



# Machine Learning

(Course Code: 18B1WCI634/18B11BI611)

**Entropy, Cross Entropy, KL Divergence**

**Mr. Sandeep Kumar Patel**

Assistant Professor

Jaypee University of Information Technology

Waknaghat, Solan, HP-173234

# Why Information Theory in AI?

- Machine learning models deal with **uncertainty**
- Outputs are **probability distributions**, not hard decisions
- We must answer:
  - How uncertain is the world? -Use **entropy**, which measures randomness or uncertainty in data
  - How wrong is our model?- Use **cross-entropy**, which measures the difference between predicted probabilities and true outcomes.
  - How different is belief from reality?-Use **KL divergence**, which quantifies how one probability distribution differs from another.

**Information theory provides mathematical answers.**

# Big Picture

All concepts answer one question:

*How well does my model explain the real world?*

- Negative Log-Likelihood → model surprise
- Entropy → world uncertainty
- Cross-Entropy → model error
- KL Divergence → belief mismatch
- Importance Sampling → bias correction

# Likelihood: Explaining Observations

Assume:

- Real world follows unknown distribution  $p(x)$
- We observe data:  $x = (x_1, x_2, \dots, x_n)$
- We propose a model  $q(x)$

Probability of observing data under model  $q$ :

$$P(x | q) = \prod_{i=1}^n q(x_i)$$

**Likelihood answers:**

*“If the world behaved like  $q$ , how likely is this data?”*

## Numerical Example: Likelihood

Assume a biased coin model:

$$q(H) = 0.8, \quad q(T) = 0.2$$

Observed data:

$$x = (H, H, T)$$

Likelihood:

$$P(x | q) = 0.8 \times 0.8 \times 0.2 = 0.128$$

**Interpretation:**

*If the coin behaved like  $q$ , this sequence is moderately likely.*

# Log-Likelihood

Multiplying probabilities leads to very small numbers.

Apply log:

$$\log P(x | q) = \sum_{i=1}^n \log q(x_i)$$

- Turns products into sums
- Numerically stable
- Preserves ordering

## Numerical Example: Log-Likelihood

Using the same data:

$$x = (H, H, T)$$

Log-likelihood:

$$\begin{aligned}\log P(x | q) &= \log 0.8 + \log 0.8 + \log 0.2 \\ &= -0.223 - 0.223 - 1.609 = -2.055\end{aligned}$$

### Observation:

- Easier to compute
- Same ordering as likelihood

## Negative Log-Likelihood (NLL)

Since  $\log q(x) \leq 0$ , define:

$$\text{NLL}(q) = - \sum_{i=1}^n \log q(x_i)$$

### Equivalent objectives:

- Maximize likelihood
- Maximize log-likelihood
- Minimize NLL

**Training ML models = minimizing NLL**

# Numerical Example: Negative Log-Likelihood

Negative log-likelihood:

$$\text{NLL}(q) = -\log P(x | q) = 2.055$$

**Lower NLL  $\Rightarrow$  better model fit**

Training a model means finding  $q$  that minimizes this value.

# Entropy: Uncertainty of the World

Entropy measures **randomness** in the real world.

- Deterministic system → zero entropy
- Random system → high entropy

## Examples:

- Sun rises every day → low entropy
- Dice roll → high entropy

# Entropy Formula

$$H(p) = \mathbb{E}_{x \sim p}[-\log p(x)]$$

or equivalently:

$$H(p) = - \sum_x p(x) \log p(x)$$

## Interpretation:

*Entropy is the expected surprise of reality.*

## Numerical Example: Entropy

Fair coin:

$$p(H) = 0.5, \quad p(T) = 0.5$$

Entropy:

$$H(p) = -[0.5 \log_2 0.5 + 0.5 \log_2 0.5] = 1 \text{ bit}$$

**Meaning:**

*Each toss gives one bit of uncertainty.*

## Cross-Entropy: Using the Wrong Belief

In practice:

- Samples come from true distribution  $p$
- Predictions come from model  $q$

$$H(p, q) = \mathbb{E}_{x \sim p}[-\log q(x)]$$

or equivalently:

$$H(p, q) = - \sum_x p(x) \log q(x)$$

**Meaning:**

*How surprised am I if I trust my model while the world follows reality?*

## Numerical Example: Cross-Entropy

True distribution:

$$p = (1, 0)$$

Model prediction:

$$q = (0.8, 0.2)$$

Cross-entropy:

$$H(p, q) = -\log_2(0.8) \approx 0.322 \text{ bits}$$

**Interpretation:**

*Surprise caused by trusting an imperfect model.*

# Cross-Entropy in Machine Learning

We do not know  $p$ , but we have samples.

Empirical estimate:

$$H(p, q) \approx -\frac{1}{n} \sum_{i=1}^n \log q(x_i)$$

**This is exactly the NLL loss.**

**Cross-Entropy loss = Negative Log-Likelihood**

# KL Divergence: Cost of Wrong Beliefs

KL divergence is defined as:

$$D_{KL}(p||q) = H(p, q) - H(p)$$

## **Interpretation:**

*Extra surprise caused by using  $q$  instead of  $p$ .*

# Derivation of KL Divergence

**Entropy:**

$$H(P) = - \sum_x P(x) \log P(x)$$

**Cross-Entropy:**

$$H(P, Q) = - \sum_x P(x) \log Q(x)$$

**KL Divergence:**

$$D_{KL}(P \parallel Q) = H(P, Q) - H(P)$$

$$D_{KL}(P \parallel Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

# Properties and Interpretation

**Continuous form:**

$$D_{KL}(P \parallel Q) = \int P(x) \log \frac{P(x)}{Q(x)} dx$$

**Expectation form:**

$$D_{KL}(P \parallel Q) = \mathbb{E}_{x \sim P} \left[ \log \frac{P(x)}{Q(x)} \right]$$

**Key properties:**

- $D_{KL}(P \parallel Q) \geq 0$
- $D_{KL}(P \parallel Q) = 0$  iff  $P = Q$
- Not symmetric

## Numerical Example: KL Divergence

True distribution:

$$p = (0.5, 0.5), \quad q = (0.6, 0.4)$$

KL divergence:

$$\begin{aligned} D_{KL}(p||q) &= 0.5 \log_2 \frac{0.5}{0.6} + 0.5 \log_2 \frac{0.5}{0.4} \\ &= 0.5(-0.263) + 0.5(0.322) \approx 0.03 \text{ bits} \end{aligned}$$

**Meaning:**

*Small but nonzero cost of wrong beliefs.*