信息论（Information Theory）是研究信息的度量、传输、存储和压缩的一门科学，由克劳德·香农（Claude Shannon）在 1948 年提出的奠基论文中首次系统化。它是现代通信、数据压缩和机器学习的重要理论基础。

---

1. 核心概念

1.1 信息量（Self-Information）

- 定义：事件 x 的信息量 $I(x)$ 表示事件发生时获取的信息大小。
- 数学公式：

$$I(x) = -\log P(x)$$

  - $P(x)$：事件 x 的概率。
  - $-\log$：概率越小，信息量越大。

- 直观解释：
  - 罕见事件包含更多信息。
  - 例子：抛硬币 $P(\text{正面}) = 0.5$，信息量 $I = -\log_2(0.5) = 1$ 比特。

1.2 熵（Entropy）

- 定义：熵 $H(X)$ 是离散随机变量 X 的期望信息量，表示不确定性或信息平均量。
- 数学公式：

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

  - 单位：比特（以 2 为底）或纳特（以 e 为底）。

- 直观解释：
  - 熵越大，不确定性越高。
  - 均匀分布（每个事件的概率相等）具有最大熵。

1.3 **相对熵**（KL **散度**）

- 定义：相对熵（KL 散度）衡量两个分布 P 和 Q 的相似程度。
- 数学公式：
$$D_{\mathrm{KL}}(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$
- 直观解释：
  - D_{KL} 越小，分布 P 和 Q 越接近。
  - 它是非对称的，即 $D_{\mathrm{KL}}(P||Q) \neq D_{\mathrm{KL}}(Q||P)$。

1.4 **交叉熵**（Cross-Entropy）

- 定义：交叉熵是两个分布之间的信息差异，用于度量预测分布 $Q(x)$ 和真实分布 $P(x)$ 的不匹配程度。
- 数学公式：
$$H(P, Q) = -\sum_x P(x) \log Q(x)$$
- 应用：
  - 在机器学习中，交叉熵是分类任务中的常用损失函数。

1.5 **信息增益**（Information Gain, IG）

- 定义：信息增益表示某个属性（或特征）对目标变量的不确定性减少的程度。
- 数学公式：
$$IG(T, A) = H(T) - H(T|A)$$
  - $H(T)$：原始熵。
  - $H(T|A)$：在属性 A 已知的条件下的熵。
- 应用：
  - 决策树中，用于选择最佳分割特征。

---

2. 信息论在机器学习中的应用

2.1 交叉熵损失

- 用于分类任务中，衡量预测分布与真实分布的差异。

2.2 KL 散度

- 在生成对抗网络（GAN）中衡量生成模型分布与真实分布的差异。
- 在变分自编码器（VAE）中，用于正则化潜在分布。

2.3 信息增益

- 决策树（如 ID3、C4.5）的特征选择标准。

2.4 熵和数据压缩

- Shannon 熵是理论上的数据压缩极限。
- 哈夫曼编码（Huffman Coding）实现了接近熵的压缩效率。

---

3. 示例代码

3.1 熵计算

```
import numpy as np

# 概率分布
P = np.array([0.2, 0.5, 0.3])

# 计算熵
H = -np.sum(P * np.log2(P))
print(f"Entropy: {H} bits")
```

3.2 KL 散度计算

```
# 两个分布
P = np.array([0.1, 0.4, 0.5])
Q = np.array([0.2, 0.3, 0.5])

# KL 散度
```

```
D_KL = np.sum(P * np.log2(P / Q))
print(f"KL Divergence: {D_KL} bits")
```

3.3 交叉熵计算

```
# 真实分布和预测分布
P = np.array([1, 0, 0])  # One-hot 编码
Q = np.array([0.7, 0.2, 0.1])

# 交叉熵
H_PQ = -np.sum(P * np.log2(Q))
print(f"Cross-Entropy: {H_PQ} bits")
```

---

4. 信息论的意义

1. **度量不确定性**：熵衡量系统的平均不确定性。
2. **优化模型**：通过最大化信息增益或最小化交叉熵，提高模型性能。
3. **压缩与通信**：信息论为数据压缩和通信提供理论基础。