COMP2610 / COMP6261 - Information Theory

Lecture 6: Entropy

Australian National University

7 August 2018

Last time

- The Bernoulli and Binomial distributions

- Ma

- Bayesian parameter etimation

This time

- Information content and entropy
- Exa

- Some basic properties of entropy

Outline

1. Information Content & Entropy
   - Entropy of a Random Variable
   - Some Basic Properties

2. Exa
   - M

3. Entro
   - Average Code Length
   - Minimum Number of Binary Questions

4. Joint Entropy, Conditional Entropy and Chain

5. An Axiomatic Characterisation

6. Wrapping up

Recap: A General Communication System

How informative is a message?

Information Content: Informally

Say that a message comprises a single bit (one binary random variable)

- Whether or not a coin comes up heads
- Whether or not my favourite horse wins a race

Informall

- Ho
  -
    much more informative than “Heads”
  - If I believe my favourite horse will win with 99.9
    surprising to find out it did not

Information Content: Informally

Say that a message comprises a single bit (one binary random variable)

- Whether or not a coin comes up heads
- Whether or not my favourite horse wins a race

Informall

- Ho

  - much more informative than “Heads”
  - If I believe my favourite horse will win with 99.9
    surprising to find out it did not

- How predictable a random variable is
  - If a coin comes up Heads 99.99% of the time, we can predict the next message as “Heads” and be right most of the time
  - If I believe my favourite horse will win with 99.99% probability, then I believe predicting so to be right most of the time

Information Content: Formally

Intuitively, we measure information of a message in relation to the other messages we could have seen

- For binary messages, we either see 0 or 1
- The hav

How can we *formalise* and thus *measure* i

- Information content of an outcome must dep

- Information content of a random variable must depend on its probability distribution

Information Content of an Outcome: Definition

Let X be a discrete r.v. with possible outcomes $\mathcal{X}$

- e.g. $\mathcal{X} = \{0, 1\}$
- e.g.

Let $p(x)$

The information content of an outcome x

$$h(x) = \log_2 \frac{}{p(x)}$$

Information Content of an Outcome: Properties

The information content of x grows as $p(x)$ shrinks

- Out

Choice of l

- If we use $\log_2$ we measure information in *bits*

What about other functions of $p(x)$, e.g. $\frac{}{p(}$

Entropy of a Random Variable: Definition

Let X be a discrete r.v. with possible outcomes $\mathcal{X}$.

The entropy of the random variable X is the average information content of the outcomes:

$$= \quad (\ ) \cdot \log_2 \text{——}$$

$$x$$

$$= -\sum_x p(x)$$

where we define $0 \log 0 \equiv 0$, as $\lim_{p \to 0} p \log p = 0$.

Entropy of a Random Variable

Some Basic Properties

- Non-negativity:

$$0 \leq p(x) \leq 1 \Rightarrow \log \frac{1}{p(x)} \geq 0$$

$$p(x) \log \frac{1}{} \quad 0$$

Entropy of a Random Variable

Some Basic Properties

- Non-negativity:

$$0 \leq p(x) \leq 1 \Rightarrow \log \frac{1}{p(x)} \geq 0$$

$$p(x) \log \frac{1}{\quad} \quad 0$$

- Change of base:

$$H_b(X) = -\sum_x p(x$$

$$= \sum_x p(x) \log$$

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

  - If we use $\log_2$ the units are called *bits*
  - If we use natural logarithm the units are called *nats*

Unrolling the Definition

The entropy of X is

$$H(X) = -\sum_x p(x) \log_2 p(x).$$

Pick a rand

- Ave

Does not depend on the values of the outcomes

- Only on their probabilities

- Contrast with expectation $\mathbb{E}[X] = \sum_x x \cdot p(X = x)$.

What Does Entropy “Mean”?

Not a well posed question.

Entropy does match some intuitive properties of our informal notion of “information content”

- Rar

But other f

$$G(X) = \sum_x p(x) \frac{1}{p(x)^2}$$

We will see some examples where our definition of entropy arises naturally
The main justification is the results we can obtain with it.

Entropy of a Random Variable
Example 1 — Bernoulli Distribution

Let $X \in \{0, 1\}$ with $X \sim \text{Bern}(X|\theta)$

Then,

So, the entropy of a Bernoulli random variable is

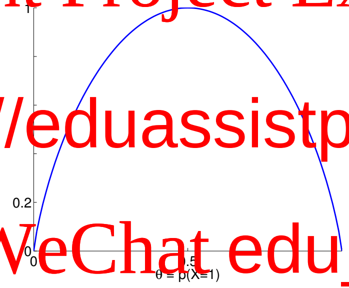
$$
\begin{aligned}
H(X) &= -\sum_{x\in\{0,1\}} p(x) \cdot \\
&= -\theta \log \theta - (1-\theta)\log(1-\theta)
\end{aligned}
$$

Entropy of a Random Variable

Example 1 — Bernoulli Distribution

Let $X \in \{0, 1\}$ with $X \sim \text{Bern}(X|\theta)$ and $\theta = p(X = 1)$

$$H(X) = -\theta \log \theta - (1 - \theta) \log(1 - \theta)$$

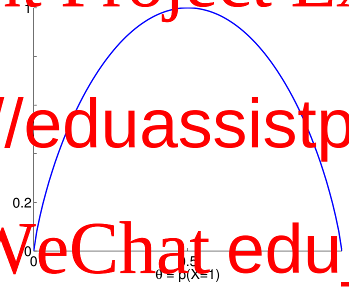


Entropy of a Random Variable

Example 1 — Bernoulli Distribution

Let $X \in \{0, 1\}$ with $X \sim \text{Bern}(X|\theta)$ and $\theta = p(X = 1)$

$$H(X) = -\theta \log \theta - (1 - \theta) \log(1 - \theta)$$

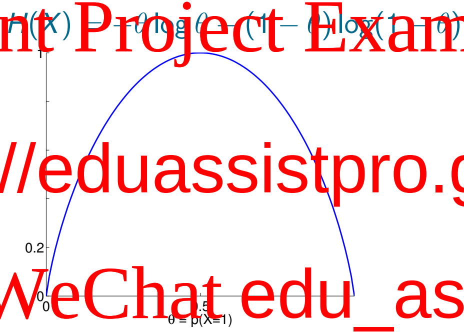


- *Concave* function of the distribution

Entropy of a Random Variable

Example 1 — Bernoulli Distribution

Let $X \in \{0, 1\}$ with $X \sim \text{Bern}(X|\theta)$ and $\theta = p(X = 1)$



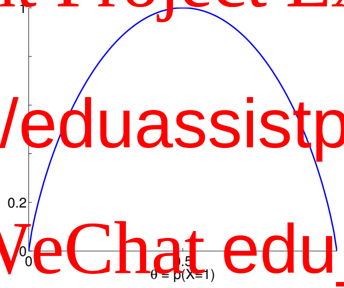
- *Concave* function of the distribution
- Minimum entropy $\rightarrow$ no uncertainty about X, i.e. $\theta = 1$ or $\theta = 0$

Entropy of a Random Variable

Example 1 — Bernoulli Distribution

Let $X \in \{0, 1\}$ with $X \sim \text{Bern}(X|\theta)$ and $\theta = p(X = 1)$

$$H(X) = -\theta \log \theta - (1-\theta) \log(1-\theta)$$



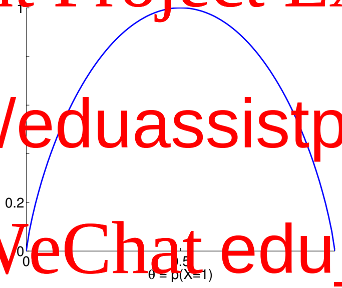
- *Concave* function of the distribution
- Minimum entropy $\rightarrow$ no uncertainty about X, i.e. $\theta = 1$ or $\theta = 0$
- Maximum when $\rightarrow$ complete uncertainty about X, i.e. $\theta = 0.5$

Entropy of a Random Variable

Example 1 — Bernoulli Distribution

Let $X \in \{0, 1\}$ with $X \sim \text{Bern}(X|\theta)$ and $\theta = p(X = 1)$

$$H(X) = -\theta \log \theta - (1 - \theta) \log(1 - \theta)$$



- *Concave* function of the distribution
- Minimum entropy $\rightarrow$ no uncertainty about X, i.e. $\theta = 1$ or $\theta = 0$
- Maximum when $\rightarrow$ complete uncertainty about X, i.e. $\theta = 0.5$
- For $\theta = 0.5$ (e.g. a fair coin) $H_2(X) = 1$ bit.

Entropy of a Random Variable

Example 2

Conside mes:

The entro

$$H(X) = -\sum_{i=1} p(i)\log_2 p(i) = -\sum_{i=1} \frac{}{3} \quad \text{—} \quad \text{its.}$$

Categorical distributions with 30 different states:

Figure from Bishop, PRML, 2006)

Entropy of a Random Variable

Example 3 — Categorical Distribution

Categorical distributions with 30 different states:

Figure from Bishop, PRML, 2006)

- The more sharply peaked the lower the entropy

Entropy of a Random Variable

Example 3 — Categorical Distribution

Categorical distributions with 30 different states:

Figure from Bishop, PRML, 2006)

- The more sharply peaked the lower the entropy
- The more evenly spread the higher the entropy

Entropy of a Random Variable

Example 3 — Categorical Distribution

Categorical distributions with 30 different states:

Figure from Bishop, PRML, 2006)

- The more sharply peaked the lower the entropy
- The more evenly spread the higher the entropy
- Maximum for *uniform* distribution: $H(X) = -\log \frac{1}{30} \approx 3.40$ nats
  - When will the entropy be minimum?

Consider a discrete variable X taking on values from the set $\mathcal{X}$

- Let p_i be the probability of each state, with $i = 1, \ldots, |\mathcal{X}|$

Entropy of a Random Variable

Maximum Entropy

Consider a discrete variable X taking on values from the set $\mathcal{X}$

- Let p_i be the probability of each state, with $i = 1, \ldots, |\mathcal{X}|$
- De

Consider a discrete variable X taking on values from the set $\mathcal{X}$

- Let p_i be the probability of each state, with $i = 1, \ldots, |\mathcal{X}|$
- De

The entro

$$H(X) \leq \log_2$$

with equality iff $p_i = \frac{1}{|\mathcal{X}|}$ for all i

Note $\log_2 |\mathcal{X}|$ is the number of bits needed to describe an outcome of X

1. Information Content & Entropy
   - Entropy of a Random Variable
   - Some Basic Properties
2. Examples: Bernoulli and Categorical Random Variables
   - Maximum Entropy
3. Entro
   - Av
   - Mi
4. Joint Entropy, Conditional Entropy and Chain
5. An Axiomatic Characterisation
6. Wrapping up

Entropy of a Random Variable

Example 4 (from Cover & Thomas, 2006) — 1 of 3

Consider a horse race with 8 horses participating:

$$\{acer, babe, cactus, daisy, epic, fancy, gem, hairy\}$$

Entropy of a Random Variable

Example 4 (from Cover & Thomas, 2006) — 1 of 3

Consider a horse race with 8 horses participating:

$$\{acer, babe, cactus, daisy, epic, fancy, gem, hairy\}$$

- Eac
  tran ..., 111

Entropy of a Random Variable

Example 4 (from Cover & Thomas, 2006) — 1 of 3

Consider a horse race with 8 horses participating:

$$\{acer, babe, cactus, daisy, epic, fancy, gem, hairy\}$$

- Eac
  tran ..., 111

  Not X, is:

$$H(X) = 8 \times \frac{1}{8} \text{lo}$$

Entropy of a Random Variable

Example 4 (from Cover & Thomas, 2006) — 1 of 3

Consider a horse race with 8 horses participating:

$$\{acer, babe, cactus, daisy, epic, fancy, gem, hairy\}$$

- Eac
  tran ..., 111
  Not X, is:

$$H(X) = 8 \times \frac{1}{8} \text{lo}$$

- Now say that the probabilities of each horse wi

$$\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64} \right)$$

Entropy of a Random Variable

Example 4 (from Cover & Thomas, 2006) — 1 of 3

Consider a horse race with 8 horses participating:

$$\{acer, babe, cactus, daisy, epic, fancy, gem, hairy\}$$

- Eac
  tran ..., 111

  Not X, is:

$$H(X) = 8 \times \frac{1}{8} \text{lo}$$

- Now say that the probabilities of each horse wi

$$\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right)$$

  What is the average code-length to transmit the identity of the winning horse?

Entropy of a Random Variable

Example 4 (from Cover & Thomas, 2006) — 2 of 3

We see that some horses have higher probability of winning:

- We can still use a 3-bit representation
  - However, this would be wasteful as some horses are more likely to win

Entropy of a Random Variable

Example 4 (from Cover & Thomas, 2006) — 2 of 3

We see that some horses have higher probability of winning:

- We can still use a 3-bit representation
  - However, this would be wasteful as some horses are more likely to win
- Idea: Use shorter codes for most probable horses and longer codes for th

Entropy of a Random Variable

Example 4 (from Cover & Thomas, 2006) — 2 of 3

We see that some horses have higher probability of winning:

- We can still use a 3-bit representation
  - However, this would be wasteful as some horses are more likely to win

- Idea: Use shorter codes for most probable horses and longer codes for th
- Let u

Decode 010 into ’aba’ or ’ac’? Ambiguous.

Entropy of a Random Variable

Example 4 (from Cover & Thomas, 2006) — 2 of 3

We see that some horses have higher probability of winning:

- We can still use a 3-bit representation
  - However, this would be wasteful as some horses are more likely to win

- Idea: Use shorter codes for most probable horses and longer codes for th
- Let u

  Decode 010 into ’aba’ or ’ac’? Ambiguous.

- We should be able to disambiguate a concate
  into the corresponding components.

Entropy of a Random Variable

Example 4 (from Cover & Thomas, 2006) — 2 of 3

We see that some horses have higher probability of winning:

- We can still use a 3-bit representation
  - However, this would be wasteful as some horses are more likely to win
- Idea: Use shorter codes for most probable horses and longer codes for th
- Let u

  Decode 010 into 'aba' or 'ac'? Ambiguous.
- We should be able to disambiguate a concate
  into the corresponding components.
- Represent the horses (states) using the following codes:

$$\{0, 10, 110, 1110, 111100, 111101, 111110, 111111\}$$

  - E.g. 11001110 $\rightarrow$??

What is the average code length that has to be transmitted?

What is the average code length that has to be transmitted?

Average $\frac{1}{\;} \quad \frac{1}{\;} \quad \frac{1}{\;} \quad \frac{1}{\;} \quad \frac{1}{\;} \times 6 = 2$ bits

Entropy of a Random Variable

Example 4 (from Cover & Thomas, 2006) — 3 of 3

What is the average code length that has to be transmitted?

Average $\frac{1}{\ }$ $\frac{1}{\ }$ $\frac{1}{\ }$ $\frac{1}{\ }$ $\frac{1}{\ } \times 6 = 2$ bits

What is the

Entropy of a Random Variable

Example 4 (from Cover & Thomas, 2006) — 3 of 3

What is the average code length that has to be transmitted?

Average $\frac{1}{\;} \quad \frac{1}{\;} \quad \frac{1}{\;} \quad \frac{1}{\;} \quad \frac{1}{\;} \times 6 = 2$ bits

What is the

$$H(X) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{4}\log_2\frac{1}{4} + \frac{1}{8}\log_2 - \quad \cdots \quad {}_2\frac{1}{64}\right)$$

$$= 2 \text{ bits}$$

Entropy of a Random Variable:

Example 5 (from Cover & Thomas, 2006)

Let $X \in \{1, 2, 3\}$ and $p(X = 1) = p(X = 2) = p(X = 3) = \frac{1}{3}$

Given the corresponding codeword:

Then $H(X) = 1.58$, and average code length = 1

Entropy of a Random Variable:

Example 5 (from Cover & Thomas, 2006)

Let $X \in \{1, 2, 3\}$ and $p(X = 1) = p(X = 2) = p(X = 3) = \frac{1}{3}$

Given the corresponding codeword:

Then $H(X) = 1.58$, and average code length = 1

In general, Entropy is a lower bound on the average n
transmit the state of a random variable.

As we shall see later, we can construct descriptors with average length within 1 bit of the entropy.

Entropy of a Random Variable:

What Questions Should We Ask? (From Cover & Thomas, 2006)

Assume that only the following horses participated in the last race: {*acer*, *babe*, *cactus*, *daisy*}.

The corre

$p(X$ … $d) = \frac{1}{8}$.

You want to determine which horse won the race wit
number of yes/no questions:

(a) What binary questions should you ask?

(b) What is the minimum expected number of binary questions for this?

Entropy of a Random Variable:

What Questions Should We Ask? (From Cover & Thomas, 2006) — Cont'd

As acer is more likely to have won the race I first ask about him: has $X = a$ won the race?

If the answ

$X = b$ w

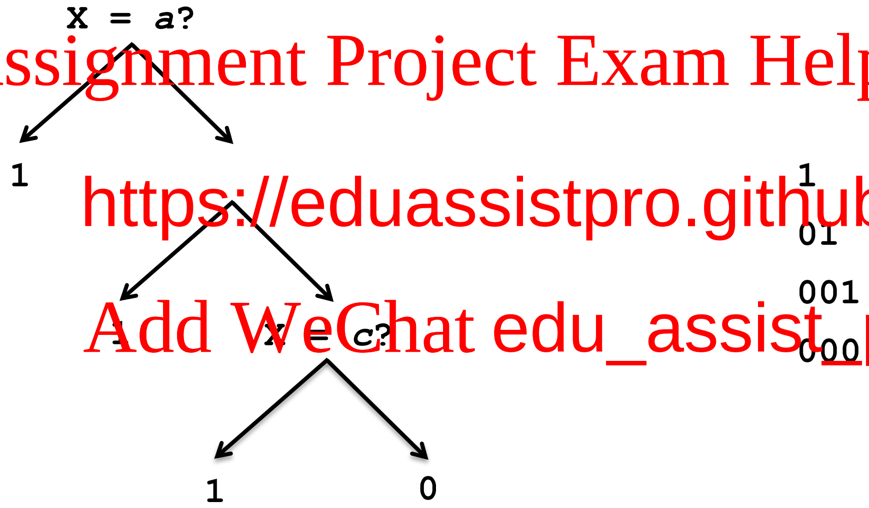
Then $X = c$?, and $X = d$?

Note that the series of questions corresponding to a

seen as a code!

Entropy of a Random Variable:

What Questions Should We Ask? (From Cover & Thomas, 2006) — Cont'd



Entropy of a Random Variable:

What Questions Should We Ask? (From Cover & Thomas, 2006) — Cont'd

The entropy of this random variable determines a lower bound for the minimum number of binary questions:

$$H_2(X) \quad \cdots \quad = 1.75 \text{ bits.}$$

This is in fact the minimum expected number of binar
general, this number lies between $H(X)$ a

Intuitively, each question reduces our amount of u
outcome by attempting to eliminate (or validate) the hard to predict outcomes

Joint Entropy

The joint entropy $H(X, Y)$ of a pair of discrete random variables with joint distribution $p(X, Y)$ is given by:

$$= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,$$

Joint Entropy:

Independent Random Variables

If X and Y are statistically independent we have that:

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

Joint Entropy:

Independent Random Variables

If X and Y are statistically independent we have that:

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

$$= p(x)p(y)$$

Joint Entropy:

Independent Random Variables

If X and Y are statistically independent we have that:

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

$$= p(x)p(y)$$

$$p(x)$$

$$x\in\mathcal{X} \qquad \underbrace{\quad y\in\mathcal{Y} \quad}_{1}$$

Joint Entropy:

Independent Random Variables

If X and Y are statistically independent we have that:

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

$$= p(x)p(y)$$

$$p(x)$$

$$x\in\mathcal{X} \qquad y\in\mathcal{Y}$$

$$= \sum_{x\in\mathcal{X}} p(x)\log\frac{1}{p(x)} + \sum_{y\in\mathcal{Y}} p(y)\log\frac{1}{p(y)}$$

Joint Entropy:

Independent Random Variables

If X and Y are statistically independent we have that:

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

$$= p(x)p(y)$$

$$p(x)$$

$$x \in \mathcal{X} \qquad y \in \mathcal{Y}$$

$$1$$

$$= \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)} + \sum_{y \in \mathcal{Y}} p(y) \log \frac{1}{p(y)}$$

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

Joint Entropy:

Independent Random Variables

If X and Y are statistically independent we have that:

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

$$= p(x)p(y)$$

$$p(x)$$

$$x\in\mathcal{X} \qquad y\in\mathcal{Y}$$

$$1$$

$$= \sum_{x\in\mathcal{X}} p(x) \log \frac{1}{p(x)} + \sum_{y\in\mathcal{Y}} p(y) \log \frac{1}{p(y)}$$

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

Entropy is additive for independent random variables

Conditional Entropy

The conditional entropy of Y given $X = x$ is the entropy of the probability distribution $p(Y|X = x)$:

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

Conditional Entropy

The conditional entropy of Y given $X = x$ is the entropy of the probability distribution $p(Y|X = x)$:

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

The condi
conditio

$$H(Y|X) = \sum_{x \in \mathcal{X}} p(x) H(Y|$$

$$= \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p \qquad \frac{}{p(y|x)}$$

Conditional Entropy

The conditional entropy of Y given $X = x$ is the entropy of the probability distribution $p(Y|X = x)$:

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

The condi

conditio

$$H(Y|X) = \sum_{x \in \mathcal{X}} p(x) H(Y|$$

$$= \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p \qquad \frac{}{p(y|x)}$$

Average uncertainty that remains about Y when X is known.

Conditional Entropy — Cont'd

We can re-write the conditional entropy as follows:

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

Conditional Entropy — Cont'd

We can re-write the conditional entropy as follows:

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

Conditional Entropy — Cont'd

We can re-write the conditional entropy as follows:

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

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

Conditional Entropy — Cont'd

We can re-write the conditional entropy as follows:

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

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

$$= \mathbb{E}_{X,Y}\left[\log \frac{1}{p(Y|X)}\right]$$

Conditional Entropy — Cont'd

We can re-write the conditional entropy as follows:

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

Note the expectation is not wrt the conditional distribution but wrt the joint distribution $p(X, Y)$

Chain Rule

The joint entropy can be written as:

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

Chain Rule

The joint entropy can be written as:

$$
\begin{aligned}
H(X, Y) &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y) \\
&= \qquad\qquad p(x, y) \;\; \log p(x) + \log p(y \;\; x)
\end{aligned}
$$

Chain Rule

The joint entropy can be written as:

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

$$= \quad p(x,y)\left(\log p(x) + \log p(y|x)\right)$$

$$p(y|x)$$

$$\underbrace{\qquad\qquad}_{p(x)}$$

Chain Rule

The joint entropy can be written as:

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

$$= \quad p(x, y) \left(\log p(x) + \log p(y \, x) \right)$$

$$p(y|x)$$

$$\underbrace{\qquad\qquad}_{p(x)}$$

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

Chain Rule

The joint entropy can be written as:

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

$$= \qquad p(x,y)\left[\log p(x) + \log p(y|x)\right]$$

$$p(y|x)$$

$$\underbrace{\qquad\qquad}_{p(x)}$$

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

The joint uncertainty of X and Y is the uncertainty of X plus the uncertainty of Y given X

An Axiomatic Characterisation

Suppose we want a measure H of "information" in a random variable X such that

1. H depends on the distribution of X, and not the outcomes themselves
2. The H for the combination of two variables X, Y is at most the sum of the c
3. The ; Y is the su
4. Adding outcomes with probability zero does not affect H
5. The H for an unbiased Bernoulli is 1
6. The H for a Bernoulli with parameter

Then, the only possible choice for H is

$$H(X) = -\sum_{x} p(x) \log_2 p(x)$$

Outline

Summary

- Entropy as a measure of information content

- Computation of entropy of discrete random variables

- Ent

- Entropy and minimum expected number of bi

- Joint and conditional entropies, chain rule

- Reading: Mackay § 1.2 – § 1.5, § 8.1; Cover & Thomas § 2.1; Bishop § 1.6

Next time

- More properties of entropy
- Rel

- Mutual information