Algebra of Information Measures

Louis Wehenkel

Institut Montefiore, University of Liège, Belgium

ELEN060-2
Information and coding theory
February 2021

Outline

- Entropies and information measures
- Chain rules for entropy and information
- More about independence, and conditional independence
- Translation of these properties into properties of information measures
- Data processing inequality
- Bayesian networks and decision trees

Conditional (a posteriori) entropy

$$H(\mathcal{X}|\mathcal{Y}) = -\sum_{i=1}^{n}\sum_{j=1}^{m} P(X_i \cap Y_j)\log P(X_i|Y_j). \tag{1}$$

The entropy of $\mathcal{X}$ knowing that $\mathcal{Y} = Y_j$ is

$$H(\mathcal{X}|Y_j) = -\sum_{i=1}^{n} P(X_i|Y_j)\log P(X_i|Y_j), \tag{2}$$

it is positive (it is an entropy) and one has

$$H(\mathcal{X}|\mathcal{Y}) = \sum_{j=1}^{m} P(Y_j)H(\mathcal{X}|Y_j), \tag{3}$$

hence this latter is also positive.

And concavity of H_n implies: $H(\mathcal{X}|\mathcal{Y}) \leq H(\mathcal{X})$, which is a fundamental property!

$$
\begin{aligned}
H(\mathcal{X},\mathcal{Y}) &\triangleq -\sum_{i=1}^{n}\sum_{j=1}^{m} P(X_i \cap Y_j)\log P(X_i \cap Y_j) \\
&= -\sum_{i=1}^{n}\sum_{j=1}^{m} P(Y_j)P(X_i|Y_j)\log(P(Y_j)P(X_i|Y_j)) \\
&= -\sum_{i=1}^{n}\sum_{j=1}^{m} P(Y_j)P(X_i|Y_j)\log P(Y_j) \\
&\quad -\sum_{i=1}^{n}\sum_{j=1}^{m} P(Y_j)P(X_i|Y_j)\log P(X_i|Y_j) \\
&= -\sum_{j=1}^{m} P(Y_j)(\sum_{i=1}^{n} P(X_i|Y_j))\log P(Y_j) + H(\mathcal{X}|\mathcal{Y}) \\
&= H(\mathcal{Y}) + H(\mathcal{X}|\mathcal{Y}) \\
&= H(\mathcal{X}) + H(\mathcal{Y}|\mathcal{X}).
\end{aligned}
$$

One deduces the following inequalities :

$H(\mathcal{X},\mathcal{Y}) \geq \max\left(H(\mathcal{X}), H(\mathcal{Y})\right)$

$H(\mathcal{X},\mathcal{Y}) \leq H(\mathcal{X}) + H(\mathcal{Y})$

Conclusion :

$$H(\mathcal{X},\mathcal{Y}) \leq H(\mathcal{X}) + H(\mathcal{Y}) \leq 2H(\mathcal{X},\mathcal{Y}) \tag{4}$$

Particular cases :

$\mathcal{X}$ and $\mathcal{Y}$ independent : $H(\mathcal{X},\mathcal{Y}) = H(\mathcal{X}) + H(\mathcal{Y})$

(because then $P(X_i \cap Y_j) = P(X_i)P(Y_j)$) ($\Rightarrow H(\mathcal{X}|\mathcal{Y}) = H(\mathcal{X})$)

$\mathcal{X}$ function of $\mathcal{Y}$: $H(\mathcal{X},\mathcal{Y}) = H(\mathcal{Y})$.

(because then $H(\mathcal{X}|\mathcal{Y}) = 0$) (since $H(\mathcal{X}|Y_j) = 0, \forall j = 1,\ldots,m$)

$$I(\mathcal{X};\mathcal{Y}) = +\sum_{i=1}^{n}\sum_{j=1}^{m} P(X_i \cap Y_j)\log\frac{P(X_i \cap Y_j)}{P(X_i)P(Y_j)}. \tag{5}$$

One can derive :

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

and hence

$$I(\mathcal{X};\mathcal{Y}) = H(\mathcal{X}) + H(\mathcal{Y}) - H(\mathcal{X},\mathcal{Y})$$

which we may also write as

$$H(\mathcal{X},\mathcal{Y}) = H(\mathcal{X}) + H(\mathcal{Y}) - I(\mathcal{X};\mathcal{Y})$$

Main conclusion :

$$0 \leq I(\mathcal{X};\mathcal{Y}) \leq \min\{H(\mathcal{X}), H(\mathcal{Y})\}$$

Exercises.

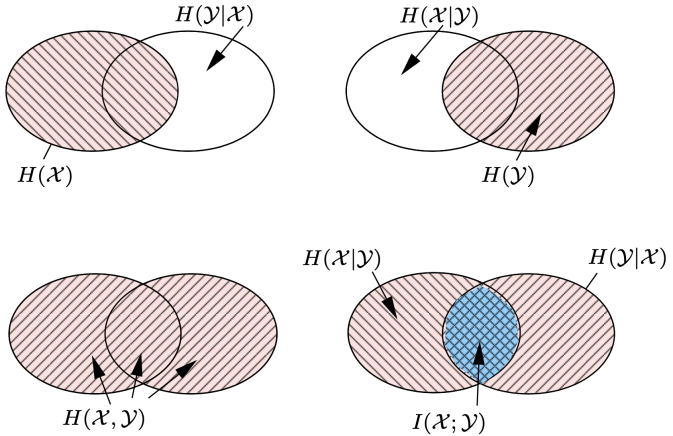
1. Show that indeed (and in the given order)
   1. $H(\mathcal{X},\mathcal{Y}) = H(\mathcal{Y}) + H(\mathcal{X}|\mathcal{Y}) = H(\mathcal{X}) + H(\mathcal{Y}|\mathcal{X})$
   2. $H(\mathcal{X},\mathcal{Y}) \geq \max\{H(\mathcal{X}), H(\mathcal{Y})\}$
   3. $H(\mathcal{X}|\mathcal{Y}) \leq H(\mathcal{X})$
   4. $H(\mathcal{X},\mathcal{Y}) \leq H(\mathcal{X}) + H(\mathcal{Y})$
   5. $I(\mathcal{X};\mathcal{Y}) = H(\mathcal{X}) + H(\mathcal{Y}) - H(\mathcal{X},\mathcal{Y}) = H(\mathcal{X}) - H(\mathcal{X}|\mathcal{Y})$

2. A tournament between two teams consists of a sequence of at most 5 games which stops as soon as one of the two teams has won three games. Let a and b denote the two teams and $\mathcal{X}$ a r.v. which represents the issue of a tournament between a and b. For example, $X = aaa, babab, bbaaa$ are possible values of $\mathcal{X}$ (there are other possible values). Let $\mathcal{Y}$ denote the random variable which denotes the number of games played (thus $\mathcal{Y} = \{3, 4, 5\}$).

   Suppose that the teams are of the same strength and the outcomes of the successive games are independent, and compute $H(\mathcal{X})$,$H(\mathcal{Y})$, $H(\mathcal{X}|\mathcal{Y})$ and $H(\mathcal{Y}|\mathcal{X})$.

   Let $\mathcal{Z} = \{a, b\}$ denote the random variable which identifies the team winning the tournament. Determine $H(\mathcal{X}|\mathcal{Z})$, compare with $H(\mathcal{X})$ and justify the result. Determine $H(\mathcal{Z}|\mathcal{X})$, and justify.

Summary



Particular cases

$\mathcal{X}$ and $\mathcal{Y}$ independent : $I(\mathcal{X};\mathcal{Y}) = 0$ (necessary and sufficient).

$\mathcal{X}$ function of $\mathcal{Y}$: $I(\mathcal{X};\mathcal{Y}) = H(\mathcal{X})$.

$\mathcal{X}$ one-to-one function of $\mathcal{Y}$: $I(\mathcal{X};\mathcal{Y}) = H(\mathcal{X}) = H(\mathcal{Y})$

Exercises.

1. Consider the following contingency table

	Y_1	Y_2
X_1	$\frac{1}{3}$	$\frac{1}{3}$
X_2	0	$\frac{1}{3}$

Compute (logarithms in base 2) :

1. $H(\mathcal{X}), H(\mathcal{Y})$
2. $H(\mathcal{X}|\mathcal{Y}), H(\mathcal{Y}|\mathcal{X})$
3. $H(\mathcal{X}, \mathcal{Y})$
4. $H(\mathcal{Y}) - H(\mathcal{Y}|\mathcal{X})$
5. $I(\mathcal{X}; \mathcal{Y})$
6. Draw a Venn diagram.

2. Consider three random variables $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$.

Prove that $H(\mathcal{X}, \mathcal{Y}|\mathcal{Z}) = H(\mathcal{X}|\mathcal{Z}) + H(\mathcal{Y}|\mathcal{X}, \mathcal{Z})$.

Other important properties (1)

1. Chain rules

A. Entropies

$$H(\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n) = \sum_{i=1}^{n} H(\mathcal{X}_i | \mathcal{X}_{i-1}, \ldots, \mathcal{X}_1)$$

B. Informations

$$I(\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n; \mathcal{Y}) = \sum_{i=1}^{n} I(\mathcal{X}_i; \mathcal{Y} | \mathcal{X}_{i-1}, \ldots, \mathcal{X}_1)$$

NB: Conditional mutual information of $\mathcal{X}$ and $\mathcal{Y}$ given $\mathcal{Z}$ is defined by

$$I(\mathcal{X}; \mathcal{Y} | \mathcal{Z}) \triangleq H(\mathcal{X} | \mathcal{Z}) - H(\mathcal{X} | \mathcal{Y}, \mathcal{Z}).$$

Almost same as before but one uses $P(\cdot|\mathcal{Z})$ (and averaging w.r.t. Z_i).

Outline of proofs.

Chain rule for entropies, by repeated application of the two variable expansion rule :

$$\begin{aligned} H(\mathcal{X}_1, \mathcal{X}_2) &= H(\mathcal{X}_1) + H(\mathcal{X}_2|\mathcal{X}_1) && (6)\\ H(\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3) &= H(\mathcal{X}_1) + H(\mathcal{X}_2, \mathcal{X}_3|\mathcal{X}_1) && (7)\\ &= H(\mathcal{X}_1) + H(\mathcal{X}_2|\mathcal{X}_1) + H(\mathcal{X}_3|\mathcal{X}_2, \mathcal{X}_1) && (8)\\ &\vdots && (9)\\ H(\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n) &= H(\mathcal{X}_1) + H(\mathcal{X}_2|\mathcal{X}_1) + \ldots + H(\mathcal{X}_n|\mathcal{X}_{n-1}, \ldots, \mathcal{X}_1) && (10) \end{aligned}$$

Chain rule for information :

$$\begin{aligned} I(\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n; \mathcal{Y}) &= H(\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n) - H(\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n|\mathcal{Y}) && (11)\\ &= \sum_{i=1}^{n} H(\mathcal{X}_i|\mathcal{X}_{i-1}, \ldots, \mathcal{X}_1) - \sum_{i=1}^{n} H(\mathcal{X}_i|\mathcal{X}_{i-1}, \ldots, \mathcal{X}_1, \mathcal{Y}) && (12)\\ &= \sum_{i=1}^{n} I(\mathcal{X}_i; \mathcal{Y}|\mathcal{X}_{i-1}, \ldots, \mathcal{X}_1) && (13) \end{aligned}$$

Equivalent definition of $I(\mathcal{X}; \mathcal{Y}|\mathcal{Z})$

$$I(\mathcal{X}; \mathcal{Y}|\mathcal{Z}) = \sum_{i,j,k} P(X_i, Y_j, Z_k) \log \frac{P(X_i, Y_j|Z_k)}{P(X_i|Z_k)P(Y_j|Z_k)} = \sum_k P(Z_k) I(\mathcal{X}; \mathcal{Y}|Z_k) \quad (14)$$

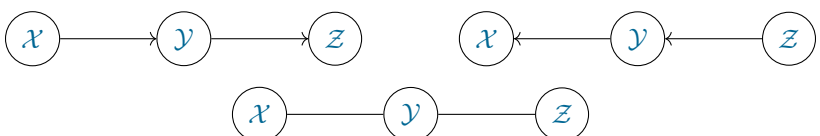
Other important properties (2)

2. Conditional independence and data processing inequality

Consider three discrete random variables : $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$
They are said to form a Markov chain if $\mathcal{Z}$ is conditionally indep. of $\mathcal{X}$ given $\mathcal{Y}$.
Notation : $\mathcal{Z} \perp \mathcal{X}|\mathcal{Y} \quad \Leftrightarrow \quad Z_i \perp X_j|Y_k, \quad \forall i, j, k.$
In other words $P(\mathcal{Z}|\mathcal{X}, \mathcal{Y}) = P(\mathcal{Z}|\mathcal{Y})$

Interpretation :

Conditioning : suppose $\mathcal{Y} = Y_k$ given $\Rightarrow P(\cdot) \to P(\cdot|Y_k)$
The probability measure becomes a conditional probability measure.
Cond. indep. $\equiv$ independence under the conditional measure, for any Y_k.
Independence is a symmetric relation : $\mathcal{Z} \perp \mathcal{X}|\mathcal{Y} \quad \Leftrightarrow \quad \mathcal{X} \perp \mathcal{Z}|\mathcal{Y}$.
$\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ form a Markov chain which is denoted by $\mathcal{X} \leftrightarrow \mathcal{Y} \leftrightarrow \mathcal{Z}$



The graphical representation is again a particular case of a Bayesian belief network, which will be introduced more precisely later on.

Bayesian belief networks provide a general and very powerful tool in order to handle conditional independence. Conditional independence is very important as a notion, because for many physical problems it may be used to represent causal relationships. Thus, the structure of conditional independence of stochastic models may be deduced from physical causality and structure.

Consider a communication system composed of two channels in series : X represents messages chosen by a source, Y messages at the receiving end of the first channel, and Z the messages at the receiving end of the second channel. These three random variables obviously represent a Markov chain.

Similarly, look at an industrial two stage process : X represents the characteristics of the input material; Y the characteristics of the output of the first stage and Z the characteristics of the output of the second stage. If Y is a precise enough description, then again we have a Markov chain. This means that if we are able to observe the output of the first stage, and want to predict what will happen during the second stage, the history $\mathcal{X}$ of the material is irrelevant.

This notion of *sufficiently precise* description of a process at an intermediate stage, is what we call in system theory the *state* of the system.

NB : these ideas may be applied to sets of variables :

$$\mathcal{X}_1, \mathcal{X}_2, \ldots \leftrightarrow \mathcal{Y}_1, \mathcal{Y}_2, \ldots \leftrightarrow \mathcal{Z}_1, \mathcal{Z}_2, \ldots$$

$$\mathcal{X}_1 \leftrightarrow \mathcal{X}_2 \leftrightarrow \cdots \leftrightarrow \mathcal{X}_k \leftrightarrow \cdots \leftrightarrow \mathcal{X}_{n-1} \leftrightarrow \mathcal{X}_n$$

Remarks.

If $\mathcal{X} \leftrightarrow \mathcal{Y} \leftrightarrow \mathcal{Z}$ then

$$P(\mathcal{X}, \mathcal{Y}, \mathcal{Z}) = P(\mathcal{X})P(\mathcal{Y}|\mathcal{X})P(\mathcal{Z}|\mathcal{Y}) = P(\mathcal{Z})P(\mathcal{Y}|\mathcal{Z})P(\mathcal{X}|\mathcal{Y}).$$

Data processing inequality

If $\mathcal{X} \leftrightarrow \mathcal{Y} \leftrightarrow \mathcal{Z}$ form a Markov chain then $I(\mathcal{X}; \mathcal{Y}) \geq I(\mathcal{X}; \mathcal{Z})$.
Indeed : chain rule of information applied in two ways to $I(\mathcal{X}; \mathcal{Y}, \mathcal{Z})$:

$$I(\mathcal{X}; \mathcal{Z}) + I(\mathcal{X}; \mathcal{Y}|\mathcal{Z}) = I(\mathcal{X}; \mathcal{Y}, \mathcal{Z}) = I(\mathcal{X}; \mathcal{Y}) + I(\mathcal{X}; \mathcal{Z}|\mathcal{Y}).$$

Since $\mathcal{X}$ et $\mathcal{Z}$ are conditionally independent, we have $I(\mathcal{X}; \mathcal{Z}|\mathcal{Y}) = 0$, and hence $I(\mathcal{X}; \mathcal{Z}) \leq I(\mathcal{X}; \mathcal{Y})$.

Examples

If $\mathcal{Z}$ is a function of $\mathcal{Y}$ it is conditionally independent of $\mathcal{X}$.

(Hence also $\mathcal{X} \leftrightarrow \mathcal{Y} \leftrightarrow \mathcal{Y}$)

If $\mathcal{Z}$ is a function of $\mathcal{Y}$ and another r.v. independent of $\mathcal{X}$ and $\mathcal{Y}$, it is also conditionally independent of $\mathcal{X}$.

Interpretation

The theorem tells us that whatever we do with $\mathcal{Y}$ in terms of data processing, there is no hope to gain more information about $\mathcal{X}$ than what is provided by $\mathcal{Y}$:
$\Rightarrow$ no way to create information by data processing.

Questions:

If A is an event of positive probability, what is the value of $P(A|A)$?
What is the meaning (value) of $P(\mathcal{X}, \mathcal{Y}, \mathcal{Y})$?
Is it true that $P(\mathcal{Y}|\mathcal{X}, \mathcal{Y}) = P(\mathcal{Y}|\mathcal{Y})$?

Another consequence

If $\mathcal{X} \leftrightarrow \mathcal{Y} \leftrightarrow \mathcal{Z}$ then $I(\mathcal{X};\mathcal{Y}|\mathcal{Z}) \leq I(\mathcal{X};\mathcal{Y})$.

In other words, in a Markov chain conditioning decreases mutual information.

This property is not true in general.

In other words, it is possible that $I(\mathcal{X};\mathcal{Y}|\mathcal{Z}) > I(\mathcal{X};\mathcal{Y})$ when $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ do not form a Markov chain.

For example

Consider the double coin flipping experiment.

Compute $I(\mathcal{H}_1;\mathcal{S})$ and $I(\mathcal{H}_1;\mathcal{S}|\mathcal{H}_2)$.

This finishes our study of information measures (algebra).

We will come back later to these notions for continuous random variables.

Exercises

1. Let $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ be three binary random variables. One gives the following information :
   - $P(\mathcal{X} = 0) = P(\mathcal{Y} = 0) = 0.5$,
   - $P(\mathcal{X}, \mathcal{Y}) = P(\mathcal{X})P(\mathcal{Y})$
   - $\mathcal{Z} = (\mathcal{X} + \mathcal{Y}) \text{mod} 2$ (i.e. $\mathcal{Z} = 1 \Leftrightarrow \mathcal{X} \neq \mathcal{Y}$).

   (a) What is the value of $P(\mathcal{Z} = 0)$?

   (b) What is the value of $H(\mathcal{X}), H(\mathcal{Y}), H(\mathcal{Z})$?

   (c) What is the value of $H(\mathcal{X}, \mathcal{Y}), H(\mathcal{X}, \mathcal{Z}), H(\mathcal{Y}, \mathcal{Z}), H(\mathcal{X}, \mathcal{Y}, \mathcal{Z})$?

   (d) What is the value of $I(\mathcal{X}; \mathcal{Y}), I(\mathcal{X}; \mathcal{Z}), I(\mathcal{Y}; \mathcal{Z})$?

   (e) What is the value of $I(\mathcal{X}; \mathcal{Y}, \mathcal{Z}), I(\mathcal{Y}; \mathcal{X}, \mathcal{Z}), I(\mathcal{Z}; \mathcal{X}, \mathcal{Y})$?

   (f) What is the value of $I(\mathcal{X}; \mathcal{Y}|\mathcal{Z}), I(\mathcal{Y}; \mathcal{X}|\mathcal{Z}), I(\mathcal{Z}; \mathcal{X}|\mathcal{Y})$?

   (g) Can you draw a Venn diagram which summarizes the situation ?

2. Let $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ be three discrete random variables. Show that

   (a) $H(\mathcal{X}, \mathcal{Y}|\mathcal{Z}) \geq H(\mathcal{X}|\mathcal{Z})$;

   (b) $I(\mathcal{X}, \mathcal{Y}; \mathcal{Z}) \geq I(\mathcal{X}; \mathcal{Z})$;

   (c) $H(\mathcal{X}, \mathcal{Y}, \mathcal{Z}) - H(\mathcal{X}, \mathcal{Y}) \leq H(\mathcal{X}, \mathcal{Z}) - H(\mathcal{X})$;

   (d) $I(\mathcal{X}; \mathcal{Z}|\mathcal{Y}) \geq I(\mathcal{Z}; \mathcal{Y}|\mathcal{X}) - I(\mathcal{Z}; \mathcal{Y}) + I(\mathcal{X}; \mathcal{Z})$.

Graphical models for probabilistic inference

Classical logic :

- Start with a theory : set of axioms which are supposed to hold in the physical world (if X has wings then X is a bird)

- Add observations from the real world : facts (Tweety has wings)

- Infer conclusions about other properties of the real world : Tweety is a bird.

Probabilistic logic :

Same, but statements and axioms are of probabilistic nature.

Inference : from a probabilistic model and observations from the real world, draw conclusions about unobserved variables.

Graphical models : represent relationships among variables by a graph.

NB.: not all models are graphical...

Main questions 1. How to build models : from first principles, from observations of nature, from both

2. How to use models : *deductive inference*

Now, we focus on probabilistic (deductive) inference with graphical models :

$\Rightarrow$ Bayesian networks, decision trees.

Model probabilistic relationships among a set of variables

- We will consider only discrete variables, but theory extends to continuous variables
- Bayesian networks : models for joint probability distributions $P(\mathcal{A}, \mathcal{B}, \ldots, \mathcal{U})$
- Decision trees : models for conditional probability distributions $P(\mathcal{A}|\mathcal{B}, \ldots, \mathcal{U})$

Bayesian networks : models for $P(\mathcal{A}, \mathcal{B}, \ldots, \mathcal{U})$

NB:

We consider only the case where $\mathcal{A}, \mathcal{B}, \ldots, \mathcal{U}$ take a finite number of value. Thus, the number of possible combinations of values is also finite.

Thus $P(\mathcal{A}, \mathcal{B}, \ldots, \mathcal{U})$ can be represented explicitly as a multidimensional table of numbers in $[0; 1]$: contingency table

But :

1. Explicit representation becomes quickly intractable (when the number of variables increases).

2. Explicit representation says nothing about structural relationships of variables (e.g. conditional independence)

Bayesian networks : compact representation, tractable, and interpretable (explicitly).

Example of inference using an explicit representation :

Given $P(\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}, \mathcal{E}, \mathcal{F})$ (model) and the fact (observation or hypothesis) that $\mathcal{B} = B_j$ and $\mathcal{C} = C_k$, what is the probability of event $\mathcal{A} = A_i$?

In other words compute : $P(A_i | B_j, C_k)$

Answer :

1. $P(A_i | B_j, C_k) = \frac{P(A_i, B_j, C_k)}{P(B_j, C_k)}$.
2. $P(A_i, B_j, C_k) = \sum_{D \in \mathcal{D}} \sum_{E \in \mathcal{E}} \sum_{F \in \mathcal{F}} P(A_i, B_j, C_k, D, E, F)$
3. $P(B_j, C_k) = \sum_{A \in \mathcal{A}} \sum_{D \in \mathcal{D}} \sum_{E \in \mathcal{E}} \sum_{F \in \mathcal{F}} P(A, B_j, C_k, D, E, F)$

Comments :

Suppose that the variables assume three values each, then $P(\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}, \mathcal{E}, \mathcal{F})$ is given by $3^6 - 1 = 728$ numbers.

The two sums concerns respectively $3^3 = 27$ and $3^4 = 81$ terms.

In applications (e.g. coding) : thousands of variables $\Rightarrow$ trivial method breaks down.

Same problem : we add some structural knowledge

Suppose we know (e.g. because of physical knowledge about the problem that :

$$P(\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}, \mathcal{E}, \mathcal{F}) = P(\mathcal{A}, \mathcal{B}, \mathcal{C}) P(\mathcal{D}, \mathcal{E}, \mathcal{F} | \mathcal{A})$$

and that

$$P(\mathcal{A}, \mathcal{B}, \mathcal{C}) = P(\mathcal{B}) P(\mathcal{C}) P(\mathcal{A} | \mathcal{B}\mathcal{C})$$

Now we need to specify the model :

- For $P(\mathcal{B})$ and $P(\mathcal{C})$ we need $4 = 2 + 2$
- For $P(\mathcal{A}|\mathcal{B}\mathcal{C})$ we need $2 \times 3 \times 3 = 18$.
- For $P(\mathcal{D}, \mathcal{E}, \mathcal{F}|\mathcal{A})$ we need $3 \times (3^3 - 1) = 78$.

$\Rightarrow$ Structural knowledge reduces the size of our model from 728 to $4 + 18 + 78 = 100$.

Computation of $P(A_i|B_j, C_k)$: trivial (table lookup)

What about computation of $P(B_j, C_k|A_i)$? (with and without structural knowledge)

Models are useful to provide not only accurate but also compact representations of the reality. In general, there is a tradeoff between model complexity and accuracy. Models are useful only if we are able to exploit them in order to understand or predict behavior of reality : in most situations tractability is possible only at the expense of accuracy.

Next week, when we will focus on channel coding, we will see that in order to efficiently exploit noisy channels it is necessary to manipulate very long sequences of symbols (long messages). For example, in the context of Turbo-codes typical message lengths which are manipulated are in the interval $[1000 \dots 100000]$. This means that we need to manipulate joint probability distributions of more than 1000 to 100000 binary variables, which would be totally impossible if we were to use explicit table-lookup models.

For those who are not yet convinced, let us make the explicit calculation : if $N = 1000$, a channel code will comprise $2^{1000} \approx 10^{301}$ code words. If every electron of the Universe (there are about 10^{80}) was a 1000 GHz processor able to store and retrieve the probability of such a code word in a single instruction, one could handle $10^{12} \times 10^{80} \approx 10^{92}$ code words per second, and in a period equal to the age of the Universe (3×10^{17} seconds), these computers would handle 3×10^{109} code words. To handle all words, we would still need to wait for a period equal to 10^{190} times the age of our Universe !

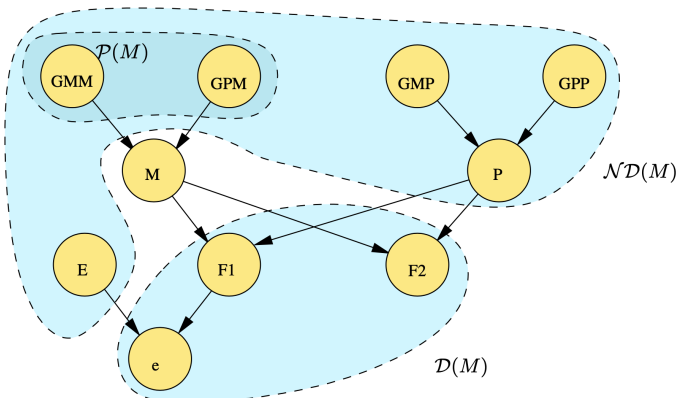
Nevertheless, by using compact models it is possible to handle the channel encoding and decoding tasks efficiently (in linear time with respect to the message length).

Later we will introduce stochastic process models. A stochastic process is a sequence of random variables corresponding to successive time instants (we will only consider discrete time models in this course). As time can grow indefinitely, a stochastic process is actually an infinite collection of random variables. Still, it is possible to devise very compact probabilistic models of such processes : actually with a few numbers it is possible to characterize the joint probability distribution of any finite sub-collection of random variables of the process.

Bayesian network : definition

Directed acyclic graph :

- Nodes model variables (one node for each variable)
- Arcs model causal relations among variables (conditional independence relations)



The figure illustrates an example Bayesian network, which is supposed to model the relationships among the color of eyes of different people in a family. The network actually models the ancestral relationships among the persons of this family. Each node represents the color (blue or brown) of one person. The arcs indicate which are the children of a person.

Note that this model does not pretend to be a correct view of Mendelean genetics. We will see later that this model is slightly more complex, but can still be easily represented by a Bayesian network. For the time being, we will use this naive picture of genetics as our running example to explain main concepts in Bayesian networks.

Terminology and notation

We use the same notation (round uppercase) to represent variables and nodes, since they are in one-to-one correspondance.

Let $\mathcal{X}_k$ denote a node in the graph $\mathcal{G}$. Then we denote by :

- $\mathcal{P}(\mathcal{X}_k)$ the set of parent nodes of $\mathcal{X}_k$, i.e. the origins of the arcs pointing towards $\mathcal{X}_k$.
- $\mathcal{F}(\mathcal{X}_k)$ the children of $\mathcal{X}_k$, i.e. the set $\{\mathcal{X}_j \in \mathcal{G} | \mathcal{X}_k \in \mathcal{P}(\mathcal{X}_j)\}$
- $\mathcal{D}(\mathcal{X}_k)$ the descendents of $\mathcal{X}_k$, i.e. the set of nodes which are in $\mathcal{F}(\mathcal{X}_k)$, or descendents of a node in $\mathcal{F}(\mathcal{X}_k)$.
- $\mathcal{ND}(\mathcal{X}_k)$ the nondescendents of $\mathcal{X}_k$, i.e. the set $\mathcal{G} \cap \neg(\{\mathcal{X}_k\} \cup \mathcal{D}(\mathcal{X}_k))$

The figure illustrates these notions for the node $\mathcal{M}$, and shows how this node partitions the graph.

Defining property of Bayesian networks

For any variable $\mathcal{X} \in \mathcal{G}$, and any subset of variables $\mathcal{W} \in \mathcal{ND}(\mathcal{X})$, we have $P(\mathcal{X}|\mathcal{P}(\mathcal{X}), \mathcal{W}) = P(\mathcal{X}|\mathcal{P}(\mathcal{X}))$, i.e. once the parents of a variable are given, it becomes independent of all its other non-descendents

Factorisation property

Suppose we are given a Bayesian network $\mathcal{G} = \{\mathcal{X}_1, \dots \mathcal{X}_n\}$ and for each variable $\mathcal{X}_i \in \mathcal{G}$ we are also given $P(\mathcal{X}_i|\mathcal{P}(\mathcal{X}_i))$, then

$$P(\mathcal{X}_1, \dots, \mathcal{X}_n) = \prod_{i=1}^{n} P(\mathcal{X}_i|\mathcal{P}(\mathcal{X}_i))$$

Note that for those variables for which $\mathcal{P}(\mathcal{X}_i) = \emptyset$ we are given the prior $P(\mathcal{X}_i)$.

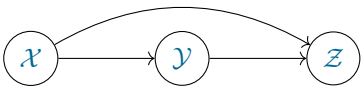
Comments :

As long as the $P(\mathcal{X}_i|\mathcal{P}(\mathcal{X}_i))$ are not specified, a Bayesian network is meant to represent all distributions which can be factorized in this way.

Any probability distribution may be represented in many ways by a Bayesian network, but not necessarily all conditional independence structures may be derived explicitly from a Bayesian network structure.

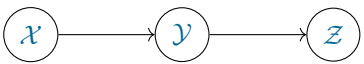
There exist probability distributions leading to independence relations which can not be represented completely by any Bayesian network.

Simple examples : some (all ?) three variable networks

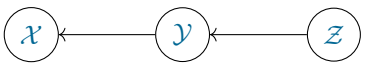


$P(\mathcal{X}, \mathcal{Y}, \mathcal{Z}) = P(\mathcal{X})P(\mathcal{Y}|\mathcal{X})P(\mathcal{Z}|\mathcal{X}, \mathcal{Y})$

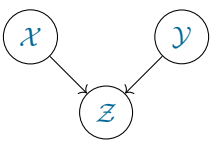
$P(\mathcal{X}, \mathcal{Y}, \mathcal{Z}) = P(\mathcal{X})P(\mathcal{Y})P(\mathcal{Z})$



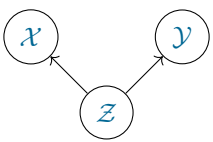
$P(\mathcal{X}, \mathcal{Y}, \mathcal{Z}) = P(\mathcal{X})P(\mathcal{Y}|\mathcal{X})P(\mathcal{Z}|\mathcal{Y})$



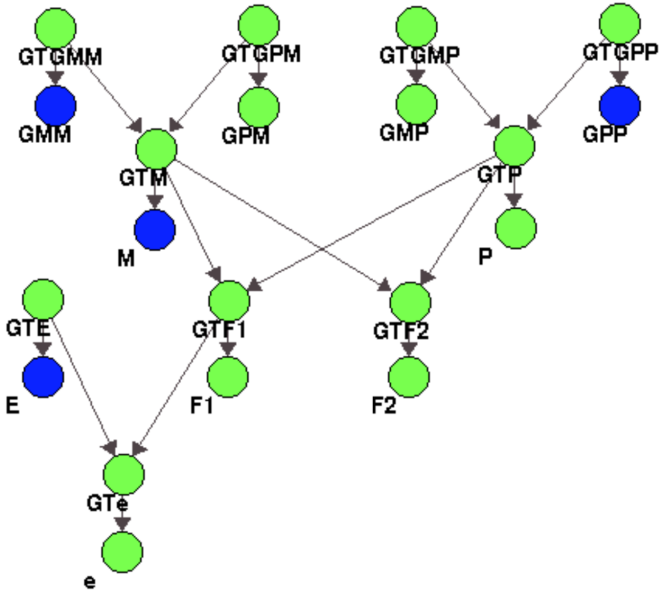
$P(\mathcal{X}, \mathcal{Y}, \mathcal{Z}) = P(\mathcal{Z})P(\mathcal{Y}|\mathcal{Z})P(\mathcal{X}|\mathcal{Y})$



$P(\mathcal{X}, \mathcal{Y}, \mathcal{Z}) = P(\mathcal{Z}|\mathcal{X}, \mathcal{Y})P(\mathcal{Y})P(\mathcal{X})$



$P(\mathcal{X}, \mathcal{Y}, \mathcal{Z}) = P(\mathcal{Z})P(\mathcal{X}|\mathcal{Z})P(\mathcal{Y}|\mathcal{Z})$



The correct Mendel model of eye colors.
Produced by JavaBayes tool.

Here is the complete model of our earlier example related to genetic.

We have added new variables denoted by $GT...$ for each individual which denote the two versions of the gene which determine the eye color for each individual. The variables may take on three values bb, bB, and BB, where b stands for blue and B for $brown$. We assume that the prior (or marginal) probability of these three values for the grand-parent generation are $0.25, 0.5, 0.25$; all the other (conditional) probability distributions are deduced from the Mendel model : the relation between parent and children genotypes assume that one of the two chromosomes is chosen at random (0.5 probability) and the relation between phenotype and genotype is deterministic, assuming that B (brown) is dominant character.

Notice that this network models the *genotype* (genes) of the individuals, and the relationship between the genotype and the observed variables (eye colors). In spite of the fact that genotypes can not be observed directly, it is possible to use this model to infer unobserved genotypes and phenotypes from the observed phenotypes (eye colors).

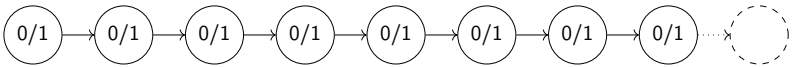
One particularity of this network is that all the conditional probabilities are identical (all people behave in the same way from the viewpoint of our model). The network can be extended to a whole population and be used to model relationships between successive generations and how one can observe genetic drift.

The present example is available on the web page http://www.montefiore.ulg.ac.be/~lwh/javabayes, where you can use a Java applet to simulate the network and see how it reacts to observations. On the same page you can also try out the earlier naive version of the same problem, and compare the differences.

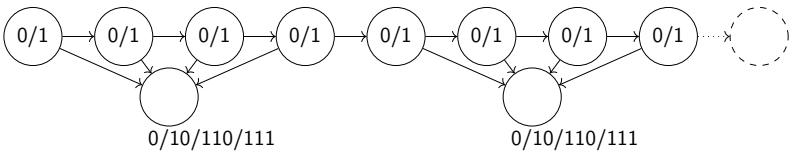
The prior probability distributions have been chosen so that without any observations all individuals have the same marginal probability distribution of genotypes (and hence phenotypes). This is what we will later on denote by *stationary* conditions. It turns out that, even if in the earlier generations the prior distribution is different from the stationary distribution, after a large enough number of generations the system converges to the stationary distribution.

Graphical models of communication systems

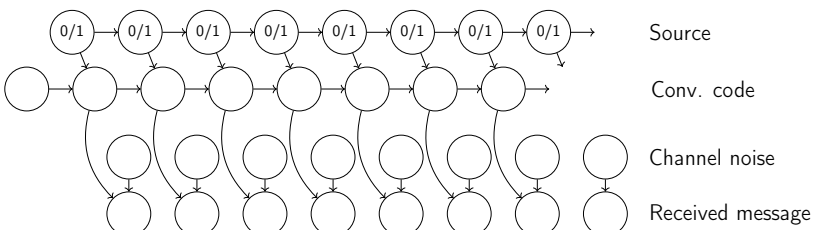
First order markov model of a black&white scanner



Same compressed by a 4 bit block code



Same, encoded and sent through a noisy, memoryless communication channel



D-separation and conditional independance relations induced by a BN

Some comments on the notion of independance of sets of random variables

Let $\mathcal{A} = \{\mathcal{X}_1, \ldots, \mathcal{X}_l\}$ and $\mathcal{B} = \{\mathcal{Y}_1, \ldots, \mathcal{Y}_m\}$ two sets of random variables.

- What is the meaning of $\mathcal{A} \perp \mathcal{B}$?

- Is it true that $\mathcal{A} \perp \mathcal{B} \Rightarrow (\forall i, j : \mathcal{X}_i \perp \mathcal{X}_j)$?

- And/or is the converse true, i.e. $(\forall i, j : \mathcal{X}_i \perp \mathcal{X}_j) \Rightarrow \mathcal{A} \perp \mathcal{B}$?

D-separation: definition

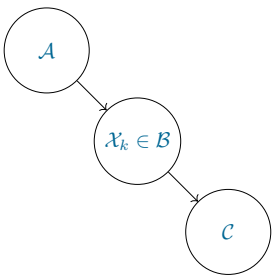
Let us denote by $\mathcal{A}, \mathcal{B}, \mathcal{C}$ three disjoint subsets of r.v. of a BN, and let us assume that $\mathcal{A}$ and $\mathcal{C}$ are non empty.

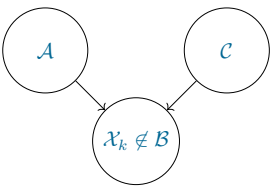
Let us consider paths over the undirected version of the DAG, from $\mathcal{A}$ to $\mathcal{C}$.

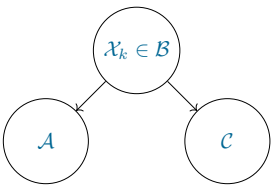
We say that $\mathcal{A}$ and $\mathcal{C}$ are *d-separated by* $\mathcal{B}$ if all paths from $\mathcal{A}$ to $\mathcal{C}$ are blocked by $\mathcal{B}$.

By definition, a path is *blocked* if it goes through a variable, say $\mathcal{X}_k$,

1. The pattern $\rightarrow \mathcal{X}_k \rightarrow$ appears in the path and $\mathcal{X}_k \in \mathcal{B}$
2. The pattern $\rightarrow \mathcal{X}_k \leftarrow$ appears in the path and $(\{\mathcal{X}_k\} \cup \mathcal{D}(\mathcal{X}_k)) \cap \mathcal{B} = \emptyset$
3. The pattern $\leftarrow \mathcal{X}_k \rightarrow$ appears in the path and $\mathcal{X}_k \in \mathcal{B}$







All paths with are not blocked, are said to be active (w.r.t. $\mathcal{A}$ to $\mathcal{C}$ and $\mathcal{B}$).

D-separation: fundamental property

If $\mathcal{A}, \mathcal{B}, \mathcal{C}$ are three disjoint sets of variables ($\mathcal{B}$ may be empty) of a bayesian network, then "$\mathcal{A}$ and $\mathcal{C}$ are d-separated by $\mathcal{B}$" $\Rightarrow \mathcal{A} \perp \mathcal{C} | \mathcal{B}$.

Notice that, if $\mathcal{A}$ and $\mathcal{B}$ are d-separated by $\mathcal{C}$ then any subset of $\mathcal{A}$ is d-separated from any subset of $\mathcal{C}$ by $\mathcal{B}$.

Notice also that we can change directions of some arrows in the graph without changing d-separations, provided that we don't change the set of $\rightarrow \mathcal{X}_k \leftarrow$ structures.

Thus, to represent the conditional independances one often uses so-called essential graphs, obtained from a DAG by replacing arrows which do not participate in a V-structure by lines.

Belief propagation

D-separation also leads to the design of effective belief propagation algorithms.

(See course notes and subsequent lessons).

There is much more to say about Bayesian belief networks, but limited time in the context of this course does not allow to go further in depth.

Bayesian networks where proposed in the eighties by Judea Pearl, in order to provide modelling tools for reasoning under uncertainty in artificial intelligence (e.g. expert systems for medical diagnosis).

In the meanwhile, both theory and practice have progressed significantly, and although the field has not yet reached full maturity there are already many significant real applications.

One of the complex questions, as regards inference, is to devise efficient algorithms to propagate evidence through the network. If the network has a tree structure this is rather easy task (a generalization of the forward-backward algorithm used for hidden markov chains, leading to an efficient algorithm). If the network is not a tree, one approach consists of grouping variables so as to yield a tree (so-called junction tree algorithm); another approach is to use approximate (but efficient) algorithms for probability propagation.

The other main problem under consideration in research concerns the automatic design of probabilistic models from data. Here also, there is still a lot to do.

Probabilistic reasoning and questionnaires

Let us consider a medical diagnostic problem and its probabilistic model. Let us denote by $\mathcal{D}$ a variable which is true when the patient under consideration has a certain disease (say hepatitis).

Ω set of all possible patients which will visit a M.D.

In order to make a diagnosis, the doctor will typically try to look at the symptoms (concentration of various types of blood cells, eye color, skin color, temperature ...) and ask questions about antecedents (factors, such as age, nutrition, smoking, addiction to heroin,...).

Note that not all questions have same relevance, and also in general the relevance of a question is dependent on already observed variables. Anyhow, typically the doctor would like to reach conclusions about the diagnostic by asking on relevant and informative questions.

Problem : how to design an efficient strategy for the diagnosis ?

Probabilistic model

Suppose that we have a model for $P(\mathcal{D}, \mathcal{A}_1, \ldots, \mathcal{A}_n, \mathcal{S}_1, \ldots, \mathcal{S}_m)$.

We can measure the residual uncertainty of the diagnosis problem by

$$H(\mathcal{D}|\mathcal{A}_1, \ldots, \mathcal{A}_n, \mathcal{S}_1, \ldots, \mathcal{S}_m)$$

i.e. the uncertainty which can not be reduced by observations.

If the disease is well known, hopefully this quantity will be small.

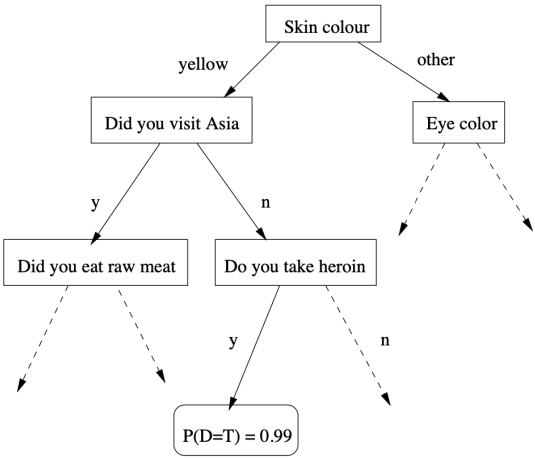
Note that, if we forbid the use of one of the possible observations, say $\mathcal{A}_1$, then the residual uncertainty increases

$$H(\mathcal{D}|\mathcal{A}_2, \ldots, \mathcal{A}_n, \mathcal{S}_1, \ldots, \mathcal{S}_m) \geq H(\mathcal{D}|\mathcal{A}_1, \ldots, \mathcal{A}_n, \mathcal{S}_1, \ldots, \mathcal{S}_m)$$

but this does not mean that all questions are relevant in all cases.

Suppose we are allowed only to ask one single question (observe one of $\mathcal{A}_i$ or $\mathcal{S}_j$), then we would choose $\mathcal{X} \in \{\mathcal{A}_1, \ldots, \mathcal{A}_n, \mathcal{S}_1, \ldots, \mathcal{S}_m\}$ maximizing $I(\mathcal{X}; \mathcal{D})$.

Strategy : same as decision tree



The test nodes of the tree (square boxes on the figure) represent essentially questions or observations that may be made by the doctor. The terminal nodes represent conclusions that will be drawn : the doctor stops to ask questions and decides that it is either very likely or very unlikely that the patient has hepatitis, or possibly decides that he is still uncertain and the patient should go to a specialist (who will ask more questions).

The tree structure defines the strategy that the doctor will use to reach a decision : the top-node (root of the tree) defines the first question, and successors define the substrategies depending on the obtained answer. Note that the terminal nodes of the tree are a function $\mathcal{T}$ of the test variables : using the tree is equivalent to observing this variable.

Tree construction algorithms :

A good decision tree is one that minimizes the average conditional entropy at the leaf nodes and at the same time minimizes complexity of the tree (different measures). If we know $P(\mathcal{D}, \mathcal{A}_1, \ldots, \mathcal{A}_n, \mathcal{S}_1, \ldots, \mathcal{S}_m)$ (say we have a Bayesian network) we can try to find an optimal tree, say one which minimizes

$$H(\mathcal{D}|\mathcal{T}) + \beta \text{Complexity}$$

Brute force :
- generate all possible trees (there is only a finite number of trees)
- for each tree compute $H(\mathcal{D}|\mathcal{T}) + \beta$Complexity (can be done using $P(\mathcal{D}, \mathcal{A}_1, \ldots, \mathcal{A}_n, \mathcal{S}_1, \ldots, \mathcal{S}_m)$)
- keep the best one.

Hill climbing :
- select the variable maximizing $I(\mathcal{X}; \mathcal{D})$ at the root node
- for each value X_i of $\mathcal{X}$ use $P(\mathcal{D}, \mathcal{A}_1, \ldots, \mathcal{A}_n, \mathcal{S}_1, \ldots, \mathcal{S}_m | X_i)$ to build subtree.
- stop when $H(\mathcal{D}|\mathcal{T}) + \beta$Complexity starts to increase.

Further reading

- D. MacKay, *Information theory, inference, and learning algorithms*
  - Chapter 2

Frequently asked questions

- State and graphically represent the algebraic relations among these quantities and their main properties (bounds, inequalities and equalities among quantities), and explain the main steps of the mathematical proofs of these relations.
- State the chaining rule for joint entropies. Define the notion of conditional mutual information and state the chaining rule for mutual informations. Explain the main steps of the mathematical proofs of these two chaining rules.
- Define the notion of Markov chain (over three discrete random variables). State, prove, discuss and illustrate the data processing inequality. Give an example where the data processing inequality can not be applied and where it is also not satisfied. State and discuss the corollaries of the data processing inequality.