Introduction to Information Theory

Lecture 1

Lecturer: Haim Permuter *Scribe: Haim Permuter and Dor Tsur*

I. NOTATION

Throughout these lecture notes we will use the following notation which is similar to the one used in the seminal book[1] by Cover and Thomas [1]:

- X - random variable.
- $\mathcal{X}$ - alphabet of X. The Alphabet of X is the set of all possible outcomes. For instance, if X is a Binary random variable then $\mathcal{X} = \{0, 1\}$. We denote sets by calligraphic letters, such as $\mathcal{A}, \mathcal{B}, ...$
- $|\mathcal{X}|$ - cardinality of the alphabet. Unless it is said otherwise, we assume that the alphabet is finite, i.e., $|\mathcal{X}| < \infty$.
- x - an observation or a specific value. Clearly, $x \in \mathcal{X}$.
- $P_X(x)$ - the probability that the random variable X gets the value x, i.e., $P_X(x) = \Pr\{X = x\}$.
- P_X or $P_X(\cdot)$ - denotes the whole vector of probabilities, also known as probability mass function (pmf).
- $P(x)$ - this is a short notation for $P_X(x)$.
- $\mathbb{E}[X]$ - expectation, i.e.,

$$\mathbb{E}[X] \triangleq \sum_{x \in \mathcal{X}} xP(x). \tag{1}$$

  Similarly $\mathbb{E}[g(X)]$ is

$$\mathbb{E}[g(X)] \triangleq \sum_{x \in \mathcal{X}} g(x)P(x). \tag{2}$$

Example 1 (A fair dice) Consider a fair dice with six faces. The random variable X is the dice. The alphabet $\mathcal{X}$ is the set $\{1, 2, 3, 4, 5, 6\}$. The cardinality of the alphabet

[1]The book [1] is an excellent book which I highly recommend to read in addition to the lecture notes. Specifically, the material of this lecture appears in [1, Chapter 2].

$|\mathcal{X}| = 6$ and $\forall x = 1, 2, ..., 6$, $P_X(x) = \frac{1}{6}$. The whole probability vector is denoted by P_X and is equal to $[\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}]$.

II. Information measures and their basic properties

A. Entropy

Definition 1 (Entropy) The *Entropy* of a random variable X with a finite alphabet is defined as

$$\begin{aligned} H(X) &\triangleq \mathbb{E}[-\log_2 P_X(X)] \\ &= -\sum_{x\in\mathcal{X}} P_X(x)\log_2 P_X(x). \end{aligned} \tag{3}$$

Note that $H(X)$ is a function of X only through P_X. The logarithm is of base 2 and the units of entropy are bits[2].

Example 2 (Entropy of a dice) Let X denotes a fair dice with 6 faces.

$$\begin{aligned} H(X) &= -\sum_{x=1}^{6} P(x)\log_2 P(x) \\ &= -\sum_{x=1}^{6} \frac{1}{6}\log_2 \frac{1}{6} \\ &= \log_2 6. \end{aligned} \tag{4}$$

Usually entropy is measured in bits, and throughout the lecture notes we would use $\log x$ as notation for base 2 logarithm, namely, $\log_2(x)$ unless otherwise is mentioned.

Example 3 (Entropy of a Bernoulli random variable) Let $X \sim Bern(p)$, i.e., $\Pr\{X = 1\} = p$ and $\Pr\{X = 0\} = 1 - p$.

$$H(X) = -p\log p - (1-p)\log(1-p). \tag{5}$$

[2]The name Bit is a contraction of binary digit binary information. Claude E. Shannon first used the word bit in his seminal 1948 paper A Mathematical Theory of Communication. He attributed its origin to John W. Tukey, who had written a Bell Labs memo on 9 January 1947 in which he contracted "binary digit" to simply "bit". Interestingly, Vannevar Bush had written in 1936 of "bits of information" that could be stored on the punched cards used in the mechanical computers of that time.

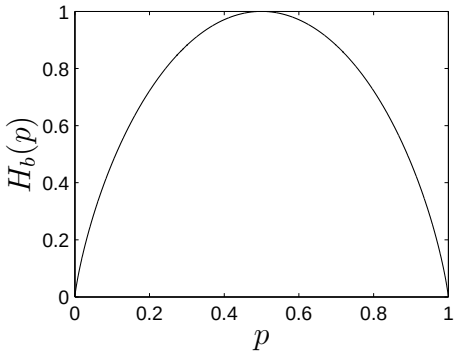

Fig. 1. Binary Entropy $H_b(p)$. The figure depicts the entropy of a Bernoulli random variable with parameter $0 \leq p \leq 1$.

We denote the entropy of a Bernoulli random variable with parameter p as $H_b(p)$. Fig. 1 depicts $H_b(p)$ as a function of p. Note that $H_b(0) = H_b(1) = 0$, the maximum is obtained at $p = 0.5$, $H_b(p) = H_b(1-p)$ and the curve is concave, a property, which we will explain later in the course.

Exercise 1 (Basic properties of entropy) Prove the following properties of entropy

- $H(X) \geq 0$.
- Let $X \sim Unif(\frac{1}{m})$, i.e., $P(x) = \frac{1}{m}$ for $x = 1, 2, ..., m$, then $H(X) = \log m$.
- $H(X) = 0 \iff X$ is a constant.

Definition 2 (Joint Entropy) Let X and Y be two random variables with a joint distribution $P(x, y)$. The *joint entropy* $H(X, Y)$ is defined as

$$\begin{aligned} H(X,Y) &\triangleq \mathbb{E}[-\log P(X,Y)] \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P(x,y)\log P(x,y). \end{aligned} \tag{6}$$

Exercise 2 (Entropy of independent random variable) Consider two random variables X, Y that are independent of each other, i.e., $P(x, y) = P(x)P(y)$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Show that

$$H(X,Y) = H(X) + H(Y). \tag{7}$$

Definition 3 (Conditional Entropy) Let X and Y be two random variables with a joint distribution $P(x, y)$. The *conditional entropy* $H(X|Y)$ is defined as

$$\begin{aligned} H(X|Y) &\triangleq \mathbb{E}[-\log P(X|Y)] \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P(x,y)\log P(x|y). \end{aligned} \tag{8}$$

The conditional entropy of X given a specific value y is denoted as $H(X|Y = y)$ or $H(X|y)$ and is defined as

$$\begin{aligned} H(X|y) &\triangleq \mathbb{E}[-\log P(X|y)] \\ &= -\sum_{x\in\mathcal{X}} P(x|y)\log P(x|y). \end{aligned} \tag{9}$$

Exercise 3 (Conditional entropy) Prove the following equality

$$H(X|Y) = \sum_{y\in\mathcal{Y}} P(y)H(X|y). \tag{10}$$

B. Chain Rule

From the definition of conditional distribution $P(y|x) = \frac{P(y,x)}{P(x)}$, which holds when $P(x) > 0$, follows the chain rule

$$P(y, x) = P(x)P(y|x). \tag{11}$$

Using this rule we obtain the chain rule for entropy.

Lemma 1 (Chain rule for entropy) For any discrete random variables X, Y

$$H(X, Y) = H(X) + H(Y|X). \tag{12}$$

Proof:

$$\begin{aligned} H(X,Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P(x,y)\log P(x,y) \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P(x,y)\log\left(P(x)P(y|x)\right) \\ &\overset{(a)}{=} -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P(x,y)\log P(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P(x,y)\log P(y|x) \end{aligned}$$

$$
\begin{aligned}
&= -\sum_{x\in\mathcal{X}} P(x)\log P(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P(x,y)\log P(y|x) \\
&\overset{(b)}{=} H(X) + H(Y|X). \qquad (13)
\end{aligned}
$$

where

(a) follows from the identity $\log ab = \log a + \log b$.

(b) follows from the definition of entropy (Def. 1) and conditional entropy (Def. 3).

A shorter proof can be written using the expectation operator

$$
\begin{aligned}
H(X,Y) &= -\mathbb{E}[\log P(X,Y)] \\
&= -\mathbb{E}[\log P(X)P(Y|X)] \\
&\overset{(a)}{=} -\mathbb{E}[\log P(X)] - \mathbb{E}[\log P(Y|X)] \\
&= H(X) + H(Y|X). \qquad (14)
\end{aligned}
$$

where Step (a) follows from the linearity of expectation, i.e., $\mathbb{E}[A+B] = \mathbb{E}[A] + \mathbb{E}[B]$.

■

We use the notation

- x^n - is the vector $(x_1, x_2, ..., x_n)$ for $n \geq 1$. If $n = 0$ then the vector is empty.
- x_i^j - is the vector $(x_i, x_{i+1}, ..., x_j)$, for $j > i$. If $j = i$, then the vector has only one element x_i and if $j < i$, the vector is empty.

Exercise 4 (Chain Rule) Show by induction the following chain rule

$$P(x^n) = \prod_{i=1}^{n} P(x_i|x^{i-1}). \qquad (15)$$

$$H(X^n) = \sum_{i=1}^{n} H(X_i|X^{i-1}). \qquad (16)$$

C. Mutual Information

Definition 4 (Mutual information) The *mutual information* between two random variables, X, Y with finite alphabet is

$$I(X;Y) \triangleq \sum_{x,y} P(x,y)\log\frac{P(x,y)}{P(x)P(y)}. \qquad (17)$$

Exercise 5 (Identity of mutual information) Prove the following equalities

$$I(X;Y) = H(X) - H(X|Y) \tag{18}$$

$$I(X;Y) = H(Y) - H(Y|X) \tag{19}$$

$$I(X;Y) = H(Y) + H(X) - H(Y,X) \tag{20}$$

$$I(Y;X) = I(X;Y) \tag{21}$$

$$I(X;X) = H(X) \tag{22}$$

Fig. 2 depicts the relation from Exercise 5 in a diagram. This diagram is often called Venn diagram.

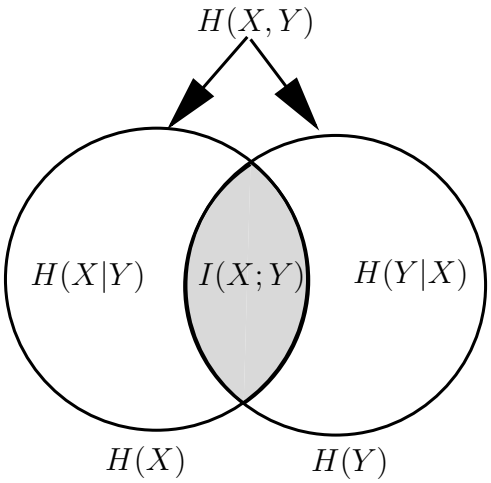

Fig. 2. Diagram that represents the relation between entropy and mutual information.

Definition 5 (Conditional mutual information) The *mutual information* between two random variables, X, Y conditioned on Z, where all the random variables have finite alphabet is

$$I(X;Y|Z) \triangleq \sum_{x,y,z} P(x,y,z) \log \frac{P(x,y|z)}{P(x|z)P(y|z)}. \tag{23}$$

Exercise 6 (Chain rule of mutual information) First prove that

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z), \tag{24}$$

then using the chain rule of entropy show that

$$I(X;Y,Z) \;\; = \;\; I(X;Y) + I(X;Z|Y). \tag{25}$$

The property in (25) is called the *chain rule of mutual information.*

D. Divergence

Definition 6 (Kullback Liebler divergence) The *Kullback Liebler divergence* a.k.a. (also know as) *relative entropy* is

$$D(P_X||Q_X) \triangleq \sum_x P(x) \log \frac{P(x)}{Q(x)} = \mathbb{E}_P \left[\log \frac{P(X)}{Q(X)} \right]. \tag{26}$$

Exercise 7 (Representing mutual information via divergence) Show that

$$I(X;Y) = D(P_{X,Y}||P_X P_Y). \tag{27}$$

E. Cross Entropy

Definition 7 (Cross Entropy) Let P and Q by two probability distributions of a random variable X. The *Cross Entropy* of P and Q is defined as

$$H(P,Q) \triangleq -\sum_x P(x) \log Q(x). \tag{28}$$

Consider some RV X with distribution P. Cross Entropy can be thought of as trying to calculating the entropy of X while assuming it's distribution is Q.

Remark 1 (Cross entropy is not joint entropy) Note that Cross entropy is not related to joint entropy. Cross entropy is a measure between two probability distributions and not two random variables, e.g.

$$H(X,Y) \neq H(P,Q). \tag{29}$$

Exercise 8 (Representing information measures using cross entropy) Show that

$$H(P,Q) = H(P) + D(P||Q). \tag{30}$$

$$H(P) = H(P, P). \tag{31}$$

Now, express all information measures, divergence, mutual-information and entropy using only cross-entropy.

Remark 2 (Cross entropy is a fundamental measure also in machine learning) If you look in most Information Theory Books [1], [2] cross entropy appears as a result to some important problems but not as an information measure by itself. As can be seen from the previous exercise cross entropy is a fundamental measure that can be used to express all other information measures such as entropy, divergence and mutual information. Furthermore, cross entropy plays an important role in machine learning which we will learn in this course later on.

REFERENCES

[1] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, New-York, 2nd edition, 2006.

[2] R. G. Gallager. *Information theory and reliable communication*. Wiley, New York, 1968.