Lecture 1. Information Measurements

Lin Zhang

Tsinghua-Berkeley Shenzhen Institute

Shenzhen, China, 2017

1. Entropy
   - Preliminaries Re-cap
   - Information Entropy
   - Joint and Conditional Entropy
   - Properties of Entropy
2. Mutual Information and K-L Divergence
   - Mutual information
   - Properties of Mutual information
   - Likelihood Ratio and Relative Entropy
   - Information Divergence is Universal
3. Optimizing over the Measurements
   - Relations between the Information Measurements
   - Convexity and Concavity of Entropy and Mutual Information
   - Bounding the Error Probabilities
4. Generalize to Continuous RVs
   - Differential Entropy
   - Properties of Differential Entropy
   - Mutual Information for Continuous RVs
5. Summary and Reference
   - Ideas in a nutshell

Outline

1. Entropy
   - Preliminaries Re-cap
   - Information Entropy
   - Joint and Conditional Entropy
   - Properties of Entropy
2. Mutual Information and K-L Divergence
3. Optimizing over the Measurements
4. Generalize to Continuous RVs
5. Summary and Reference

Basic Concepts and Intuitions

- A discrete random variable $X \sim p_X(x)$ and $p_X(a) = \Pr\{X = a\}$.
- We call $p_X(x)$ the probability mass function (PMF)
- What is the amount of information that a random event provides?
  - The *self-information* of a random event is

$$I(a) = \log \frac{1}{p(a)}. \tag{1}$$

  - $I(a) \geq 0$
  - $I_\alpha(a) = \log_\alpha \beta I_\beta(a)$

The definition of H

- **Definition 1.1** The entropy $H(X)$ is defined as

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x). \tag{2}$$

- $H(\mathbf{p})$ is a function of $\mathbf{p}$.
- $H(\mathbf{p}) = 0$, when X is deterministic.
- Another interpretation of entropy is $H(\mathbf{p}) = \mathbb{E}\left[\log \frac{1}{p(x)}\right]$

The Uniqueness of the Form of Entropy

- Three conditions given by Shannon [1948].
    1. Continuity.
    2. Monotonousity.
    3. Additivity.
- **Theorem 1.1** The form of entropy is unique, defined as

$$f(p_1, p_2, \ldots, p_n) = -C \sum_{i=1}^{n} p_i \log p_i, \tag{3}$$

where C is a scalar constant.

The Uniqueness of the Form of Entropy

- Conditions given by A.I. Khinchin.
  1. Continuity
  2. Additivity
  3. Maximum achievable at the uniform distribution.

$$\max_{\mathbf{p}} f(p_1, p_2, \ldots, p_n) = f(\underbrace{\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}}_{n}) \tag{4}$$

  4. Zero-probability event does not change entropy

$$f(p_1, p_2, \ldots, p_n) = f(p_1, p_2, \ldots, p_n, 0) \tag{5}$$

- Note that these conditions are equivalent to the Shannon conditions.

Joint and Conditional Entropy

- **Definition 1.2** Joint entropy of (X, Y)

$$H(X,Y) = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x,y) \log \frac{1}{p(x,y)} \tag{6}$$

- **Definition 1.3** Conditional entropy

$$\begin{aligned} H(X|Y) &= \mathbb{E}\left[\log \frac{1}{p(x|y)}\right] && (7) \\ &= \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x,y) \log \frac{1}{p(x|y)} && (8) \end{aligned}$$

Chain Rule of Entropy

- **Theorem 1.2** Chain rule of entropy.

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) \tag{9}$$

- **Corollary 1.3** For $(X_1, X_2, \ldots, X_n) \sim p(x_1, x_2, \ldots, x_n)$

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1) \tag{10}$$

Basic Properties of H

1. Symmetry w.r.t $\mathbf{p} = (p_1, p_2, \dots, p_n) \in \mathbf{R}^n$
2. Non-negativity $H(X) \geq 0$
3. Additivity $H(p, q, 1-p-q) = H(p) + (1-p)H(\frac{q}{1-p})$
4. Conditioning reduces entropy $H(X|Y) \leq H(X)$
5. Maximum entropy achievable when p_i is uniformly distributed

$$H(\mathbf{p}) = H(p_1, p_2, \dots, p_n) \leq H(\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n}) = \log n = \log |\mathcal{X}|. \quad (11)$$

Outline

1. Entropy

2. Mutual Information and K-L Divergence
   - Mutual information
   - Properties of Mutual information
   - Likelihood Ratio and Relative Entropy
   - Information Divergence is Universal

3. Optimizing over the Measurements

4. Generalize to Continuous RVs

5. Summary and Reference

Definition of Mutual Information

- **Definition 1.4** The mutual information between X and Y is defined as

$$I(X;Y) = H(X) - H(X|Y) \tag{12}$$

1. $I(X;Y) = \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$
2. $I(X;Y) = H(X) + H(Y) - H(X,Y)$
3. $I(X;Y) = 0$ when X and Y are independent
4. $I(X;Y) = H(X) = H(Y)$, one-to-one mapping between X and Y

- **Definition 1.5** Mutual information between multi-variables

$$I(X;Y,Z) = H(X) - H(X|Y,Z) = H(Y,Z) - H(Y,Z|X) \tag{13}$$

Definition of Conditional Mutual Information

- **Definition 1.6** Conditional mutual information of multi-variables

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = H(Y|Z) - H(Y|X,Z) \quad (14)$$

- **Definition 1.7** Mutual information amongst three random variables is

$$I(X;Y;Z) = I(X;Y) - I(X;Y|Z). \quad (15)$$

- Note that it can be a negative value.

Basic Properties of $I(X;Y)$

1. Symmetry $I(X;Y) = I(Y;X)$
2. Non-negativity $I(X;Y) \geq 0$ and $I(X;Y|Z) \geq 0$
3. $I(X;Y) \leq \min(H(X), H(Y))$
4. Additivity

$$I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y | X_1, X_2, \ldots, X_{i-1}) \tag{16}$$

Likelihood ratio $\Lambda(x)$

- MAP (maximum a posterior probability) estimation with observed x
  if $\frac{p(\theta_0|x)}{p(\theta_1|x)} > 1$, then $H_0 : \theta = \theta_0$ is inferred.
  Otherwise, if $\frac{p(\theta_0|x)}{p(\theta_1|x)} < 1$, then $H_1 : \theta = \theta_1$ is inferred.
- By using Bayesian rule

$$p(\theta_i|x) = \frac{p(\theta_i)p(x|\theta_i)}{p(x)}, \tag{17}$$

  we reformulate MAP as a Likelihood Ratio Test (LRT).
- If the likelihood ratio

$$\Lambda(x) = \frac{p(x|\theta_0)}{p(x|\theta_1)} > \frac{p(\theta_1)}{p(\theta_0)}, \tag{18}$$

  hypotheses H_0 is true and H_1 is true otherwise. Note that $p(x|\theta_i)$ and $p(\theta_i)$ denote likelihood and prior distribution, respectively.
- The log-likelihood ratio equals $\log \Lambda(x) = \log \frac{p(x|\theta_0)}{p(x|\theta_1)}$.

Relative Entropy $D(\mathbf{p}\|\mathbf{q})$

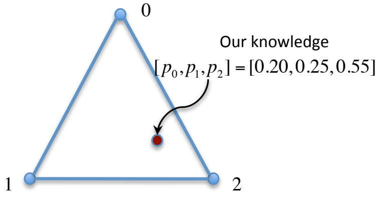
- **Definition 1.8** The Relative Entropy between $p(x)$ and $q(x)$ is defined as

$$D(\mathbf{p}\|\mathbf{q}) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} \tag{19}$$

- Other names of relative entropy are Kullback-Leibler Divergence and Cross Entropy.
- Symmetric property and triangle-inequality do NOT hold.
- $D(\mathbf{p}\|\mathbf{q}) \geq 0$ with equality if and only if $\mathbf{p} = \mathbf{q}$.

Information Divergence

- The Universe is Bayesian.
- The Mathematical Simplification.
  - Data space and space of distribution.
  - How to measure distance between distributions.
- Information as movement of knowledge.
- Divergence: A measure of volume of information



Different Version of Divergence

- Kullback-Leibler Divergence

$$D(\mathbf{p}\|\mathbf{q}) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} \tag{20}$$

- Renyi Divergence

$$D_\alpha(\mathbf{p}\|\mathbf{q}) = \frac{1}{\alpha-1}\log\left(\sum_{x\in\mathcal{X}}\frac{p^\alpha(x)}{q^{\alpha-1}(x)}\right) \tag{21}$$

- f-divergence, for convex function f

$$D_f(\mathbf{p}\|\mathbf{q}) = \sum_{x\in\mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}\right) \tag{22}$$

- Hellinger, Bregman, total variation, chi-square, alpha, etc.

Demonstration of Information Divergence

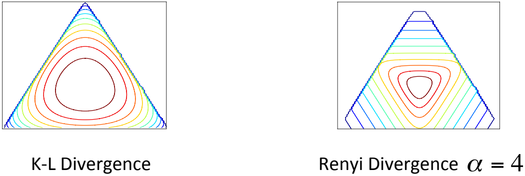


Figure: A Ternary Example of the Information Divergence

Outline

1. Entropy

2. Mutual Information and K-L Divergence

3. Optimizing over the Measurements
   - Relations between the Information Measurements
   - Convexity and Concavity of Entropy and Mutual Information
   - Bounding the Error Probabilities

4. Generalize to Continuous RVs

5. Summary and Reference

Entropy, Mutual Information and Relative entropy

- **Theorem 1.4** Entropy and K-L Divergence

$$H(X) = \log |\mathcal{X}| - D(\mathbf{p} \| \mathbf{u}), \tag{23}$$

  where $\mathbf{u}$ denotes the uniform distribution and $D(\mathbf{p} \| \mathbf{u})$ measures the divergence from $\mathbf{p}$ to $\mathbf{u}$.

- **Theorem 1.5** Mutual Information and Relative Entropy

$$I(X;Y) = D(p(x,y) \| p(x)p(y)) \tag{24}$$

The Convex Set, Convex Function and Two Lemmas

- Convex Set.
- Convex Functions.
- **Lemma 1.6 Jensen's inequality:** For any convex function f, $E(f(X)) \geq f(E(X))$ holds.
  If f is strictly convex, the equality holds if and only if X is a constant.
- **Lemma 1.7 Log-sum inequality:** For non-negative real values $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$ and

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq (\sum_{i=1}^{n} a_i) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}, \tag{25}$$

  and the equality holds if and only if $\forall i = 1, 2, \ldots n,\ \frac{a_i}{b_i}$ equals to a constant.

Then Convexity of K-L Divergence

- **Theorem 1.8** $D(\mathbf{p}\|\mathbf{q})$ is convex over $(\mathbf{p},\mathbf{q})$, that is, for pmf $(\mathbf{p}_1,\mathbf{q}_1)$ and $(\mathbf{p}_2,\mathbf{q}_2)$,

  $$D(\lambda\mathbf{p}_1+(1-\lambda)\mathbf{p}_2\|\lambda\mathbf{q}_1+(1-\lambda)\mathbf{q}_2) \le \lambda D(\mathbf{p}_1\|\mathbf{q}_1)+(1-\lambda)D(\mathbf{p}_2\|\mathbf{q}_2) \tag{26}$$

  for all $0 \le \lambda \le 1$.
- **Theorem 1.9** Entropy $H(X)=H(\mathbf{p})$ is concave over $\mathbf{p}$.
- **Theorem 1.10** Mutual information $I(X;Y)=I(\mathbf{p},\mathbf{Q})$ is concave over $\mathbf{p}$ and convex over channel transition matrix $\mathbf{Q}$, respectively.

Fano's inequality and Estimation

- **Theorem 1.11 Fano's inequality:** For any estimator $\hat{X}$ such that $X \longrightarrow Y \longrightarrow \hat{X}$ with $P_e = \Pr(X \neq \hat{X})$, we have

$$H(P_e) + P_e \log |\mathcal{X}| \geq H(X|\hat{X}) \geq H(X|Y). \tag{27}$$

- **Corollary 1.12** $\forall X, Y$ and let $p = \Pr(X \neq Y)$,

$$H(p) + p \log |\mathcal{X}| \geq H(X|Y). \tag{28}$$

- **Corollary 1.13** If $\hat{X} = Y$, Fano's inequality can be strengthened as

$$H(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(X|Y). \tag{29}$$

Outline

1. Entropy

2. Mutual Information and K-L Divergence

3. Optimizing over the Measurements

4. Generalize to Continuous RVs
   - Differential Entropy
   - Properties of Differential Entropy
   - Mutual Information for Continuous RVs

5. Summary and Reference

The definition of Differential entropy

- **Definition 1.14** The differential entropy $h(X)$ of a continuous random variable X is

$$h(X) = -\int_{x \in S} f(x) \log f(x) dx, \tag{30}$$

where $f(x)$ is the p.d.f (probability density function) and S is the support set.

- For variables X and Y, we have
  - joint differential entropy $h(X,Y) = -\int\int_{x,y} f(x,y) \log f(x,y) dxdy$,
  - conditional differential entropy $h(X|Y) = -\int\int_{x,y} f(x,y) \log f(x|y) dxdy$,
  - $h(X,Y) = h(X) + h(Y|X) = h(Y) + h(X|Y)$,
    $h(X|Y) \leq h(X)$ and $h(X,Y) \leq h(X) + h(Y)$.
  - Note that $h(X)$ is not necessarily positive.

Properties of Differential Entropy

- **Theorem 1.15** The transform of $h(X)$ has

$$h(aX) = h(X) + \log|a|. \tag{31}$$

- The differential entropy of a gaussian random variable $X \sim \mathcal{N}(m, \sigma^2)$ is

$$h(X) = \frac{1}{2}\log 2\pi e\sigma^2. \tag{32}$$

- **Theorem 1.16** Let the random variable X have variance σ^2, then

$$h(X) \leq \frac{1}{2}\log 2\pi e\sigma^2 \tag{33}$$

with equality if and only if $X \sim \mathcal{N}(m, \sigma^2)$.

Mutual Information for Continuous RVs

- **Definition 1.15** The mutual information between continuous random variables X and Y is defined as

$$I(X;Y) = \int\int_{x,y} f(x,y) \log \frac{f(x,y)}{f(x)f(y)} dxdy, \tag{34}$$

where $f(x,y)$ and $f(x)$ are joint p.d.f and marginal p.d.f, respectively.

Outline

1. Entropy

2. Mutual Information and K-L Divergence

3. Optimizing over the Measurements

4. Generalize to Continuous RVs

5. Summary and Reference
   Ideas in a nutshell

Ideas in a nutshell

- The definitions of Entropy, Mutual Information and K-L Divergence.
- The relations between these measurements
- Jensen's inequality and log-sum inequality.
- The convexity/concavity of entropy, mutual information and relative entropy
- Fano's inequality
- Differential entropy and mutual information for continuous random variables.

The whole story revisited

- The K-L Divergence is fundamental.
- K-L Divergence induces Shannon Entropy and Mutual Information.
- Convexity and concavity enables global optimizability.
- A trailer for three main results of Shannon theory.

Reference

- Lin Zhang, Lecture notes on Fundamentals of applied information theory, 2014-spring, in Chinese.
- Claude E. Shannon: A Mathematical Theory of Communication, Bell System Technical Journal, 1948.
- Cover T M, Thomas J A. Elements of information theory[M]. John Wiley and Sons, 2012.
- Xuelong Zhu, Fundamentals of applied information theory, Tsinghua Univ. Press, 2001, in Chinese.