

# Entropy and information gain



## Entropy

**Use when:** Quantifying the amount of uncertainty in a probability distribution.

For a discrete random variable  $Y$  with outcomes  $y_1, y_2, \dots, y_k$  and probabilities  $p(y_i)$ , entropy measures average uncertainty:

$$H(Y) = - \sum_{i=1}^k p(y_i) \log_2 p(y_i)$$

The sum is over all outcomes where  $p(y_i) > 0$  (since  $0 \log 0$  is conventionally taken as 0). Entropy is measured in **bits** (using logarithm base 2).

**Interpretation:**

- $H(Y) = 0$  when all probability is on one outcome (no uncertainty; the outcome is certain).
- $H(Y)$  is highest when the distribution is uniform — all outcomes equally likely.
- High entropy means the outcome is hard to predict; low entropy means it is easily predictable.

**Example:** A dataset has 10 items of class A and 5 items of class B.

Probabilities:  $p(A) = \frac{10}{15} = \frac{2}{3}$  and  $p(B) = \frac{5}{15} = \frac{1}{3}$

$$\begin{aligned} H(Y) &= -\frac{2}{3} \log_2 \frac{2}{3} - \frac{1}{3} \log_2 \frac{1}{3} \\ &= -\frac{2}{3}(-0.585) - \frac{1}{3}(-1.585) = 0.390 + 0.528 = 0.918 \text{ bits} \end{aligned}$$

**Example:** A fair coin ( $p(\text{heads}) = 0.5, p(\text{tails}) = 0.5$ ).

$$H(Y) = -0.5 \log_2(0.5) - 0.5 \log_2(0.5) = -0.5(-1) - 0.5(-1) = 1.0 \text{ bit}$$

A fair coin has maximum entropy (1 bit) because the outcome is completely unpredictable.

## Conditional entropy

When you know the value of variable  $X$ , the remaining uncertainty about  $Y$  is:

$$H(Y | X = x) = - \sum_{i=1}^k p(y_i | x) \log_2 p(y_i | x)$$

Student Learning Support, Teaching and Learning Centre

[studentlearning@ontariotechu.ca](mailto:studentlearning@ontariotechu.ca)

[ontariotechu.ca/studentlearning](https://ontariotechu.ca/studentlearning)



This document is licensed under Attribution-NonCommercial 4.0 International (CC BY-NC 4.0).

The **average** conditional entropy over all values of  $X$  is:

$$H(Y | X) = \sum_x p(x) \cdot H(Y | X = x)$$

**Example:** Suppose in the original dataset, class A always occurs with feature  $X = 1$ , and class B always occurs with  $X = 2$ .

- If  $X = 1$ : all 10 items are class A, so  $H(Y | X = 1) = 0$
- If  $X = 2$ : all 5 items are class B, so  $H(Y | X = 2) = 0$

$$H(Y | X) = \frac{10}{15}(0) + \frac{5}{15}(0) = 0 \text{ bits}$$

Knowing  $X$  completely removes the uncertainty about  $Y$ .

## Information gain

**Use when:** Deciding which feature or split most reduces uncertainty about a target variable (e.g., choosing splits in a decision tree).

Information gain is the reduction in entropy when you learn the value of variable  $X$ :

$$\text{IG}(Y | X) = H(Y) - H(Y | X)$$

### Interpretation:

- $\text{IG} = 0$  if  $X$  tells you nothing about  $Y$  (the variables are independent).
- $\text{IG} = H(Y)$  if  $X$  completely determines  $Y$  (knowing  $X$  removes all uncertainty).
- Larger information gain means  $X$  is more informative about  $Y$ .

In decision trees: at each node, choose the feature/split that maximizes information gain.

**Example:** From above,  $H(Y) = 0.918$  bits and  $H(Y | X) = 0$  bits.

$$\text{IG}(Y | X) = 0.918 - 0 = 0.918 \text{ bits}$$

Knowing feature  $X$  gives full information (0.918 bits of gain) about the class label.

## Properties and cautions

**Caution:** Do not confuse entropy with “disorder” in a loose sense. Entropy is a precise quantity defined by the formula. It depends only on the probability distribution, not on the actual values or categories.

Also: when computing entropy, remember to use logarithm base 2 (or state if using natural log).  
Results differ by a constant factor otherwise.

**Key properties:**

- $H(Y) \geq 0$  always.
- $H(Y | X) \leq H(Y)$  (conditioning cannot increase uncertainty).
- $IG(Y | X) \geq 0$  (information gain is always non-negative).
- If  $X$  and  $Y$  are independent:  $IG(Y | X) = 0$ .