Application of Information Theory, Lecture 2
Joint & Conditional Entropy, Mutual Information

Handout Mode

Iftach Haitner

Tel Aviv University.

Nov 4, 2014
Part I

Joint and Conditional Entropy
Joint entropy

- Recall that the entropy of rv $X$ over $\mathcal{X}$, is defined by
  \[
  H(X) = - \sum_{x \in \mathcal{X}} P_X(x) \log P_X(x)
  \]
- Shorter notation: for $X \sim p$, let $H(X) = - \sum x p(x) \log p(x)$ (where the summation is over the domain of $X$).
- The joint entropy of (jointly distributed) rvs $X$ and $Y$ with $(X, Y) \sim p$, is
  \[
  H(X, Y) = - \sum_{x,y} p(x, y) \log p(x, y)
  \]
  This is simply the entropy of the rv $Z = (X, Y)$.
- Example:

<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y$</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>$\frac{1}{4}$</td>
<td>$\frac{1}{4}$</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>$\frac{1}{2}$</td>
<td>0</td>
</tr>
</tbody>
</table>

\[
H(X, Y) = - \frac{1}{2} \log \frac{1}{2} - \frac{1}{4} \log \frac{1}{4} - \frac{1}{4} \log \frac{1}{4}
\]

\[
= \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 2 = 1 \frac{1}{2}
\]
The joint entropy of \((X_1, \ldots, X_n) \sim p\), is

\[
H(X_1, \ldots, X_n) = - \sum_{x_1, \ldots, x_n} p(x_1, \ldots, x_n) \log p(x_1, \ldots, x_n)
\]
Conditional entropy

- Let \((X, Y) \sim p\).
- For \(x \in \text{Supp}(X)\), the random variable \(Y|X = x\) is well defined.
- The entropy of \(Y\) conditioned on \(X\), is defined by

\[
H(Y|X) := \mathbb{E}_{x \leftarrow X} H(Y|X = x) = \mathbb{E}_X H(Y|X)
\]

- Measures the uncertainty in \(Y\) given \(X\).
- Let \(p_X\) & \(p_{Y|X}\) be the marginal & conditional distributions induced by \(p\).

\[
H(Y|X) = \sum_{x \in \mathcal{X}} p_X(x) \cdot H(Y|X = x) = - \sum_{x \in \mathcal{X}} p_X(x) \sum_{y \in \mathcal{Y}} p_{Y|X}(y|x) \log p_{Y|X}(y|x)
\]

\[
= - \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x, y) \log p_{Y|X}(y|x)
\]

\[
= - \mathbb{E}_{(X, Y)} \log p_{Y|X}(Y|X)
\]

\[
= - \mathbb{E}_{Z=p_{Y|X}(Y|X)} \log Z
\]
Example

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1/4</td>
<td>1/4</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1/2</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

What is $H(Y|X)$ and $H(X|Y)$?

$$H(Y|X) = \mathbb{E}_{x \leftarrow X} H(Y|X = x)$$

$$= \frac{1}{2} H(Y|X = 0) + \frac{1}{2} H(Y|X = 1)$$

$$= \frac{1}{2} H(\frac{1}{2}, \frac{1}{2}) + \frac{1}{2} H(1, 0) = \frac{1}{2}.$$ 

$$H(X|Y) = \mathbb{E}_{y \leftarrow Y} H(X|Y = y)$$

$$= \frac{3}{4} H(X|Y = 0) + \frac{1}{4} H(X|Y = 1)$$

$$= \frac{3}{4} H(\frac{1}{3}, \frac{2}{3}) + \frac{1}{4} H(1, 0) = 0.6887 \neq H(Y|X).$$
Conditional entropy, cont..

\[
H(X|Y,Z) = \mathbb{E}_{(y,z)\leftarrow(Y,Z)} H(X|Y=y, Z=z) \\
= \mathbb{E}_{y\leftarrow Y} \mathbb{E}_{z\leftarrow Z|Y=y} H(X|Y=y, Z=z) \\
= \mathbb{E}_{y\leftarrow Y} \mathbb{E}_{z\leftarrow Z|Y=y} H((X|Y=y)|Z=z) \\
= \mathbb{E}_{y\leftarrow Y} H(X_y|Z_y)
\]

for \((X_y, Z_y) = (X, Z)|Y = y\)
Relating mutual entropy to conditional entropy

- What is the relation between $H(X)$, $H(Y)$, $H(X,Y)$ and $H(Y|X)$?
- Intuitively, $0 \leq H(Y|X) \leq H(Y)$
  
  Non-negativity is immediate. We prove upperbound later.
- $H(Y|X) = H(Y)$ iff $X$ and $Y$ are independent.
- In our example, $H(Y) = H(\frac{3}{4}, \frac{1}{4}) > \frac{1}{2} = H(Y|X)$
- Note that $H(Y|X = x)$ might be larger than $H(Y)$ for some $x \in \text{Supp}(X)$.
- Chain rule (proved next). $H(X,Y) = H(X) + H(Y|X)$
- Intuitively, uncertainty in $(X,Y)$ is the uncertainty in $X$ plus the uncertainty in $Y$ given $X$.
- $H(Y|X) = H(X,Y) - H(X)$ is as an alternative definition for $H(Y|X)$.
Chain rule (for the entropy function)

Claim 1
For rvs $X, Y$, it holds that $H(X, Y) = H(X) + H(Y|X)$.

- Proof immediately follow by the grouping axiom:

  Let $q_i = \sum_{j=1}^{n} p_{i,j}$

  
  $H(P_{1,1}, \ldots, P_{n,n})$
  
  $= H(q_1, \ldots, q_n) + \sum q_i H(\frac{P_{i,1}}{q_i}, \ldots, \frac{P_{i,n}}{q_i})$
  
  $= H(X) + H(Y|X)$.

- Another proof. Let $(X, Y) \sim p$.

  - $p(x, y) = p_X(x) \cdot p_{Y|X}(x|y)$.

  $\Rightarrow$ log $p(x, y) = \log p_X(x) + \log p_{Y|X}(x|y)$
  
  $\Rightarrow$ E log $p(X, Y) = E \log p_X(X) + E \log p_{Y|X}(Y|X)$
  
  $\Rightarrow$ $H(X, Y) = H(X) + H(Y|X)$. 
$H(Y|X) \leq H(Y)$

Jensen inequality: for any concave function $f$, values $t_1, \ldots, t_k$ and $\lambda_1, \ldots, \lambda_k \in [0, 1]$ with $\sum_i \lambda_i = 1$, it holds that $\sum_i \lambda_i f(t_i) \leq f(\sum_i \lambda_i t_i)$.

Let $(X, Y) \sim p$. 

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

$= \sum_{x, y} p(x, y) \log \frac{p_X(x)}{p(x, y)}$

$= \sum_{x, y} p_Y(y) \cdot \frac{p(x, y)}{p_Y(y)} \log \frac{p_X(x)}{p(x, y)}$

$= \sum_{y} p_Y(y) \sum_{x} \frac{p(x, y)}{p_Y(y)} \log \frac{p_X(x)}{p(x, y)}$

$\leq \sum_{y} p_Y(y) \log \sum_{x} \frac{p(x, y)}{p_Y(y)} \frac{p_X(x)}{p(x, y)}$

$= \sum_{y} p_Y(y) \log \frac{1}{p_Y(y)} = H(Y).$
$H(Y|X) \leq H(Y)$ cont.

- Assume $X$ and $Y$ are independent (i.e., $p(x, y) = p_X(x) \cdot p_Y(y)$ for any $x, y$)

$$\implies p_{Y|X} = p_Y$$

$$\implies H(Y|X) = H(Y)$$
Other inequalities

- \( H(X), H(Y) \leq H(X, Y) \leq H(X) + H(Y) \).
  
  Follows from \( H(X, Y) = H(X) + H(Y|X) \).
  
  - Left inequality since \( H(Y|X) \) is non negative.
  - Right inequality since \( H(Y|X) \leq H(Y) \).

- \( H(X, Y|Z) = H(X|Z) + H(Y|X, Z) \) (by chain rule)

- \( H(X|Y, Z) \leq H(X|Y) \)

Proof:

\[
H(X|Y, Z) = \mathbb{E}_{Z, Y} H(X \mid Y, Z) \\
= \mathbb{E}_{Z, Y} \mathbb{E}_{Y} H(X \mid Y, Z) \\
= \mathbb{E}_{Z, Y} \mathbb{E}_{Y} H((X \mid Y) \mid Z) \\
\leq \mathbb{E}_{Z, Y} \mathbb{E}_{Y} H(X \mid Y) \\
= \mathbb{E}_{Y} H(X \mid Y) \\
= H(X|Y).
\]
Chain rule (for the entropy function), general case

Claim 2

For rvs $X_1, \ldots, X_k$, it holds that
\[ H(X_1, \ldots, X_k) = H(X_i) + H(X_2 | X_1) + \ldots + H(X_k | X_1, \ldots, X_{k-1}). \]

Proof: ?

- Extremely useful property!
- Analogously to the two variables case, it also holds that:
  - $H(X_i) \leq H(X_1, \ldots, X_k) \leq \sum_i H(X_i)$
  - $H(X_1, \ldots, X_k | Y) \leq \sum_i H(X_i | Y)$
Examples

- (from last class) Let $X_1, \ldots, X_n$ be Boolean iid with $X_i \sim (\frac{1}{3}, \frac{2}{3})$. Compute $H(X_1, \ldots, X_n)$

- As above, but under the condition that $\bigoplus_i X_i = 0$?
  - Via chain rule?
  - Via mapping?
Applications

- Let $X_1, \ldots, X_n$ be Boolean i.i.d. with $X_i \sim (p, 1-p)$ and let $X = X_1, \ldots, X_n$. Let $f$ be such that $\Pr[f(X) = z] = \Pr[f(X) = z']$, for every $k \in \mathbb{N}$ and $z, z' \in \{0, 1\}^k$. Let $K = |f(X)|$.
  
  Prove that $E K \leq n \cdot h(p)$.

- Interpretation

- Positive results
How many comparisons it takes to sort $n$ elements?

Let $A$ be a sorter for $n$ elements algorithm making $t$ comparisons. What can we say about $t$?

Let $X$ be a uniform random permutation of $[n]$ and let $Y_1, \ldots, Y_t$ be the answers $A$ gets when sorting $X$.

$X$ is determined by $Y_1, \ldots, Y_t$.

Namely, $X = f(Y_1, \ldots, Y_t)$ for some function $f$.

$H(X) = \log n!$

\[
H(X) = H(f(Y_1, \ldots, Y_n)) \\ \leq H(Y_1, \ldots, Y_n) \\ \leq \sum_i H(Y_i) \\ = t
\]

$\implies t \geq \log n! = \Theta(n \log n)$
Concavity of entropy function

Let $p = (p_1, \ldots, p_n)$ and $q = (q_1, \ldots, q_n)$ be two distributions, and for $\lambda \in [0, 1]$ consider the distribution $\tau_\lambda = \lambda p + (1 - \lambda) q$.
(i.e., $\tau_\lambda = (\lambda p_1 + (1 - \lambda) q_1, \ldots, \lambda p_n + (1 - \lambda) q_n)$).

Claim 3

$$H(\tau_\lambda) \geq \lambda H(p) + (1 - \lambda) H(q)$$

Proof:

- Let $Y$ over $\{0, 1\}$ be $1$ wp $\lambda$
- Let $X$ be distributed according to $p$ if $Y = 0$ and according to $q$ otherwise.
- $H(\tau_\lambda) = H(X) \geq H(X \mid Y) = \lambda H(p) + (1 - \lambda) H(q)$

We are now certain that we drew the graph of the (two-dimensional) entropy function right...
Part II

Mutual Information
Mutual information

- $I(X; Y)$ — the “information" that $X$ gives on $Y$

\[
I(X; Y) := H(Y) - H(Y|X)
= H(Y) - (H(X, Y) - H(X))
= H(X) + H(Y) - H(X, y)
= I(Y; X).
\]

- The mutual information that $X$ gives about $Y$ equals the mutual information that $Y$ gives about $X$.

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

- $I(X; f(X)) = H(f(X))$ (and smaller then $H(X)$ is $f$ is non-injective)

- $I(X; Y, Z) \geq I(X; Y), I(X; Z)$ (since $H(X | Y, Z) \leq H(X | Y), H(X | Z)$)

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

- $I(X; Y|Z) = I(Y; X|Z)$ (since $I(X'; Y') = I(Y'; X')$)
Numerical example

Example

<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y$</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td>$\frac{1}{4}$</td>
<td>$\frac{1}{4}$</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>$\frac{1}{2}$</td>
<td>0</td>
</tr>
</tbody>
</table>

\[ I(X; Y) = H(X) - H(X|Y) \]
\[ = 1 - \frac{3}{4} \cdot h\left(\frac{1}{3}\right) \]
\[ = I(Y; X) \]
\[ = H(Y) - H(Y|X) \]
\[ = h\left(\frac{1}{4}\right) - \frac{1}{2} h\left(\frac{1}{2}\right) \]
Claim 4 (Chain rule for mutual information)

For rvs $X_1, \ldots, X_k, Y$, it holds that

$$I(X_1, \ldots, X_k; Y) = I(X; Y) + I(X_2; Y|X_1) + \ldots + I(X_k; Y|X_1, \ldots, X_{k-1}).$$

Proof: ? HW
Examples

Let $X_1, \ldots, X_n$ be iid with $X_i \sim (p, 1 - p)$, under the condition that $\bigoplus_i X_i = 0$. Compute $I(X_1, \ldots, X_{n-1}; X_n)$.

By chain rule

$$I(X_1, \ldots, X_{n-1}; X_n) = H(X_1; X_n) + H(X_2; X_n|X_1) + \ldots + H(X_{n-1}; X_n|X_1, \ldots, X_{n-2}) = 0 + 0 + \ldots + 1 = 1.$$

Let $T$ and $F$ be the top and front side, respectively, of a 6-sided fair dice. Compute $I(T; F)$.

$$I(T; F) = H(T) - H(T|F) = \log 6 - \log 4 = \log 3 - 1.$$
Part III

Data processing
Data processing Inequality

**Definition 5 (Markov Chain)**

Rvs \((X, Y, Z) \sim p\) form a **Markov chain**, denoted \(X \rightarrow Y \rightarrow Z\), if

\[
p(x, y, z) = p_X(x) \cdot p_{Y|X}(y|x) \cdot p_{Z|Y}(z|y),
\]

for all \(x, y, z\).

Example: random walk on graph.

**Claim 6**

If \(X \rightarrow Y \rightarrow Z\), then \(I(X; Y) \geq I(X; Z)\).

- By Chain rule, \(I(X; Y, Z) = I(X; Z) + I(X; Y|Z) = I(X; Y) + I(X; Z|Y)\).
- \(I(X; Z|Y) = 0\)
  - \(p_{Z|Y=y} = p_{Z|Y=y,X=x}\) for any \(x, y\)
  - \[I(X; Z|Y) = H(Z|Y) - H(Z|Y, X)\]
    \[= \mathbb{E}_Y H(p_{Z|Y=y}) - \mathbb{E}_{Y,X} H(p_{Z|Y=y,X=x})\]
    \[= \mathbb{E}_Y H(p_{Z|Y=y}) - \mathbb{E}_Y H(p_{Z|Y=y}) = 0.\]
- Since \(I(X; Y|Z) \geq 0\), we conclude \(I(X; Y) \geq I(X; Z)\).
Fano’s Inequality

- How well can we guess $X$ from $Y$?
- Could with no error if $H(X|Y) = 0$. What if $H(X|Y)$ is small?

Theorem 7 (Fano’s inequality)

For any rvs $X$ and $Y$, and any (even random) $g$, it holds that

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

for $\hat{X} = g(Y)$ and $P_e = \Pr[\hat{X} \neq X]$.

- Note that $P_e = 0$ implies that $H(X|Y) = 0$
- The inequality can be weekend to $1 + P_e \log |\mathcal{X}| \geq H(X|Y)$,
- Alternatively, to $P_e \geq \frac{H(X|Y) - 1}{\log |\mathcal{X}|}$
- Intuition for $\propto \frac{1}{\log |\mathcal{X}|}$
- We call $\hat{X}$ an estimator for $X$ (from $Y$).
Proving Fano’s inequality

Let $X$ and $Y$ be rvs, let $\hat{X} = g(Y)$ and $P_e = \Pr[\hat{X} \neq X]$.

- Let $E = \begin{cases} 1, & \hat{X} \neq X \\ 0, & \hat{X} = X \end{cases}$.

\[
H(E, X|\hat{X}) = H(X|\hat{X}) + H(E|X, \hat{X}) = H(E|\hat{X}) + H(X|E, \hat{X}) \leq H(E)=h(P_e) \leq P_e \log |\mathcal{X}|(?)
\]

- It follows that $h(P_e) + P_e \log |\mathcal{X}| \geq H(X|\hat{X})$

- Since $X \rightarrow Y \rightarrow \hat{X}$, it holds that $I(X; Y) \geq I(X; \hat{X}) \implies H(X|\hat{X}) \geq H(X|Y)$