Advanced Data Analysis and Machine Learning

Lecture: Bayesian Networks

Lasse Lensu

2015-10-26

Outline

LUT
Lappeenranta
University of Technology

Bayesian inference

- Simple Bayesian inference is unsuitable for complex modelling problems with a large number of variables.
- High dimensionality $\Rightarrow$ a huge amount of data would be required to learn the joint probability distributions of several random variables.
- Independence of variables would be handy, but this is uncommon.
- In real-world modeling problems it is typical that the majority of the model variables depend on each other to some extent.

Conditional probability and chain rule

- A priori probability $P(A)$ is the probability of event A (similarly event B and its a priori probability $P(B)$).
- Conditional probability $P(B|A)$ means the probability of observing event B when A has happened and it can be defined

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \tag{1}$$

where $A \cap B$ says that both A and B happen.

- The chain rule follows from Eq. 1 directly

$$\begin{aligned} P(A \cap B) &= P(A)P(B|A) && (2) \\ \Rightarrow P(A_1 \cap \ldots \cap A_k) &= P(A_1)P(A_2|A_1)\ldots \\ & \quad P(A_k|A_1 \cap \ldots \cap A_{k-1}). && (3) \end{aligned}$$

Conditional probability and Bayes

- Bayes' rule is a consequence of the conditional probability and chain rule:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{4}$$

- In the case of a background event C affecting the occurence of other events, Bayes' rule can be written in a more general form

$$P(A|B \cap C) = \frac{P(B|A \cap C)P(A|C)}{P(B|C)}. \tag{5}$$

- (Bayes' rule allows to compute the conditional probability $P(A|B)$ from the "inverse" conditional probability $P(B|A)$.)

Random variables and distributions

- A random variable is defined by a function associating a value with each outcome of a random experiment.
- Depending on the experiment outcome space and the events a random variable can describe, the values arise from a continuous or discrete distribution.
- When several random variables are involved, the interesting problem is that what is the joint distribution which represents the probabilities of co-occurring events.
- The problem can be solved by marginalising the joint distribution as needed (for example, summing up the rows or columns of the joint distribution in the case of two random variables).

Independence and conditional independence

- An event A is independent of event B in P (can be denoted as $P \models (A \perp B)$), if

$$P(A|B) = P(A), \text{ or if} \quad (6)$$
$$P(B) = 0. \quad (7)$$

- Alternatively, a distribution P satisfies $(A \perp B)$ if and only if

$$P(A \cap B) = P(A)P(B) \quad (8)$$
$$\Rightarrow P(A_1, A_2, \ldots, A_n) = P(A_1)P(A_2)\cdots P(A_n). \quad (9)$$

Conditional independence in practice

- Two events A and B are conditionally independent of a third event C if and only if, given knowledge of whether C occurs, knowledge of whether A occurs provides no information on the likelihood of B occurring and vice versa.
- Example:
  - John and Jeff usually use the same train to go to work.
  - Event A: "John is late", B: "Jeff is late", C: "the train is late".
  - A and B are conditionally independent (given C) since if we already know that the train is late (C), knowing that John is late (A) does not give any new knowledge whether Jeff is late (B).
- Independence of variables would be handy, but this is uncommon. It is more common that two events are independent given a background event.

Bayesian networks

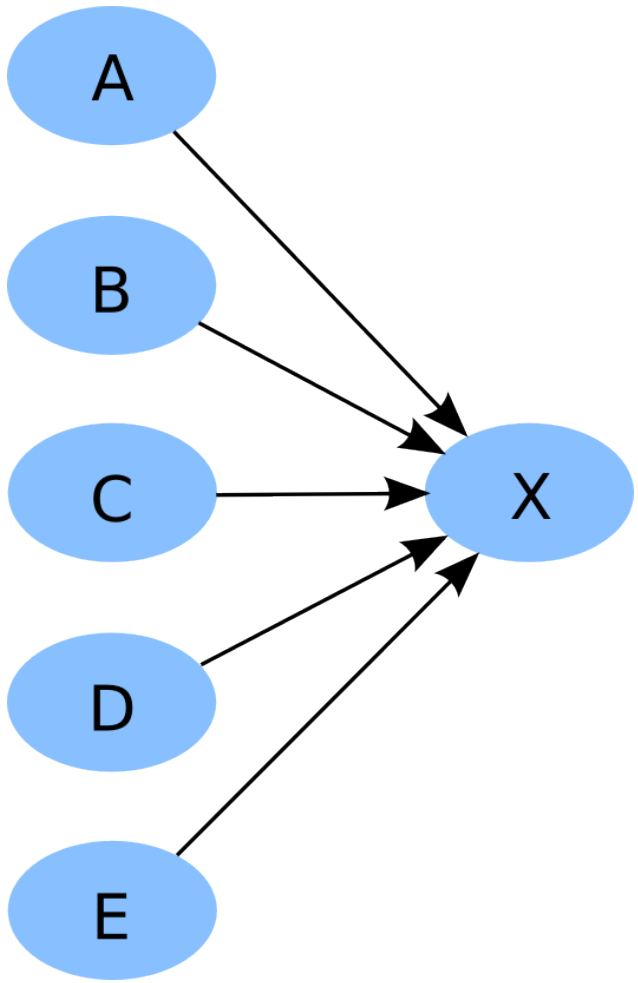
LUT Lappeenranta University of Technology

- A probabilistic graphical model.
- Representation of a set of random variables $\mathcal{X} = X_1, \ldots, X_n$ as nodes and their conditional dependences as edges $\mathcal{E}$ by using a directed acyclic graph (DAG) $\mathcal{G} = (\mathcal{X}, \mathcal{E})$.
- If there is a directed edge from X_i to X_j, X_i is X_j's parent.

Bayesian networks

LUT Lappeenranta University of Technology

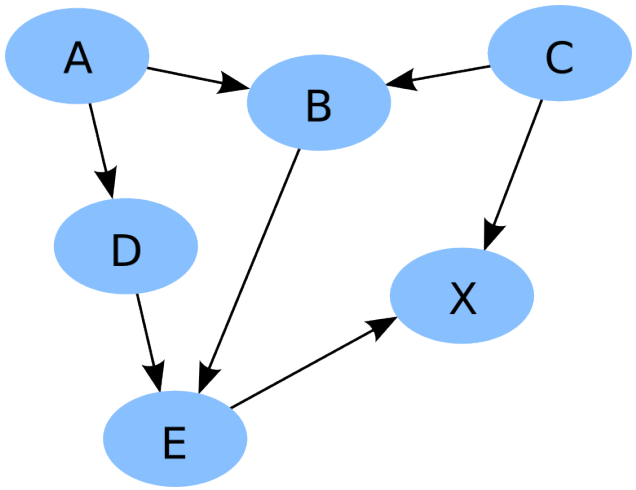
- Instead of defining $P(X|A, B, C, D, E)$ ...
- The network can be seen as i) a data structure representing a joint distribution in a factorised way or ii) a representation of the conditional independence assumptions of a distribution.
- Once network is constructed and parameters estimated, it can be used to answer queries related to the joint distribution, for example, compute the posterior probability distribution for any node.



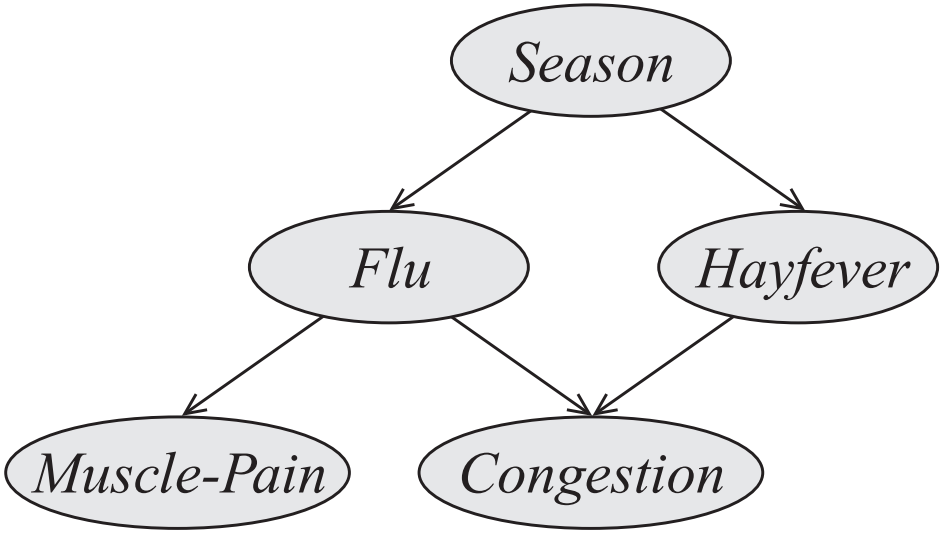
Bayesian networks

LUT Lappeenranta University of Technology

- Instead of defining $P(X|A, B, C, D, E)$ we need $P(A), P(B|A, C), P(C), P(D|A), P(E|B, D), P(X|C, E)$.
- The network can be seen as i) a data structure representing a joint distribution in a factorised way or ii) a representation of the conditional independence assumptions of a distribution.
- Once network is constructed and parameters estimated, it can be used to answer queries related to the joint distribution, for example, compute the posterior probability distribution for any node.

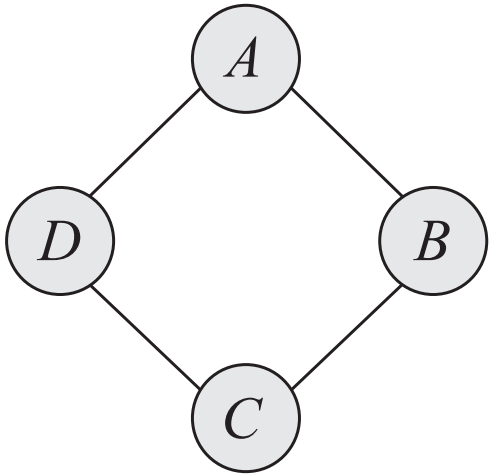


Bayesian and Markov networks



Independences

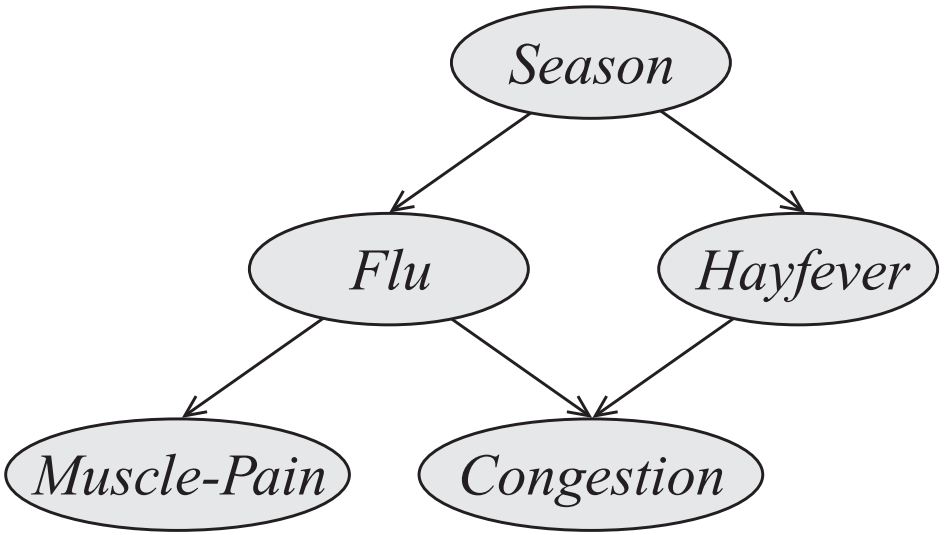
$$(F \perp H | S)$$
$$(C \perp S | F, H)$$
$$(M \perp H, C | F)$$
$$(M \perp C | F)$$



Independences

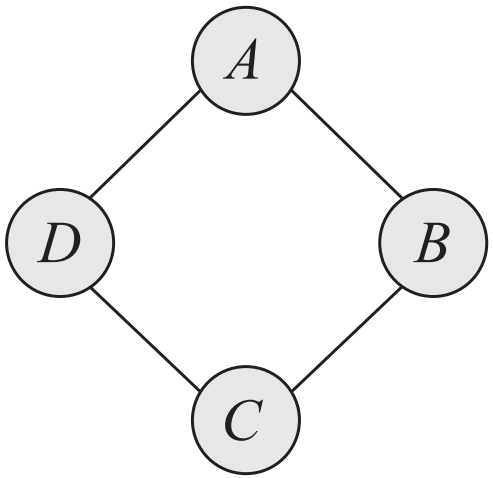
$$(A \perp C | B, D)$$
$$(B \perp D | A, C)$$

Bayesian and Markov networks



Factorisation

$$P(S,F,H,C,M) = P(S)P(F|S)$$
$$P(H|S)P(C|F,H)P(M|F)$$



Factorisation

$$P(A,B,C,D) = \frac{1}{Z}\phi_1(A,B)$$
$$\phi_2(B,C)\phi_3(C,D)\phi_4(A,D)$$

Network interpretation

LUT
Lappeenranta
University of Technology

- Causality
  - Arrows and their directionality implies causal relationship.
  - However, this does not need to be the case.

Network interpretation

LUT
Lappeenranta
University of Technology

- Causality
  - Arrows and their directionality implies causal relationship.
  - However, this does not need to be the case.

Inference

- Answering queries of interest based on a joint probability distribution over multiple random variables.
- Inference can be realised by using the following [$\mathcal{X} = X_1, \ldots, X_n$ is a set of random variables whose joint distribution is $P(X_1, \ldots, X_n)$]:
  - Computing the marginal distribution of one or more nodes $\boldsymbol{Y} \subset \mathcal{X}$ (model output) based on evidence $\boldsymbol{E} \subset \mathcal{X}$ (model inputs): $P(\boldsymbol{Y}|\boldsymbol{E} = \boldsymbol{e})$, that is, the posterior probability distribution of $\boldsymbol{y} \in \boldsymbol{Y}$ conditioned on the fact that $\boldsymbol{E} = \boldsymbol{e}$.
  - Finding the most probable explanations, that is the maximum a posteriori (MAP) assignment of the values for several nodes:

$$\text{MAP}(\boldsymbol{W}|\boldsymbol{e}) = \arg\max_{\boldsymbol{w}} P(\boldsymbol{w}, \boldsymbol{e}) \tag{10}$$

where $\boldsymbol{W} = \mathcal{X} - \boldsymbol{E}$.

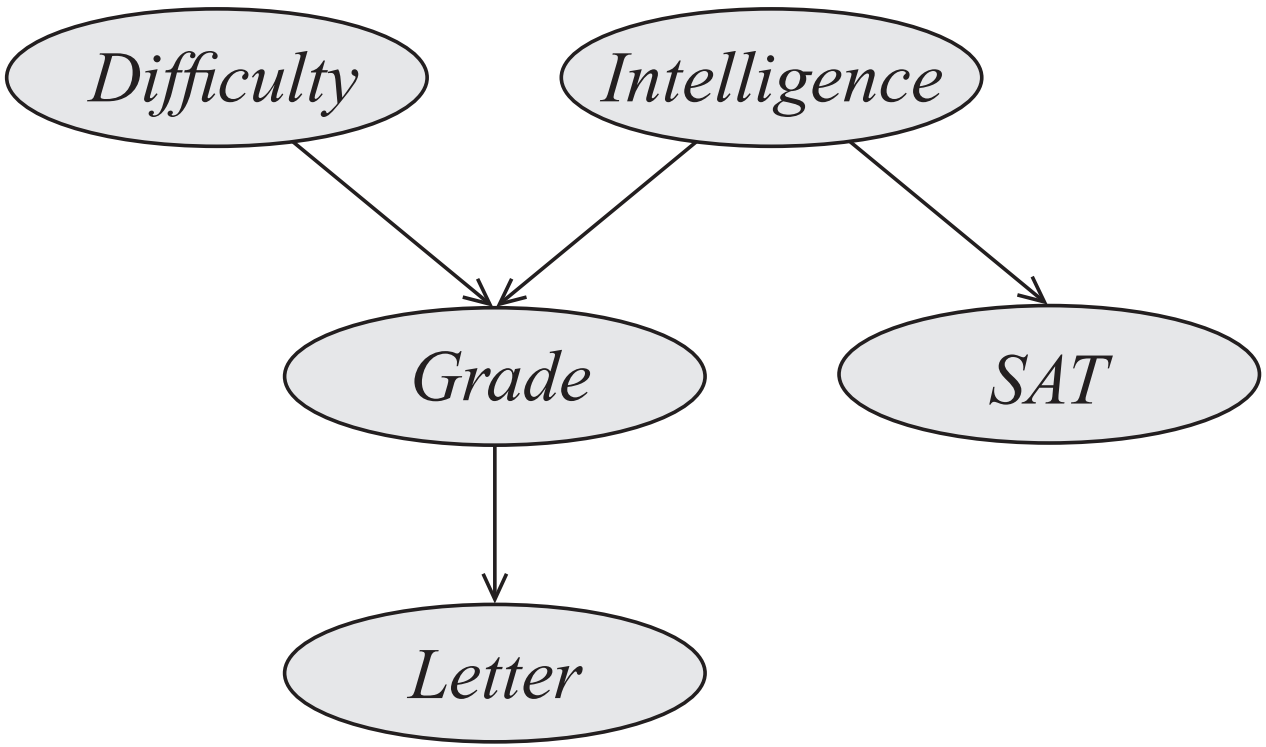
Learning Bayesian networks

- Parameter learning
  - Learning probability distributions for nodes conditional upon their parents.
  - Efficient algorithms do exist.
- Structure learning
  - Learning conditional dependencies between variables.
  - NP-complete problem.

Bayesian network example

LUT Lappeenranta University of Technology

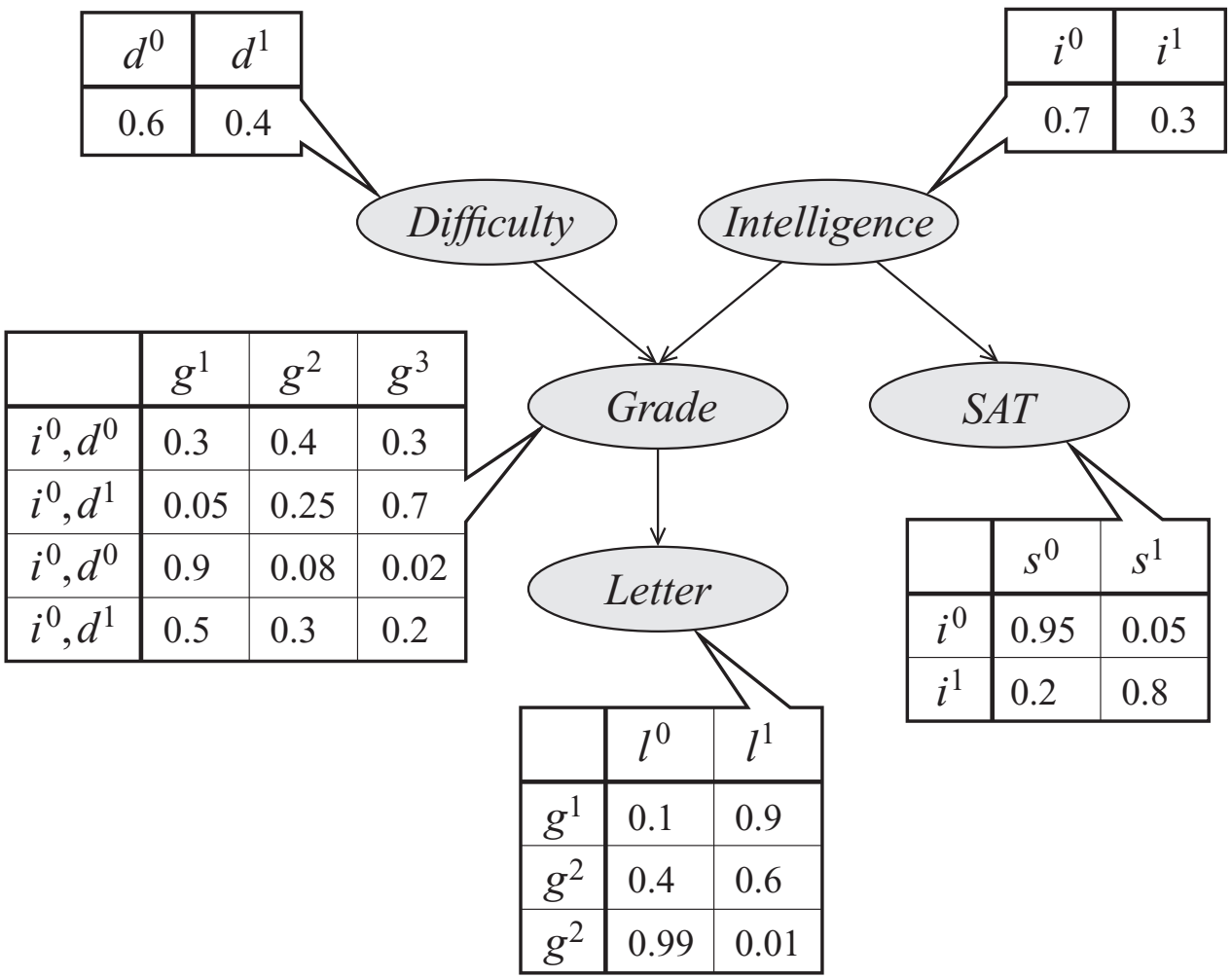
- Student example [1] :



Bayesian network example

LUT Lappeenranta University of Technology

- Student example [1] (Can you find the misspelled symbols?):



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^0, d^0	0.9	0.08	0.02
i^0, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^2	0.99	0.01

Summary

- Simple Bayesian inference is unsuitable for complex modelling problems with a large number of variables.
- Independence of variables would be handy, but in real-world modeling problems it is typical that the model variables depend on each other to some extent.
- Bayesian networks are probabilistic graphical models that represent the random variable dependencies (and independences) as a directed acyclic graph (DAG).
- The conditional joint distrubutions can be parameterised and learnt to perform inference.

References

LUT
Lappeenranta
University of Technology


Daphne Koller and Nir Friedman.
Probabilistic Graphical Models: Principles and Techniques.
The MIT Press, Cambridge, Massachusetts, 2009.