Information Theory: Principles and Applications

Tiago T. V. Vinhoza

March 19, 2010

1. Course Information
2. What is Information Theory?
3. Review of Probability Theory
4. Information Measures

Information Theory: Principles and Applications

- **Prof. Tiago T. V. Vinhoza**
  - Office: FEUP Building I, Room I322
  - Office hours: Wednesdays from 14h30-15h30.
  - Email: tiago.vinhoza@ieee.org
- Prof. José Vieira
- Prof. Paulo Jorge Ferreira

Information Theory: Principles and Applications

- http://paginas.fe.up.pt/∼vinhoza (link for Info Theory)
  - Homeworks
  - Other notes
- My Evaluation: (Almost) Weekly Homeworks + Final Exam
- References:
  - Elements of Information Theory, Cover and Thomas, Wiley
  - Information Theory and Reliable Communication, Gallager
  - Information Theory, Inference, and Learning Algorithms, McKay (available online)

What is Information Theory?

- IT is a branch of math (a strictly deductive system). (C. Shannon, The bandwagon)
- General statistical concept of communication. (N. Wiener, What is IT?)
- It was build upon the work of Shannon (1948)
- It answers to two fundamental questions in Communications Theory:
  - What is the fundamental limit for information compression?
  - What is the fundamental limit on information transmission rate over a communications channel?

What is Information Theory?

- Mathematics: Inequalities
- Computer Science: Kolmogorov Complexity
- Statistics: Hypothesis Testings
- Probability Theory: Limit Theorems
- Engineering: Communications
- Physics: Thermodynamics
- Economics: Portfolio Theory

Communications Systems

- The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. (Claude Shannon: A Mathematical Theory of Communications, 1948)

Digital Communications Systems

- Source
- Source Coder: Convert an analog or digital source into bits.
- Channel Coder: Protection against errors/erasures in the channel.
- Modulator: Each binary sequence is assigned to a waveform
- Channel: Physical Medium to send information from transmitter to receiver. Source of randomness.
- Demodulator, Channel Decoder, Source Decoder, Sink.

Digital Communications Systems

- Modulator + Channel = Discrete Channel.
- Binary Symmetric Channel.
- Binary Erasure Channel.

Review of Probability Theory

- Axiomatic Approach
- Relative Frequency Approach

Axiomatic Approach

- Application of a mathematical theory called *M*easure Theory.
- It is based on a triplet

$$(\Omega, \mathcal{F}, P)$$

  where
  - Ω is the sample space, which is the set of all possible outcomes.
  - $\mathcal{F}$ is the $\sigma-$algebra, which is the set of all possible events (or combinations of outcomes).
  - P is the probability function, which can be any set function, whose domain is Ω and the range is the closed unit interval [0,1]. It must obey the following rules:
    - $P(\Omega) = 1$
    - Let A be any event in $\mathcal{F}$, then $P(A) \geq 0$.
    - Let A and B be two events in $\mathcal{F}$ such that $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B)$.

Axiomatic Approach: Other properties

- Probability of complement: $P(\overline{A}) = 1 - P(A)$.
- $P(A) \leq 1$.
- $P(\emptyset) = 0$.
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Conditional Probability

- Let A and B be two events, with $P(A) > 0$. The conditional probability of B given A is defined as:

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

- Hence, $P(A \cap B) = P(B|A)P(A) = P(A|B)P(B)$
- If $A \cap B = \emptyset$ then $P(B|A) = 0$.
- If $A \subset B$, then $P(B|A) = 1$.

Bayes Rule

- If A and B are events

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Total Probabilty Theorem

- A set of B_i, $i = 1, \ldots, n$ of events is a partition of Ω when:
  - $\bigcup_{i=1}^{n} B_i = \Omega$.
  - $B_i \cap B_j = \emptyset$, if $i \neq j$.
- Theorem: If A is an event and B_i, $i = 1, \ldots, n$ of is a partition of Ω, then:

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$

Independence between Events

- Two events A and B are statistically independent when

$$P(A \cap B) = P(A)P(B)$$

- Supposing that both $P(A)$ and $P(B)$ are greater than zero, from the above definition we have that:

$$P(A|B) = P(A) \qquad P(B|A) = P(B)$$

- Independent events and mutually exclusive events are different!

Independence between events

- N events are statistically independent if the intersection of the events contained in any subset of those N events have probability equal to the product of the individual probabilities
- Example: Three events A, B and C are independent if:

$$P(A \cap B) = P(A)P(B), \; P(A \cap C) = P(A)P(C), \; P(B \cap C) = P(B)P(C)$$

$$P(A \cap B \cap C) = P(A)P(B)P(C)$$

Random Variables

- A random variable (rv) is a function that maps each $\omega \in \Omega$ to a real number.

$$\begin{aligned} X \quad : \quad & \Omega \rightarrow \mathbb{R} \\ & \omega \rightarrow X(\omega) \end{aligned}$$

- Through a random variable, subsets of Ω are mapped as subsets (intervals) of the real numbers.

$$P(X \in I) = P(\{w | X(\omega) \in I\})$$

Random Variables

- A real random variable is a function whose domain is Ω and such that
  - for all real number x, the set $A_x = \{\omega | X(\omega) \leq x\}$ is an event.
  - $P(w | X(w) = \pm\infty) = 0$.

Cumulative Distribution Function

$$
\begin{aligned}
F_X \quad &: \quad \mathbb{R} \rightarrow [0,1] \\
& \quad X \rightarrow F_X(x) = P(X \leq x) = P(\omega | X(\omega) \leq x)
\end{aligned}
$$

- $F_X(\infty) = 1$
- $F_X(-\infty) = 0$
- If $x_1 < x_2$, $F_X(x_2) \geq F_X(x_1)$.
- $F_X(x^+) = \lim_{\epsilon \rightarrow 0} F_X(x + \epsilon) = F_X(x)$. (continuous on the right side).
- $F_X(x) - F_X(x^-) = P(X = x)$.

Types of Random Variables

- Discrete: Cumulative function is a step function (sum of unit step functions)

$$F_X(x) = \sum_i P(X = x_i)u(x - x_i)$$

  where $u(x)$ is the unit step funtion.
- Example: X is the random variable that describes the outcome of the roll of a die. $X \in \{1, 2, 3, 4, 5, 6\}$

Types of Random Variable

- Continous: Cumulative function is a continous function.
- Mixed: Neither discrete nor continous.

Probability Density Function

- It is the derivative of the cumulative distribution function:

$$p_X(x) = \frac{d}{dx} F_X(x)$$

- $\int_{-\infty}^{x} p_X(x)dx = F_X(x)$.
- $p_X(x) \geq 0$.
- $\int_{-\infty}^{\infty} p_X(x)dx = 1$.
- $\int_{a}^{b} p_X(x)dx = F_X(b) - F_X(a) = P(a \leq X \leq b)$.
- $P(X \in I) = \int_I p_X(x)dx$, $I \subset \mathbb{R}$.

Discrete Random Variables

- Let us now focus only on discrete random variables.
- Let X be a random variable with sample space $\mathcal{X}$
- The probability mass function (probability distribution function) of X is a mapping $p_X(x) : \mathcal{X} \rightarrow [0, 1]$ satisfying:

$$\sum_{X \in \mathcal{X}} p_X(x) = 1$$

- The number $p_X(x) := P(X = x)$

Discrete Random Vectors

- Let $Z = [X, Y]$ be a random vector with sample space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$
- The joint probability mass function (probability distribution function) of Z is a mapping $p_Z(z) : \mathcal{Z} \to [0, 1]$ satisfying:

$$\sum_{Z \in \mathcal{Z}} p_Z(z) = \sum_{x,y \times \mathcal{Y}} p_{XY}(x, y) = 1$$

- The number $p_Z(z) := p_{XY}(x, y) = P(Z = z) = P(X = x, Y = y)$.

Discrete Random Vectors

- Marginal Distributions

$$p_X(x) = \sum_{y \in \mathcal{Y}} p_{XY}(x, y)$$

$$p_Y(y) = \sum_{x \in \mathcal{X}} p_{XY}(x, y)$$

Discrete Random Vectors

- Conditional Distributions

$$p_{X|Y=y}(x) = \frac{p_{XY}(x, y)}{p_Y(y)}$$

$$p_{Y|X=x}(y) = \frac{p_{XY}(x, y)}{p_X(x)}$$

Discrete Random Vectors

- Random variables X and Y are independent if and only if

$$p_{XY}(x, y) = p_X(x)p_Y(y)$$

- Consequences:

$$p_{X|Y=y}(x) = p_X(x)$$

$$p_{Y|X=x}(y) = p_Y(y)$$

Moments of a Discrete Random Variable

- The $n-$th order moment of a discrete random variable X is defined as:

$$E[X^n] = \sum_{x \in \mathcal{X}} x^n p_X(x)$$

- if $n = 1$, we have the mean of X, $m_X = E[X]$.
- The $m-$th order central moment of a discrete random variable X is defined as:

$$E[(X - m_X)^m] = \sum_{x \in \mathcal{X}} (x - m_X)^m p_X(x)$$

- if $m = 2$, we have the variance of X, σ_X^2.

Moments of a Discrete Random Vector

- The joint moment $n-$th order with relation to X and $k-$th order with relation to Y:

$$m_{nk} = E[X^n Y^k] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} x^n y^k p_{XY}(x, y)$$

- The joint central $n-$th order with relation to X and $k-$th order with relation to Y:

$$\mu_{nk} = E[(X - m_X)^n (Y - m_Y)^k] == \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} (x - m_X)^n (y - m_Y)^k p_{XY}(x, y)$$

Correlation and Covariance

- The correlation of two random variables X and Y is the expected value of their product (joint moment of order 1 in X and order 1 in Y):

$$Corr(X, Y) = m_{11} = E[XY]$$

- The covariance of two random variables X and Y is the joint central moment of order 1 in X and order 1 in Y:

$$Cov(X, Y) = \mu_{11} = E[(X - m_X)(Y - m_Y)]$$

- $Cov(X, Y) = Corr(X, Y) - m_X m_Y$
- Correlation Coefficient:

$$\rho_{XY} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} \quad \rightarrow \quad -1 \leq \rho_{XY} \leq 1$$

What is Information?

- It is a measure that quantifies the *uncertainty* of an event with given probability - Shannon 1948.
- For a discrete source with finite alphabet $\mathcal{X} = \{x_0, x_1, \ldots, x_{M-1}\}$ where the probability of each symbol is given by $P(X = x_k) = p_k$

$$I(x_k) = \log \frac{1}{p_k} = -\log(p_k)$$

- If logarithm is base 2, information is given in bits.

What is Information?

- It represents the *surprise* of seeing the outcome (a highly probable outcome is not surprising).

event	probability	surprise
one equals one	1	0 bits
wrong guess on a 4-choice question	3/4	0.415 bits
correct guess on true-false question	1/2	1 bit
correct guess on a 4-choice question	1/4	2 bits
seven on a pair of dice	6/36	2.58 bits
win any prize at Euromilhões	1/24	4.585 bits
win Euromilhões Jackpot	$\approx$ 1/76 million	$\approx$ 26 bits
gamma ray burst mass extinction today	$< 2.7 \cdot 10^{-12}$	$>$ 38 bits

Entropy

- Expected value of information from a source.

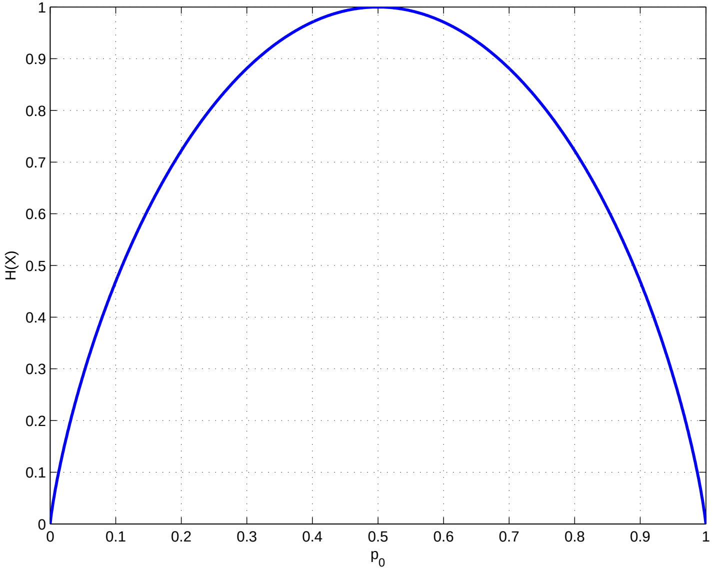
$$
\begin{aligned}
H(X) = E[I(x_k)] &= \sum_{x \in \mathcal{X}} p_x(x) I(x_k) \\
&= -\sum_{x \in \mathcal{X}} p_x(x) \log p_x(x)
\end{aligned}
$$

Entropy of binary source

- Let X be a binary source with p_0 and p_1 being the probabililty of symbols x_0 and x_1 respectively.

$$\begin{aligned} H(X) &= -p_0 \log p_0 - p_1 \log p_1 \\ &= -p_0 \log p_0 - (1 - p_0) \log(1 - p_0) \end{aligned}$$

Entropy of binary source



Joint Entropy

- The joint entropy of a pair of random variables X and Y is given by:.

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

Conditional Entropy

- Average amount of information of a random variable given the occurence of other.

$$
\begin{aligned}
H(X|Y) &= \sum_{y\in\mathcal{Y}} p_Y(y) H(X|Y=y) \\
&= -\sum_{y\in\mathcal{Y}} p_Y(y) \sum_{x\in\mathcal{X}} p_{X|Y=y}(x) \log p_{x|Y=y}(x) \\
&= -\sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} p_{XY}(x,y) \log p_{X|Y=y}(x)
\end{aligned}
$$

Chain Rule of Entropy

- The entropy of a pair of random variables is equal to the entropy of one of them plus the conditional entropy.

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

- Corollary

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

Chain Rule of Entropy: Generalization

$$H(X_1, X_2, \ldots, X_M) = \sum_{j=1}^{M} H(X_j | X_1, \ldots, X_{j-1})$$

Relative Entropy: Kullback-Leibler Distance

- Is a measure of the distance between two distributions.
- The relative entropy between two probability density functions $p_X(x)$ and $q_X(x)$ is defined as:

$$D(p_X(x)||q_X(x)) = \sum_{x \in \mathcal{X}} p_X(x) \log \frac{p_X(x)}{q_X(x)}$$

Relative Entropy: Kullback-Leibler Distance

- $D(p_X(x)||q_X(x)) \geq 0$ with equality if and only if $p_X(x) = q_X(x)$.
- $D(p_X(x)||q_X(x)) \neq D(q_X(x)||p_X(x))$

Mutual Information

- The mutual information of two random variables X and Y is defined as the relative entropy between the joint probability density $p_{XY}(x,y)$ and the product of the marginals $p_X(x)$ and $p_Y(y)$

$$\begin{aligned} I(X;Y) &= D(p_{XY}(x,y)||p_X(x)p_Y(y)) \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{XY}(x,y)\log\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)} \end{aligned}$$

Mutual Information: Relations with Entropy

- Reducing uncertainty of X due to the knowledge of Y:

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

- Symmetry of the relation above: $I(X;Y) = H(Y) - H(Y|X)$
- Sum of entropies:

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

- "Self" Mutual Information:

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

Mutual Information: Other Relations

- Conditional Mutual Information:

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

- Chain Rule for Mutual Information

$$I(X_1, X_2, \ldots, X_M; Y) = \sum_{j=1}^{M} I(X_j; Y|X_1, \ldots, X_{j-1})$$

Convex and Concave Functions

- A function $f(\cdot)$ is convex over ain interval (a, b) if for every $x_1, x_2 \in [a, b]$ and $0 \leq \lambda \leq 1$, if :

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2)$$

- A function $f(\cdot)$ is convex over an interval (a, b) if its second derivative is non-negative over that interval (a, b).
- A function $f(\cdot)$ is concave if $-f(\cdot)$ is convex.
- Examples of convex functions: x^2, $|x|$, e^x, $x \log x, x \geq 0$.
- Examples of concave functions: $\log x$ and $\sqrt{x}$, for $x \geq 0$.

Jensen's Inequality

- If $f(\cdot)$ is a convex function and X is a random variable

$$E[f(X)] \geq f(E[X])$$

- Used to show that relative entropy and mutual information are greater than zero.
- Used also to show that $H(X) \leq \log |\mathcal{X}|$.

Log-Sum Inequality

- For n positive numbers $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots b_n$

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i \right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$$

  with equality if and only if $a_i / b_i = c$.

- This inequality is used to prove the convexity of the relative entropy and the concavity of the entropy.
- Convexity/Concavity of mutual information

Data Processing Inequality

- Random variables X, Y, Z are said to form a Markov chain in that order $X \to Y \to Z$, if the conditional distribution of Z depends only on Y and is onditionally independent of X.

$$p_{XYZ}(x,y,z) = p_X(x)p_{Y|X=x}(y)p_{Z|Y=y}(y)$$

- If $X \to Y \to Z$, then

$$I(X;Y) \geq I(X;Z)$$

- Let $Z = g(Y)$, $X \to Y \to g(Y)$, then $I(X;Y) \geq I(X;g(Y))$

Fano's Inequality

- Suppose we know a random variable Y and we wish to guess the value of a correlated random variable X.
- Fano's inequality relates the probability of error in guessing X from Y to its conditional entropy $H(X|Y)$.
- Let $\hat{X} = g(Y)$, if $P_e = P(\hat{X} \neq X)$, then

$$H(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(X|Y)$$

where $H(P_e)$ is the binary entropy function evaluated at P_e.