EE 231A: Information Theory

Professor Wesel
Cell: (310) 922-7831
wesel@ucla.edu

EE 231E: Information Theory
Lecture 1

A. Introduction
B. Entropy
C. Relative Entropy
D. Mutual Information

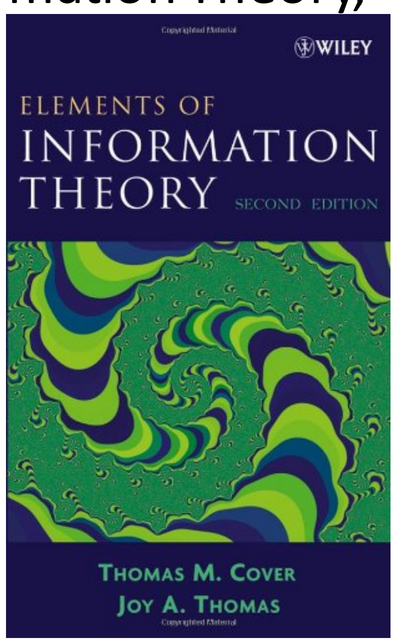
EE 231E: Information Theory
Lecture 1

A. Introduction
B. Entropy
C. Relative Entropy
D. Mutual Information

Elements of Information Theory,

Second Edition

Tom Cover
and
Joy Thomas



Course Structure

- Two video lectures each week, Tuesday and Thursday. Each lecture is broken into modules that will be provided as separate recordings.
- There are about eight homework assignments (50% of grade) due on Tuesdays. There are seven parts to the class as described on the syllabus. The other 50% of the grade will be based on assessment of mastery of these seven parts through quizzes or projects.
- Often, you can start homework after watching a single module, so you can do a little work each night.

5

Course Support

- Professor Richard Wesel is the instructor
  - Email wesel@ucla.edu
  - Cell (310) 922-7831
- Hengjie Yang is the TA.
  - Email: hengjie.yang@ucla.edu
  - Cell: (310) 746-6950.
- Hengjie and I are planning to hold (recorded) office hours all Monday afternoon based on historical demand. Please attend.
- We are using Piazza for online Q/A. Please join.
- Call our cells or email us if you need to arrange help before the next office hour and Piazza has not been sufficient. Either Hengjie or I can hold a pop-up (recorded) office hour.

6

Two major themes

- 1) How much can we compress Data (with a known probabilistic distribution)?
  - For lossless compression, the answer is entropy.
  - For lossy compression, the answer is the rate-distortion function $R(D)$.
- 2) How much "information" (i.e. fully compressed data) can we send reliably over a channel (with a known probabilistic structure)?
  - The answer is capacity.

7

Practical Significance

- The results on channel capacity and lossless compression have a direct quantitative impact on real systems.
- Communication channels and text have a well-defined probabilistic structure.
- The results on lossy compression provide valuable insight on how to do compression in many cases.
- More complex data sources such as images, music, and videos also have plenty of structure.
- A complication of directly applying lossy information theory is that simple distortion metrics do not always match well with subjective notions of quality.

8

An application of Ergodic Theory

- From a mathematical or probabilistic point of view, most results in information theory may be thought of as an application of ergodic theory (i.e. the law of large numbers).
- At the end of the quarter we will provide some results that apply to short transmission lengths that are of particular interest to Hengjie and to me right now.

EE 231E: Information Theory Lecture 1

A. Introduction
B. Entropy
C. Relative Entropy
D. Mutual Information

Entropy

- Entropy $H(X)$
  - number of bits required to describe X on the average.

$$H(X) = -\sum_{x \in \chi} p(x) \log p(x) \qquad \log \triangleq \log_2$$

$$= H(p)$$ If for the same X multiple distributions are possible…

- $H(X) = E_p[-\log p(X)]$
- Entropy has units of bits

11

Example of entropy

x	1	2	3	4
P(x)	1/2	1/4	1/8	1/8

$$-\log p = \log \frac{1}{p}$$

$$H(X) = \frac{1}{2}\log 2 + \frac{1}{4}\log 4 + 2 \cdot \frac{1}{8}\log 8$$

$$= 1\frac{3}{4} \quad \text{bits}$$

12

The “names” don’t matter

x	a	b	c	d
$P(x)$	1/2	1/4	1/8	1/8

$$-\log p = \log \frac{1}{p}$$

$$H(X) = \frac{1}{2}\log 2 + \frac{1}{4}\log 4 + 2 \cdot \frac{1}{8}\log 8$$

$$= 1\frac{3}{4} \quad \text{bits}$$

13

H(X) as average description length

x	a	b	c	d
$P(x)$	1/2	1/4	1/8	1/8

$$-\log p = \log \frac{1}{p}$$

An efficient description length

$$H(X) = E_p\left[\log \frac{1}{p(X)}\right]$$

Best possible average description length

$$H(X) = \frac{1}{2}\log 2 + \frac{1}{4}\log 4 + 2 \cdot \frac{1}{8}\log 8 = 1\frac{3}{4} \quad \text{bits}$$

14

Properties of entropy

- Lemma 2.1.1 $H(X) \geq 0$ Why?

$$\sum_x p(x)(-\log p(x)) \geq 0$$

- Lemma 2.1.2 (changing bases)
  – Define $H_b(X) = -\sum_x p(x) \log_b p(x)$

$$H_b(X) = (\log_b a) H_a(X)$$

since $\log_b p(x) = (\log_b a) \log_a p(x)$

15

Entropy using Natural Logarithm

- Sometimes we might compute entropy using base-e logarithm instead of base-2.

$$H_e(X) = -\sum_{x \in \mathcal{X}} p(x) \ln p(x)$$

- In this case, the entropy is in units of nats.

16

Joint entropy

- Joint entropy is the number of bits to describe both X and Y on the average.

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

$$= E_{P(x,y)}[-\log p(X,Y)]$$

17

Dimensionality doesn't change entropy for discrete distributions

- Note that for discrete alphabets, whether we describe the probabilities with one or two dimensions doesn't really matter. It's still the negative of the sum of p log p.

	X=1	X=2
Y=1	1/2	1/4
Y=2	1/8	1/8

18

Dimensionality doesn't change entropy for discrete distributions

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

$$= \frac{1}{2}\log 2 + \frac{1}{4}\log 4 + 2\cdot\frac{1}{8}\log 8$$

$$= 1\frac{3}{4} \quad \text{bits}$$

	X=1	X=2
Y=1	1/2	1/4
Y=2	1/8	1/8

19

Conditional entropy

- Conditional entropy is the number of bits to describe Y given that X is already known exactly, averaged over possible X values.

$$\begin{aligned} H(Y\mid X) &= \sum_x p(x) H(Y\mid X=x) \\ &= -\sum_x p(x)\sum_y p(y\mid x)\log p(y\mid x) \\ &= -\sum_x\sum_y p(x,y)\log p(y\mid x) \\ &= E_{p(x,y)}\left[-\log p(Y\mid X)\right] \end{aligned}$$

20

Chain rule for entropy

$$H(X,Y) = H(X) + H(Y \mid X)$$

Information required to describe X on the average

Information required to describe Y on the average Given X is known (averaged over X's)

Proof of chain rule

$$H(X)+H(Y \mid X) = -\sum_x p(x)\log p(x) - \sum_x\sum_y p(x,y)\log p(y \mid x)$$

$$= -\sum_x p(x)\sum_y p(y \mid x)\log p(x) - \sum_x\sum_y p(x,y)\log p(y \mid x)$$

$$= -\sum_x\sum_y p(x,y)\log p(x) + p(x,y)\log p(y \mid x)$$

$$= -\sum_x\sum_y p(x,y)[\log p(x) + \log p(y \mid x)]$$

$$= -\sum_x\sum_y p(x,y)\log p(x,y)$$

$$= H(X,Y) \; = H(Y) + H(X \mid Y)$$

General chain rule

$$H(X_1, X_2, ..., X_n) = \sum_{i=1}^{n} H(X_i \mid X_{i-1}, ..., X_1)$$

Example:

$$H(X_1, X_2, X_3) = H(X_1) + H(X_2 \mid X_1) + H(X_3 \mid X_2, X_1)$$

23

Proof of General chain rule

- Proof by induction
  - Base case: $H(X_1, X_2) = H(X_1) + H(X_2 \mid X_1)$
  - Suppose $H(X_1, X_2, ..., X_{n-1}) = \sum_{i=1}^{n-1} H(X_i \mid X_{i-1}, ..., X_1)$

$$H(X_1, X_2, ..., X_n) \overset{(a)}{=} H(X_1, X_2, ..., X_{n-1}) + H(X_n \mid X_{n-1}, ..., X_1)$$

$$= \sum_{i=1}^{n-1} H(X_i \mid X_{i-1}, ..., X_1) + H(X_n \mid X_{n-1}, ..., X_1)$$

$$= \sum_{i=1}^{n} H(X_i \mid X_{i-1}, ..., X_1)$$

For (a) the proof goes exactly like $H(X,Y)=H(X)+H(Y|X)$ with X replaced by $X_1,...,X_{n-1}$ and Y replaced by X_n.

24