UNIVERSITY OF MARYLAND

MSML610: Advanced Machine Learning

# Information Theory

**Instructor**: GP Saggese, PhD - gsaggese@umd.edu

**References**:

# Information theory

- **Information theory**
  - Entropy

# Entropy

- Information theory
    - **Entropy**

# Entropy and Uncertainty

- **Entropy** $H(X)$ of a discrete random variable $X$ is defined as:

$$H(X) \stackrel{def}{=} -\sum_{x} p(x) \log p(x)$$

- **Intuition**
  - Entropy quantifies the average level of information / surprise / uncertainty inherent in the variable's possible outcomes
    - High entropy = more unpredictability
    - Low entropy = more certainty
  - Usually the log is base 2 $\log_2$ so unit of entropy is bits
- **Examples**
  - Fair coin
    - A fair coin toss has two equally likely outcomes, heads or tails, leading to maximum uncertainty $H = 1$
  - Biased coin
    - If a coin lands on heads 90% of the time, it creates less uncertainty and thus less entropy $H < 1$

# Why Entropy is Defined in That Way

- **Entropy as expected information content**
  - Information of outcome $x_i$ is $-\log p_i$
    - Rare events ($p_i \to 0$) yield more information ($-\log p_i \to \infty$)
    - Common events ($p_i \to 1$) yield little information ($-\log p_i \to 0$)
  - Entropy is the expected information content:

$$H(X) = \mathbb{E}[-\log p_i] = -\sum_i p_i \log p_i$$

- **Axiomatic derivation of entropy**
  - Shannon's criteria for $H(p_1, ..., p_n)$:
    1. Continuity: $H()$ is continuous in $p_i$
    2. Maximality: $H()$ is maximal when outcomes are equally likely
    3. Additivity: For composite outcomes, $H(X, Y) = H(X) + H(Y|X)$
  - The only solution is:

$$H(X) = -K \sum_i p_i \log p_i$$

# Entropy and PDF

- Entropy is related to variance but is not the same
  - If a distribution has more spread, typically its entropy is larger
  - It is possible that variance increases, but entropy doesn't
- Entropy is related to information and uncertainty
  - The flatter the prior distribution, the less informative it is

# Joint Entropy

- **Joint entropy** $H(X, Y)$ of two variables $X$ and $Y$ is defined as:

$$H(X, Y) \stackrel{def}{=} -\sum_{x,y} p(x, y) \log p(x, y)$$

  - Describes the information needed for the joint distribution of $X$ and $Y$
- **Properties**
  - Non-negative and zero if $X$ and $Y$ are perfectly determined
  - For two independent binary variables $X$ and $Y$, the joint entropy is the sum of the entropy
- **Applications**
  - Identifies dependencies or correlations in datasets
  - Aids in feature selection by finding informative variable combinations
  - E.g., in sensor network data, joint entropy can highlight overlapping sensor information

# Conditional Entropy

- **Conditional entropy** $H(Y|X)$ is defined as

$$H(Y|X) = -\sum_{x,y} p(x,y) \log p(y|x)$$

- **Intuition**
  - Represents the average uncertainty in $Y$ after observing $X$
  - Measures the effectiveness of $X$ in determining $Y$
- **Properties**
  - Low $H(Y|X)$ implies stronger predictive power of $X$ on $Y$
    - There is less uncertainty about $Y$ after knowing $X$
    - I.e., $Y$ has predictive power on $X$
  - If $Y = X$, then $H(Y|X) = 0$
    - No uncertainty about $Y$ once $X$ is known
    - $X$ completely determines $Y$
  - If $X$ and $Y$ are independent, then $H(Y|X) = H(Y)$
    - Knowledge of $X$ provides no new information about $Y$
- **Applications**
  - In feature selection assess the predictive power of independent variables on dependent variables

# Mutual Information

- The **mutual information** $I(X;Y)$ between $X$ and $Y$ is defined as:

$$I(X;Y) \stackrel{def}{=} H(X) - H(X|Y)$$

- **Intuition**
  - Measures how much knowing $X$ reduces uncertainty about $Y$
  - Gauges the shared information between two variables
- **Properties**
  - Non-negative: $I(X;Y) \geq 0$
  - Symmetric: $I(X;Y) = I(Y;X)$
  - If $X$ and $Y$ are independent, then $I(X;Y) = 0$
  - Higher mutual information indicates greater relation between $X$ and $Y$
  - Related to the joint entropy but symmetric:

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

- **Applications**
  - Selects features sharing high information with the target variable
  - Used to reduce dimensionality

# Kullback-Leibler (KL) Divergence

- The **KL divergence** $D_{\text{KL}}(P\|Q)$ between distributions $P$ and $Q$ is defined as:

$$D_{\text{KL}}(P\|Q) \stackrel{def}{=} \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

- **Intuition**
  - Quantify how much one distribution deviates from another distribution
- **Properties**
  - Not symmetric: $D_{\text{KL}}(P\|Q) \neq D_{\text{KL}}(Q\|P)$
  - $D_{\text{KL}}(P\|Q) = 0$ iff $P = Q$, i.e., $Q(x) = P(x)$ for all $x$
  - It is a distance, but not a metric (not symmetric, no triangle inequality)
- **Applications**
  - In optimization of machine learning models by minimizing divergence
  - E.g., variational autoencoders use KL divergence to ensure that the learned distribution is close to the true distribution

# Cross-Entropy

- The **cross-entropy** $H(P, Q)$ between two distributions $P$ and $Q$ is defined as:

$$H(P, Q) \stackrel{def}{=} -\sum_x P(x) \log Q(x)$$

  where:
  - $P(x)$ is the true probability of the event $x$
  - $Q(x)$ is the probability assigned by the model
- **Intuition**
  - Measures the average number of bits needed to encode data from $P$ using a code optimized for $Q$
  - Indicates inefficiency when the code for $Q$ is used to represent $P$
- **Properties**
  - Cross-entropy is related to entropy and KL divergence:

$$H(P, Q) = H(P) + D_{\mathrm{KL}}(P \| Q)$$

- **Applications**
  - Used to compare the similarity of the predicted outcomes probability distribution to the true distribution
    - E.g., as loss function in logistic regression
    - A perfect model has a cross-entropy of 0

# Data Processing Inequality

- **Data processing inequality** states that:

  *Processing data cannot increase information, it can only lose information over*

- Formally: if $X \to Y \to Z$, then $I(X; Z) \leq I(X; Y)$
  - After passing through an additional stage (from $Y$ to $Z$), the mutual information with the initial stage ($X$) cannot increase

- **Examples**
  - If $X$ is a raw image, $Y$ is a compressed version
    - No additional processing $Z$ will uncover more information about $X$ than what $Y$ already represents
  - If $Y$ is a dataset derived from $X$ (e.g., summary statistics)
    - Any analysis applied to $Y$ alone cannot provide more insights into $X$ than $Y$ itself

- **Applications**
  - Compression can only lead to information loss
  - Identify "information bottlenecks" in a modeled process, ensuring model designs consider the constraints imposed by information processing

# Chain Rule for Entropy and Mutual Information

- The **entropy chain rule**

$$H(X, Y) = H(X) + H(Y|X)$$

  - In words, the joint entropy of two random variables can be decomposed into the entropy of one and the conditional entropy of the other

- Examples:

  - $X$ represents the weather (sunny, rainy)
  - $Y$ represents outdoor activity (park, cinema)
  - $H(Y|X)$ would provide information about outdoor activities given specific weather conditions

- The **mutual information chain rule**:

$$I(X, Y; Z) = I(X; Z) + I(Y; Z|X)$$

- **Applications**

  - By using chain rules, one can decompose and understand the complexity of joint distributions with multiple interacting variables
  - Useful in sequential models (e.g., speech recognition) and time-series analysis (e.g., stock market prediction)

# Source Coding Theorem

- Aka “Shannon’s first theorem”
- **Statement**
  - Compression cannot achieve an average code length less than the entropy of the source
- **Implications**
  - It asserts a limit on lossless compression:
  - E.g., if a source has entropy $H(X) = 3$ bits, you cannot on average encode it with fewer than 3 bits per symbol
- **Examples**
  - Lossless compression methods approach the entropy limit
    - E.g., Huffman coding encode data by creating variable-length codes

# Noisy-Channel Coding Theorem

- Aka “Shannon’s second theorem”
- **Statement**
  - For any discrete memoryless channel with capacity $C$, and for any desired level of reliability $\varepsilon > 0$, it is possible to transmit information at any rate $R < C$ with an arbitrarily small probability of error, using a sufficiently long encoding scheme
- **Implications**
  - Accurate communication can be achieved even with noise
  - Error correction techniques can be applied to achieve this
  - It does not construct the code, it only proves existence
- Channel capacity defines the upper limit of information that can be transmitted reliably (e.g., 10 Mbps)
- Applications:
  - Fundamental principle for designing digital communication systems
  - E.g., mobile networks, satellite communications, and the internet

# Redundancy and Compression

- **Redundancy** is the difference between actual and optimal code length
  - Measures excess information in the data
  - Redundancy implies room for compression, i.e., reduce data size without losing information
- **Compression techniques** remove redundancy while preserving information
  - Aim to make data smaller without losing meaning or important details
  - Useful for reducing storage or speeding up data transmission
- **Examples**
  - Run-length encoding
    - Compress by replacing consecutive identical elements with a single value and count
    - E.g., `AAAABBBCCDAA` becomes `4A3B2C1D2A`
  - Huffman coding
    - Uses variable-length codes for encoding
    - Frequently used symbols get shorter codes, reducing length

# Typical Set

- Set of sequences with probability close to $2^{-nH(X)}$
  - E.g., if $H(X) = 2$, then for large $n$, sequences have a probability close to $2^{-2n}$
- Central to proving coding theorems
  - The typical set is essential in demonstrating the efficiency of compression algorithms
- Almost all sequences in large samples lie in the typical set
  - E.g., for a sequence length $n$, the probability of falling outside the typical set decreases exponentially as $n$ increases
- Enables asymptotic analysis of information theory
  - Used in deriving limits related to data compression and reliable communication

# Rate-Distortion Theory

- Trade-off between compression rate $R$ vs distortion $D$
  - **Compression rate** $R$: amount of data remaining after compression
  - **Distortion** $D$: difference between the original and compressed data
  - Balancing $R$ and $D$ is crucial for effective lossy compression
- **Goal**: reduce data size (lossy compression) while maintaining an acceptable level of quality
- Rate-distortion function $R(D)$ defines the minimal rate for a given distortion
  - Describes the lower bound of the data rate necessary to achieve a specified level of distortion
  - E.g., in image compression, $R(D)$ helps in determining the lowest bitrate for a desired image quality
- **Applications**
  - Widely used in image/audio/video compression, e.g., MP3, JPEG and MPEG formats
  - Important for streaming services and storage optimization

# Fano's Inequality

- When $X$ is guessed from $Y$, it holds:

$$H(X|Y) \leq h(P_e) + P_e \log(|X| - 1)$$

  where
  - $X$ is a discrete random variable
  - $Y$ is the estimate variable of $X$ based on some observations
  - $H(X|Y)$ is the conditional entropy of $X$ given $Y$
  - $h(P_e)$ is the binary entropy function quantifying uncertainty of a binary random variable, where $h(p) = -p \log p - (1 - p) \log(1 - p)$
- **Intuition**
- It relates the uncertainty remaining about $X$ after observing $Y$ to the probability of making an error in guessing $X$
- "You cannot simultaneously have low error and low entropy (uncertainty). If the entropy is high, your probability of error must also be high."

# Differential Entropy

- **Definition**
  - The differential entropy $h(x)$ for continuous random variable $X$ with density $p(x)$ is defined as:

$$h(X) \overset{def}{=} -\int p(x) \log p(x) \, dx$$

- **Concept**
  - Extends Shannon entropy to continuous distributions
  - Measures "spread" or uncertainty of $X$
- **Key Properties**
  - Not invariant under variable change
    - E.g., scaling affects $h(X)$
  - Can be negative, unlike non-negative discrete entropy
  - Units depend on logarithm base
- **Example**
  - For Gaussian $X \sim \mathcal{N}(\mu, \sigma^2)$:

$$h(X) = \frac{1}{2} \log(2\pi e \sigma^2)$$

- **Limitations**
  - Cannot compare directly across variables with different units or scales

# Maximum entropy principle

- **Definition**
  - Use the prior with the largest entropy (i.e., the least informative) given the constraints of the problem
  - Can be solved as an optimization problem
- **Examples**
  - The distribution with largest entropy given a constraint is:
    - Without constraints: uniform
    - A positive mean: exponential
    - A given variance: normal distribution

# Minimum Description Length (MDL)

- **Definition**
  - The total description length of a dataset $MDL(H)$ is given by:

    $$MDL(H) = L(H) + L(D \mid H)$$

    where:
    - $L(H)$ is the length (in bits) of the model or hypothesis
    - $L(D \mid H)$ is the length of the data encoded using the model
- **Principle**
  - MDL selects the hypothesis $H$ that minimizes the total description length

    $$\mathrm{MDL}(H) = L(H) + L(D \mid H)$$

- **Intuition**
  - Prefers the model that gives the shortest total description of data
  - Based on Occam's Razor: simpler models are preferred
  - Balances model complexity and data fit
- **Example**
  - Given two decision trees for classifying email as spam:
    - Tree A is small and classifies 95% correctly
    - Tree B is large and classifies 96% correctly
    - MDL may prefer Tree A if the increased accuracy doesn't justify the extra complexity
- **Applications**

# Kolmogorov Complexity

- The Kolmogorov complexity $K(x)$ of a string $x$ is the length of the shortest binary program that outputs $x$ on a universal Turing machine
- **Examples**
  - A string of 1000 random bits: high Kolmogorov complexity, no compressible pattern
  - A string of 1000 repeated 0s: low Kolmogorov complexity, described by a short loop
- **Intuition**
  - Measures "algorithmic randomness" or compressibility of a string
  - A string is complex if it has no shorter description than itself
- **Formal Properties**
  - Incomputable: no algorithm computes $K(x)$ for all $x$
  - $K(x) \leq |x| + c$ for some constant $c$, showing the trivial upper bound (print $x$)
- **Relation to MDL**
  - MDL approximates Kolmogorov complexity by minimizing a practical description length

# Information Bottleneck

- Framework for extracting relevant information
- Trade-off: compression of $X$ vs retention of info about $Y$
- Optimization: minimize $I(X;T)$ while preserving $I(T;Y)$
- Used in deep learning theory and representation learning

# Multi-Information and Total Correlation

- Generalization of mutual information to multiple variables
- Total correlation: $C(X_1, \ldots, X_n) = \sum_i H(X_i) - H(X_1, \ldots, X_n)$
- Measures total dependency in a set of variables
- Used in ICA, variational inference, and dependency modeling