# Entropy & KL Divergence

Yuhan Zhang

# Information Theory

Information theory quantifies uncertainty and information, introduced by Claude Shannon in 1948

Some key concepts: entropy, KL Divergence

# Entropy

Entropy measures the randomness/uncertainty of a system

The same definition from physics

Shannon's Entropy:

$$H(X) = \sum P(X)\ log\ \frac{1}{P(X)}$$

High entropy means more randomness

Low entropy means less randomness

# Entropy: Example

Fair coin vs. biased coin

Tossed 10 times, fair coin $X$ has 5 heads, 5 tails; biased coin $Y$ has 7 heads, 3 tails

Entropy for $X$: $H(X) = \sum P(X) \log \frac{1}{P(X)} = 0.5 \log \frac{1}{0.5} + 0.5 \log \frac{1}{0.5} = 1$

Entropy for $Y$: $H(Y) = \sum P(Y) \log \frac{1}{P(Y)} = 0.7 \log \frac{1}{0.7} + 0.3 \log \frac{1}{0.3} \approx 0.8$

# Entropy: Another Example

Two random variables: $X \sim Bernoulli(\pi = 0.5)$, $Y \sim Bernoulli(\pi = 0.99)$

$$H(X) = \sum P(X) \log \frac{1}{P(X)} = 0.5 \log \frac{1}{0.5} + 0.5 \log \frac{1}{0.5} = 1$$

$$H(Y) = \sum P(Y) \log \frac{1}{P(Y)} = 0.99 \log \frac{1}{0.99} + 0.01 \log \frac{1}{0.01} \approx 0.08$$

$H(X)$ is much larger than $H(Y)$, means that $X$ has more randomness than $Y$

# Kullback-Leibler (KL) Divergence

Measures how one probability distribution diverges from another

Formula:

$$D_{KL}(P||Q) = \sum_{x} P(x) \log \frac{P(x)}{Q(x)}$$

Properties:

- Asymmetry: $D_{KL}(P||Q) \neq D_{KL}(Q||P)$
- Non-negative
- Zero iff $P = Q$

# Reminder: Likelihood Ratio Test

Likelihood Ratio Tests

Let $X_1$, $X_2$, $X_3$, ..., $X_n$ be a random sample from a distribution with a parameter $\theta$. Suppose that we have observed $X_1 = x_1$, $X_2 = x_2$, $\cdots$, $X_n = x_n$. Define

$$\lambda(x_1, x_2, \cdots, x_n) = \frac{\sup\{L(x_1, x_2, \cdots, x_n; \theta) : \theta \in S_0\}}{\sup\{L(x_1, x_2, \cdots, x_n; \theta) : \theta \in S\}}.$$

To perform a **likelihood ratio test (LRT)**, we choose a constant $c$ in $[0, 1]$. We reject $H_0$ if $\lambda < c$ and accept it if $\lambda \geq c$. The value of $c$ can be chosen based on the desired $\alpha$.

https://www.probabilitycourse.com/chapter8/8_4_5_likelihood_ratio_tests.php

# Hypothesis Testing & KL Divergence

Hypothesis testing: Null & alternative hypothesis, testing statistics, testing distribution, significant level

Likelihood Ratio Test (LRT) uses the likelihood ratio $\Lambda = \frac{P(X|H_0)}{P(X|H_1)}$.

Take the log and normalizing by sample size $n$, we have $\widehat{\Lambda}_n = \frac{1}{n}\sum \log \frac{p_0(X)}{p_1(X)}$

The $\widehat{\Lambda}_n$ has a very similar form to the KL Divergence $D(P_0||P_1) = \sum P_0(\mathrm{X}) \log \frac{P_0(X)}{P_1(X)}$

For large $n$, we will have $\widehat{\Lambda}_n = E[\Lambda] = D(q||p_0) - D(q||p_1)$. With $\lambda = 0$, we reject the null hypothesis if $D(q||p_0) \geq D(q||p_1)$

# Hypothesis Testing & KL Divergence

Suppose we now have a coin, we don't know what is the probability of the coin landing on head. We now have two hypothesis:

$$H_0: P(Head) = 0.5, H_1: P(Head) = 0.7$$

We tossed the coin for 100 times, and we have 60 heads and 40 tails

With the formula on previous slide, $D(q||p_0) = \sum q \frac{q}{p_0} = 0.029$, $D(q||p_1) = \sum q \frac{q}{p_1} = 0.03258$

Since we have $D(q||p_1) > D(q||p_0)$, we fail to reject $H_0$