ENGG5301 Information Theory Tutorial

Tutorial 5: Midterm Review

Yanxiao Liu

Department of Information Engineering

Oct 25, 2022

Note

- I will review lectures 1-3 today.
- I will focus on the big picture and some practices, and select some important proofs. Details are in lecture slides.
- **Important**: The course is for you to learn something, we are from different background, it is not a competition!
- **Important**: No cheating in midterm!!!

Lecture 1 Review: Definitions

Self-information

- Self-information: $\iota_X(x) = \log \frac{1}{p_X(x)}$
- Joint pmf $p_{X,Y}$: $\iota_{X,Y}(x,y) = \log \frac{1}{p_{X,Y}(x,y)}$

1. $\iota_X(x) \geq 0$
2. For a function f, $\iota_{f(X)}(f(x)) \leq \iota_X(x)$, equality iff f is injective.
3. (Additive) If X, Y are independent, $\iota_{X,Y}(x,y) = \iota_X(x) + \iota_Y(y)$.
4. $\iota_X(x)$ is constant iff X follows a uniform distribution
5. Weakness: information spectrum is a probability distribution, but we want a single number to summarize the amount of information.

Lecture 1 Review

Entropy

- Shannon entropy: $H(X) = \mathbf{E}[\iota_X(X)] = \sum_x p_X(x) \log \frac{1}{p_X(x)}$, which is the average of the self-information.
- Joint entropy: $H(X, Y) = \mathbf{E}[\iota_{X,Y}(X, Y)]$
  1. Positivity: $H(X) \geq 0$ with equality iff X is a constant.
  2. Uniform distribution maximizes entropy: For $|\mathcal{X}| < \infty$, $H(X) \leq \log |\mathcal{X}|$.
  3. Invariance under relabeling: $H(X) = H(f(X))$ for any bijective f.
  4. Conditioning reduces entropy: $H(X|Y) \leq H(X)$ with equality iff X, Y indpt.
  5. Full chain rule: $H(X_1, \ldots, X_n) = \sum_{i=1}^n H(X_i|X^{i-1}) \leq \sum_{i=1}^n H(X_i)$.
  6. $H(X)$ is concave in p_X.

Lecture 1 Review

Convexity

- $f : S \mapsto \mathbb{R}$ is convex if $f(\alpha x + \bar{\alpha} y) \leq \alpha f(x) + \bar{\alpha} f(y)$ for $\alpha \in [0, 1]$.
- Jensen's inequality: is f is convex, then for $X \in \mathbb{R}^n$, $f(\mathbf{E}[X]) \leq \mathbf{E}[f(X)]$.
  1. If f strictly convex, then $f(\mathbf{E}[X]) = \mathbf{E}[f(X)]$ iff X is constant.

Log sum ineq

For $a_1, \ldots, a_n, b_1, \ldots, b_n \geq 0$, $a = \sum_i a_i, b = \sum_i b_i$,

$$\sum_i a_i \log \frac{a_i}{b_i} \geq a \log \frac{a}{b}$$

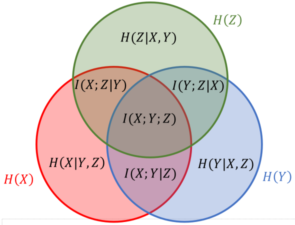
Lecture 2 Review

Overview

- Venn diagrams: Combinatorics VS information theory
- Conditional entropy
- Concavity of entropy
- Conditional Mutual Information

Lecture 2 Review

Information diagram for 3 RVs

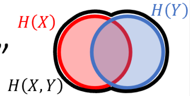


Lecture 2 Review

- Venn diagrams
  1. Shannon-type inequality: inequality implied by $I(X;Y|Z) \geq 0$
  2. $I(X;Y;Z)$ might be negative!
- Intuitively, information is similar to set.

Lecture 2 Review

Random variables as “sets”



Union $A \cup B$ of sets A, B	Joint RV (X, Y) of X, Y
$A, B \subseteq A \cup B$	Both X and Y can be determined (i.e., are functions of) (X, Y)
For any C satisfying $A, B \subseteq C$, we have $A \cup B \subseteq C$	For any Z satisfying that both X, Y are functions of Z, then (X, Y) is a function of Z
$\max\{\lvert A\rvert, \lvert B\rvert\} \leq \lvert A \cup B\rvert \leq \lvert A\rvert + \lvert B\rvert$	$\max\{H(X), H(Y)\} \leq H(X,Y) \leq H(X) + H(Y)$
If A, B disjoint, then $\lvert A \cup B\rvert = \lvert A\rvert + \lvert B\rvert$	If X, Y indep. ,then $H(X,Y) = H(X) + H(Y)$

Lecture 2 Review

Lecture 2 Review

- Conditional entropy of Y given X:

$$H(Y|X) = \sum_{x} P_X(x) H(Y|X = x)$$
$$= \mathbf{E}\left[\log \frac{1}{p_{Y|X}(Y|X)}\right]$$
$$= \sum_{x,y} p_{X,Y}(x,y) \log \frac{1}{p_{Y|X}(y|x)}$$

  1. Average amount of new info in Y if we already know X.

- Conditional entropy vs set difference:

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

- Concavity of entropy

Lecture 2 Review

Lecture 2 Review

- Mutual information: $I(X;Y) = \mathbf{E}\left[\log \frac{p_{X,Y}(X,Y)}{p_X(X)p_Y(Y)}\right]$.
  1. Measures how much information do X, Y share
  2. $I(X;Y) \geq 0$
  3. $I(X;Y) \leq \min\{H(X), H(Y)\}$
  4. $I(X;Y) = H(Y)$ iff Y is a function of X, by $I(X;Y) = H(Y) - H(Y|X)$.
- $I(X;Y)$ is convex in $p_{Y|X}$ and concave in p_X

Lecture 2 Review

Lecture 2 Review

- Conditional mutual information:

$$\begin{aligned} I(X;Y|Z) &= H(X|Z) + H(Y|Z) - H(X,Y|Z) \\ &= H(Y|Z) - H(Y|X,Z) \\ &= \sum_z p_Z(z) I(X;Y|Z=z) \\ &= \mathbf{E}\left[\log \frac{p_{X,Y|Z}(X,Y|Z)}{p_{X|Z}(X|Z)p_{Y|Z}(Y|Z)}\right] \end{aligned}$$

$I(X;Y|Z) \geq 0$ with equality iff $X \perp\!\!\!\perp Y|Z$.

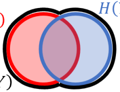
1. Condition may increase/decrease mutual information

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

- Information diagram for 4 and 5 RVs

Lecture 2 Review

Sets vs RVs



Sets	RVs
Cardinality $\lvert\cdots\rvert$	$H(\cdots)$ or $I(\cdots)$
$A \cup B$	(X, Y) (is an RV)
$A \cap B$	"$X; Y$" (not an RV!)
$A \backslash B$	"$X\|Y$" (not an RV!)
$A \cap B = \emptyset$	$X \perp Y$
$B \subseteq A$	Y is a function of X

- $|A| = |A \cap B| + |A \backslash B|$ becomes $H(X) = I(X; Y) + H(X|Y)$
- $|A \cap (B \cup C)| = |A \cap B| + |(A \cap C) \backslash B|$ becomes $I(X; Y, Z) = I(X; Y) + I(X; Z|Y)$
  - Operator precedence: "," then ";" then "|"

Lecture 2 Review

Lecture 2 Review

- Chain rule: $H(X, Y, Z) = H(X) + H(Y|X) + H(Z|X, Y)$
- Generally,

$$H(X_1, \dots, X_n) = \sum_{i=1}^{n} H(X_i|X_1, \dots, X_{i-1})$$

- For mutual information,

$$I(X_1, \dots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_1, \dots, X_{i-1})$$

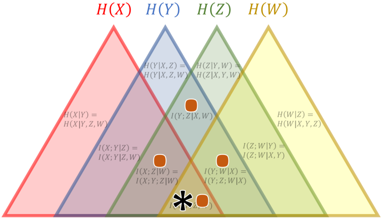
Lecture 2 Review

Markov chain

$$P(X_{i+1} = x_{i+1} | X_1 = x_1, \ldots, X_i = x_i) = P(X_{i+1} = x_{i+1} | X_i = x_i)$$

- Data processing inequality: If $X \to Y \to Z \to W$, then $I(X; W) \leq I(Y; Z)$.

- $I(Y;Z) = I(Y;W) + I(Y;Z|W)$
  $= I(X;W) + I(Y;W|X) + I(Y;Z|W)$



Lecture 2 Review

Lecture 2 Review

- Kullback-Leibler divergence: $D(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$.
  1. $D(p\|q) \geq 0$ with equality iff $p = q$.
  2. $I(X;Y) = D(p_{X,Y}\|p_X(x)p_Y(y))$: Mutual information is the divergence between the true joint distribution and the hypothetical joint distribution if X, Y were independent.
  3. It is not a distance measure! (not symmetric)
- Total variation distance: $\delta_{TV}(p,q) = \sup_{A\subseteq\mathcal{X}} |p(A) - q(A)|$.[a].
- Pinsker's inequality: $\delta_{TV}(p,q) \leq \sqrt{\frac{1}{2\log e} D(p\|q)}$.

[a] Rudin, Walter. Principles of mathematical analysis. Vol. 3. New York: McGraw-hill, 1976.

Lecture 3: Lossless Compression

Lecture 3 Review

- If X is uniformly distributed, you need $n \approx H(X) = \log_2 k$ bits to compress X.
- You can do better if you allow n to change according to value of X: **Variable-length compression**
- You should be able to **uniquely decode**: let the decoder know the boundaries of the codewords $m = f(X_1) \| \cdots \| f(X_n)$
- Prefix-free code: can be represented as a binary tree

Kraft's inequality

There exists a prefix-free code with $L(f(x)) = \ell_x$ for $x \in \mathcal{X}$ if and only if $\sum_{x \in \mathcal{X}} 2^{-\ell_x} \leq 1$

- Expected length: $\mathbb{E}[L(f(x))] = \mathbb{E}[\ell_x] = \sum_x p_X(x)\ell_x$
- Expected length must be at least $H(X)$(proved in lec3)

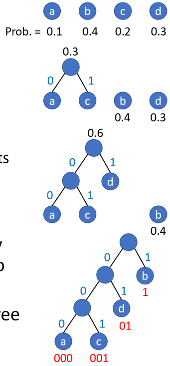
Lecture 3: Lossless Compression

Huffman coding

- An algorithm for finding the optimal prefix-free code
- Optimality: attains the smallest possible $\mathbb{E}[\ell_X]$(proved in lec3)

Huffman coding

- An algorithm for finding the optimal prefix-free code
- Maintain a collection of trees
  - Initially, each alphabet $x \in \mathcal{X}$ is its own tree
- Repeatedly find two trees with smallest total probabilities, and combine them into one tree (by adding a new root, with the two trees as left and right subtree)
- Repeat until there is only one tree



Lecture 3: Lossless Compression

Fano's inequality

$X, \hat{X}$ are r.v. over $\mathcal{X}$, $P_e = \mathbb{P}(X \neq \hat{X})$, then

$$H(X|\hat{X}) \leq H_b(P_e) + P_e \log(|\mathcal{X}| - 1) \leq 1 + P_e \log |\mathcal{X}|$$

where $H_b(a) = H(Bern(a))$) is the binary entropy function.

Lecture 3: Lossless Compression

Fano's inequality

$X, \hat{X}$ are r.v. over $\mathcal{X}$, $P_e = \mathbb{P}(X \neq \hat{X})$, then

$$H(X|\hat{X}) \leq H_b(P_e) + P_e \log(|\mathcal{X}| - 1) \leq 1 + P_e \log |\mathcal{X}|$$

where $H_b(a) = H(Bern(a))$) is the binary entropy function.

- Let $T = \mathbf{1}\{X \neq \hat{X}\}$ be the indicator of event $X \neq \hat{X}$
- $H(X|\hat{X}) = H(X, T|\hat{X}) \leq H(T) + H(X|\hat{X}, T)$
  $= H_b(P_e) + P_e H(X|\hat{X}, T = 1) + (1 - P_e) H(X|\hat{X}, T = 0)$
  $\leq H_b(P_e) + P_e \log(|\mathcal{X}| - 1)$
  since for any $\hat{x}$, $H(X|\hat{X} = \hat{x}, T = 1) \leq \log(|\mathcal{X}| - 1)$ because X can only take values in $\mathcal{X} \setminus \{\hat{x}\}$

Lecture 3: Lossless Compression

Compree $X_1, \ldots, X_n$ i.i.d. following p_X into fixed length codeword $M = \{1, \ldots, \lfloor 2^{nR} \rfloor\}$ with error probability ϵ_n.

Shannon's source coding theorem

If $R > H(X)$, then there is a code with $\epsilon_n \to 0$. If $R < H(X)$, then there does not exist code with $\epsilon_n \to 0$.

Lecture 3: Lossless Compression

Compree $X_1, \ldots, X_n$ i.i.d. following p_X into fixed length codeword $M = \{1, \ldots, \lfloor 2^{nR} \rfloor\}$ with error probability ϵ_n.

Shannon's source coding theorem

If $R > H(X)$, then there is a code with $\epsilon_n \to 0$. If $R < H(X)$, then there does not exist code with $\epsilon_n \to 0$.

- Achievability follows from Huffman coding
- For converse, assume $\epsilon_n \to 0$. By Fano's inequality,
  $H\left(X^n \middle| \hat{X}^n\right) \leq 1 + \epsilon_n \log(|\mathcal{X}|^n) = 1 + n\epsilon_n \log|\mathcal{X}|$
- $H(X) = \frac{1}{n} H(X^n) \leq \frac{1}{n}\left(H\left(\hat{X}^n\right) + H\left(X^n \middle| \hat{X}^n\right)\right)$
  $\leq \frac{1}{n}(H(M) + 1 + n\epsilon_n \log|\mathcal{X}|)$
  $\leq R + \frac{1}{n} + \epsilon_n \log|\mathcal{X}| \to R$ as $n \to \infty$

Concluding Remarks

- Def of (joint) entropy, **properties**
- **Venn diagrams**
- **Conditional entropy**
- (Conditional) Mutual Information
- Karnaugh map
- Total variation distance
- **Variable-length compression**: uniquely decodability, Prefix-free code, Kraft's inequality, Optimality
- Fano's inequality
- Shannon's source coding theorem

Remarks

- Review all the homeworks carefully!
- No cheating, and good luck!