MSML610: Advanced Machine Learning

Information Theory

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

References:

Information theory

- **Information theory**
  - Entropy
  - Connections between Information Theory and ML

Entropy

- Information theory
  - **Entropy**
  - Connections between Information Theory and ML

Entropy and Uncertainty

- Entropy quantifies the uncertainty of a random variable X and is defined as

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

  - Represents the average level of information, surprise, or uncertainty inherent in the variable's possible outcomes
- It can be considered as a measure of randomness or average amount of "information" produced
  - High entropy = more unpredictability
  - Low entropy = more certainty
- E.g.,
  - Fair coin has $H = 1$
    - A fair coin toss has two equally likely outcomes, heads or tails, leading to maximum uncertainty
  - Biased coin has $H < 1$
    - If a coin lands on heads 90% of the time, it creates less uncertainty and thus less entropy

Joint Entropy

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

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

  - Describes the information needed for the joint distribution of X and Y
  - Non-negative and zero if X and Y are perfectly determined
  - E.g., for two independent binary variables X and Y, each with a probability of 0.5 for both 0 and 1, the joint entropy would be 2 bits
- 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

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

  - Represents the average uncertainty in Y after observing X
  - Measures the effectiveness of X in determining Y
- Lower conditional entropy implies stronger predictive power of X on Y
  - Less uncertainty about Y after knowing X means more predictive power
  - Used in feature selection to assess the impact of X
- E.g., if $Y = X$, then $H(Y|X) = 0$
  - Indicates no uncertainty about Y once X is known
  - If X completely determines Y, the entropy is zero
- E.g., if X and Y are independent, then $H(Y|X) = H(Y)$
  - Knowledge of X provides no new information about Y

Mutual Information

- Mutual information $I(X;Y)$ between X and Y is defined as

$$I(X;Y) = H(X) - H(X|Y)$$

  - Measures how much knowing X reduces uncertainty about Y
  - Related to the joint entropy: $I(X;Y) = H(X) + H(Y) - H(X,Y)$
  - Gauges the shared information between two variables
- Properties:
  - Symmetric: $I(X;Y) = I(Y;X)$
  - $I(X;Y) = 0$ if X and Y are independent
  - Non-negative: $I(X;Y) \geq 0$
  - Higher mutual information indicates greater relation between X and Y
- Applications
  - Feature selection and dependency analysis
  - Selects features sharing high information with the target variable
  - Used to reduce dimensionality
  - E.g., in a dataset, selecting features maximizing mutual information with the target variable can improve model performance

Kullback-Leibler (KL) Divergence

- KL divergence between P and Q is Defined as:

$$D_{\mathrm{KL}}(P\|Q) = \sum_x \Pr(x) \log \frac{\Pr(x)}{Q(x)}$$

  - Quantifies how much one distribution deviates from another distribution
- Properties:
  - Not symmetric: $D_{\mathrm{KL}}(P\|Q) \neq D_{\mathrm{KL}}(Q\|P)$
  - $D_{\mathrm{KL}}(P\|Q) = 0$ (no divergence) iff $P = Q$, i.e., $Q(x) = \Pr(x)$ for all x
- Applications
  - Used in variational inference, information gain
  - Helps 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

- Defined as

$$H(P, Q) = -\sum_{x} \Pr(x) \log Q(x)$$

  where:
  - $\Pr(x)$ is the true probability of the event x
  - $Q(x)$ is the probability assigned by the model
- Measures the average number of bits needed to encode data from P using a code optimized for Q
  - The average here is computed over the probability distribution P
  - Indicates inefficiency when the code for Q is used to represent P
- Related to entropy and KL divergence:

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

  - $H(P)$ is the entropy of the distribution P, representing the uncertainty inherent in P
  - $D_{\mathrm{KL}}(P \| Q)$ is the KL divergence, a measure of how one probability distribution diverges from a second, expected probability distribution
- Used in classification loss functions (e.g., logistic regression)
  - Assists in optimizing the model parameters to better fit the data
  - A perfect model has a cross-entropy of 0
  - Often used in machine learning to compare the similarity of the predicted

Data Processing Inequality

- Data processing inequality states that: *"Processing data cannot increase information, it can only lose information over time"*
- 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
  - E.g., if X is a raw image, Y is a compressed version, and Z is a further processed output, no additional processing will uncover more information about X than what Y already represents
  - E.g., if Y is a dataset derived from X, any analysis or summary statistics applied to Y alone cannot provide more insights into X than Y itself
- Applications
  - Suggests that compression can only lead to information loss
  - Model Bottlenecks: identify where "information bottlenecks" may occur in a modeled process, ensuring model designs consider the constraints imposed by information processing

Chain Rule for Entropy and Mutual Information

- Entropy chain rule:

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

  - The total entropy of two random variables can be decomposed into the entropy of one and the conditional entropy of the other
  - E.g., in a system where X represents the weather (sunny, rainy) and Y represents outdoor activity (park, cinema), $H(Y|X)$ would provide information about outdoor activities given specific weather conditions

- Mutual information chain rule:

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

  - It illustrates how mutual information between a pair of variables and a third variable can be broken down
  - E.g., when assessing the influence of two different sensors X and Y on a machine state Z, this rule can help isolate the individual and combined contributions of X and Y on Z

- By using these rules, one can simplify and manage the complexity of joint distributions with multiple interacting variables

- Applications:

Source Coding Theorem

- Aka "Shannon's first theorem"
- It asserts a limit on lossless compression: compression cannot achieve an average code length less than the entropy of the source
  - E.g., if a source has entropy $H(X) = 3$ bits, you cannot, on average, encode it with fewer than 3 bits per symbol
- For a source with entropy $H(X)$, average code length $\geq H(X)$
  - Lossless compression methods like Huffman coding approach this limit
  - E.g., Huffman coding provides an efficient way to encode data by creating variable-length codes, closer to the entropy limit

Noisy Channel Coding Theorem

- Aka "Shannon's second theorem"
- Demonstrates that even with noise, accurate communication can be achieved
- Channel capacity C is the maximal achievable information rate
  - Channel capacity defines the upper limit of information that can be transmitted reliably
  - E.g., if the capacity C of a channel is 10 Mbps, it means information can be transmitted at up to 10 Mbps without errors
- If the transmission rate R is less than the channel capacity C, then the likelihood of errors in transmission can be minimized
  - Error correction techniques can be applied to achieve this
- Applications:
  - Fundamental principle for designing digital communication systems (e.g., mobile networks, satellite communications, and the internet)
  - Used in protocols and standards to ensure data integrity and efficiency

Redundancy and Compression

- Redundancy = difference between actual and optimal code length
  - Measures excess information in the data
  - High redundancy implies room for compression
  - Potential to 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
- E.g.,
  - Run-length encoding: Compresses 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

- Studies lossy compression
  - Focuses on reducing data size while maintaining an acceptable level of quality
- Trade-off: 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
- 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
  - Essential for designing efficient compression algorithms
  - E.g., in image compression, $R(D)$ helps in determining the lowest bitrate for a desired image quality
- Applications in image/audio/video compression
  - Widely used in JPEG, MP3, and MPEG formats
  - Important for streaming services and storage optimization
  - E.g., video platforms use rate-distortion theory to deliver high-quality streams at low bitrates

Fano's Inequality

- When X is guessed from Y, it holds:

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

- Lower bound on error probability in terms of entropy
  - This provides a limit on classification accuracy of the best guessing strategy
  - The term $h(P_e)$ is the binary entropy function quantifying uncertainty of a binary random variable
  - As error probability P_e decreases, conditional entropy $H(X|Y)$ also decreases, indicating better predictive accuracy
  - Provides insights into required information extraction from Y about X to achieve a certain error probability
- Used to prove converse results in information theory
  - Establishes limits on data compression and transmission efficiency
  - Part of strategies to determine ultimate limits in information-theoretic problems
  - E.g., in a communication scenario, recovering a message X from a received signal Y; this bound aids understanding minimum achievable error
  - Clarifies trade-off between data rate and reliability in communication

Differential Entropy

- Extension of entropy to continuous variables
- Defined as $h(X) = -\int p(x) \log p(x)\, dx$
- Not invariant under change of variables
- Can be negative, unlike discrete entropy

Fisher Information

- Measures sensitivity of likelihood to parameter changes
- Defined as $I(\theta) = \mathbb{E}\left[\left(\frac{\partial \log p(x|\theta)}{\partial \theta}\right)^2\right]$
- Related to the Cramér-Rao bound on estimator variance
- Important in statistical inference and learning theory

Minimum Description Length (MDL)

- Principle: best hypothesis compresses data most
- Inspired by Occam's razor
- Connects compression with learning and model selection
- Used in unsupervised learning, clustering, and structural risk minimization

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

Entropy: definition

- Entropy is defined as

$$H(p) = -\sum_i p_i \log(p_i)$$

Entropy and variance

- 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 and information

- Entropy is related to information and uncertainty
- The flatter the prior distribution, the less informative it is

Maximum entropy principle

- Use the prior with the largest entropy (i.e., the least informative) given the constraints of the problem
- This can be solved as an optimization problem
- E.g., the distribution with largest entropy given a constraint is:
  - without constraints: uniform
  - a positive mean: exponential
  - a given variance: normal distribution

KL divergence

- It measures how close two distributions p and q are
- Defined as

$$D_{KL}(p, q) = \sum_i p_i \log(\frac{p_i, q_i}{)}$$

- It is the difference of entropy of p and the cross-entropy of p and q

$$= \sum_i p_i \log p_i - \sum_i p_i \log q_i$$

- It can be interpreted as the extra entropy that is introduced to approximate the distribution p using q
- It is a measure of "how surprised we are to see q, when we expect p"
- Note that it is not symmetric and thus is not a distance

Connections between Information Theory and ML

- Information theory
  - Entropy
  - **Connections between Information Theory and ML**