Decision and Risk Analysis

Lecture 9: **Mutual Information**

(not in the textbook!)

DAEN 427 / ISEN 427 / ISEN 627 - Fall 2026

Alexander Abuabara



Question

Suppose we observe two variables:

- X: weather (rainy or sunny)
- Y: whether a person carries an umbrella (yes or no)

How much does observing X reduce our uncertainty about Y?

Mutual information

Measures *statistical dependence* rather than just a linear association.

Example: 100 observations

	Umbrella: yes	Umbrella: no	Total
Rainy	40	10	50
Sunny	10	40	50
Total	50	50	100

- $P(Y = \text{yes}) = 50/100 = 0.5$.
- $P(Y = \text{yes} \mid X = \text{rainy}) = 40/50 = 0.8$.
- $P(Y = \text{yes} \mid X = \text{sunny}) = 10/50 = 0.2$.

The weather is informative, but not perfectly predictive.

Probability terms

Marginal $p(y)$: chance of an umbrella, ignoring weather.

Joint $p(x, y)$: chance of a particular weather–umbrella pair.

Conditional $p(y \mid x) = p(x, y)/p(x)$: chance of an umbrella after weather is known.

Independence $p(x, y) = p(x)p(y)$ for every pair (x, y).

For example:

$$p(\text{rainy, yes}) = 0.40, \qquad p(\text{yes} \mid \text{rainy}) = \frac{0.40}{0.50} = 0.80.$$

What is information?

Less likely events carry more information because they reduce our uncertainty more when they occur.

Consider two events:

- Event A: The sun rises tomorrow. Probability is nearly 1.

→ If Event A occurs, you learn almost nothing new because you already expected it.

- Event B: You win a lottery with a one-in-a-million chance.

→ If Event B occurs, you learn something highly unexpected. The observation rules out a much larger set of possibilities.

Thus, the less predictable an event is,
the more information its occurrence provides.

What is information in bits?

The *surprisal* of an event with probability p is

$$\mathcal{I}(\text{event}) = -\log_2 p \quad \text{bits.}$$

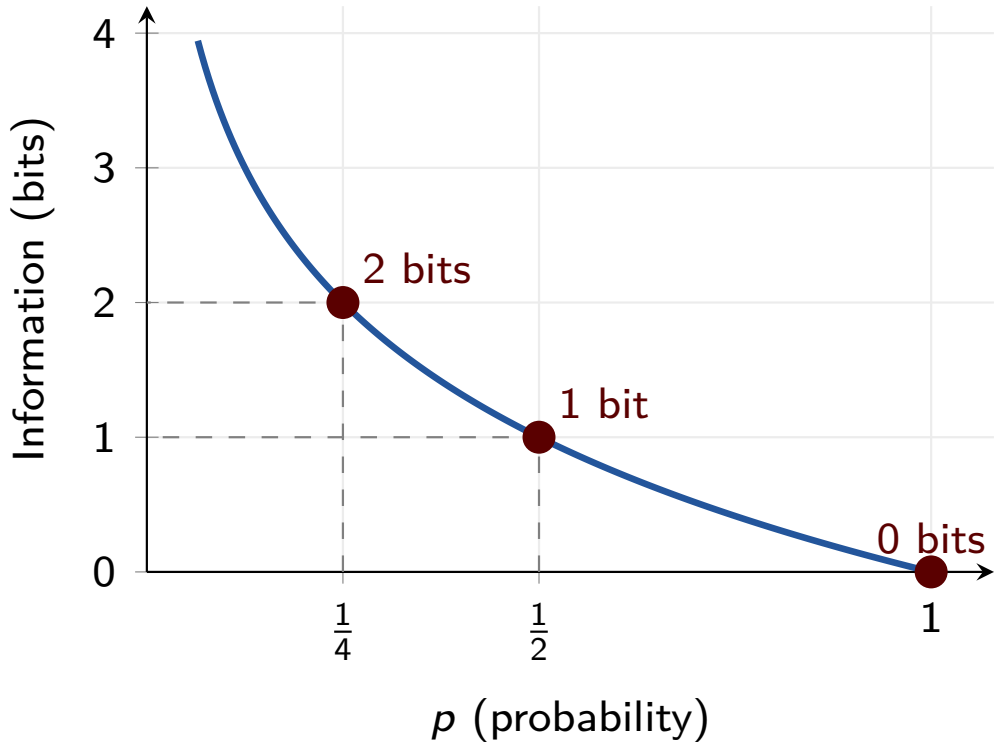
- Probability $1/2 \to 1$ bit of surprisal.
- Probability $1/4 \to 2$ bits of surprisal.
- Probability $1 \to 0$ bits of surprisal.

A less likely event is more surprising.
A base-2 logarithm expresses information in *bits*.

The *surprisal* of an event with probability p is $\boxed{\mathcal{I}(p) = -\log_2 p \quad \text{bits}}$

Examples:

- $p = \frac{1}{2} \Rightarrow 1$ bit
- $p = \frac{1}{4} \Rightarrow 2$ bits
- $p = 1 \Rightarrow 0$ bits



Less likely events carry more information.

What Is Entropy?

Entropy is a foundational concept in physics that measures the dispersal of energy, randomness, or disorder within a closed system.

Here, entropy measures the average uncertainty associated with a random variable.

- **High entropy:** outcomes are difficult to predict.
- **Low entropy:** outcomes are easier to predict.
- **Zero entropy:** the outcome is certain.

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

Coin toss	P(**Heads**)	**Entropy**
Fair coin	0.50	1.000 bit
Biased coin	0.90	0.469 bits
Certain heads	1.00	0 bits

More uncertainty before observing the outcome means more information gained when it is revealed, on average.

Entropy: average uncertainty

For a discrete random variable Y,

$$H(Y) = -\sum_y p(y) \log_2 p(y).$$

- Entropy is the *expected surprisal* before observing Y.
- By convention, $0 \log_2 0 = 0$.
- For two equally likely outcomes, entropy is 1 bit.

In our example, umbrella use is evenly split:

$$H(Y) = -0.5 \log_2 0.5 - 0.5 \log_2 0.5 = \boxed{1 \text{ bit}}.$$

Conditional entropy: uncertainty after X

$$H(Y \mid X) = \sum_x p(x) H(Y \mid X = x),$$

where

$$H(Y \mid X = x) = -\sum_y p(y \mid x) \log_2 p(y \mid x).$$

Rainy days: 80% carry umbrellas, 20% do not:

$$H(Y \mid \text{rainy}) = -0.8 \log_2 0.8 - 0.2 \log_2 0.2 \approx 0.722.$$

Sunny days have the probabilities reversed, so the entropy is also 0.722.

$$H(Y \mid X) = 0.5(0.722) + 0.5(0.722) = \boxed{0.722 \text{ bits}}.$$

Mutual information (reduction)

$$\boxed{I(X;Y) = H(Y) - H(Y \mid X)}$$

For the weather–umbrella data:

$$I(X;Y) = 1 - 0.722 = \boxed{0.278 \text{ bits}}.$$

- Knowing the weather removes, on average, 0.278 bits of the original 1 bit of uncertainty about umbrella use.
- Here this is a 27.8% reduction *because* the initial entropy is exactly 1 bit.
- MI is not generally a percentage.

The joint-probability formula

An equivalent definition is

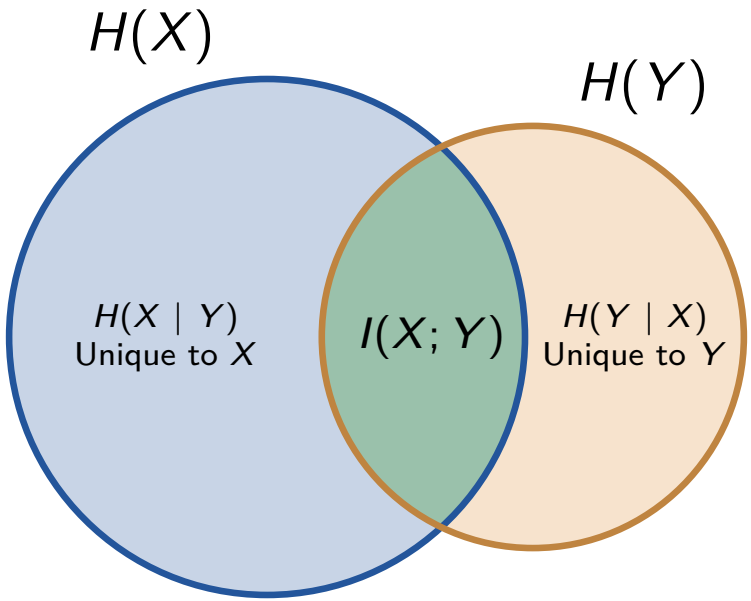
$$\boxed{I(X;Y) = \sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}}.$$

For our 2×2 table, the joint probabilities are $0.4, 0.1, 0.1, 0.4$; every product $p(x)p(y)$ equals 0.25:

$$\begin{aligned} I(X;Y) &= 2(0.4) \log_2 \frac{0.4}{0.25} + 2(0.1) \log_2 \frac{0.1}{0.25} \\ &\approx \boxed{0.278 \text{ bits}}. \end{aligned}$$

This also equals the KL divergence $D_{\mathrm{KL}}(p(x,y) \,\|\, p(x)p(y))$.

Kullback-Leibler divergence measures how different one probability distribution is from another. Mutual information is a special case of KL divergence. `https://en.wikipedia.org/wiki/KullbackLeibler_divergence`



Shared information is the overlap

So, mutual information cannot exceed the entropy of either variable:

$$\underbrace{I(X;Y) = H(Y) - H(Y \mid X)}_{\leq H(Y)}$$

This means that knowing X cannot tell us more about Y than the total uncertainty in Y, and vice versa.

$$\boxed{I(X;Y) \leq \min\{H(X), H(Y)\}}$$

Important mathematical properties

- **Nonnegative:** $I(X;Y) \geq 0$.
- **Symmetric:** $I(X;Y) = I(Y;X)$.
- **Zero exactly at independence:** $I(X;Y) = 0$ if and only if X and Y are independent (for the true distribution).
- **Upper bound for discrete variables:**

$$I(X;Y) \leq \min\{H(X), H(Y)\}.$$

- **Perfect knowledge:** if Y is fully determined by X, then $H(Y \mid X) = 0$ and $I(X;Y) = H(Y)$.

Mutual information vs. correlation

	Pearson correlation r	Mutual information I
Captures	Linear association	Any statistical dependence
Sign	Negative, zero, positive	Nonnegative only
Scale	-1 to 1	0 to $\min(H(X), H(Y))$ for discrete variables
Units	Unitless	Bits (base 2) or nats (natural log)
Direction	Sign shows slope direction	No sign or causal direction

- Our binary weather–umbrella table has $|r| = 0.60$ (with compatible 0/1 coding), but $I = 0.278$ bits.
- They are *different measures*, not numerically comparable.

A nonlinear relationship correlation can miss

Let X be equally likely to be $-2, -1, 0, 1, 2$, and let $Y = X^2$.

X	-2	-1	0	1	2
Y	4	1	0	1	4

- Pearson's $r = 0$ because the positive and negative linear effects cancel.
- But Y is completely determined by X.

Let X be uniformly distributed over $\{-2, -1, 0, 1, 2\}$, with $Y = X^2$.

Step 1: Find the probability distribution of Y

Y	Corresponding X	$P(Y = y)$
0	0	$\frac{1}{5}$
1	$-1, 1$	$\frac{2}{5}$
4	$-2, 2$	$\frac{2}{5}$

Step 2: Calculate the entropy of Y

$$H(Y) = -\sum_{y} P(y) \log_2 P(y)$$

$$H(Y) = -\frac{1}{5} \log_2 \frac{1}{5} - 2\left(\frac{2}{5} \log_2 \frac{2}{5}\right) \approx 1.522 \text{ bits}$$

Step 3: Calculate mutual information (next slide)
Since $Y = X^2$, knowing X determines Y exactly. Therefore, $H(Y \mid X) = 0$.

$$\boxed{I(X; Y) = H(Y) - H(Y \mid X) = 1.522 \text{ bits}}$$

Knowing X eliminates all uncertainty about Y.

Let X be uniformly distributed over $\{-2, -1, 0, 1, 2\}$, with $Y = X^2$.

Step 1: Examine the conditional probabilities

Known X	$Y = X^2$	$P(Y = X^2 \mid X)$
-2	4	1
-1	1	1
0	0	1
1	1	1
2	4	1

Step 2: Calculate conditional entropy for each X

For example, when $X = -2$, we know $Y = 4$ with certainty:

$$H(Y \mid X = -2) = -1 \log_2(1) = 0.$$

The same result holds for every other value of X.

Step 3: Average over all possible values of X

$$H(Y \mid X) = \sum_x P(x) H(Y \mid X = x) = \frac{1}{5}(0 + 0 + 0 + 0 + 0) = \boxed{0 \text{ bits}}$$

Since $Y = X^2$ is deterministic, knowing X eliminates all uncertainty about Y.

Recall that X is uniformly distributed over $\{-2, -1, 0, 1, 2\}$ and $Y = X^2$.

Entropy of X:

$$H(X) = \log_2 5 \approx 2.322 \text{ bits}$$

Mutual information:

$$I(X; Y) = H(Y) \approx 1.522 \text{ bits}$$

Knowing X determines Y exactly,
but knowing Y does not always determine X.

Observed Y	Possible X	Remaining uncertainty
0	0	0 bits
1	$-1, 1$	1 bit
4	$-2, 2$	1 bit

Remaining uncertainty about X:

$$\begin{aligned} H(X \mid Y) &= H(X) - I(X; Y) \\ &= 2.322 - 1.522 \\ &= \boxed{0.8 \text{ bits}} \end{aligned}$$

The relationship is deterministic in one direction, but not invertible.

A nonlinear relationship correlation can miss

Let X be equally likely to be $-2, -1, 0, 1, 2$, and let $Y = X^2$.

X	-2	-1	0	1	2
Y	4	1	0	1	4

- Pearson's $r = 0$ because the positive and negative linear effects cancel.
- But Y is completely determined by X.
- $I(X;Y) = H(Y)$, probabilities of $Y = 0, 1, 4$ are $1/5, 2/5, 2/5$.

$$I(X;Y) = -\tfrac{1}{5}\log_2\tfrac{1}{5} - 2\left(\tfrac{2}{5}\log_2\tfrac{2}{5}\right) \approx \boxed{1.522 \text{ bits}}.$$

Assumptions: theory vs. estimation

Theoretical definition:

- Requires a well-defined joint probability distribution.
- Does *not* require a linear relationship, Gaussian variables, or causality.
- Formulas so far describe *discrete* variables; continuous variables require probability densities and integrals.

Estimating MI from observations:

- Observations should represent the target population or process.
- The frequency-table estimator generally relies on sufficiently many observations in each relevant category.
- Independence or an appropriate dependence-aware sampling model matters for uncertainty estimates and significance tests.

Limitations and common pitfalls

- **Not causal:** MI cannot establish whether weather causes umbrella use, or vice versa.
- **No direction or sign:** $I(X;Y) = I(Y;X)$; it cannot say whether a relationship increases or decreases.
- **Finite-sample bias:** plug-in estimates are often positive even for truly independent variables.
- **High dimensionality:** sparse counts and continuous-density estimation become difficult; binning choices affect results.
- **No universal maximum:** raw MI depends on entropy and alphabet size, so values across different feature pairs may not be directly comparable.
- **Redundant predictors:** a feature's MI with the target does not measure its additional value given other features.

Does 1 information bit equal 1 computer bit?

Yes, in the unit of measurement: both use base-2 logarithms.

- A **binary digit** is a storage symbol, either 0 or 1.
- A **Shannon bit** measures uncertainty or expected information.
- An equally likely random 0/1 value has entropy 1 bit.
- A constant 0/1 value still occupies one binary digit if stored, but has entropy 0 bits.

Connection to compression

For ideal lossless coding of long sequences, entropy in bits per symbol gives a fundamental lower limit on *average* coding rate (under appropriate source assumptions). Actual storage also reflects coding overhead and representation.

Putting the "bit" interpretation together

- Before weather: umbrella use is a balanced yes/no question $\rightarrow H(Y) = 1$ bit.
- After weather: uncertainty averages 0.722 bits.
- The gain from weather is 0.278 bits on average.

$$\underbrace{1}_{\text{before}} = \underbrace{0.278}_{\text{information gained}} + \underbrace{0.722}_{\text{uncertainty left}} \quad \text{bits.}$$

This does *not* mean that each weather observation physically stores 0.278 of a binary digit. It is an *average reduction in uncertainty*.

1. Entropy is average uncertainty;
   conditional entropy is uncertainty remaining after observing another variable.
2. Mutual information is their difference,
   or equivalently the divergence between the joint distribution and independence.
3. MI detects nonlinear dependence that Pearson correlation can miss.
4. MI is noncausal and sensitive to how probabilities are estimated from finite data.
5. One Shannon bit uses the same mathematical unit as a digital bit, but information content is not the same thing as a physical storage location.