbf{V}^{-1} \mathbf{A} \mathbf{V} \\ \text{trace}(\mathbf{\Lambda}) &= \sum_i^N \lambda_i \\ &= \text{trace}(\mathbf{V}^{-1} \mathbf{A} \mathbf{V}) \\ &= \text{trace}((\mathbf{V}^{-1} \mathbf{V}) \mathbf{A}) \\ &= \text{trace}(\mathbf{I} \mathbf{A}) \\ &= \text{trace}(\mathbf{A}) \\ &\therefore \\ \text{trace}(\mathbf{A}) &= \sum_{n=1}^N \lambda_n \end{aligned}$$

- (b) What is the result of the multiplication $\mathbf{V}^T \mathbf{V}$? Show your work or present an argument. [2pts]

Solution:

\mathbf{A} is a symmetric matrix, so since \mathbf{V} is formed from the eigenvectors of \mathbf{A} , \mathbf{V} is guaranteed to be an orthogonal matrix, therefore $\mathbf{V}^T = \mathbf{V}^{-1}$.

$$\therefore \mathbf{V}^T \mathbf{V} = \mathbf{I}$$

1.3.3 Eigenvalue and Eigenvector Calculations [10pts]

Given a matrix

$$\mathbf{A} = \begin{bmatrix} x & 5 \\ 5 & x \end{bmatrix}$$

- (a) Calculate the eigenvalues of \mathbf{A} as a function of x . (Calculation process is required). [3pts]

Solution:

$$\begin{aligned} \det \begin{bmatrix} x - \lambda & 5 \\ 5 & x - \lambda \end{bmatrix} &= 0 \\ (x - \lambda)^2 - 25 &= 0 \\ (x - \lambda)^2 &= 25 \\ (x - \lambda) &= \pm 5 \\ \lambda &= x + 5, x - 5 \end{aligned}$$

- (b) Find the normalized eigenvectors of matrix \mathbf{A} . (Calculation process is required). [7pts]

Solution:

For $\lambda = x - 5$:

$$\begin{aligned}
 (A - \lambda I)x &= 0 \\
 \begin{bmatrix} x - \lambda & 5 \\ 5 & x - \lambda \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} &= \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\
 \begin{bmatrix} x - (x - 5) & 5 \\ 5 & x - (x - 5) \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} &= \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\
 \begin{bmatrix} 5 & 5 \\ 5 & 5 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} &= \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\
 5a + 5b &= 0 \\
 a = -b &\rightarrow \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}
 \end{aligned}$$

For $\lambda = x + 5$:

$$\begin{aligned}
 (A - \lambda I)x &= 0 \\
 \begin{bmatrix} x - \lambda & 5 \\ 5 & x - \lambda \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} &= \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\
 \begin{bmatrix} x - (x + 5) & 5 \\ 5 & x - (x + 5) \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} &= \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\
 \begin{bmatrix} -5 & 5 \\ 5 & -5 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} &= \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\
 5a + -5b &= 0 \\
 a = b &\rightarrow \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}
 \end{aligned}$$

2 Covariance, Correlation, and 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 distribution function is

$$p(x) = \begin{cases} 0.5 & x = c \\ 0.5 & x = -c. \end{cases}$$

c is a constant here. Let Y obey the Standard Normal (Gaussian) Distribution, which can be written as $Y \sim N(0, 1)$. X and Y are independent. Meanwhile, let $Z = XY$.

Calculate the covariance of Y and Z ($Cov(Y, Z)$) and determine whether values of c would affect the correlation.

Solution:

$$\begin{aligned} Cov(Y, Z) &= E(YZ) - E(Y)E(Z) \\ &= E(XY^2) - E(Y)E(XY) \\ &= E(X)E(Y^2) - [E(Y)]^2E(X) \\ &= E(X)[E(Y^2) - [E(Y)]^2] \\ E(X) &= \sum_{x \in S} xp(x) \\ &= 0.5c - 0.5c \\ &= 0 \\ \therefore Cov(Y, Z) &= 0[E(Y^2) - [E(Y)]^2] \\ &= 0 \end{aligned}$$

No, the value of c will not affect the correlation because the distribution is always centered at 0, and the covariance is also 0 which is the numerator when calculating the correlation.

2.2 Correlation Coefficient [4pts]

Let X and Y be independent random variables with $\text{var}(X) = 4$ and $\text{var}(Y) = 12$. We do not know $E[X]$ or $E[Y]$. Let $Z = 3X + Y$. What is the correlation coefficient $\rho(X, Z) = \frac{\text{cov}(X, Z)}{\sqrt{\text{var}(X)\text{var}(Z)}}$?

Solution:

$$\begin{aligned}\text{Cov}(X, Z) &= E(XZ) - E(X)E(Z) \\&= E(X(3X + Y)) - E(X)E(3X + Y) \\&= E(3X^2 + YX) - E(X)E(3X + Y) \\&= E(3X^2) + E(YX) - E(X)(E(3X) + E(Y)) \\&= 3E(X^2) + E(YX) - E(X)(3E(X) + E(Y)) \\&= 3E(X^2) + E(YX) - 3E(X)^2 - E(X)E(Y) \\&= 3[E(X^2) - E(X)^2] + [E(YX) - E(X)E(Y)] \\&= 3[\text{var}(X)] + [\text{Cov}(X, Y)] \\&= 3[4] + 0 \\&= 12\end{aligned}$$

$$\begin{aligned}\text{var}(Z) &= \text{var}(3X + Y) \\&= \text{var}(3X) + 2\text{Cov}(3X, Y) + \text{var}(Y) = 9\text{var}(X) + 6\text{Cov}(X, Y) + \text{var}(Y) = 9(4) + 6(0) + 12 = 36 + 12 = 48\end{aligned}$$

$$\begin{aligned}\rho(X, Z) &= \frac{\text{cov}(X, Z)}{\sqrt{\text{var}(X)\text{var}(Z)}} \\&= \frac{12}{\sqrt{4 * 48}} \\&= \frac{12}{\sqrt{192}}\end{aligned}$$

3 Optimization [15pts]

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:

$$\begin{aligned} \min_{x,y} \quad & f(x, y) = 7x^2 + 4y \\ \text{s.t.} \quad & g_1(x, y) = x^2 + y^2 \leq 2 \\ & g_2(x, y) = x \leq 1 \end{aligned}$$

HINT 1: Click [here](#) for a maximization example.

HINT 2: Click [here](#) to determine how to set up the problem for minimization in parts (a) and (b).

(a) Specify the Lagrange function [2pts]

Solution:

$$L(x, y) = 7x^2 + 4y - \lambda_1(x^2 + y^2 - 2) - \lambda_2(x - 1)$$

(b) List the equations for the KKT conditions [2pts]

Solution:

Equality Constraints:

$$\frac{\partial L}{\partial x} = 14x - 2\lambda_1 x - \lambda_2 = 0 \quad (\text{a})$$

$$\frac{\partial L}{\partial y} = -2\lambda_1 y + 4 = 0 \quad (\text{b})$$

Complimentary Slack:

$$\lambda_1(x^2 + y^2 - 2) = 0 \quad (\text{c})$$

$$\lambda_2(x - 1) = 0 \quad (\text{d})$$

(c) Solve for 4 possibilities formed by each constraint being active or inactive [5pts]

Solution:

(1) When both constraints are active:

$$x^2 + y^2 = 2, \lambda_1 < 0$$

$$x = 1, \lambda_2 < 0$$

$$\rightarrow y^2 = 1$$

$$y = \pm 1$$

i) if $y = 1$

$$2\lambda_1 = 4$$

$$\lambda_1 = 2$$

$$\lambda_1 < 0 \rightarrow \text{fail}$$

ii) if $y = -1$

$$2\lambda_1 = -4$$

$$\lambda_1 = -2$$

$$\lambda_1 < 0 \rightarrow \text{pass}$$

apply KKT stationary condition from equation (a):

$$14x - 2\lambda_1 x - \lambda_2 = 0$$

$$\Rightarrow 14(1) - 2(-2)(1) - \lambda_2 = 0$$

$$14 = \lambda_2$$

$$\lambda_2 < 0 \rightarrow \text{fail}$$

(2) Only λ_1 is active

$$x^2 + y^2 = 2, \lambda_1 < 0$$

$$x < 1, \lambda_2 = 0$$

$$\rightarrow x = \pm\sqrt{2 - y^2}$$

i) if $x = -\sqrt{2 - y^2}$

apply KKT stationary condition from equation (a):

$$14x - 2\lambda_1 x - \lambda_2 = 0$$

$$\Rightarrow 14(-\sqrt{2 - y^2}) - 2\lambda_1(-\sqrt{2 - y^2}) - \lambda_2 = 0$$

$$-14\sqrt{2 - y^2} + 2\lambda_1\sqrt{2 - y^2} - \lambda_2 = 0$$

$$-14\sqrt{2 - y^2} + 2\lambda_1\sqrt{2 - y^2} = 0$$

$$2\lambda_1\sqrt{2 - y^2} = 14\sqrt{2 - y^2}$$

$$2\lambda_1 = 14$$

$$\lambda_1 = 7$$

$$\lambda_1 < 0 \rightarrow \text{fail}$$

ii) if $x = 0$

$$y = \pm\sqrt{2}$$

apply KKT stationary condition from equation (b):
when $y = \sqrt{2}$:

$$-2\sqrt{2}\lambda_1 = -4$$

$$\lambda_1 = \frac{2}{\sqrt{2}}$$

$$\lambda_1 < 0 \rightarrow \text{fail}$$

when $y = -\sqrt{2}$:

$$2\sqrt{2}\lambda_1 = -4$$

$$\lambda_1 = -\frac{2}{\sqrt{2}}$$

$$\lambda_1 < 0 \rightarrow \text{pass}$$

(3) When only λ_2 is active

$$\begin{aligned} x^2 + y^2 &< 2, \lambda_1 = 0 \\ x &= 1, \lambda_2 < 0 \end{aligned}$$

apply KKT stationary condition from equation (b)

$$4 - 2(0)y = 0 \rightarrow \text{fail}$$

(4) When both are inactive

$$\begin{aligned} x^2 + y^2 &< 2, \lambda_1 = 0 \\ x &< 1, \lambda_2 = 0 \end{aligned}$$

apply KKT stationary condition from equation (b)

$$4 - 2(0)y = 0 \rightarrow \text{fail}$$

- (d) List the candidate point(s) (there may be 0, 1, 2, or 4 candidate points) [4pts]

Solution:

The candidate points are $(0, -\sqrt{2})$; these points satisfy the KKT conditions.

- (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]

Solution:

$$|H(x,y)| = \left| \begin{bmatrix} L_{xx}(x,y) & L_{xy}(x,y) \\ L_{yx}(x,y) & L_{yy}(x,y) \end{bmatrix} \right|$$

$$L(x,y) = 7x^2 + 4y - \lambda_1(x^2 + y^2 - 2) - \lambda_2(x - 1)$$

$$L_{xx}(x,y) = 2(7 + \sqrt{2})$$

$$L_{yy}(x,y) = 2\sqrt{2}$$

$$L_{xy}(x,y) = 0$$

$$L_{yx}(x,y) = 0$$

$$|H(0, -\sqrt{2})| = \begin{vmatrix} 2(7 + \sqrt{2}) & 0 \\ 0 & 2\sqrt{2} \end{vmatrix}$$

Since $H > 0$ and $L_{xx} > 0$, by the second partial derivative test, $L(0, -\sqrt{2})$ is a local minimum. L is also convex at this point by definition.

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

4.1 Discrete Example [10pts]

Devshree and Angana are arguing over where they should go for winter break. Devshree's argument is that they should go to Boston because Megha, their childhood friend, lives there and thus they will get to have a good time. Angana's argument is that they should go to Florida because Florida is very near to Atlanta, it has the mildest winter and on the other hand, Boston's weather is very rough during the winter.

To resolve this conflict, their other friend Yusuf makes a proposition that they should leave it to chance to decide where they should spend their winter break. Devshree then proposes that Angana will toss a 6-sided die 10 times, and Angana must get anything except 2 during the first 9 times and must get 2 during the 10th time. Any other combination will make Devshree the winner. But Angana is also allowed to tamper with the die in any manner she likes to increase her odds.

Now, Angana needs you to help her have her way. If the probability of getting a 2 is θ and the probabilities of landing on any other number is the same, what value of θ is most likely to ensure that they will have to go to Florida? Use your expertise of Maximum Likelihood Estimation and probability distribution function to convince Angana.

NOTE: You may assume that the log-likelihood function is concave for this question

Solution:

This problem is best modeled by a geometric distribution. The distribution is given by the following function, where $x = 10$ (success on 10th trial), and $p = \theta$ (probability of success, which requires rolling a 2.)

$$f(x, p) = p(1 - p)^{x-1}$$

The likelihood function:

$$L(\theta) = \theta^n (1 - \theta)^{\sum_1^n x_i - n}$$

Now we maximize the likelihood.

$$\begin{aligned} \ln(\theta) &= n \ln(\theta) + \left(\sum_1^n x_i - n \ln(1 - \theta) \right) \\ \frac{\partial \ln(\theta)}{\partial \theta} &= \frac{n}{\theta} - \frac{\sum_1^n x_i - n}{1 - \theta} \\ 0 &= \frac{n}{\theta} - \frac{\sum_1^n x_i - n}{1 - \theta} \\ \frac{\sum_1^n x_i - n}{1 - \theta} &= \frac{n}{\theta} \\ \theta &= \frac{n}{\sum_1^n x_i} \\ \theta &= \frac{1}{X} = \frac{1}{10} = 0.1 \end{aligned}$$

4.2 Poisson distribution [15pts Bonus for Undergrads]

The Poisson distribution is defined as

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

- (a) Assume that we have one observed data x_1 , and $X_1 \sim \text{Poisson}(\lambda)$. What is the likelihood if λ is given?
[2 pts]

Solution:

YOUR ANSWER HERE

- (b) Now, assume that we are given n such values (x_1, \dots, x_n) , $(X_1, \dots, X_n) \sim \text{Poisson}(\lambda)$. Here X_1, \dots, X_n are i.i.d. random variables. What is the likelihood of this data given λ ? [3 pts]

Solution:

YOUR ANSWER HERE

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

NOTE: You may assume that the log-likelihood function is concave for this question

Solution:

YOUR ANSWER HERE

5 Information Theory [6pts + 19pts + 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	$\frac{1}{8}$	$\frac{3}{8}$
1	0	$\frac{1}{2}$

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

Solution:

Marginal distribution of X :

$$P(X = 0) = 1/8 + 3/8 = 1/2$$

$$P(X = 1) = 0 + 1/2 = 1/2$$

Marginal distribution of Y :

$$P(Y = 0) = 1/8 + 0 = 1/8$$

$$P(Y = 1) = 3/8 + 1/2 = 7/8$$

- (b) Find mutual information for the joint probability distribution in the previous question [3pts]

Solution:

$$\begin{aligned} I(x_i, y) &= H(y) - H(y|x_i) \\ &= \sum_x P(x) \log_2\left(\frac{1}{P(x)}\right) - \sum_{x,y} P(x,y) \log_2\left(\frac{1}{P(x|y)}\right) \\ &= \sum_{x,y} P(x,y) \log_2\left(\frac{P(x,y)}{P(x)P(y)}\right) \\ &= 0.138 \end{aligned}$$

5.2 Mutual Information and Entropy [19pts]

Given a dataset as below.

<i>Sr.No.</i>	<i>Age</i>	<i>Immunity</i>	<i>Travelled?</i>	<i>UnderlyingConditions</i>	<i>Self – quarantine?</i>
1	<i>young</i>	<i>high</i>	<i>no</i>	<i>yes</i>	<i>yes</i>
2	<i>young</i>	<i>high</i>	<i>no</i>	<i>no</i>	<i>no</i>
3	<i>midleaged</i>	<i>high</i>	<i>no</i>	<i>yes</i>	<i>yes</i>
4	<i>senior</i>	<i>medium</i>	<i>no</i>	<i>yes</i>	<i>yes</i>
5	<i>senior</i>	<i>low</i>	<i>yes</i>	<i>yes</i>	<i>yes</i>
6	<i>senior</i>	<i>low</i>	<i>yes</i>	<i>no</i>	<i>no</i>
7	<i>midleaged</i>	<i>low</i>	<i>yes</i>	<i>no</i>	<i>yes</i>
8	<i>young</i>	<i>medium</i>	<i>no</i>	<i>yes</i>	<i>no</i>
9	<i>young</i>	<i>low</i>	<i>yes</i>	<i>yes</i>	<i>no</i>
10	<i>senior</i>	<i>medium</i>	<i>yes</i>	<i>yes</i>	<i>yes</i>
11	<i>young</i>	<i>medium</i>	<i>yes</i>	<i>no</i>	<i>yes</i>
12	<i>midleaged</i>	<i>medium</i>	<i>no</i>	<i>no</i>	<i>yes</i>
13	<i>midleaged</i>	<i>high</i>	<i>yes</i>	<i>yes</i>	<i>yes</i>
14	<i>senior</i>	<i>medium</i>	<i>no</i>	<i>no</i>	<i>no</i>

We want to decide whether an individual working in an essential services industry should be allowed to work or self-quarantine. Each input has four features (x_1, x_2, x_3, x_4): Age, Immunity, Travelled, Underlying Conditions. The decision (quarantine vs does not quarantine) is represented as Y .

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

Solution:

$$\begin{aligned}
 H(Y) &= - \sum_{k=1}^k P(y=k) \log_2 P(y=k) \\
 &= - \frac{5}{14} \log_2 \left(\frac{5}{14} \right) + \frac{9}{14} \log_2 \left(\frac{9}{14} \right) \\
 &= 0.940
 \end{aligned}$$

- (b) Find conditional entropy $H(Y|x_1)$, $H(Y|x_3)$, respectively, to at least 3 decimal places. [8pts]

Solution:

$$H(Y|x_1) = H(Y|X = age) = P(x = young) * H(Y|x = young) + P(x = middleaged) * H(Y|x = middleaged) + P(x = senior) * H(Y|x = senior)$$

$$H(Y|x = young) = H(2/5, 3/5) = 0.971$$

$$H(Y|x = middleaged) = H(1, 0) = 0$$

$$H(Y|x = senior) = H(3/5, 2/5) = 0.971$$

$$P(X = young) = 5/14$$

$$P(X = senior) = 5/14$$

$$H(Y|x_1) = (5/14)(0.971) + (5/14)(0) + (5/14)(0.971) = \mathbf{0.693}$$

$$H(Y|Travelled = yes) = H(5/7, 2/7) = 0.863$$

$$H(Y|Travelled = no) = H(4/7, 3/7) = 0.985$$

$$P(Travelled = yes) = 1/2$$

$$P(Travelled = no) = 1/2$$

$$= (1/2) * 0.863 + (1/2) * 0.985 = \mathbf{0.924}$$

- (c) Find mutual information $I(x_1, Y)$ and $I(x_3, Y)$ and determine which one (x_1 or x_3) is more informative.
[4pts]

Solution:

$$I(x_1, Y) = H(Y) - H(Y|x_1) = 0.940 - 0.693 = \mathbf{0.247}$$

$$I(x_3, Y) = H(Y) - H(Y|x_3) = 0.940 - 0.924 = \mathbf{0.016}$$

x_1 is more informative since it contains more information.

- (d) Find joint entropy $H(Y, x_2)$ to at least 3 decimal places. [4pts]

Solution:

$$\begin{aligned}
 H(Y, x_2) &= P(\text{high}, \text{yes}) * \log_2\left(\frac{1}{P(\text{high}, \text{yes})}\right) + P(\text{high}, \text{no}) * \left(\frac{1}{P(\text{high}, \text{no})}\right) \\
 &\quad + P(\text{med}, \text{yes}) * \log_2\left(\frac{1}{P(\text{med}, \text{yes})}\right) + P(\text{med}, \text{no}) * \left(\frac{1}{P(\text{med}, \text{no})}\right) \\
 &\quad + P(\text{low}, \text{yes}) * \log_2\left(\frac{1}{P(\text{low}, \text{yes})}\right) + P(\text{low}, \text{no}) * \left(\frac{1}{P(\text{low}, \text{no})}\right) \\
 &= \frac{3}{14} \log_2\left(\frac{1}{3/14}\right) + \frac{1}{14} \log_2\left(\frac{1}{1/14}\right) + \frac{4}{14} \log_2\left(\frac{1}{4/14}\right) + \frac{2}{14} \log_2\left(\frac{1}{2/14}\right) + \\
 &\quad \frac{2}{14} \log_2\left(\frac{1}{2/14}\right) + \frac{2}{14} \log_2\left(\frac{1}{2/14}\right) = \mathbf{2.467}
 \end{aligned}$$

5.3 Entropy Proofs [10pts]

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

Solution:

$$H(x|y) = \sum_{x,y} P(x,y) \log_2(1/P(x|y))$$

$$H(x) = \sum_x P(x) \log_2(1/P(x))$$

- (b) Using the mathematical definition of $H(X)$ and $H(X|Y)$, prove that $H(X|Y) = H(X)$ if X and Y are independent. (Note: you have to prove it and cannot use the visualization shown in class [found here](#) [7pts])

Solution:

$$\begin{aligned} H(x|y) &= \sum_{x,y} P(x,y) \log(1/P(x|y)) \\ &= \sum_{x,y} P(x,y) \log\left(\frac{P(y)}{P(x,y)}\right) \\ &= \sum_{x,y} P(x)P(y) \log\left(\frac{P(y)}{P(x)P(y)}\right) \\ &= \sum_{x,y} P(x)P(y) \log\left(\frac{1}{P(x)}\right) \\ &= \sum_x P(x) \log\left(\frac{1}{P(x)}\right) \sum_y P(y) \\ &= \sum_x P(x) \log\left(\frac{1}{P(x)}\right) = H(x) \end{aligned}$$

6 Bonus for All [15pts]

- (a) If a random variable X has a Poisson distribution with mean 5, then calculate the expectation $E[(2X - 3)^2]$
[5 pts]

Solution:

YOUR ANSWER HERE

(b) Suppose that X and Y have joint pdf given by [5 pts]

$$f_{X,Y}(x,y) = \begin{cases} 2e^{-2y}, & 0 \leq x \leq 1, y \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

What are the marginal probability density functions for X and Y ?

Solution:

YOUR ANSWER HERE

- (c) A person decides to toss a biased coin with $P(heads) = 0.4$ repeatedly until he gets a head. He will make at most 3 tosses. Let the random variable Y denote the number of heads. Find the variance of Y . [5 pts]

Solution:

YOUR ANSWER HERE