

Information theory

Advanced topics in AI-I (CK0146)

Francesco Corona

UFC/DC
ATAI-I (CK0146)
2017.1

Information theory

Information content

Entropy

Relative entropy

Mutual information

- 1 Information theory
 - Information content
 - Entropy
 - Relative entropy
 - Mutual information

Information theory

We introduce some concepts from **information theory**

Consider a discrete random variable x and ask:

- How much information is received when we observe a specific value of this variable?

The **amount of information** can be understood as the **degree of surprise** on learning the value of x

If we are told that a highly improbable event has occurred, we receive more information than if we are told that some very likely event has occurred

- If we knew that the event was certain we would receive no information

Information content

Information theory

A measure of information content depends on the probability distribution $p(x)$

- We look for a quantity $h(x)$ that is a monotonic function of the probability $p(x)$ and that expresses the information content

The form of $h(\cdot)$ can be found considering two events x and y

If x and y are unrelated, observing them both will lead to an information gain

- The gain should be the sum of information gained from each one alone

$$h(x, y) = h(x) + h(y)$$

Two unrelated events are statistically independent and $p(x, y) = p(x)p(y)$

- Information $h(x)$ must be given by the logarithm of $p(x)$

$$h(x) = -\log_2 p(x) \tag{1}$$

Definition

Shannon information content of an outcome

$$h(x) = -\log_2 p(x)$$

The negative sign ensures that information is positive or zero, with low probability events x corresponding to high information content

The choice of basis for the logarithm is arbitrary

- Prevalent convention is to use the base of 2
- The units of $h(x)$ are thus bits (binary digits)

Entropy

Information theory

Suppose a sender sends the value of a random variable to a receiver

Definition

- The average amount of information that they transmit in the process is obtained by taking the expectation of $h(x)$, wrt the distribution $p(x)$

$$H[x] = \sum_x p(x)h(x) = - \sum_x p(x) \log_2 p(x) \quad (2)$$

- This important quantity is called **entropy** of x

Example

An English language book (C. R. Darwin: *On the origin of species*, 1859)

Information theory

Information content

Entropy

Relative entropy

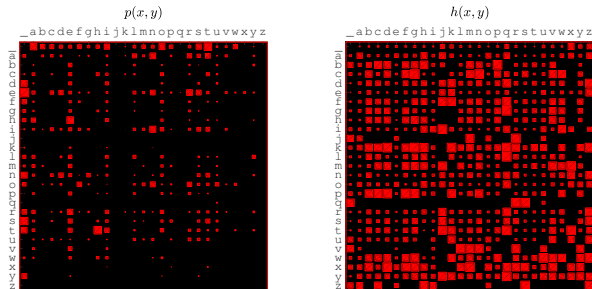
Mutual information

	$p(x)$			$h(x)$
1	0.15754	a	1	2.66621
2	0.06715	b	2	3.89656
3	0.01420	c	3	6.13787
4	0.02950	c	4	5.08324
5	0.03120	d	5	5.00238
6	0.11100	e	6	3.17136
7	0.02335	f	7	5.42046
8	0.01536	g	8	6.02446
9	0.04191	h	9	4.57660
10	0.06259	i	10	3.99791
11	0.00060	i	11	10.70148
12	0.00310	k	12	8.33417
13	0.03530	l	13	4.82423
14	0.02115	m	14	5.56303
15	0.06032	n	15	4.05128
16	0.06091	o	16	4.03726
17	0.01601	p	17	5.96511
18	0.00077	q	18	10.33494
19	0.05287	r	19	4.24148
20	0.05785	s	20	4.11162
21	0.07597	t	21	3.71841
22	0.02158	u	22	5.53415
23	0.00997	v	23	6.64832
24	0.01347	w	24	6.21431
25	0.00209	x	25	8.90438
26	0.01387	y	26	6.17231
27	0.00039	z	27	11.31352

- Assuming the distribution $p(x)$, the average Shannon information content of an outcome (entropy of a randomly selected letter) is

$$H[x] = 4.12 \text{ [bits]}$$

Example



- Assuming the joint distribution $p(x, y)$, the average Shannon information content of an outcome (joint entropy of a randomly selected digram) is

$$H[x, y] = \sum_{x, y} p(x, y) \log \frac{1}{p(x, y)} = 7.49 \text{ [bits]}$$

Example

Consider a random variable x having 8 possible states, each equally likely

To communicate the value of x , we need to transmit a 3-bit long message

- Notice that the entropy of x is $H[x] = -8 \times \frac{1}{8} \log_2 \frac{1}{8} = 3$ [bits]

Example

Consider a random variable x having 8 possible states $\{a, b, c, d, e, f, g, h\}$

- The respective probabilities of the states are $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64})$
- The entropy of x is

$$H[x] = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{4} \log_2 \frac{1}{4} - \frac{1}{8} \log_2 \frac{1}{8} - \frac{1}{16} \log_2 \frac{1}{16} - \frac{4}{64} \log_2 \frac{1}{64} = 2 \text{ [bits]}$$

The nonuniform distribution has a smaller entropy than the uniform one

How would we transmit the identity of the variable's state to a receiver?

- As before, we could use a 3-bit number

However, to keep the average code length shorter, we could take advantage of the nonuniform distribution by using shorter codes for more probable events

- We could set the following set of code strings

$$\begin{array}{cccccccc}
 a & b & c & d & e & f & g & h \\
 \underbrace{0} & \underbrace{10} & \underbrace{110} & \underbrace{1110} & \underbrace{111100} & \underbrace{111101} & \underbrace{111110} & \underbrace{111111} \\
 1/2 & 1/4 & 1/8 & 1/16 & 1/64 & 1/64 & 1/64 & 1/64
 \end{array}$$

to represent the states $\{a, b, c, d, e, f, g, h\}$

- The **average length of code** that has to be transmitted is

$$\text{average code length} = \frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + \frac{1}{16} \times 4 + 4 \times \frac{1}{64} \times 6 = 2 \text{ [bits]}$$

Which is, again, equal to the entropy of the random variable x

Note that shorter code strings cannot be used because it must be possible to disambiguate a concatenation of such strings into its component parts

Example

- 11001110 decodes uniquely into the state sequence $\{c, a, d\}$

$$\underbrace{11001110}_{\{c,a,d\}} \equiv \underbrace{110}_c \underbrace{0}_a \underbrace{1110}_d$$

There is a theorem (**noiseless coding theorem**) that states that the entropy is a lower bound on the number of bits needed to transmit the state of a RV

From now on, we switch to the use of natural logarithms in defining entropy

- The entropy will be measured in units of *nats*, instead of bits
- They differ by a factor of $\ln(2)$

Entropy (cont.)

UFC/DC
ATAI-I (CK0146)
2017.1

For a discrete random variable X with states x_i such that $P(X = x_i) = p(x_i)$

$$H = - \sum_i p(x_i) \ln(p(x_i)) \quad (3)$$

Information theory

Information content

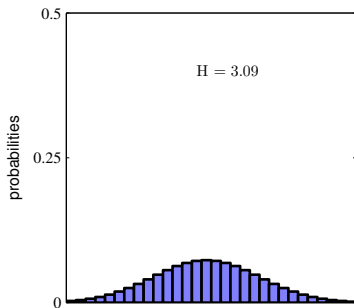
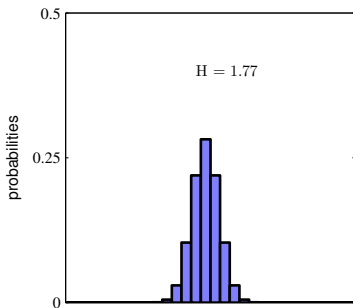
Entropy

Relative entropy

Mutual information

Example

Sharply peaked distributions will have a relatively low entropy



Because $0 \leq p(x_i) \leq 1$, the entropy is nonnegative, and it has a minimum

- It will be equal to 0, when one $p(x_i)$ is 1 and all other $p(x_{j \neq i}) = 0$

Remark

The maximum entropy configuration can be found by maximising H

- A Lagrange multiplier enforces normalisation on probabilities

$$\tilde{H} = - \sum_i p(x_i) \ln(p(x_i)) + \lambda \left(\sum_i p(x_i) - 1 \right) \quad (4)$$

- It is found when all of the $p(x_i)$ are equal, with $p(x_i) = 1/M$
- The maximum value of entropy is $H = \ln(M)$
- M is the total number of states of X

Entropy is also defined for distributions $p(x)$ over continuous variables x

We divide x into bins of width Δ and, assuming that $p(x)$ is continuous, we use the **mean value theorem** which tells us that there must exist a value x_i s.t.

$$\int_{(i)\Delta}^{(i+1)\Delta} p(x)dx = p(x_i)\Delta \quad (5)$$

- We can now quantise the continuous variable x by assigning any value x to the value x_i whenever x falls into the i -th bin
- The probability of observing a value x_i is given by $p(x_i)\Delta$

The entropy of this (still discrete) distribution takes the form

$$H_{\Delta} = - \sum_i p(x_i)\Delta \ln \left(p(x_i)\Delta \right) = - \sum_i p(x_i)\Delta \ln p(x_i) - \ln \Delta \quad (6)$$

- We used $\sum_i p(x_i)\Delta = 1$, which follows from Equation 5

$$H_{\Delta} = - \sum_i p(x_i) \ln (p(x_i)\Delta) = - \sum_i p(x_i)\Delta \ln p(x_i) - \ln \Delta \quad (7)$$

Omitting the second term $-\ln \Delta$ and considering the limit for $\Delta \rightarrow 0$, we have that the first term approaches the integral of $p(x) \ln p(x)$ in the limit, so that

Definition

$$- \lim_{\Delta \rightarrow 0} \left(\sum_i p(x_i) \ln (p(x_i)\Delta) \right) = - \int p(x) \ln p(x) dx \quad (8)$$

The quantity on the right-hand side is called **differential entropy**

The discrete and continuous forms of the entropy differ by a quantity $\ln \Delta$

Definition

For a density over multiple continuous variables \mathbf{x} , the differential entropy is

$$H[\mathbf{x}] = - \int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x} \quad (9)$$

For discrete distributions, maximum entropy configuration corresponded to an equal distribution of probabilities across all possible states of the variable

For the maximum entropy configuration of a continuous variable x to be well-defined, we must constrain the 1-st and 2-nd order moments of $p(x)$

- and, preserve normalisation

$$\int_{-\infty}^{+\infty} p(x) dx = 1 \quad (10)$$

$$\int_{-\infty}^{+\infty} xp(x) dx = \mu \quad (11)$$

$$\int_{-\infty}^{+\infty} (x - \mu)^2 p(x) dx = \sigma^2 \quad (12)$$

The constrained maximisation can be performed using Lagrange multipliers

- We maximise the following functional with respect to $p(x)$:

$$\begin{aligned} & - \int_{-\infty}^{+\infty} p(x) \ln p(x) dx + \lambda_1 \left(\int_{-\infty}^{+\infty} p(x) dx - 1 \right) \\ & + \lambda_2 \left(\int_{-\infty}^{+\infty} xp(x) dx - \mu \right) + \lambda_3 \left(\int_{-\infty}^{+\infty} (x - \mu)^2 p(x) dx - \sigma^2 \right) \end{aligned}$$

- We set the derivative to zero to get

$$p(x) = \exp \left(-1 + \lambda_1 + \lambda_2 x + \lambda_3 (x - \mu)^2 \right) \quad (13)$$

The result of the maximisation is given by the following functional of $p(x)$

$$p(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (14)$$

So, the distribution $p(x)$ that maximises differential entropy is the Gaussian

- Note that we did not constraint $p(x)$ to be nonnegative
- The result is already nonnegative, so there is no need

Remark

The Lagrange multipliers are found by back substitution into the constraints

If we evaluate the differential entropy for the Gaussian, we get

$$H[x] = \frac{1}{2} \left(1 + \ln(2\pi\sigma^2) \right) \quad (15)$$

which shows that entropy increases as the distribution gets fat

Remark

Differential entropy can be negative, for $\sigma^2 < 1/(2\pi e)$

We have a joint distribution $p(\mathbf{x}, \mathbf{y})$, and we draw pairs of values of \mathbf{x} and \mathbf{y}

If a value of \mathbf{x} is already known, then the additional information needed to specify the corresponding value of \mathbf{y} is given by $-\ln p(\mathbf{y}|\mathbf{x})$

Definition

The average information needed to specify \mathbf{y} given \mathbf{x} is

$$H[\mathbf{y}|\mathbf{x}] = - \int \int p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{y}|\mathbf{x}) d\mathbf{y}d\mathbf{x} \quad (16)$$

This quantity is called **conditional entropy** of \mathbf{y} given \mathbf{x}

Using the product rule, we see that conditional entropy satisfies

$$H[\mathbf{x}, \mathbf{y}] = H[\mathbf{y}|\mathbf{x}] + H[\mathbf{x}] \quad (17)$$

which is the differential entropy of the joint distribution $p(\mathbf{x}, \mathbf{y})$

- $H[\mathbf{x}]$ is the differential entropy of the marginal distribution $p(\mathbf{x})$

Relative entropy

Information theory

We can start relating the ideas of information theory to pattern recognition

Consider some unknown distribution $p(\mathbf{x})$ and suppose that we have modelled $p(\mathbf{x})$ using an approximating distribution $q(\mathbf{x})$

- We use $q(\mathbf{x})$ to construct a coding scheme for transmitting values of \mathbf{x}
- As a result of using $q(\mathbf{x})$ instead of the true distribution $p(\mathbf{x})$, additional amount of information (in nats) is required to specify the value of \mathbf{x}

Definition

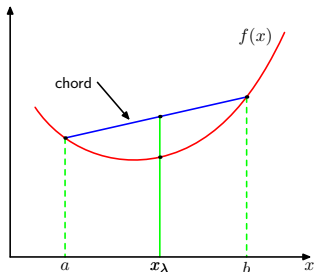
The average amount of additional information needed is given by

$$\begin{aligned} KL[p||q] &= - \int p(\mathbf{x}) \ln q(\mathbf{x}) d\mathbf{x} - \left(- \int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x} \right) \\ &= - \int p(\mathbf{x}) \ln \left(\frac{q(\mathbf{x})}{p(\mathbf{x})} \right) d\mathbf{x} \end{aligned} \quad (18)$$

Relative entropy or **Kullback-Leibler divergence** between the distributions $p(\mathbf{x})$ and $q(\mathbf{x})$, and it is not symmetrical quantity (i.e., $KL[p||q] \neq KL[q||p]$)

Remark

The KL divergence satisfies $KL(p||q) \geq 0$, with equality iff $p(x) = q(x)$



A function $f(x)$ is convex if every chord lies on or above the function

- Any $x \in [a, b]$ is $x_\lambda = \lambda a + (1 - \lambda)b$, with $0 \leq \lambda \leq 1$
- The corresponding point on the chord is $\lambda f(a) + (1 - \lambda)f(b)$
- The corresponding value of the function is $f(x_\lambda) = f(\lambda a + (1 - \lambda)b)$

Convexity of function $f(x)$ implies $f(\lambda a + (1 - \lambda)b) \leq \lambda f(a) + (1 - \lambda)f(b)$

- $f(x)$ is strictly convex if the equality is satisfied only for $\lambda \in \{0, 1\}$

Using convexity conditions, we can show that a convex function $f(x)$ satisfies

$$f\left(\sum_{i=1}^M \lambda_i x_i\right) \leq \sum_{i=1}^M \lambda_i f(x_i) \quad (19)$$

where $\lambda_i \geq 0$ and $\sum_i \lambda_i = 1$ for any set of points $\{x_i\}$ (**Jensen's inequality**)

Interpret the λ_i as the probability distribution over a discrete variable x taking values in $\{x_i\}$, Jensen's inequality is written with $\mathbb{E}[\cdot]$ denoting expectations as

$$f(\mathbb{E}[x]) \leq \mathbb{E}[f(x)] \quad (20)$$

For continuous variables, Jensen's inequality takes the form

$$f\left(\int \mathbf{x} p(\mathbf{x}) d\mathbf{x}\right) \leq \int f(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \quad (21)$$

$$f\left(\int \mathbf{x}p(\mathbf{x})d\mathbf{x}\right) \leq \int f(\mathbf{x})p(\mathbf{x})d\mathbf{x}$$

We can apply Jensen's inequality above to the Kullback-Liebler divergence

$$KL[p||q] = - \int p(\mathbf{x}) \ln\left(\frac{q(\mathbf{x})}{p(\mathbf{x})}\right) d\mathbf{x}$$

Using that $-\ln(x)$ is a convex function and the normalisation $\int q(\mathbf{x})d\mathbf{x} = 1$

$$KL(q||p) = - \int p(\mathbf{x}) \ln\left(\frac{q(\mathbf{x})}{p(\mathbf{x})}\right) d\mathbf{x} \geq - \ln \int q(\mathbf{x})d\mathbf{x} = 0 \quad (22)$$

In fact $-\ln(x)$ is strictly convex, and the equality holds only if $q(\mathbf{x}) = p(\mathbf{x})$

Suppose that we have some data generated from an unknown distribution $p(\mathbf{x})$

- Our goal is to model (approximate) $p(\mathbf{x})$
- We use a parametric distribution $q(\mathbf{x}|\theta)$

The set of adjustable parameters θ govern the approximating distribution $q(\mathbf{x})$

Parameters θ could be determined by minimising the KL divergence $KL[p||q]$

- Not directly though, because we do not know $p(\mathbf{x})$
- We only have observed a finite set of data $\{\mathbf{x}_n\}_{n=1}^N$

We approximate the expectation wrt to $p(\mathbf{x})$ by a finite sum over the data

$$KL[p||q] \simeq \frac{1}{N} \sum_{n=1}^N \left(-\ln q(\mathbf{x}_n|\theta) + \ln p(\mathbf{x}_n) \right) \quad (23)$$

$$KL[p||q] \simeq \sum_{n=1}^N \left(-\ln q(\mathbf{x}_n|\theta) + \ln p(\mathbf{x}_n) \right)$$

The second term is independent of θ and the first term is negative log likelihood for θ under the distribution $q(\mathbf{x}|\theta)$, evaluated from the data

Remark

- **Minimising KL divergence is maximising the likelihood function**

Mutual information

Information theory

Consider the joint distribution $p(\mathbf{x}, \mathbf{y})$ between two sets of variables \mathbf{x} and \mathbf{y}

- If the sets are independent the joint distribution will factorise into the product of the respective marginal distributions

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$$

- If the sets are not independent, we can evaluate how close they are to being independent

Definition

KL divergence between the joint distribution and the product of the marginals

$$\begin{aligned} I[\mathbf{x}, \mathbf{y}] &\equiv KL[p(\mathbf{x}, \mathbf{y}) || p(\mathbf{x})p(\mathbf{y})] \\ &= - \int \int p(\mathbf{x}, \mathbf{y}) \ln \left(\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} \right) d\mathbf{x}d\mathbf{y} \end{aligned} \quad (24)$$

which is a quantity called **mutual information** between the variables \mathbf{x} and \mathbf{y}

$$\begin{aligned} I[\mathbf{x}, \mathbf{y}] &\equiv KL[p(\mathbf{x}, \mathbf{y}) || p(\mathbf{x})p(\mathbf{y})] \\ &= - \int \int p(\mathbf{x}, \mathbf{y}) \ln \left(\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} \right) d\mathbf{x}d\mathbf{y} \end{aligned}$$

From the properties of the KL divergence, we have that $I(\mathbf{x}, \mathbf{y}) \geq 0$

- with equality iff \mathbf{x} and \mathbf{y} are independent

Remark

We have that mutual information is related to conditional entropy

$$I[\mathbf{x}, \mathbf{y}] = H[\mathbf{x}] - H[\mathbf{x}|\mathbf{y}] = H[\mathbf{y}] - H[\mathbf{y}|\mathbf{x}] = I[\mathbf{y}, \mathbf{x}] \quad (25)$$

Mutual information is the reduction in the uncertainty about \mathbf{x}

- by the virtue that the value of \mathbf{y} is given, and viceversa