CS7641: Machine Learning **(Due: 05/29/2020)**

Homework Assignment #1

Instructor: Mahdi Roozbahani *Name:* Chen-Yang(Jim), Liu *GTID:* 90345****

Problem 1: Linear Algebra

(25+8=33 points)

1.1: Determinant and Inverse of Matrix [11pts]

Given a matrix M:

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

1. Calculate the determinant of M. [2pts] (Calculation process required.)
2. Calculate M^{-1}. [5pts] (Calculation process required)
   (**Hint:** Please double check your answer and make sure $MM^{-1} = I$)
3. What is the relationship between the determinant of M and the determinant of M^{-1}? [2pts]
4. When does a matrix not have an inverse? Provide an example. [2pts]

Solution:

1. Calculate the determinant of M.

$$\begin{aligned} \det(M) &= \sum_{j=1}^{d} (-1)^{i+j} a_{ij} M_{ij} \ for\ fixed\ i \\ &= \sum_{j=1}^{3} (-1)^(1+j) a_{1j} M_{1j} \\ &= (-1)^{1+1} a_{11} M_{11} + (-1)^{1+2} a_{12} M_{12} + (-1)^{1+3} a_{13} M_{13} \\ &= 1 \times 2 \times 1 + (-1) \times 4 \times (-2) + 1 \times 2 \times (2-1) \\ &= 2 + 8 + 2 \\ &= 12 \end{aligned}$$

2. Calculate M^{-1}.
   We performed Gauss–Jordan elimination to get an inverse matrix.

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

$$= \begin{bmatrix} 1 & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 1 & -\frac{4}{3} & -\frac{2}{3} & \frac{1}{3} & 0 \\ 0 & 0 & 1 & -\frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} 1 & 0 & -\frac{1}{6} & \frac{1}{6} & \frac{1}{6} & 0 \\ 0 & 1 & -\frac{4}{3} & -\frac{2}{3} & \frac{1}{3} & 0 \\ 0 & 0 & 1 & -\frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \frac{1}{12} & \frac{1}{6} & \frac{1}{12} \\ 0 & 1 & 0 & -\frac{4}{3} & \frac{1}{3} & \frac{2}{3} \\ 0 & 0 & 1 & -\frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix}$$

As a result, $M^{-1} = \begin{bmatrix} \frac{1}{12} & \frac{1}{6} & \frac{1}{12} \\ -\frac{4}{3} & \frac{1}{3} & \frac{2}{3} \\ -\frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix}$

Finally,

$$M \times M^{-1} = \begin{bmatrix} 2 & -1 & 1 \\ 4 & 1 & -2 \\ 2 & -1 & 3 \end{bmatrix} \begin{bmatrix} \frac{1}{12} & \frac{1}{6} & \frac{1}{12} \\ -\frac{4}{3} & \frac{1}{3} & \frac{2}{3} \\ -\frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} \frac{1}{6} + \frac{4}{3} - \frac{1}{2} & 0 & 0 \\ 0 & \frac{2}{3} + \frac{1}{3} + 0 & 0 \\ 0 & 0 & \frac{1}{6} - \frac{2}{3} + \frac{3}{2} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

3. What is the relationship between the determinant of M and the determinant of M^{-1}?

$$\begin{aligned} det(M^{-1}) &= \frac{1}{12} \begin{vmatrix} \frac{1}{3} & \frac{2}{3} \\ 0 & \frac{1}{2} \end{vmatrix} - \frac{1}{6} \begin{vmatrix} -\frac{4}{3} & \frac{2}{3} \\ -\frac{1}{2} & \frac{1}{2} \end{vmatrix} + \frac{1}{12} \begin{vmatrix} -\frac{4}{3} & \frac{1}{3} \\ -\frac{1}{2} & 0 \end{vmatrix} \\ &= \frac{1}{12} \times \frac{1}{6} - \frac{1}{6} \times (\frac{-2}{3} + \frac{1}{3}) + \frac{1}{12} \times \frac{1}{6} \\ &= \frac{1}{72} + \frac{1}{18} + \frac{1}{72} \\ &= \frac{1+1+4}{72} = \frac{1}{12} \end{aligned}$$

From the above calculation, we could see

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

4. When does a matrix not have an inverse? Provide an example.
When $det(M) = 0$, then the inverse matrix does not exist. And, if one row in a square matrix is all zero, then the determinant of this matrix must be 0. For example,

Given $a, b, c, d, e, f \in \mathbb{R}^1$, $M = \begin{bmatrix} 0 & 0 & 0 \\ a & b & c \\ d & e & f \end{bmatrix}$

$$\begin{vmatrix} 0 & 0 & 0 \\ a & b & c \\ d & e & f \end{vmatrix} = 0 \times \begin{vmatrix} b & c \\ e & f \end{vmatrix} + 0 \times \begin{vmatrix} a & c \\ d & f \end{vmatrix} + 0 \times \begin{vmatrix} a & b \\ d & e \end{vmatrix} = 0$$

So, M^{-1} does not exist.

1.2 Characteristic Equation [8pts] (Bonus)

Consider the eigenvalue problem:

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

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

Solution:

$$\begin{aligned}
A \in \mathbf{R}^{d \times d}, Ax &= \lambda x, x \neq 0 \\
Ax - \lambda x &= 0 \\
(A - \lambda I)x &= 0 \\
(A - \lambda I)x &= 0x \\
&\Rightarrow dim(Null(A - \lambda I)) \neq 0 \; since \; x \neq 0 \\
&(Rank(A - \lambda I) + dim(Null(A - \lambda I)) = d) \\
&\Rightarrow rank(A - \lambda I) < d \\
&\Rightarrow det(A - \lambda I) = 0
\end{aligned}$$

From the above result, $(A - \lambda I)_{d \times d}$ matrix must contain one eigenvalue that is 0.

That is, according to Rank theorem, $rank(A-\lambda I)+dim(Null(A-\lambda I)) = d$, since $dim(Null(A-\lambda I)) \neq 0$, it means that $rank(A-\lambda I) < d$. Then, $A-\lambda I$ matrix is not a full-rank matrix. Therefore, $det(A - \lambda I) = 0$

1.3 Singular Value Decomposition [14pts]

Given a matrix A:

$$A = \begin{bmatrix} 3 & 3 & 0 \\ -2 & 2 & 0 \end{bmatrix}$$

Compute the Singular Value Decomposition (SVD) by following the steps below. Your full calculation process is required.

1. Calculate all eigenvalues of AA^T and A^TA. The square roots of the positive eigenvalues make up the singular values, the diagonal entries in Σ. They will be arranged in descending order, all other values in Σ are 0. [4pts]
2. Calculate all eigenvectors of AA^T normalized to unit length. These will make up the left singular vectors, or the columns of U. [4pts]
3. Calculate all eigenvectors of A^TA normalized to unit length. These will make up the right singular vectors, or the rows of V^T. [4pts]
4. Put it all together. Write out the SVD of matrix A in the following form: $A = U\Sigma V^T$[2pts]
   Hint: Reconstruct matrix A from the SVD to check your answer.

Solution:

1. Calculate all eigenvalues of AA^T and A^TA. The square roots of the positive eigenvalues make up the singular values, the diagonal entries in Σ. They will be arranged in descending order, all other values in Σ are 0.

$$AA^T = \begin{bmatrix} 3 & 3 & 0 \\ -2 & 2 & 0 \end{bmatrix} \begin{bmatrix} 3 & -2 \\ 3 & 2 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 18 & 0 \\ 0 & 8 \end{bmatrix}$$

$$A^TA = \begin{bmatrix} 3 & -2 \\ 3 & 2 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 3 & 3 & 0 \\ -2 & 2 & 0 \end{bmatrix} = \begin{bmatrix} 13 & 5 & 0 \\ 5 & 13 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

- Eigenvalue of AA^T

$$\begin{aligned} AA^Tx &= \lambda x \\ &\Rightarrow (AA^T - \lambda I)x = 0 \\ &\Rightarrow \begin{bmatrix} 18-\lambda & 0 \\ 0 & 8-\lambda \end{bmatrix} \qquad det(AA^T - \lambda I) = 0 \Rightarrow \lambda = 18 \; or \; 8 \end{aligned}$$

- Eigenvalue of A^TA

$$
\begin{aligned}
(A^TA - \lambda I)x &= 0 \\
det(A^TA - \lambda I) = 0 \Rightarrow \begin{vmatrix} 13-\lambda & 5 & 0 \\ 5 & 13-\lambda & 0 \\ 0 & 0 & -\lambda \end{vmatrix} &= (13-\lambda)(13-\lambda) - \lambda + 25\lambda = 0 \\
&= -169\lambda + 26\lambda^2 - \lambda^3 + 25\lambda = 0 \\
&= \lambda^3 + 26\lambda^2 - 144\lambda = 0 \\
&= \lambda(\lambda^2 + 26\lambda - 144) = 0 \\
&= \lambda(\lambda - 8)(\lambda - 18) \\
\lambda &= 0 \ or \ 8 \ or \ 18
\end{aligned}
$$

Then,

$$\Sigma = \begin{bmatrix} \sqrt{18} & 0 & 0 \\ 0 & \sqrt{8} & 0 \end{bmatrix}$$

2. Calculate all eigenvectors of AA^T normalized to unit length. These will make up the left singular vectors, or the columns of U.

$AA^Tx = \lambda x$

- $\lambda = 8$

$$\begin{bmatrix} 18 & 0 \\ 0 & 8 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = 8 \begin{bmatrix} a \\ b \end{bmatrix} \Rightarrow \begin{cases} 18a = 8a \\ 8b = 8b \end{cases} \Rightarrow a = 0, b = t, t \in \mathbb{R}$$

Let t = 1, we get a unit-length eigenvector, that is, $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$

- $\lambda = 18$

$$\begin{bmatrix} 18 & 0 \\ 0 & 8 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = 18 \begin{bmatrix} a \\ b \end{bmatrix} \Rightarrow \begin{bmatrix} 0 & 0 \\ 0 & -10 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow \begin{cases} 0a + 0b = 0 \\ -10b = 0 \end{cases}$$

$b = 0, a = t, t \in \mathbb{R}$, let t = 1, then we get a unit-length eigenvector, that is, $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$

Therefore, $U = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

3. Calculate all eigenvectors of A^TA normalized to unit length. These will make up the right singular vectors, or the rows of V^T.

$A^TAx = \lambda x$

- $\lambda = 0$

$$\begin{bmatrix} 13 & 5 & 0 \\ 5 & 13 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow \begin{cases} 13a + 5b + 0c = 0 \\ 5a + 13b + 0c = 0 \end{cases} \quad \Rightarrow a = b = 0$$

  Let $c = t, t \in \mathbb{R}$, then we get a unit-length eigenvector if $t = 1$, that is, $\begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$

- $\lambda = 18$

$$\begin{bmatrix} -5 & 5 & 0 \\ 5 & -5 & 0 \\ 0 & 0 & -18 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \begin{cases} -5a + 5b + 0c = 0 \\ 5a - 5b + 0c = 0 \\ -18c = 0 \end{cases}$$

$$\Rightarrow c = 0, \begin{cases} -5a + 5b = 0 \\ 5a - 5b = 0 \end{cases}$$

  Let $a = t \Rightarrow b = t, t \in \mathbb{R}$ If t = 1, then an unit-length eigenvector is, $\begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{bmatrix}$

- $\lambda = 8$

$$\begin{bmatrix} 5 & 5 & 0 \\ 5 & 5 & 0 \\ 0 & 0 & -8 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \begin{cases} 5a + 5b + 0c = 0 \\ 5a + 5b + 0c = 0 \\ -8c = 0 \end{cases}$$

$$\Rightarrow c = 0, \begin{cases} 5a + 5b = 0 \\ 5a + 5b = 0 \end{cases}$$

  Let $a = t \Rightarrow b = -t, t \in \mathbb{R}$ If t = 1, then an unit-length eigenvector is, $\begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \\ 0 \end{bmatrix}$

Finally,

$$V = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \end{bmatrix}, V^T = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

4. Put it all together. Write out the SVD of matrix A in the following form: $A = U\Sigma V^T$

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \sqrt{18} & 0 & 0 \\ 0 & \sqrt{8} & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 3 & 3 & 0 \\ -2 & 2 & 0 \end{bmatrix}$$

Problem 2: Expectation, Co-variance and Independence (25 points)

Suppose X, Y and Z are three different random variables. Let X obeys Bernouli Distribution. The probability disbribution function is

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

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

1. What is the Expectation and Variance of X?(in terms of c) [4pts]
2. Show that Z also follows a Normal (Gaussian) distribution. Calculate the Expectation and Variance of Z. [9pts]
3. How should we choose c such that Y and Z are uncorrelated(which means $Cov(Y,Z) = 0$)? [5pts]
4. Determine whether the following probability is greater than or equal to 0: (1) $P(Y = 0)$; (2) $P(Z = c)$; (3) $P(Y \in (-1,0))$; (4) $P(Z \in (2c, 3c))$; (5) $P(Y \in (-1,0), Z \in (2c,3c))$; (6) $P(Y \in (-2,-1), Z \in (c, 2c))$.[3pts]
5. Are Y and Z independent? Make use of the above probabilities to show your conclusion.[4pts]

Solution:

1. What is the Expectation and Variance of X?(in terms of c)

$$E(x) = 0.5 \times c + 0.5 \times (-c) = 0$$
$$Var(x) = E(x - \mu_x)^2 = c^2 \times 0.5 + (-c)^2 \times 0.5 = c^2$$

2. Show that Z also follows a Normal (Gaussian) distribution. Calculate the Expectation and Variance of Z.

$$\begin{aligned}
Z &= XY \\
F_Z(z) &= P(Z \le z) \\
&= P(Z \le z \& x = c) + P(Z \le z \& x = -c)\ (Sum\ Rule) \\
&= P(Z \le z | x = c)P(x = c) + P(Z \le z | x = -c)P(x = -c)\ (Product\ Rule) \\
&= \frac{1}{2}P[cY \le z] + \frac{1}{2}P[-cY \le z] \\
&= \frac{1}{2}[P(Y \le \frac{z}{c} + P(Y \ge \frac{z}{c}] \\
&= P(Y \le \frac{z}{c})\ sinceY\ \sim N(0,1),\ it\ is\ symmetric.\ i.e.P(Y \le -y) = P(Y \ge y)
\end{aligned}$$

Thus, $F_Z(z) = \Phi(\frac{z}{c}), \Phi : CDF\ of\ N(0,1)$. Then, $Z \sim Normal\ Distribution.$
Now assume $Z \sim N(\mu, \sigma^2)$,

$$P(\frac{Z-\mu}{\sigma} \le \frac{z-\mu}{\sigma}) = \Phi(\frac{z-\mu}{\sigma}) = \Phi(\frac{z-0}{c}) \Rightarrow E(Z) = 0, Var(Z) = c^2$$

3. How should we choose c such that Y and Z are uncorrelated(which means $Cov(Y, Z) = 0$)? First, we will use $E(Y^2)$ later ,so we derive its property to get the result. Since $Y \sim N(0, 1)$, $Y^2 \sim \chi_1^2$. Then, the expected value of chi-square distribution is $E(Y^2) = 1$

$$
\begin{aligned}
cov(Y, Z) &= E(Y - \mu_y)(Z - \mu_z) \\
&\Rightarrow E(YZ) = 0 \ (E(Z) = E(Y) = 0) \\
P(YZ) &= \begin{cases} 0.5, \ YZ = cY^2 \\ 0.5, \ YZ = -cY^2 \end{cases} \\
E(YZ) &= E(XY^2) = |c|E(Y^2) \ \textit{since c is a constant.} \\
E(Y^2) &= 1 \\
&\Rightarrow c = 0
\end{aligned}
$$

4. Determine whether the following probability is greater than or equal to 0: (1) $P(Y = 0)$; (2) $P(Z = c)$; (3) $P(Y \in (-1, 0))$; (4) $P(Z \in (2c, 3c))$; (5) $P(Y \in (-1, 0), Z \in (2c, 3c))$; (6) $P(Y \in (-2, -1), Z \in (c, 2c))$.

   (a) $P(Y = 0) = 0$ since Y is a continuous distribution.

   (b) $P(Z = c) = 0$ since Z is a continuous distribution.

   (c) $P(Y \in (-1, 0)) = \Phi(0) - \Phi(-1) \approx 0.34 > 0$

   (d) $P(Z \in (2c, 3c)) = P(\frac{2c-0}{c} \leq \frac{Z-0}{c} \leq \frac{3c-0}{c}) = \Phi(3) - \Phi(2) \approx 0.02 > 0$

   (e) $P(Y \in (-1, 0), Z \in (2c, 3c)) = P(Y \in (-1, 0))P(Z \in (2c, 3c)|Y \in (-1, 0))$ since $P(Y\&Z) = P(Y)P(Z|Y), P(Y) > 0, P(Z|Y) > 0$.
   That is, the intersection of two sets $P(Z \in (2c, 3c)|Y \in (-1, 0)) > 0$. So, $P(Y \in (-1, 0), Z \in (2c, 3c)) > 0$

   (f) $P(Y \in (-2, -1), Z \in (c, 2c)) = P(Y \in (-2, -1)P(Z \in (c, 2c)|Y \in (-2, -1))$ since $P(Y \in (-2, -1)) = \Phi(-1) - \Phi(-2) > 0$, the intersection of set, that is, $P(Z \in (c, 2c)|Y \in (-2, -1)) > 0$ So, $P(Y \in (-2, -1), Z \in (c, 2c)) > 0$

5. Are Y and Z independent? Make use of the above probabilities to show your conclusion.
   If Y and Z are independent, $P(YZ) = P(Y)P(Z|Y) = P(Y)P(Z) = 0$ But, from(4), we know that $P(Y \in (-1, 0), Z \in (2c, 3c)) > 0$. So they are not independent.

Problem 3: Maximum Likelihood (25+10=35 points)

3.1 Discrete Example [10 pts] Suppose we have two types of coins, A and B. The probability of a Type A coin showing heads is θ. The probability of a Type B coin showing heads is 2θ. Here, we have a bunch of coins of either type A or B. Each time we choose one coin and flip it. We do this experiment 10 times and the results are shown in the chart below.

Coin Type	Result
A	Tail
A	Head
A	Tail
B	Head
A	Tail
A	Tail
B	Head
B	Head
B	Head
A	Tail

1. What is the likelihood of the result given θ? [4pts]
2. What is the maximum likelihood estimation for θ? [6pts]

Solution:

1. What is the likelihood of the result given θ?

$$\begin{aligned} L(\theta) &= \frac{1}{2} \times (1-\theta) \times \frac{1}{2} \times \theta \times \frac{1}{2} \times (1-\theta) \times \frac{1}{2} \times 2\theta \times \frac{1}{2} \times (1-\theta) \\ &\times \frac{1}{2} \times (1-\theta) \times \frac{1}{2} \times 2\theta \times \frac{1}{2} \times 2\theta \times \frac{1}{2} \times 2\theta \times \frac{1}{2} \times (1-\theta) \\ &= \frac{1}{2}^{10} (1-\theta)^5 \theta^5 \times 2^4 \\ &= \frac{1}{2}^{6} (1-\theta)^5 \theta^5 \end{aligned}$$

2. What is the maximum likelihood estimation for θ?

$$\begin{aligned} l(\theta) &= logL(\theta) = -6log(2) + 5log(1-\theta) + 5log(\theta) \\ \frac{l(\theta)}{\theta} &= \frac{5}{1-\theta}(-1) + \frac{5}{\theta} = 0 \\ &\Rightarrow \frac{-5\theta + 5(1-\theta)}{\theta(1-\theta)} = 0 \\ &\Rightarrow -5\theta + 5 - 5\theta = 0 \\ &\Rightarrow \hat{\theta} = \frac{1}{2} \end{aligned}$$

3.2 CDF Example [10 pts] The C.D.F of independent random variables $X_1, X_2, ..., X_n$ is

$$P(X_i \leq x|\alpha,\beta) = \begin{cases} 0, & x < 0 \\ (\frac{x}{\beta})^{\alpha}, & 0 \leq x \leq \beta \\ 1, & x > \beta \end{cases}$$

where $\alpha \geq 0$, $\beta \geq 0$. Find the MLEs of α and β.

Solution: PDF: $P(X_i = x|\alpha,\beta) = \alpha(\frac{x}{\beta})^{\alpha-1}\frac{1}{\beta}$ where $\alpha \geq 0$, $\beta \geq x \geq 0$. We will use an indicator function in this problem and then define:

$$I = \begin{cases} 1 & \beta \geq x_i \geq 0 \\ 0 & otherwise \end{cases}$$

$$\begin{aligned} L(\alpha,\beta) &= \prod_{i=1}^{n} P(X_i = x_i|\alpha,\beta) \\ &= \alpha^n(\frac{1}{\beta})^{n\alpha-n}\prod_{i=1}^{n} x_i(\frac{1}{\beta})^n I(\beta \geq x_i) \\ &= \alpha^n(\frac{1}{\beta})^{n\alpha}\prod_{i=1}^{n} x_i I(\beta \geq x_i) \\ l = l(\alpha,\beta) &= nlog(\alpha) - n\alpha log(\beta) + log(\prod_{i=1}^{n} x_i) \\ \frac{\partial l}{\partial \beta} &= -n\alpha\frac{1}{\beta}\prod_{i=1}^{n} I(\beta \geq x_i) \\ &\Rightarrow \hat{\beta} = X_{(n)} \end{aligned}$$

From the above calculation, we would like to maximize the function, so we decide when β is the smallest value and is from x_i where $X_{(n)}$ means the largest value among $x_i, i = 1, 2, \ldots, n$
Now, to get α

$$\begin{aligned} L(\alpha,\beta) &= \alpha^n(\frac{1}{\beta})^{n\alpha}\prod_{i=1}^{n} x_i \\ l = l(\alpha,\beta) &= nlog(\alpha) - n\alpha log(\beta) + log(\prod_{i=1}^{n} x_i) \\ \frac{\partial l}{\partial \alpha} &= \frac{n}{\alpha} - nlog(\beta) \\ &= \frac{n}{\alpha} - nlog(\hat{\beta}) = 0 \\ &\Rightarrow \hat{\alpha} = \frac{1}{log(X_{(n)})} \end{aligned}$$

3.3 Poisson distribution [5 pts]

The Poisson distribution is defined as

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

What is the maximum likelihood estimator of λ?

Solution:

$$\begin{aligned}
L(\lambda) &= \prod_{i=1}^{n} P(x_i = k_i) \\
&= \lambda^{\sum_{i=1}^{n} k_i} e^{-n\lambda} / (\prod_{i=1}^{n} k_i) \\
l = l(\lambda) &= \sum_{i=1}^{n} k_i log(\lambda) - n\lambda - log(\prod_{i=1}^{n} k_i) \\
\frac{\partial l}{\partial \lambda} &= \sum_{i=1}^{n} k_i/\lambda - n = 0 \\
\hat{\lambda} &= \sum_{i=1}^{n} k_i/n
\end{aligned}$$

3.4 Bonus [10 pts]

Given n i.i.d. observations $\{(x_i, y_i)\}_{i=1}^n \in \mathbb{R}^d \times \{-1, 1\}$, we assume

$$\mathbb{P}(y_i = 1|x_i) = h(x_i^T\theta) \text{ and } \mathbb{P}(y_i = -1|x_i) = 1 - h(x_i^T\theta)$$

where $h(x) = \frac{1}{1+\exp(-x)}$ and θ is the model parameter and $\theta = (\theta_1, \theta_2, \ldots, \theta_d)^T$.

Write out the likelihood function $L(\theta)$ given (x_i, y_i). Then formulate the log-likelihood function.

Solution:

- Likelihood Function:
  First, we define indicator function as follows:

$$I(y_i) = \begin{cases} 1 & y_i = 1 \\ 0 & y_i = -1 \end{cases}$$

$$\begin{aligned} L(\theta) &= \prod_{i=1}^n h(x_i^T\theta)^{I(y_i=1)}(1 - h(x_i^T\theta))^{1-I(y_i=1)} \\ &= h(x_i^T\theta)^{\sum_{i=1}^n I(y_i=1)}(1 - h(x_i^T\theta))^{n-\sum_{i=1}^n I(y_i=1)} \\ &= (\frac{1}{1 + exp(-x_i^T\theta)})^{\sum_{i=1}^n I(y_i=1)}(\frac{exp(-x_i^T\theta)}{1 + exp(-x_i^T\theta)})^{n-\sum_{i=1}^n I(y_i=1)} \end{aligned}$$

- Formulate the log-likelihood function

$$l(\theta) = -\sum_{i=1}^n I(y_i = 1)log(1 + exp(-x_i^T\theta)) + (n - \sum_{i=1}^n I(y_i = 1))[(-x_i^T\theta) - log(1 + exp(-x_i^T\theta))]$$

Problem 4: Information Theory (25+7=32 points)

4.1 Marginal Distribution [6pts] Suppose the joint probability distribution of two binary random variables X and Y are given as follows.

$X \mid Y$	1	2
0	$\frac{1}{4}$	$\frac{1}{4}$
1	$\frac{1}{2}$	0

1. Show the marginal distribution of X and Y, respectively. [3pts]
2. Find mutual information for the joint probability distribution in the previous question [3pts]

Solution:

1. Show the marginal distribution of X and Y, respectively.

$$
\begin{aligned}
f_X(x) &= \sum_y f_{x,y}(x,y) \\
&= \begin{cases} \frac{1}{2} & , x = 0 \\ \frac{1}{2} & , x = 1 \end{cases} \\
f_Y(y) &= \sum_x f_{x,y}(x,y) \\
&= \begin{cases} \frac{3}{4} & , y = 1 \\ \frac{1}{4} & , y = 2 \end{cases}
\end{aligned}
$$

2. Find mutual information for the joint probability distribution in the previous question.

$$
\begin{aligned}
I(X,Y) &= \sum_{X \in x} \sum_{Y \in y} P(X,Y) log(\frac{P(X,Y)}{P(X)P(Y)}) \\
&= I(x=0,y=1) + I(x=0,y=2) + I(x=1,y=1) + I(x=1,y=2) \\
&= \frac{1}{4} \times log(\frac{1/4}{3/8}) + \frac{1}{4} \times log(\frac{1/4}{1/8}) + \frac{1}{2} \times log(\frac{1/2}{3/8}) \\
&= 0.311
\end{aligned}
$$

4.2 Mutual Information and Entropy [19pts] Given a dataset as below.

$Player$	$Experience$	$NumUtilities$	$BuysBoardwalk?$	$Hunger$	$Outcome$
1	$novice$	2	no	low	$lose$
2	$intermediate$	0	no	$high$	$lose$
3	$novice$	1	no	low	win
4	$expert$	0	no	$medium$	win
5	$intermediate$	0	yes	$high$	win
6	$expert$	0	yes	$high$	$lose$
7	$intermediate$	2	yes	low	win
8	$intermediate$	1	no	$medium$	win
9	$expert$	1	no	low	$lose$
10	$novice$	0	no	$medium$	$lose$
11	$novice$	2	yes	low	win
12	$intermediate$	1	no	$medium$	$lose$
13	$intermediate$	0	yes	$high$	win
14	$novice$	0	yes	$high$	$lose$

You are analyzing data from your last few Monopoly games in hopes of becoming a world champion. We want to determine what makes a player win or lose. Each input has four features (x_1, x_2, x_3, x_4): Experience, NumUtilities, BuysBoardwalk, Hunger. The outcome (win vs lose) is represented as Y.

1. Find entropy $H(Y)$. [3pts]
2. Find conditional entropy $H(Y|x_1)$, $H(Y|x_4)$, respectively. [8pts]
3. Find mutual information $I(x_1, Y)$ and $I(x_4, Y)$ and determine which one (x_1 or x_4) is more informative. [4pts]
4. Find joint entropy $H(Y, x_3)$. [4pts]

Solution:

1. Find entropy $H(Y)$.
   Lose: 7, Win: 7

$$\begin{aligned} H(Y) &= -\sum_i P_i \times logP_i \\ &= -\frac{1}{2}log(\frac{1}{2}) - \frac{1}{2}log(\frac{1}{2}) \\ &= -log(\frac{1}{2}) = 1 \end{aligned}$$

2. Find conditional entropy $H(Y|x_1)$, $H(Y|x_4)$, respectively.
For $H(Y|x_1)$,

	novice	intermediate	expert
Lose	$\frac{3}{5}$	$\frac{2}{6}$	$\frac{2}{3}$
Win	$\frac{2}{5}$	$\frac{4}{6}$	$\frac{1}{3}$
	1.0	1.0	1.0

$$
\begin{aligned}
H(Y|X_1 = novice) &= H(\frac{3}{5}, \frac{2}{5}) \\
&= \sum_{Y=lose,win} P(x = novice, y)log(\frac{P(x = novice)}{P(x = novice, y)}) \\
&= \frac{3}{5}log(\frac{5}{3}) + \frac{2}{5}log(\frac{5}{2}) \\
&= 0.97 \\
H(Y|X_1 = intermediate) &= H(\frac{2}{6}, \frac{4}{6}) \\
&= \sum_{Y=lose,win} P(x = intermediate, y)log(\frac{P(x = intermediate)}{P(x = intermediate, y)}) \\
&= \frac{2}{6}log(\frac{6}{2}) + \frac{4}{6}log(\frac{6}{4}) \\
&= 0.91 \\
H(Y|X_1 = expert) &= H(\frac{2}{3}, \frac{1}{3}) \\
&= \sum_{Y=lose,win} P(x = expert, y)log(\frac{P(x = expert)}{P(x = expert, y)}) \\
&= \frac{2}{3}log(\frac{3}{2}) + \frac{1}{3}log(\frac{3}{1}) \\
&= 0.91
\end{aligned}
$$

Thus, $H(Y|x_1) = 5/14 * 0.97 + 6/14 * 0.91 + 3/14 * 0.91 = 0.9314285714285715$
For $H(Y|x_4)$,

	low	medium	high
Lose	$\frac{2}{5}$	$\frac{2}{4}$	$\frac{3}{5}$
Win	$\frac{3}{5}$	$\frac{2}{4}$	$\frac{2}{5}$
	1.0	1.0	1.0

$$
\begin{aligned}
H(Y|X_1 = low) &= H(\frac{2}{5}, \frac{3}{5}) \\
&= \sum_{Y=lose,win} P(x = low, y) log(\frac{P(x = low)}{P(x = low, y)}) \\
&= \frac{2}{5} log(\frac{5}{2}) + \frac{3}{5} log(\frac{5}{3}) \\
&= 0.97 \\
H(Y|X_1 = medium) &= H(\frac{2}{4}, \frac{2}{4}) \\
&= \sum_{Y=lose,win} P(x = medium, y) log(\frac{P(x = medium)}{P(x = medium, y)}) \\
&= \frac{2}{4} log(\frac{4}{2}) + \frac{2}{4} log(\frac{4}{2}) \\
&= 1 \\
H(Y|X_1 = high) &= H(\frac{3}{5}, \frac{2}{5}) \\
&= \sum_{Y=lose,win} P(x = high, y) log(\frac{P(x = high)}{P(x = high, y)}) \\
&= \frac{3}{5} log(\frac{5}{3}) + \frac{2}{5} log(\frac{5}{2}) \\
&= 0.97
\end{aligned}
$$

Thus, $H(Y|x_4) = 5/14 * 0.97 + 4/14 * 1 + 5/14 * 0.97 = 0.9785714285714285$

3. Find mutual information $I(x_1, Y)$ and $I(x_4, Y)$ and determine which one (x_1 or x_4) is more informative.

   $I(x_1, Y) = H(Y) - H(Y|x_1) = 1 - 0.9314285714285715 = 0.0685714285714285$
   $I(x_4, Y) = H(Y) - H(Y|x_4) = 1 - 0.9785714285714285 = 0.021428571428571463$
   Since $I(x_1, Y) > I(x_4, Y)$, the more reduction in this means that the feature is more informative. So the feature x_1(Experience) is more informative.

4. Find joint entropy $H(Y, x_3)$.

	no	yes	
Lose	$\frac{5}{14}$	$\frac{2}{14}$	$\frac{1}{2}$
Win	$\frac{3}{14}$	$\frac{4}{14}$	$\frac{1}{2}$
	$\frac{8}{14}$	$\frac{6}{14}$	

$$
\begin{aligned}
H(Y, x_3) &= \sum_{y=lose,win,x=no,yes} P(x,y)log(1/P(x,y)) \\
&= \frac{5}{14}log(\frac{14}{5}) + \frac{2}{14}log(\frac{14}{2}) + \frac{3}{14}log(\frac{14}{3}) + \frac{4}{14}log(\frac{14}{4}) \\
&= 1.92
\end{aligned}
$$

4.3 Bonus Question [7pts]

1. Suppose X and Y are independent. Show that $H(X|Y) = H(X)$. [2pts]
2. Suppose X and Y are independent. Show that $H(X,Y) = H(X) + H(Y)$. [2pts]
3. Prove that the mutual information is symmetric, i.e., $I(X,Y) = I(Y,X)$ and $x_i \in X, y_i \in Y$ [3pts]

Solution:

1. Suppose X and Y are independent. Show that $H(X|Y) = H(X)$.

$$\begin{aligned}
H(X|Y) &= \sum_{Y \in Y} P(y)H(X|Y=y) \\
&= \sum_{x \in X, y \in Y} P(x,y)log(\frac{P(y)}{P(x,y)}) \\
&= \sum_{x \in X, y \in Y} P(x,y)log(\frac{1}{P(x)}) \ (since\ X \perp\!\!\!\perp Y \Rightarrow P(x,y) = P(x)P(y)) \\
&= \sum_{x \in X} P(x)log(\frac{1}{P(x)}) \ by\ Sum\ Rule \\
&= H(x)
\end{aligned}$$

2. Suppose X and Y are independent. Show that $H(X,Y) = H(X) + H(Y)$.

$$\begin{aligned}
H(X,Y) &= \sum_{x \in X, y \in Y} P(x,y)log(\frac{1}{P(x,y)}) \\
&= -\sum_{x \in X, y \in Y} P(x)P(y)log(P(x)) + log(P(y)) \\
&= -\sum_{x \in X, y \in Y} P(x)P(y)log(P(x) + P(x)P(y)log(P(y)) \ by\ distributive\ rule \\
&= -\sum_{x \in X, y \in Y} P(x,y)log(P(x) + P(x,y)log(P(y)) \ (since\ X \perp\!\!\!\perp Y \Rightarrow P(x,y) = P(x)P(y)) \\
&= -\sum_{x \in X} P(x)log(P(x) - \sum_{y \in Y} P(y)log(P(y)) \ by\ Sum\ Rule \\
&= \sum_{x \in X} P(x)log(\frac{1}{P(x)}) + \sum_{y \in Y} P(y)log(\frac{1}{P(y)}) \\
&= H(x) + H(y)
\end{aligned}$$

3. Prove that the mutual information is symmetric, i.e., $I(X,Y) = I(Y,X)$ and $x_i \in X, y_i \in Y$

$$
\begin{aligned}
I(X,Y) &= H(Y) - H(Y|X) \\
&= \sum_{x\in X} P(y)log(\frac{1}{P(y)}) - \sum_{x\in X, y\in Y} log(\frac{1}{P(y|x)}) \\
&= \sum_{y\in Y} P(y)log(\frac{1}{P(y)}) - \sum_{x\in X, y\in Y} P(x,y)log(\frac{P(x)}{P(x,y)}) \\
&= \sum_{x\in X, y\in Y} P(x,y)log(\frac{1}{P(y)}) - \sum_{x\in X, y\in Y} P(x,y)log(\frac{P(x)}{P(x,y)}) \\
&= \sum_{x\in X, y\in Y} P(x,y)log(\frac{P(x,y)}{P(y)P(x)}) \\
&= \sum_{x\in X} P(x)log(\frac{1}{P(x)}) - \sum_{x\in X, y\in Y} log(\frac{P(y)}{P(x,y)}) \\
&= \sum_{x\in X} P(x)log(\frac{1}{P(x)}) - \sum_{x\in X, y\in Y} log(\frac{1}{P(x|y)}) \\
&= H(X) - H(X|Y) \\
&= I(Y,X)
\end{aligned}
$$

Problem 5: Bonus for All (10 points)

Due to the recent social distancing requirement, Wal-Mart is re-evaluating their delivery policies. In order to properly update their policy, Wal-Mart is analyzing data from previous records. Delivery time can be classified as early, on time or late. Delivery distance can be classified as within 5 miles, between 5 and 10 miles and over 10 miles. From the previous records, 15% of deliveries arrive early, and 55% arrive on time. 70% of orders are within 5 miles and 25% of orders are between 5 and 10 miles. The probability for arriving on time if delivery distance is over 10 miles is 0. The probability of a shipment arriving on time and having a delivery distance between 5 and 10 miles is 10%. The probability for arriving early if delivery distance is within 5 miles is 20%.

1. What is the probability that the delivery will arrive on time if the distance is between 5 and 10 miles? [2 pts]

2. What is the probability that the delivery will arrive on time if the distance is within 5 miles? [4 pts]

3. What is the probability that the delivery will arrive late if the distance is within 5 miles? [4 pts]

Solution:

Delivery	< 5 miles	5∼10 miles	>10 miles	Sub-total
Early	20%			15%
On time	45%	10%	0%	55%
Late	5%			30%
Sub-total	70%	25%	5%	100%

1. What is the probability that the delivery will arrive on time if the distance is between 5 and 10 miles?

   $$P(On\ time|5 \sim 10\ miles) = P(On\ Time \cap 5 \sim 10\ miles)/P(5 \sim 10\ miles) = 0.1/0.25 = 2/5$$

2. What is the probability that the delivery will arrive on time if the distance is within 5 miles?

   $$P(On\ time| < 5\ miles) = P(On\ Time \cap < 5\ miles)/P(< 5\ miles) = 0.45/0.70 = 9/14$$

3. What is the probability that the delivery will arrive late if the distance is within 5 miles?

   $$P(Late| < 5\ miles) = P(Late \cap < 5\ miles)/P(< 5\ miles) = 0.05/0.70 = 1/14$$