CS 7641 Machine Learning Fall 2019 HW1

Grade: 129/130

1 Linear Algebra (25pts + 8pts)

1.1 Determinant and Inverse of Matrix [11pts]

Given a matrix M:

$$M = \begin{bmatrix} 5 & 0 & 1 \\ 6 & 1 & 2 \\ 0 & 4 & 3 \end{bmatrix}$$

- Calculate the determinant of M. [5pts] (Calculation process required)
- Does the inverse of M exist? If so, calculate M^{-1}. [6pts] (Calculation process required)

  (**Hint:** please double check your answer and make sure $MM^{-1} = I$)

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 $|A - \lambda I| = 0$.

(**Hint**: If a matrix is not full-rank (has linearly dependent columns), it is singular and non-invertible)

1.3 Eigenvalue [7pts]

Following 1.2, given a matrix A:

$$A = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$$

Calculate all the eigenvalues of A. (Calculation process required. Your answer should be expressed as a function of r.)

1.4 Eigenvector [7pts]

Following 1.3, given that the l_2 norm of each eigenvector is 1, what are the eigenvectors of matrix A? For example, if an eigenvector is $v = \begin{bmatrix} x1 \\ x2 \end{bmatrix}$, then $||v||_2 = \sqrt{x_1^2 + x_2^2} = 1$ (Calculation process required.)

Answers begin here

1.1

1.1.1 $|M| = 5\times(1\times3-2\times4)-0(6\times3-2\times0)+1\times(6\times4-1\times0)$

$= 5\times(3-8)-0+24$

$= 5\times(-5)+24$

$= \boxed{-1}$ the determinant of M.

1.1.2 By observation, it looks like that both M's rows and columns are linearly independent, so M is invertible. Furthermore, since $|M| = -1 \neq 0$, $\boxed{\text{M is invertible}}$. Write M like this and calculate M^{-1}:

$$\begin{array}{c} ① \\ ② \\ ③ \end{array}\left[\begin{array}{ccc|ccc} 5 & 0 & 1 & 1 & 0 & 0 \\ 6 & 1 & 2 & 0 & 1 & 0 \\ 0 & 4 & 3 & 0 & 0 & 1 \end{array}\right] \times 4 - ③ \Rightarrow \left[\begin{array}{ccc|ccc} 5 & 0 & 1 & 1 & 0 & 0 \\ 24-0 & 4-4 & 8-3 & 0 & 4 & -1 \\ 0 & 4 & 3 & 0 & 0 & 1 \end{array}\right]$$

$$= \left[\begin{array}{ccc|ccc} 5 & 0 & 1 & 1 & 0 & 0 \\ 24 & 0 & 5 & 0 & 4 & -1 \\ 0 & 4 & 3 & 0 & 0 & 1 \end{array}\right] - 5\times① \Rightarrow \left[\begin{array}{ccc|ccc} 5 & 0 & 1 & 1 & 0 & 0 \\ 24-25 & 0 & 5-5 & 0-5 & 4 & -1 \\ 0 & 4 & 3 & 0 & 0 & 1 \end{array}\right]$$

$$= \left[\begin{array}{ccc|ccc} 5 & 0 & 1 & 1 & 0 & 0 \\ -1 & 0 & 0 & -5 & 4 & -1 \\ 0 & 4 & 3 & 0 & 0 & 1 \end{array}\right] \times -1 \Rightarrow \left[\begin{array}{ccc|ccc} 5 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 5 & -4 & 1 \\ 0 & 4 & 3 & 0 & 0 & 1 \end{array}\right] - 5\times②$$

$$\Rightarrow \left[\begin{array}{ccc|ccc} 5-5 & 0 & 1 & 1-25 & +20 & -5 \\ 1 & 0 & 0 & 5 & -4 & 1 \\ 0 & 4 & 3 & 0 & 0 & 1 \end{array}\right] = \left[\begin{array}{ccc|ccc} 0 & 0 & 1 & -24 & 20 & -5 \\ 1 & 0 & 0 & 5 & -4 & 1 \\ 0 & 4 & 3 & 0 & 0 & 1 \end{array}\right] - 3\times①$$

$$\Rightarrow \left[\begin{array}{ccc|ccc} 0 & 0 & 1 & -24 & 20 & -5 \\ 1 & 0 & 0 & 5 & -4 & 1 \\ 0 & 4 & 0 & 72 & -60 & 16 \end{array}\right] \div 4 = \left[\begin{array}{ccc|ccc} 0 & 0 & 1 & -24 & 20 & -5 \\ 1 & 0 & 0 & 5 & -4 & 1 \\ 0 & 1 & 0 & 18 & -15 & 4 \end{array}\right]$$

Therefore, $\boxed{M^{-1} = \begin{bmatrix} 5 & -4 & 1 \\ 18 & -15 & 4 \\ -24 & 20 & -5 \end{bmatrix}}$

Check: $MM^{-1} = \begin{bmatrix} 5 & 0 & 1 \\ 6 & 1 & 2 \\ 0 & 4 & 3 \end{bmatrix} \begin{bmatrix} 5 & -4 & 1 \\ 18 & -15 & 4 \\ -24 & 20 & -5 \end{bmatrix}$

$$= \begin{bmatrix} 25+0+(-24) & -20+0+20 & 5+0-5 \\ 30+18-48 & -24-15+40 & 6+4-10 \\ 0+72-72 & 0-60+60 & 0+16-15 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = I$$

1.2

proof: Since $Ax = \lambda x$, we know that $(A-\lambda I)x = 0$. By definition, eigenvector x is in the nullspace of the matrix $(A-\lambda I)$. (The nullspace of a matrix M is the set of all vectors $\vec{v} \in \mathbb{R}^n$ such that $M\vec{v} = 0$. By definition, a matrix's columns are linearly independent if and only if the nullspace of this matrix contains only $\vec{0}$. Since $\vec{x}$ is non-zero, we know that $(A-\lambda I)$ is linearly dependent. $\Rightarrow$ $(A-\lambda I)$ is not full rank $\Rightarrow$ $(A-\lambda I)$ is singular and non-invertible. $\Rightarrow$ For non-invertible matrix, its determinant is 0 $\Rightarrow$ $\det(A-\lambda I) = 0$. $\square$

1.3

$$A = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}, \quad \lambda I = \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}$$

$\Rightarrow$ $\det(A-\lambda I) = 0 \Rightarrow \det \begin{pmatrix} 1-\lambda & r \\ r & 1-\lambda \end{pmatrix} = 0 \Rightarrow (1-\lambda)^2 - r^2 = 0$

$\Rightarrow (1-\lambda)^2 = r^2 \Rightarrow 1-\lambda = \pm r \Rightarrow \boxed{\lambda = 1 \pm |r|} \Rightarrow$ which reduces to $\boxed{\lambda = 1 \pm r}$

1.4

Since $\lambda = 1 \pm r$ from 1.3. We have $\begin{cases} \lambda_1 = 1+r \\ \lambda_2 = 1-r \end{cases}$

λ_1: $(A - \lambda_1 I)x = 0 \Rightarrow \begin{pmatrix} 1-(1+r) & r \\ r & 1-(1+r) \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0$

$\Rightarrow \begin{pmatrix} -r & r \\ r & -r \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0 \Rightarrow \begin{cases} -rx_1 + rx_2 = 0 \\ rx_1 - rx_2 = 0 \end{cases} \Rightarrow r(x_1 - x_2) = 0$

* Case 1

$\Rightarrow$ if $r=0$, then A becomes identity matrix, $\vec{v}$ can be any 2×1 vector with its l_2 norm $=1$.

* Case 2 : if $r \neq 0$, $x_1 = x_2$, $\sqrt{x_1^2 + x_2^2} = 1$, this solves to

$$\boxed{\vec{v} = \begin{pmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{pmatrix} \text{ or } \vec{v} = \begin{pmatrix} -\frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{pmatrix}}$$

λ_2: $(A - \lambda_2 I)x = 0 \Rightarrow \begin{pmatrix} 1-(1-r) & r \\ r & 1-(1-r) \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0 \Rightarrow \begin{pmatrix} r & r \\ r & r \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0$

$\Rightarrow r(x_1 + x_2) = 0$.

* Case 1: If $r=0$, the answer is the same, see above case 1.

* Case 2 : If $r \neq 0$, then $x_1 = -x_2$, $\sqrt{x_1^2 + x_2^2} = 1$, this solves to

$$\boxed{\vec{v} = \begin{pmatrix} -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{pmatrix} \text{ or } \vec{v} = \begin{pmatrix} \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{pmatrix}}$$

2 Expectation, Co-variance and Independence [25pts + 5pts]

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 $Z = XY$.

- What is the Expectation and Variance of X? (in terms of c) [4pts]
- Show that when $c = 1$, Z is a standard Normal (Gaussian) distribution, which means $Z \sim N(0, 1)$. [9pts]
- How should we choose c such that Y and Z are uncorrelated (which means $Cov(Y, Z) = 0$)? [9pts]
- Are Y and Z independent? (Just clarify) [3pts]
- Show your conclusion for the above question with an example. **(Bouns)** [5pts]

Answers begin here

2.1

2.1. For Bernoulli Distribution, the expectation of X is

$$E[X] = P(X=c) \times c + P(X=-c) \times (-c) = 0.5 \times c - 0.5 \times c \boxed{=0}$$

$$Var[X] = E[X^2] - E[X]^2 = P(X=c) \cdot c^2 + P(X=-c) \cdot (-c)^2 - E[X]^2$$

$$= 0.5 \times c^2 + 0.5 \times (-c^2) - 0^2$$

$$\boxed{= c^2}$$

2.2

When $c=1$, $p(z) = p(XY)$

$$= p(XY | X=1)\, p(X=1) + p(XY | X=-1)\, p(X=-1)$$

$$= p(Y | X=1)\, p(X=1) + p(Y | X=-1)\, p(X=-1)$$

$$= p(Y) \cdot 0.5 + p(Y) \cdot 0.5$$, since X. Y are independent.

$$= p(Y).$$

Therefore, Z is also a standard normal distribution, $Z \sim N(0,1)$.

2.3

$\text{cov}(Y, Z) = E[YZ] - E[Y]E[Z]$

$= E[XY^2] - E[Y] \cdot E[Z]$ since $Z = XY$

$= E[X] \cdot E[Y^2] - E[Y^2] \cdot E[Z]$

$\boxed{= 0.}$

Y & Z are uncorrelated regardless of c.

2.4

$Z = XY$. y and z $\boxed{\text{are not independent}}$. Z depends on Y, also X is just a constant ($\pm C$).

2.5

2.5. For $Z = XY$ and X, Y independent. Even though Y, Z can be uncorrelated, Z, Y are not independent and $p(Y, Z) \neq p(Y) \cdot p(Z)$.

Example: given the definition of Y, Z, $p(Y \geq 0, Z \geq 0) = 0.5$ (half of the numbers drawn are bigger than 0). However, since $Y \sim N(0,1)$ and $Z \sim N(0,1)$, we have $p(Y \geq 0) = 0.5$ and $p(Z \geq 0) = 0.5$, and so

$p(Y \geq 0) \cdot p(Z \geq 0) = 0.5 \times 0.5 = 0.25 \neq 0.5$.

(For dependent variables, $p(A, B) = p(A) \cdot p(B|A)$. $\square$

3 Maximum Likelihood [25pts + 10pts]

3.1 Discrete Example [15pts]

Suppose you are playing two unfair coins. The probability of tossing a head is 2θ for coin 1, and θ for coin 2. You toss each coin for several times, and you get the following results:

Coin No.	Result
1	head
2	head
1	tail
2	tail
1	head
2	tail

- What is the probability of tossing a tail for coin 1 (p_{t1}) and tossing a tail for coin 2 (p_{t2}) [3pts]?
- What is the likelihood of the data given θ [6pts]?
- What is maximum likelihood estimation for θ [6pts]?

3.2 Continues Example [10pts] (BONUS)

A uniform distribution in the range of $[a, b]$ is given by

$$f(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

What is maximum likelihood estimation for a and b? (You need to show the derivation of your answer.)

(**Hint**: Think of two cases, where $x < max(x_1, x_2, \dots, x_n)$ and $x \ge max(x_1, x_2, \dots, x_n)$.)

3.3 Maximum A Posteriori (MAP) [10pts]

Suppose there exists an unknown parameter θ that describe whether the sun will explode tomorrow. $\theta = 1$ means the sun will explode and $\theta = 0$ if it won't. The likelihood function is:

$$P(yes|\theta) = \begin{cases} 1/36 & \theta = 0 \\ 35/36 & \theta = 1 \end{cases}$$

- What is the maximum likelihood estimate of θ?[3pts]
- Maximum A Posteriori (MAP) estimator aims to maximize the value of θ in $p(\theta|yes)$. What is the MAP estimate of θ given that $P(\theta = 0) \gg P(\theta = 1)$? Comment on the result.[7pts]

(**Hint**: You can use Bayes Rule to get $p(\theta|yes)$ from the likelihood!)

Answers begin here

3.1.1

The probability of tossing a tail for coin 1

= 1 − the probability of tossing a head for coin 1

= $\boxed{1-2\theta}$

Similarly, the probability of tossing a tail for coin 2 is $\boxed{1-\theta.}$

3.1.2

The likelihood of the data given θ:

Coin No.	1	2	1	2	1	2
result	head	head	tail	tail	head	tail
probability	2θ	θ	$1-2\theta$	$1-\theta$	2θ	$1-\theta$

The likelihood is:

L = P(head for coin 1) · P(head for coin 2) · P(tail for coin 1) ...

...

$$= 2\theta \cdot \theta \cdot (1-2\theta) \cdot (1-\theta) \cdot 2\theta \cdot (1-\theta)$$

$$= 4\theta^3 (1-\theta)^2 (1-2\theta)$$

3.1.3

Take log of likelihood (by convention): $\log(L) = \log 4 + 3\log\theta + 2\log(1-\theta) + \log(1-2\theta)$. Take derivative with respect to θ: $\frac{3}{\theta} - \frac{2}{1-\theta} - \frac{2}{1-2\theta} = 0$, solve for θ,

we get: $\frac{(3-5\theta)(1-2\theta) - 2\theta(1-\theta)}{\theta(1-\theta)(1-2\theta)} = 0$

$$\Rightarrow \quad 12\theta^2 - 13\theta + 3 = 0$$

$$\Rightarrow \quad \theta = \frac{13 \pm \sqrt{25}}{24} = \frac{1}{3} \text{ or } \frac{3}{4}$$

However, if $\theta = \frac{3}{4}$, then $2\theta = \frac{6}{4} > 1$ which can't be true.

Therefore, $\boxed{\theta = \frac{1}{3}}$

3.2

$a \le \min(x_1, x_2, \dots x_n)$ and $b \ge \max(x_1, x_2, \dots x_n)$ otherwise we can't get such X_i samples.

$$\mathcal{L}(a,b) = \prod_{i=1}^{n} f(x;a,b) = \frac{1}{b-a} \cdot \frac{1}{b-a} \cdots \frac{1}{b-a} = \frac{1}{(b-a)^n}$$

$$\log \mathcal{L}(a,b) = \log \prod_{i=1}^{n} f(x;a,b) = \sum_{i=1}^{n} \log f(x;a,b)$$

so $$\log \mathcal{L}(a,b) = \sum_{i=1}^{n} \log \frac{1}{(b-a)} = n \log \frac{1}{(b-a)}$$

$\frac{\partial}{\partial a} \log \mathcal{L}(a,b) = \frac{n}{b-a}$, as a increases, $\frac{\partial}{\partial a} \log(\mathcal{L})$ increases.

$\frac{\partial}{\partial b} \log \mathcal{L}(a,b) = \frac{-n}{b-a}$, as b increases, $\frac{\partial}{\partial b} \log(\mathcal{L})$ decreases.

Therefore, the largest possible a is $\min\{x_1, \dots x_n\}$ and the smallest possible b is $\max\{x_1, \dots x_n\}$.

3.3

Missing comment (-1 pts)

3.3.1

If $\theta = 1$, then $p(\text{yes}|\theta) = \frac{35}{36}$, and if $\theta = 0$, $p(\text{yes}|\theta) = \frac{1}{36}$.

When $\theta = 1$, $p(\text{yes}|\theta)$ is more likely, $\boxed{\text{so } \theta = 1}$

3.3.2.

MAP aims to maximize $p(\theta|\text{yes})$, which can be calculated by

$p(\text{yes}|\theta)\, p(\theta)$ using Bayes Rules and we can ignore $p(\text{yes})$

in the denominator (conventionally).

$$p(\text{yes}|\theta)\,p(\theta) \begin{cases} p(\text{yes}|\theta)\ p(\theta=0) = \frac{1}{36} \cdot p(\theta=0) \\ p(\text{yes}|\theta)\ p(\theta=1) = \frac{35}{36}\ p(\theta=1) \end{cases}$$, when

$p(\theta=0) \gg p(\theta=1)$, $p(\text{yes}|\theta)\, p(\theta=0)$ is bigger. $\boxed{\text{so } \theta = 0.}$

4 Information Theory [25pts + 7pts]

4.1 Marginal Distribution [4pts]

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

$X \| Y$	1	2
0	$\frac{1}{5}$	$\frac{2}{5}$
1	0	$\frac{2}{5}$

- Show the marginal distribution of X and Y, respectively. [4pts]

4.2 Mutual Information and Entropy [21pts]

Given a dataset as below.

Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*Play?*
1	*overcast*	*hot*	*normal*	*medium*	*yes*
2	*sunny*	*hot*	*high*	*weak*	*no*
3	*sunny*	*mild*	*normal*	*weak*	*yes*
4	*rain*	*cool*	*high*	*strong*	*no*
5	*overcast*	*cool*	*normal*	*strong*	*yes*
6	*rain*	*mild*	*normal*	*medium*	*no*
7	*sunny*	*mild*	*high*	*medium*	*yes*
8	*overcast*	*hot*	*normal*	*strong*	*no*
9	*rain*	*hot*	*high*	*weak*	*no*
10	*sunny*	*cool*	*normal*	*strong*	*yes*

We want to decide whether to play or not to play basketball on a certain day. Each input has four features (x_1, x_2, x_3, x_4): Outlook, Temperature, Humidity, Wind. The decision (play vs no-play) is represented as Y.

- Find entropy $H(Y)$. [4pts]
- Find conditional entropy $H(Y|x_1)$, $H(Y|x_4)$, respectively. [8pts]
- Find mutual information $I(x_1, Y)$ and $I(x_4, Y)$ and determine whether which one (x_1 or x_4) is more informative. [5pts]
- Find joint entropy $H(Y, x_3)$. [4pts]

4.3 Bonus Question [7pts]

- Suppose X and Y are independent. Show that $H(X|Y) = H(X)$. [2pts]
- Suppose X and Y are independent. Show that $H(X, Y) = H(X) + H(Y)$. [2pts]
- 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]

Answers begin here

4.1

X \ Y	1	2	total
0	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{3}{5}$
1	0	$\frac{2}{5}$	$\frac{2}{5}$
total	$\frac{1}{5}$	$\frac{4}{5}$	1

$\Rightarrow$

The marginal distribution of X is:

X	
X = 0	$\frac{3}{5}$
X = 1	$\frac{2}{5}$

The marginal distribution of Y is:

Y	Y=1	Y=2
	$\frac{1}{5}$	$\frac{4}{5}$

4.2

4.2.1

$H(Y) = -\left[\frac{5}{10}\log_2\left(\frac{5}{10}\right) + \frac{5}{10}\log_2\left(\frac{5}{10}\right)\right]$, since 5/10 we decided to play, and 5/10 we decided not to play

$= -\left(-\frac{1}{2}+\left(-\frac{1}{2}\right)\right)$

$= \boxed{1}$

4.2.2

(weather)

$X_1 \rightarrow$	overcast	sunny	rain	total
play	2	3	0	5
no play	1	1	3	5
total	3	4	3	10

$H(Y|\text{overcast}) = H(2/3, 1/3) = 0.9183$

$H(Y|\text{sunny}) = H(3/4, 1/4) = 0.8112$

$H(Y|\text{rain}) = H(0/3, 3/3) = 0$

$H(Y|X_1) = P(\text{overcast}) \cdot H(Y|\text{overcast}) + P(\text{sunny}) \cdot H(Y|\text{Sunny})$

$+ P(\text{rain}) \cdot H(Y|\text{rain}) = \frac{3}{10} \times 0.9183 + \frac{4}{10} \times 0.8112 + \frac{3}{10} \times 0$

$\boxed{= 0.59997}$

(wind)

X_4	medium	weak	strong	total
play	2	1	2	5
no play	1	2	2	5
total	3	3	4	10

$H(Y \mid \text{wind} = \text{medium}) = H(2/3, 1/3) = 0.9183$

$H(Y \mid \text{wind} = \text{weak}) = H(1/3, 2/3) = 0.9183$

$H(Y \mid \text{wind} = \text{strong}) = H(2/4, 2/4) = 1$

$$H(Y \mid X_4) = p(\text{medium}) \cdot H(Y \mid \text{medium}) + p(\text{wind} = \text{weak}) \cdot H(Y \mid \text{weak}) + p(\text{wind} = \text{strong}) \cdot H(Y \mid \text{wind} = \text{strong})$$

$$= \frac{3}{10} \cdot 0.9183 + \frac{3}{10} \cdot 0.9183 + \frac{4}{10} \cdot 1$$

$\boxed{= 0.9510}$

4.2.3

$I(X_1, Y) = H(Y) - H(Y \mid X_1) = 1 - 0.59997 = \boxed{0.40003}$

$I(X_4, Y) = H(Y) - H(Y \mid X_4) = 1 - 0.9510 = \boxed{0.049}$

Since $I(X_1, Y) \geq I(X_4, Y)$, $\boxed{X_1 \text{ is the more informative feature}}$.

4.3

4.3.1 Given the definition of mutual information, we know that $I(X,Y) = H(X) - H(X|Y)$. And since X, Y are independent, $I(X,Y) = 0 \Rightarrow H(X) = H(X|Y)$. $\square$

4.3.2 By Theorem for conditional entropy we know that

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

$$\Rightarrow H(X, Y) = H(Y) + H(X|Y)$$

From 4.3.1 we obtained that $H(X) = H(X|Y)$. Therefore,

$$H(X,Y) = H(X) + H(Y). \square$$

4.3.3 $I(X,Y) - I(Y,X) = H(X) - H(X|Y) - H(Y) + H(Y|X)$

$= H(X) - [H(X,Y) - H(Y)] - H(Y) + [H(X,Y) - H(X)]$

$= H(X) - H(X,Y) + H(Y) - H(Y) + H(X,Y) - H(X)$

$= 0.$

Therefore, $I(X,Y) = I(Y,X)$. $\square$

Another proof: $I(X,Y) = H(X) - H(X|Y)$

$$= \sum_x p(x) \log \frac{1}{p(x)} - \sum_{x,y} p(x,y) \log \left(\frac{1}{p(x|y)}\right)$$

$$= \sum_{x,y} p(x,y) \log \frac{p(x|y)}{p(x)}$$

$$= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

$$= \sum_{x,y} p(x,y) \log \frac{p(y|x)}{p(y)}$$

$$= \sum_y p(y) \log \frac{1}{p(y)} - \sum_{x,y} p(x,y) \log \frac{1}{p(y|x)}$$

$$= H(Y) - H(Y|X) = I(Y,X). \square$$