Information theory

Advanced topics in AI-I (CK0146)
Pattern recognition (TIP8311)

Francesco Corona

Information theory
UFC/DC
ATAI-I (CK0146)
PR (TIP8311)
2016.2
Information theory
Relative entropy and mutual information

Outline

Information theory
UFC/DC
ATAI-I (CK0146)
PR (TIP8311)
2016.2
Information theory
Relative entropy and mutual information

Information theory

Information theory
UFC/DC
ATAI-I (CK0146)
PR (TIP8311)
2016.2
Information theory
Relative entropy and mutual information

Information theory (cont.)

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 theory (cont.)

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 $p(x)$

$$h(x) = -\log_2 p(x) \qquad (1)$$

Information theory (cont.)

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

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)

Information theory (cont.)

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

Definition

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

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

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

Information theory (cont.)

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]

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

Information theory (cont.)

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

$$\overbrace{\underbrace{0}_{1/2}}^{a}, \overbrace{\underbrace{10}_{1/4}}^{b}, \overbrace{\underbrace{110}_{1/8}}^{c}, \overbrace{\underbrace{1110}_{1/16}}^{d}, \overbrace{\underbrace{111100}_{1/64}}^{e}, \overbrace{\underbrace{111101}_{1/64}}^{f}, \overbrace{\underbrace{111110}_{1/64}}^{g}, \overbrace{\underbrace{111111}_{1/64}}^{h}$$

  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

Information theory (cont.)

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

- For instance, 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

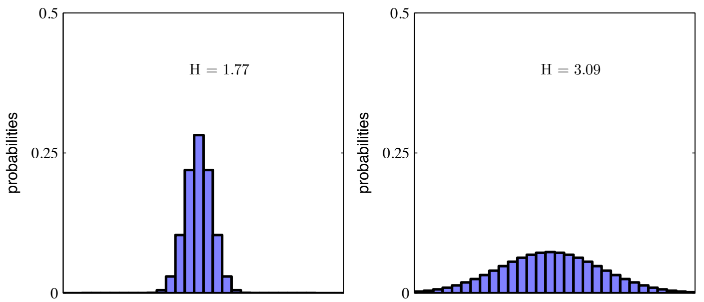
- the entropy will be measured in units of *nats*, instead of bits
- they differ simply by a factor of $\ln(2)$

Information theory (cont.)

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

$$H = -\sum_i p(x_i) \ln\Big(p(x_i)\Big) \tag{3}$$

Sharply peaked distributions will have a relatively low entropy



Information theory (cont.)

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 $p(x_{j \neq i}) = 0$

Example

The maximum entropy configuration can be found by maximising H

- A Lagrange multiplier enforces normalisation on probabilities

$$\tilde{H} = -\sum_i p(x_i) \ln\Big(p(x_i)\Big) + \lambda\Big(\sum_i p(x_i) - 1\Big) \tag{4}$$

- 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

Information theory (cont.)

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.

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

- 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

$$H_\Delta = -\sum_i p(x_i)\Delta \ln \Big(p(x_i)\Delta\Big) = -\sum_i p(x_i)\Delta \ln p(x_i) - \ln \Delta \tag{6}$$

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

Information theory (cont.)

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

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

Definition

$$-\lim_{\Delta \to 0} \Big(\sum_i p(x_i) \ln \Big(p(x_i)\Big)\Delta\Big) = -\int p(x) \ln p(x) dx \tag{8}$$

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

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

Information theory (cont.)

Definition

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

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

Information theory (cont.)

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

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

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

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

Information Theory

The constrained maximisation can be performed using Lagrange multipliers

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

$$-\int_{-\infty}^{+\infty} p(x)\ln p(x)dx + \lambda_1\Big(\int_{-\infty}^{+\infty} p(x)dx - 1\Big) + \lambda_2\Big(\int_{-\infty}^{+\infty} xp(x)dx - \mu\Big) + \lambda_3\Big(\int_{-\infty}^{+\infty}(x-\mu)^2 p(x)dx - \sigma^2\Big)$$

- We set the derivative to zero to get

$$p(x) = \exp\Big(-1 + \lambda_1 + \lambda_2 x + \lambda_3(x-\mu)^2\Big) \tag{13}$$

The Lagrange multipliers are found by back substitution into the constraints

Information Theory

Example

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

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

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

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

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

which shows that entropy increases as the distribution gets fat

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

Information theory (cont.)

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

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

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

Using the product rule, we see that conditional entropy satisfies

$$H[\mathbf{x}, \mathbf{y}] = H[\mathbf{y}|\mathbf{x}] + H[\mathbf{x}] \tag{17}$$

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 mutual information

Information theory

Relative entropy and mutual information

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 $q(\mathbf{x})$

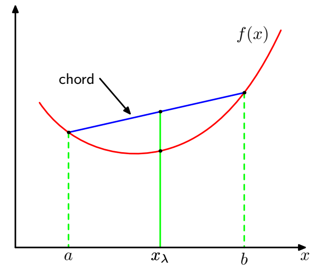
- We use $q(\mathbf{x})$ to construct a coding scheme for transmitting values of $\mathbf{x}$
- As a result of using $q(\mathbf{x})$ instead of the true distribution $p(\mathbf{x})$, additional amount of information (in nats) is required to specify the value of $\mathbf{x}$
- The average amount of additional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

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

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

Relative entropy and mutual information (cont.)

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



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 $f(x)$ then 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 $\lambda \in \{0, 1\}$

Relative entropy and mutual information (cont.)

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

$$f\Big(\sum_{i=1}^{M} \lambda_i x_i \Big) \leq \sum_{i=1}^{M} \lambda_i f(x_i) \tag{19}$$

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

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

For continuous variables, Jensen's inequality takes the form

$$f\Big(\int \mathbf{x} p(\mathbf{x}) d\mathbf{x} \Big) \leq \int f(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \tag{21}$$

Relative entropy and mutual information (cont.)

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

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

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

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

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

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

Relative entropy and mutual information (cont.)

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 $q(\mathbf{x}|\boldsymbol{\theta})$

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

Parameters $\boldsymbol{\theta}$ 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

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

Relative entropy and mutual information (cont.)

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

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

Remark

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

Relative entropy and mutual information (cont.)

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

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

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

Relative entropy and mutual information (cont.)

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

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