I232 Information Theory
Chapter 3: Mutual Information

Brian Kurkoski

Japan Advanced Institute of Science and Technology

2023 April

Outline

3.1 Mutual Information

Definition

Let X and Y be jointly distributed random variables. Then the *mutual information* $I(\mathsf{X};\mathsf{Y})$ between X and Y is:

$$I(\mathsf{X};\mathsf{Y}) = H(\mathsf{X}) - H(\mathsf{X}|\mathsf{Y})$$

- $I(\mathsf{X};\mathsf{Y})$ is the reduction in uncertainty of X, given you know Y
- $I(\mathsf{X};\mathsf{Y})$ is how much X tells you about Y
- $I(\mathsf{X};\mathsf{Y})$ is the number of bits of X relevant to Y

Understanding Mutual Information

Mutual information $I(\mathsf{X};\mathsf{Y})$ is how much X tells you about Y.

Example 1: x and y are similar — mutual information is high.

$$\begin{array}{rccccccccccccccccccccccc} x = & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \\ y = & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \end{array}$$

Example 2: x and y are quite different (in this case, independent) — mutual information is low.

$$\begin{array}{rccccccccccccccccccccccc} x = & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 \\ y = & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 \end{array}$$

Mutual Information of a Coin Flip

The Front Porch
http://bit.ly/1fU5Oqt

Consider a coin flip:

- X is the side facing up.
- Y is the side facing down.

What is $I(X;Y)$?

★Poll - Coin Flip

How to Measure Dependence Between X and Y?

Another to measure dependence is the correlation coefficient is a measure of the linear correlation between two variables.

Definition

The *covariance* between jointly distributed random variables X, Y is:

$$\mathrm{Cov}(\mathsf{X}, \mathsf{Y}) = E[\mathsf{XY}] - E[\mathsf{X}]E[\mathsf{Y}].$$

If X, Y are independent then $E[\mathsf{XY}] = E[\mathsf{X}]E[\mathsf{Y}]$ and $\mathrm{Cov}(\mathsf{X}, \mathsf{Y}) = 0$.

How to Measure Dependence Between X and Y?

Definition

The *correlation coefficient* ρ between two random variables X and Y is:

$$\rho = \frac{\mathrm{Cov}(\mathsf{X}, \mathsf{Y})}{\sqrt{\mathrm{Var}[\mathsf{X}]\mathrm{Var}[\mathsf{Y}]}}$$

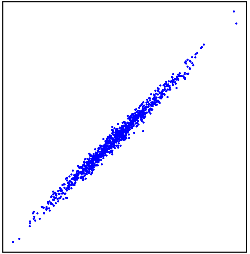
The correlation coefficient satisfies $-1 \leq \rho \leq 1$. $\rho = 0$ implies X and Y have no linear correlation.

Correlation measures only the *linear* relationship between X and Y. If X and Y are dependent but with no linear correlation, then $\rho = 0$, and the dependence is not clear.

Two Ways to Measure Dependence

- Correlation coefficient $\rho = \frac{E[XY] - E[X]E[Y]}{\sqrt{\mathrm{Var}[X]\mathrm{Var}[Y]}}$
- Mutual information $I(X;Y) = H(X) - H(X|Y)$

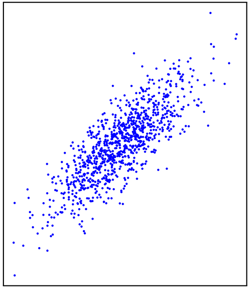
Which of A, B, C have high correlation?
Which has high mutual information?



A

$\rho =$

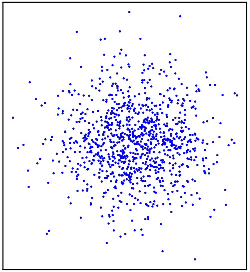
$I(X;Y) =$



B

$\rho =$

$I(X;Y) =$



C

$\rho =$

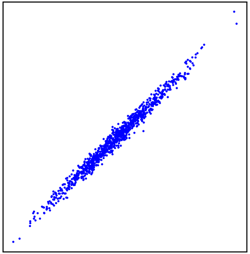
$I(X;Y) =$

★Poll - Correlation

Two Ways to Measure Dependence

- Correlation coefficient $\rho = \frac{E[\mathsf{XY}] - E[\mathsf{X}]E[\mathsf{Y}]}{\sqrt{\mathrm{Var}[\mathsf{X}]\mathrm{Var}[\mathsf{Y}]}}$
- Mutual information $I(\mathsf{X};\mathsf{Y}) = H(\mathsf{X}) - H(\mathsf{X}|\mathsf{Y})$

Which of A, B, C have high correlation?
Which has high mutual information?

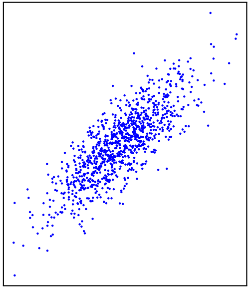


A

$\rho = 0.989$

$I(\mathsf{X};\mathsf{Y}) = 2.313$

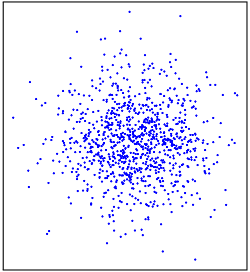
high



B

$\rho = 0.801$

$I(\mathsf{X};\mathsf{Y}) = 0.866$



C

$\rho = 0.0104$

$I(\mathsf{X};\mathsf{Y}) = 0.0095$

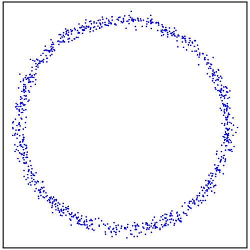
low

Two Ways to Measure Dependence — Circular Data

- Correlation coefficient ρ
- Mutual information $I(\mathsf{X};\mathsf{Y})$

Does circular data have:

- High correlation?
- High mutual information?



A

In this data set, clearly X depends on Y.

$\rho =$

$I(\mathsf{X};\mathsf{Y}) =$

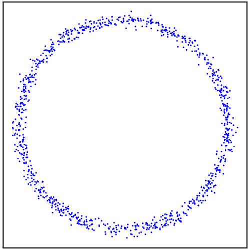
★Poll - Correlation of Circular Data

Two Ways to Measure Dependence — Circular Data

- Correlation coefficient ρ
- Mutual information $I(\mathsf{X};\mathsf{Y})$

Does circular data have:

- High correlation?
- High mutual information?



$$\rho = -0.0237 \text{ low}$$

$$I(\mathsf{X};\mathsf{Y}) = 1.414 \text{ high}$$

In this data set, clearly X depends on Y.

- Correlation coefficient is low — near 0
- Mutual information is high

Correlation only shows *linear* dependence

$\Rightarrow$ Mutual information is a better measure of dependence.

3.2 Kullback-Leiber Divergence

The KL divergence is a measure of a distance between two distributions $p(x)$ and $q(x)$.

Definition

The *KL divergence* $D\big(p(x)||q(x)\big)$ or $D(p||q)$ between the two probability distribution functions $p(x)$ and $q(x)$ is:

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

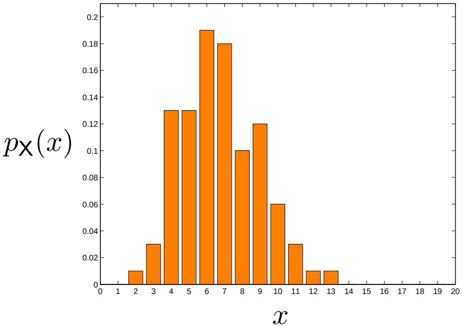
with $0 \log \frac{0}{q} = 0$ and $p \log \frac{p}{0} = \infty$.

Think of:

- $p(x)$ as a true distribution, and
- $q(x)$ as an approximation distribution.

$D(p||q)$ is the penalty of using q to approximate p.

Kullback-Leiber Divergence: Distance Between “True” and “Approximate” Distributions

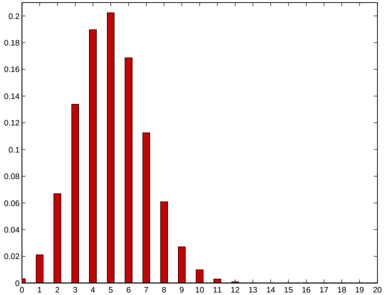


Suppose you ran an experiment:

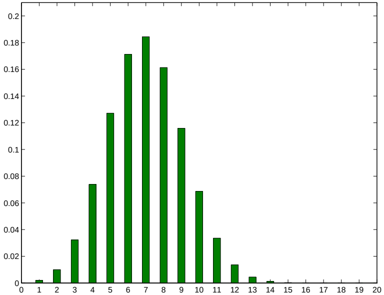
- Collected data shown in orange
- You want an analytic model of your data
- A good model might be the binomial distribution
- What model parameter p gives the best model?

Two Candidates for Model

You have two candidate model parameters $p = 0.25$ and $p = 0.35$.



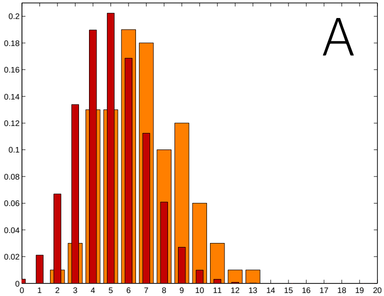
$q_1(x)$ binomial distribution $p = 0.25$



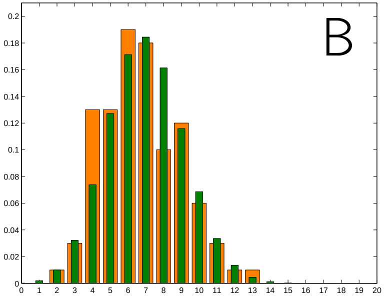
$q_2(x)$ binomial distribution $p = 0.35$

Visual Inspection of Two Candidates

If you don't have too much data, then you can visually compare the two models.



$q_1(x)$ binomial distribution $p = 0.25$

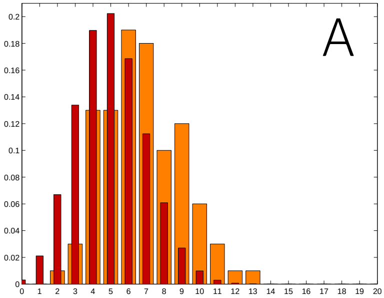


$q_2(x)$ binomial distribution $p = 0.35$

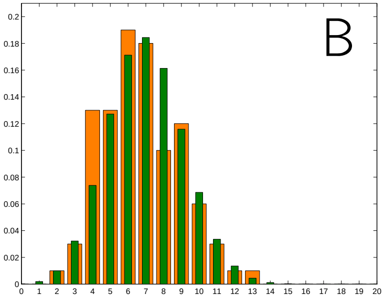
Which one is a better model?

★Poll - Best Approximation

Compare Using KL Divergence



$$D(p(x)||q_1(x)) = 0.5913$$



$$D(p(x)||q_1(x)) = 0.0571$$

Clearly, the KL divergence agrees with our visual intuition: B is a better match than A.

Visual comparison works well for a small amount of data.

If you have a large amount of data, KL divergence is more suitable.

Properties of KL Divergence

KL divergence is non-negative:

$$D(p||q) \geq 0$$

Properties of KL Divergence

KL divergence is non-negative:

$$D(p||q) \geq 0$$

KL divergence is 0 if and only if p and q are the same:

$$D(p||q) = 0 \quad \Leftrightarrow \quad p(x) = q(x) \text{ for all } x$$

KL divergence is not symmetric:

$$D(p||q) \neq D(q||p)$$

Mutual information can be expressed using KL divergence:

$$I(\mathsf{X}; \mathsf{Y}) = D(p_{\mathsf{XY}}(x, y) || p_{\mathsf{X}}(x) p_{\mathsf{Y}}(y))$$

3.2.1 Consequences of Non-Negativity of KL Divergence

The non-negativity of KL divergence allows proving three results:

(1) The non-negativity of mutual information:

$$I(\mathsf{X};\mathsf{Y}) = D(p_{\mathsf{XY}}(x,y)||p_{\mathsf{X}}(x)p_{\mathsf{Y}}(y)) \geq 0$$

(2) Conditioning reduces entropy:

$$H(\mathsf{X}|\mathsf{Y}) \leq H(\mathsf{X})$$

(3) The uniform distribution maximizes entropy:

$$H(\mathsf{X}) \leq \log|\mathcal{X}|$$

★1

3.3 Data Processing Inequality

3.3.1 Markov Chains

Let X, Y and Z be jointly distributed random variables. These random variables form a *Markov chain*, written

$$\mathsf{X} \to \mathsf{Y} \to \mathsf{Z}$$

if the conditional probability $p_{\mathsf{Z}|\mathsf{XY}}(z|x, y)$ does not change if X is dropped:

$$\Pr(\mathsf{Z} = z|\mathsf{X} = x, \mathsf{Y} = y) = \Pr(\mathsf{Z} = z|\mathsf{Y} = y) \text{ or}$$
$$p_{\mathsf{Z}|\mathsf{XY}}(z|x, y) = p_{\mathsf{Z}|\mathsf{Y}}(z|y).$$

The idea of Markovity is expressed by "the future (Z) depends on the present (Y) and not the past (X) ."

Markov chains will be handled in more detail in Chapter 6.

★2

3.3.2 Data Processing Inequality

The data processing inequality expresses the idea *processing cannot not increase information*.

Proposition

Data Processing Inequality. If $X \to Y \to Z$ is a Markov chain then

$$I(X;Y) \geq I(X;Z).$$

★3

3.4 Fano's Inequality

Let X and Y be jointly distributed. We know Y and want to estimate X.

Definition

Let $\widehat{\mathsf{X}} = g(\mathsf{Y})$ be an *estimate* of X. The function g is called the *estimator* of X.

Since $\widehat{\mathsf{X}} = g(\mathsf{Y})$, a Markov chain is formed:

$$\mathsf{X} \to \mathsf{Y} \to \widehat{\mathsf{X}}$$

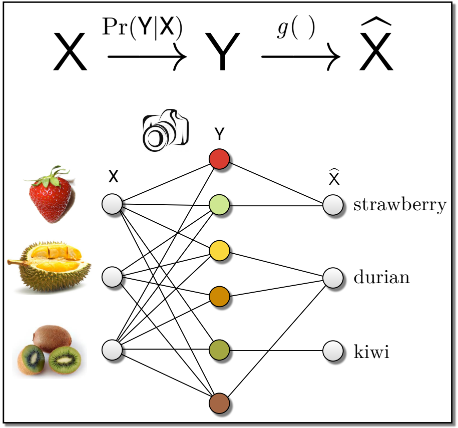
The probability of error P_e is the probability of estimation error:

$$P_e = \Pr[\widehat{\mathsf{X}} \neq \mathsf{X}].$$

Example: Durian Markov Chain

Consider the following example of a Markov chain. A camera takes a picture of a fruit, but it only outputs one pixel, a single color.

Example: Durian Markov Chain



Estimation:

- X is a fruit
- Y is the color of the observation
- $\widehat{\mathsf{X}} = g(\mathsf{Y})$ estimate of fruit from observed color
- Probability of error:

$$P_e = \Pr[\mathsf{X} \neq \widehat{\mathsf{X}}]$$

Fano's Inequality

Fano's inequality gives a lower bound on the probability of error P_e.

Proposition

Fano's Inequality For any estimator $\widehat{\mathsf{X}}$ such that $\mathsf{X} \to \mathsf{Y} \to \widehat{\mathsf{X}}$, we have:

$$h(P_e) + P_e \log|\mathcal{X}| \geq H(\mathsf{X}|\widehat{\mathsf{X}}) \;\geq\; H(\mathsf{X}|\mathsf{Y})$$

If $P_e \to 0$, then the bound $H(\mathsf{X}|\widehat{\mathsf{X}})$ and $H(\mathsf{X}|\mathsf{Y})$ must also go to zero.
Important when proving the converse to the channel coding theorem.
★4

3.5 Descriptions Using Expectation

Entropy, mutual information and KL divergence can be described using expectation.

Recall that: For a random variable X with distribution $p_{\mathsf{X}}(x)$ and a function g, the expectation of $g(\mathsf{X})$ is given by:

$$E[g(\mathsf{X})] = \sum_{x \in \mathcal{X}} g(x) p_{\mathsf{X}}(x).$$

Suppose we take $g(x) = -\log p_{\mathsf{X}}(x)$. What happens?

★5

What You Should Have Learned

- Mutual information is how much one variable tells you about the other
- In many cases, better than correlation which expresses only linear dependence
- Kullback-Leiber divergence: similarity of two probability distributions
- Data-processing inequality: processing cannot increase information
- Fano's inequality: Bound on probability of error Pe. If $P_e \to 0$, then the bound must also go to zero
- Entropy, etc. can be described using expectation. Eerily self-referential!

Class Info

- Next lecture: Monday, April 24. Source Coding for a Single Source. Lecture 4 Pop Quiz Preparation now available.
- Tutorial Hours: Monday, April 24 at 13:30. Ask questions about Homework.
- Homework 1 and 2 on LMS. Deadline: Monday, April 24 at 18:00