Probabilistic Graphical Models

Part 2: Bayesian Networks – Syntax and Semantics

M. Jaeger

Aalborg University

Bayesian Networks

Syntax

Let $X_1, \ldots, X_n$ be random variables. A *Bayesian network* for $X_1, \ldots, X_n$ consists of

- a *directed acyclic graph (DAG)* $\mathcal{G}$ whose nodes are the variables $X_1, \ldots, X_n$,
- a *conditional probability table (cpt)* for each node X_i specifying a conditional distribution $P(X_i \mid \mathrm{Pa}(X_i))$, where $\mathrm{Pa}(X_i)$ are the parents of X_i in the graph.

Notation

We write $\mathcal{N} = (\mathcal{G}, cpt(X_i)_{i=1,\ldots,n})$ for a BN, or more succinctly, $\mathcal{N} = (\mathcal{G}, \boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is a parameter vector containing all the entries of all $cpt(X_i)$.

Semantics

A Bayesian network $\mathcal{N}$ with nodes $X_1, \ldots, X_n$ defines the joint distribution

$$P_{\mathcal{N}}(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \mathrm{Pa}(X_i))$$

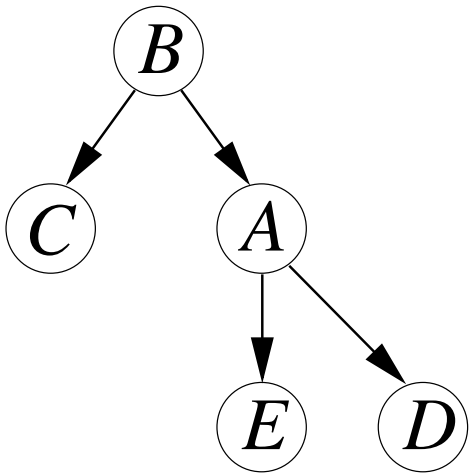
Given

- an ordered set of variables $X_1, \ldots, X_n$ with joint distribution P
- conditional independencies

$$P(X_i \mid X_1, \ldots, X_{i-1}) = P(X_i \mid \mathrm{Pa}(X_i)) \quad (\mathrm{Pa}(X_i) \subseteq \{X_1, \ldots, X_{i-1}\})$$

one obtains a Bayesian network representing P:

- the parents of X_i in the dag are the variables in $\mathrm{Pa}(X_i)$
- the conditional probability table of X_i specifies the conditional distribution $P(X_i \mid \mathrm{Pa}(X_i))$

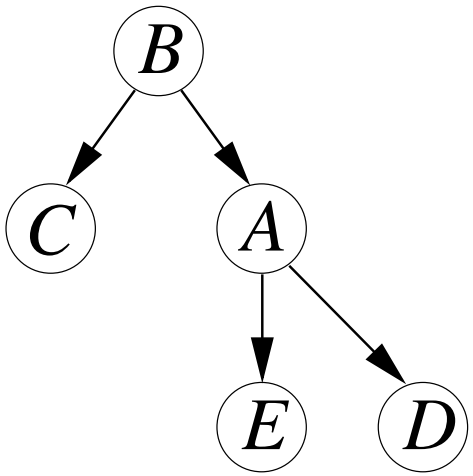


BN could have been constructed using

Variable ordering: B, C, A, E, D

Conditional Independencies:

$$P(A \mid B, C) = P(A \mid B)$$
$$P(E \mid B, C, A) = P(E \mid A)$$
$$P(D \mid B, C, A, E, D) = P(D \mid A)$$

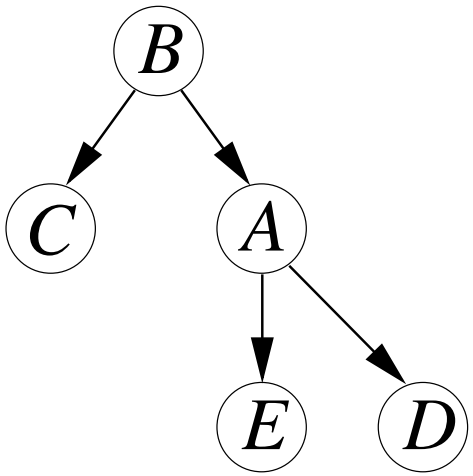


BN could have been constructed using

Variable ordering: B, C, A, E, D
Conditional Independencies:
$$P(A \mid B, C) = P(A \mid B)$$
$$P(E \mid B, C, A) = P(E \mid A)$$
$$P(D \mid B, C, A, E, D) = P(D \mid A)$$

or

Variable ordering: B, A, C, D, E
Conditional Independencies:
$$P(C \mid B, A) = P(C \mid B)$$
$$P(D \mid B, A, C) = P(D \mid A)$$
$$P(E \mid B, A, C, D) = P(E \mid A)$$



BN could have been constructed using

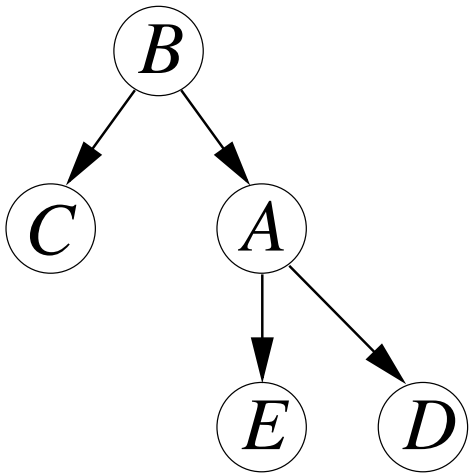
Variable ordering: B, C, A, E, D
Conditional Independencies:
$$P(A \mid B, C) = P(A \mid B)$$
$$P(E \mid B, C, A) = P(E \mid A)$$
$$P(D \mid B, C, A, E, D) = P(D \mid A)$$

or

Variable ordering: B, A, C, D, E
Conditional Independencies:
$$P(C \mid B, A) = P(C \mid B)$$
$$P(D \mid B, A, C) = P(D \mid A)$$
$$P(E \mid B, A, C, D) = P(E \mid A)$$

or

...



BN could have been constructed using

Variable ordering: B, C, A, E, D
Conditional Independencies:
$$P(A \mid B, C) = P(A \mid B)$$
$$P(E \mid B, C, A) = P(E \mid A)$$
$$P(D \mid B, C, A, E, D) = P(D \mid A)$$

or

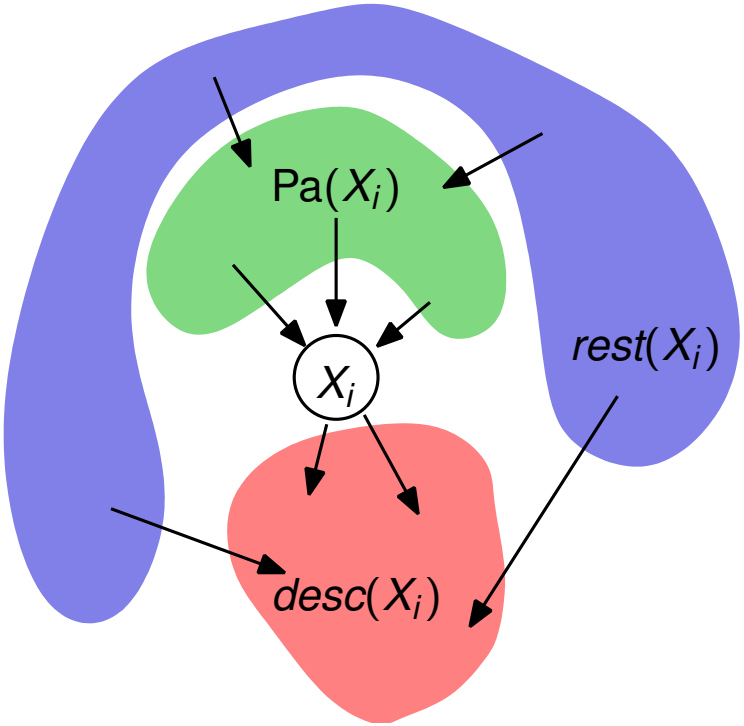
Variable ordering: B, A, C, D, E
Conditional Independencies:
$$P(C \mid B, A) = P(C \mid B)$$
$$P(D \mid B, A, C) = P(D \mid A)$$
$$P(E \mid B, A, C, D) = P(E \mid A)$$

or

...

Which of these conditional independencies are actually true in the distribution defined by BN?

All!



In the distribution P defined by the BN the following independence relation holds:

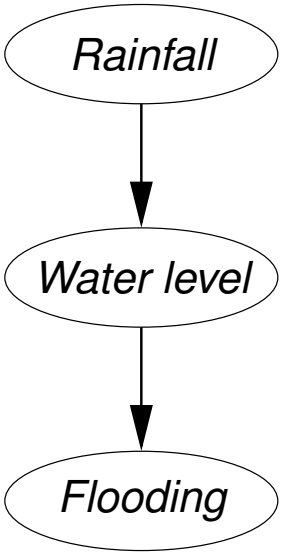
$$P(X_i \mid \mathrm{Pa}(X_i), rest(X_i)) = P(X_i \mid \mathrm{Pa}(X_i))$$

"X_i is independent of its non-descendants given its parents"

This is the independence relation corresponding to a variable ordering in which X_i is in the last possible position (just before all its descendants).

Independence and Causal Reasoning

Val(*Rainfall*)={*no,light,medium,heavy*}, *Val*(*Water level*)={*low,medium,high*}, *Val*(*Flooding*)={*yes,no*},



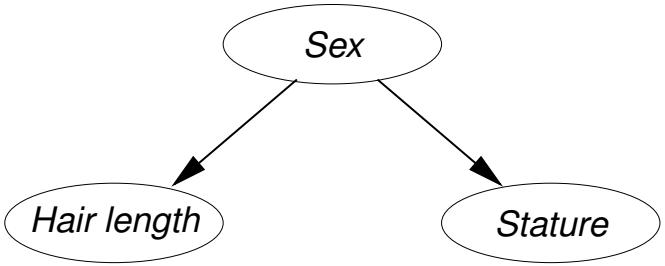
$$P(Flooding \mid Water\ level, Rainfall) = P(Flooding \mid Water\ level)$$

but not

$$P(Flooding \mid Rainfall) = P(Flooding)$$

Causal model: cause-effect chain.

If state of immediate cause is known, then state of more indirect cause does not affect probabilities for effect.



$$P(\textit{Hair length} \mid \textit{Stature}, \textit{Sex}) = P(\textit{Hair length} \mid \textit{Sex})$$

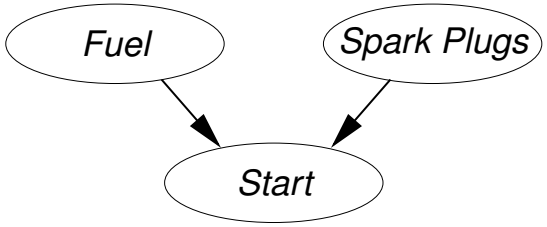
but not

$$P(\textit{Hair length} \mid \textit{Stature}) = P(\textit{Hair length})$$

Causal model: a common cause with two effects.

- if state of cause is not known:
  knowing that one effect has happened makes the second effect also more probable
- if state of cause is known:
  knowing about one effect does not affect the probabilities for the other effect

Val(*Fuel*)={*yes,no*}, *Val*(*Spark Plugs*)={*clean,dirty*}, *Val*(*Start*)={*yes,no*},



$$P(Fuel \mid Spark\ Plugs) = P(Fuel)$$

but not

$$P(Fuel \mid Spark\ Plugs, Start) = P(Fuel \mid Start)$$

Causal model: two distinct causes for one effect

- if state of effect is not known:
  state of one cause does not affect probabilities for other cause
- if state of effect is known:
  state of one cause changes probabilities of other cause (*explaining away*)

Constructing BNs using causal dependencies:

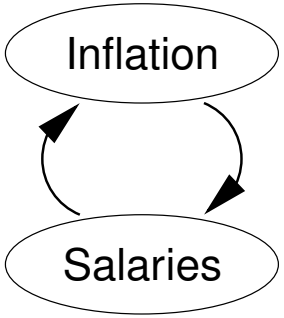
- specify dag so that edges represent direct cause-effect relations
- specify $P(X_i \mid \mathrm{Pa}(X_i))$ by assessing the probability for X_i for all combinations of its possible (direct) causes $\mathrm{Pa}(X_i)$.

Constructing BNs using causal dependencies:

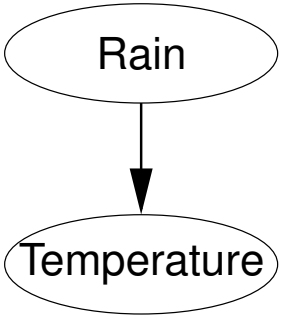
- specify dag so that edges represent direct cause-effect relations
- specify $P(X_i \mid \mathrm{Pa}(X_i))$ by assessing the probability for X_i for all combinations of its possible (direct) causes $\mathrm{Pa}(X_i)$.

This is very often how BNs are constructed. But:

Causal dependencies need not be acyclic:



There can be dependencies without causality:



The Naive Bayes Classifier

Example (Poole & Mackworth, Artificial Intelligence, 2010)

User preference data (will a user read a new posting on an online forum):

	Example	Author	Thread	Length	WhereRead	UserAction (=Class variable)
Training Data	e_1	known	new	long	home	skips
	e_2	unknown	new	short	work	reads
	e_3	unknown	follow up	long	work	skips
	e_4	known	follow up	long	home	skips
	e_5	known	new	short	home	reads
	e_6	known	follow up	long	work	skips
	e_7	unknown	follow up	short	work	skips
	e_8	unknown	new	short	work	reads
	e_9	known	follow up	long	home	skips
	e_{10}	known	new	long	work	skips
	e_{11}	unknown	follow up	short	home	skips
	e_{12}	known	new	long	work	skips
	e_{13}	known	follow up	short	home	reads
	e_{14}	known	new	short	work	reads
	e_{15}	known	new	short	home	reads
Test Data	e_{16}	known	follow up	short	work	?
	e_{17}	known	new	short	home	?
	e_{18}	unknown	new	short	work	?
	e_{19}	unknown	new	long	work	?
	e_{20}	unknown	follow up	long	home	?

Task: from the training data, learn a predictor for (future) test cases.

Classification Model

Joint probability distribution P on $Val(A_1) \times \ldots Val(A_n) \times Val(C)$

Classification Rule

$$\mathcal{C}(a_1, \ldots, a_n) \;:=\; arg\,max_{c \in Val(C)} P(C = c \mid A_1 = a_1, \ldots, A_n = a_n)$$

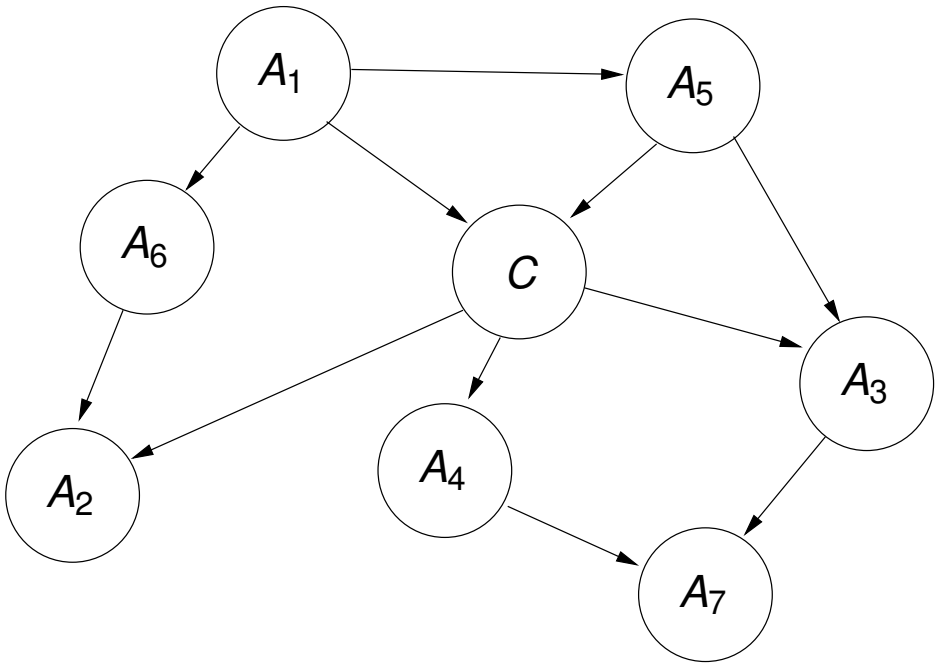
or for a specified *loss function* $Loss(c, c')$:

$$\mathcal{C}(a_1, \ldots, a_n) \;:=\; arg\,min_{c \in Val(C)} \sum_{c' \in Val(C)} P(C = c' \mid A_1 = a_1, \ldots, A_n = a_n) Loss(c, c')$$

This also works with incompletely observed instances: denote by $\tilde{\boldsymbol{A}} \subseteq \boldsymbol{A}$ the attributes for which values $A_i = a_i$ are observed. Then classify according to

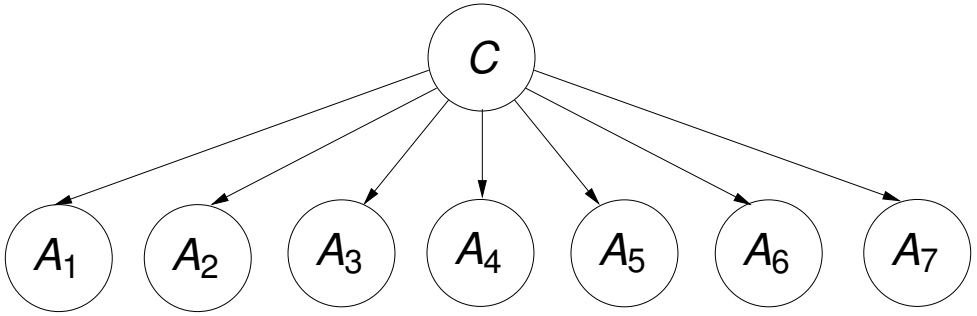
$$P(C = c \mid A_i = a_i \, (A_i \in \tilde{\boldsymbol{A}}))$$

Probabilistic classifiers differ in what classes of joint distributions P can be represented – and how. One possibility: P is represented using a (restricted type of) Bayesian network.



- Bayesian network can contain much information that is irrelevant for classification task. Only relevant: $P(C \mid A_1, \ldots, A_n)$.
- For classification typically restricted types of BNs are used.

The Naive Bayes Model is given by the BN structure:

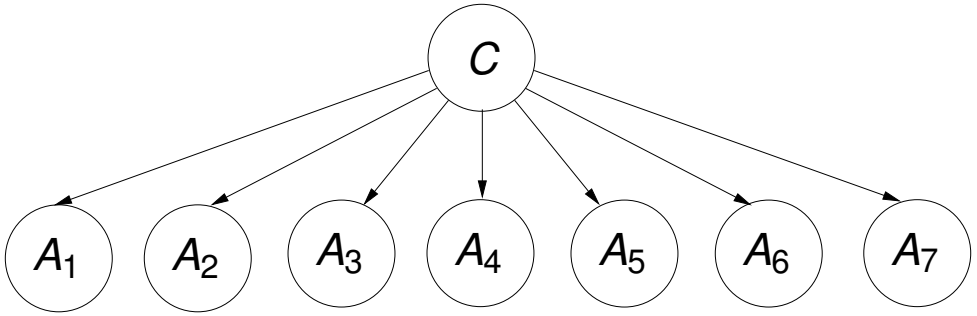


- Attributes are independent, given the class label.
- $$P(C, A_1, \ldots, A_n) = P(C) \cdot \prod_{i=1}^{n} P(A_i \mid C)$$

Learning Naive Bayes

- Model parameters simply learned by counting:

$$P(A_i = a \mid C = c) = \frac{\#\textit{training ex. with } A_i = a, C = c}{\#\textit{training ex. with } C = c}$$



Predicting with Naive Bayes

$$P(C = c \mid \boldsymbol{A} = \boldsymbol{a}) = \frac{P(C = c, \boldsymbol{A} = \boldsymbol{a})}{P(\boldsymbol{A} = \boldsymbol{a})}$$

$$arg\,max_c P(C = c \mid \boldsymbol{A} = \boldsymbol{a}) = arg\,max_c P(C = c, \boldsymbol{A} = \boldsymbol{a}) = arg\,max_c P(C) \cdot \prod_{i=1}^{n} P(A_i = a_i \mid C = c)$$

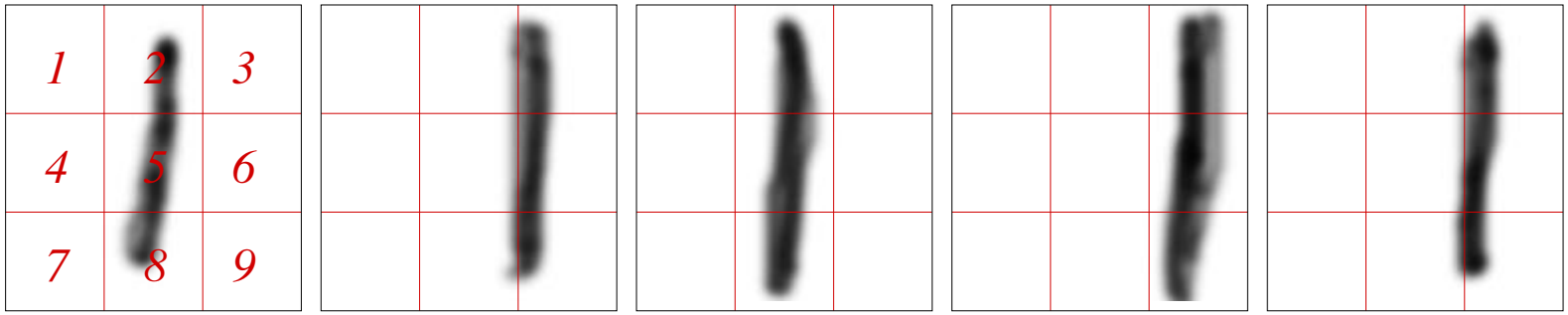
For incomplete data:

$$P(C = c, \tilde{\boldsymbol{A}} = \tilde{\boldsymbol{a}}) = P(C) \cdot \prod_{i:A_i \in \tilde{\boldsymbol{A}}} P(A_i = a_i \mid C = c)$$

In both cases: classification computable in time $O(n)$.

Class: $Symbol \in \{A, \ldots, Z, 0, \ldots, 9\}$

Predictors: $Cell\text{-}1, \ldots, Cell\text{-}9 \in \{b, w\}$.



For example:

$$P(Cell\text{-}2 = b \mid Cell\text{-}5 = b, Symbol = 1) > P(Cell\text{-}2 = b \mid Symbol = 1)$$

Attributes not independent given *Symbol*=1!

Class: $Spam \in \{y, n\}$

Predictors:
Subject-all-caps,*Known-spam-server*,. . . ,*Contains'Money'*,*Contains'Watch'*,. . . $\in \{y, n\}$.

For example:

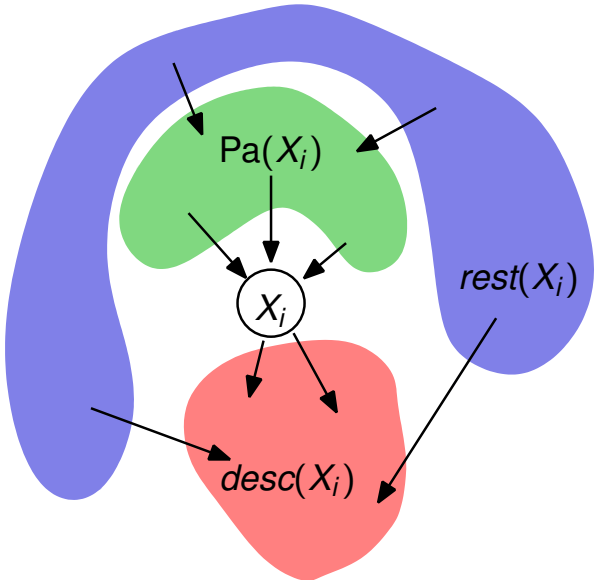
$$P(Contains'Nigeria' = y \mid Contains'Confidential' = y, Spam = y)$$
$$\gg$$
$$P(Contains'Nigeria' = y \mid Spam = y)$$

Attributes not independent given *Spam*=yes!

$\rightsquigarrow$ Naive Bayes assumption often not realistic. Nevertheless, Naive Bayes often successful.

D-Separation

The non-descendant criterion:



$$P_{\mathcal{N}}(X_i \mid \mathrm{Pa}(X_i), rest(X_i)) = P_{\mathcal{N}}(X_i \mid \mathrm{Pa}(X_i))$$

Questions:

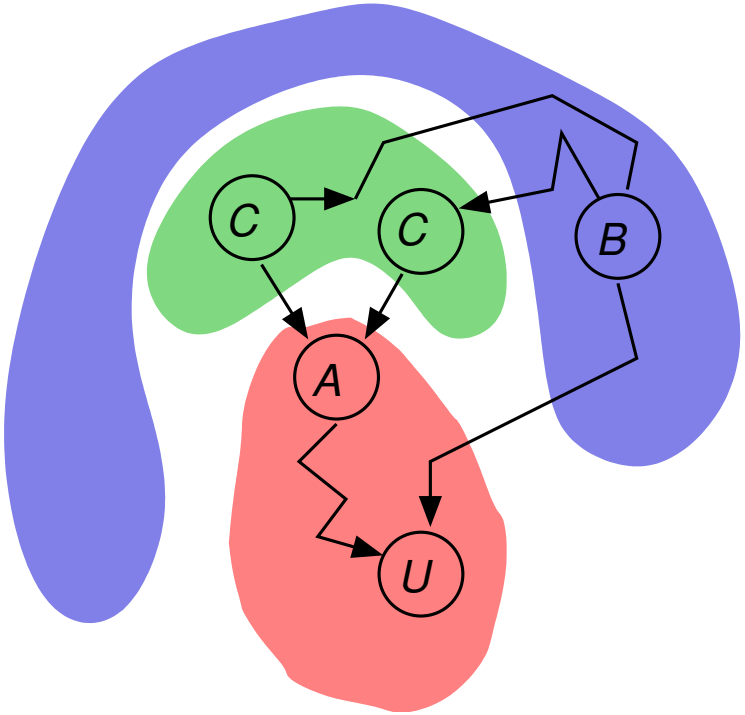
- What other conditional independence relations can be inferred from the graph structure of a Bayesian network?
- Is there a single, most general criterion?
- Given a distribution P and a BN graph $\mathcal{G}$, can P be represented via a suitable parameterization of $\mathcal{G}$?

$\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C} \subseteq V$ disjoint subsets of variables in a Bayesian network $\mathcal{G}$. $\boldsymbol{C}$ <u>d-separates</u> $\boldsymbol{A}$ from $\boldsymbol{B}$ if the following holds:

every undirected path that connects a node $A \in \boldsymbol{A}$ with a node $B \in \boldsymbol{B}$ satisfies at least one of the following two conditions:

1. the path contains a node $C \in \boldsymbol{C}$, and the edges that connect C are serial ($\ldots \rightarrow C \rightarrow \ldots$) or divergent ($\ldots \leftarrow C \rightarrow \ldots$).
2. the path contains a node U, the edges that connect U are convergent ($\ldots \rightarrow U \leftarrow \ldots$), and neither U nor any of the descendants of U belong to $\boldsymbol{C}$.

Pa(A) d-separates A from *rest*(A):



Correctness

For all pairwise disjoint sets $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ of nodes in a Bayesian network $\mathcal{N} = (\mathcal{G}, \boldsymbol{\theta})$:

$$\boldsymbol{C} \textit{ d-separates } \boldsymbol{A} \textit{ from } \boldsymbol{B} \textit{ in } \mathcal{G} \Rightarrow P_{\mathcal{N}}(\boldsymbol{A} \mid \boldsymbol{B}, \boldsymbol{C}) = P_{\mathcal{N}}(\boldsymbol{A} \mid \boldsymbol{C}).$$

(d-separation is a correct criterion to identify independencies)

Completeness

For all BN structures $\mathcal{G}$ there exists a parameterization $\boldsymbol{\theta}$, so that for $\mathcal{N} = (\mathcal{G}, \boldsymbol{\theta})$:

$$\boldsymbol{C} \textit{ d-separates } \boldsymbol{A} \textit{ from } \boldsymbol{B} \textit{ in } \mathcal{G} \Leftrightarrow P_{\mathcal{N}}(\boldsymbol{A} \mid \boldsymbol{B}, \boldsymbol{C}) = P_{\mathcal{N}}(\boldsymbol{A} \mid \boldsymbol{C})$$

A graph $\mathcal{G}$ over variables $\boldsymbol{X}$ is an *I-Map* for a distribution $P(\boldsymbol{X})$, if for all $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C} \subseteq \boldsymbol{X}$:

$$\boldsymbol{C} \textit{ d-separates } \boldsymbol{A} \textit{ from } \boldsymbol{B} \textit{ in } \mathcal{G} \Rightarrow P(\boldsymbol{A} \mid \boldsymbol{B}, \boldsymbol{C}) = P(\boldsymbol{A} \mid \boldsymbol{C}).$$

$\mathcal{G}$ is a *perfect map* for P, if

$$\boldsymbol{C} \textit{ d-separates } \boldsymbol{A} \textit{ from } \boldsymbol{B} \textit{ in } \mathcal{G} \Leftrightarrow P(\boldsymbol{A} \mid \boldsymbol{B}, \boldsymbol{C}) = P(\boldsymbol{A} \mid \boldsymbol{C}).$$

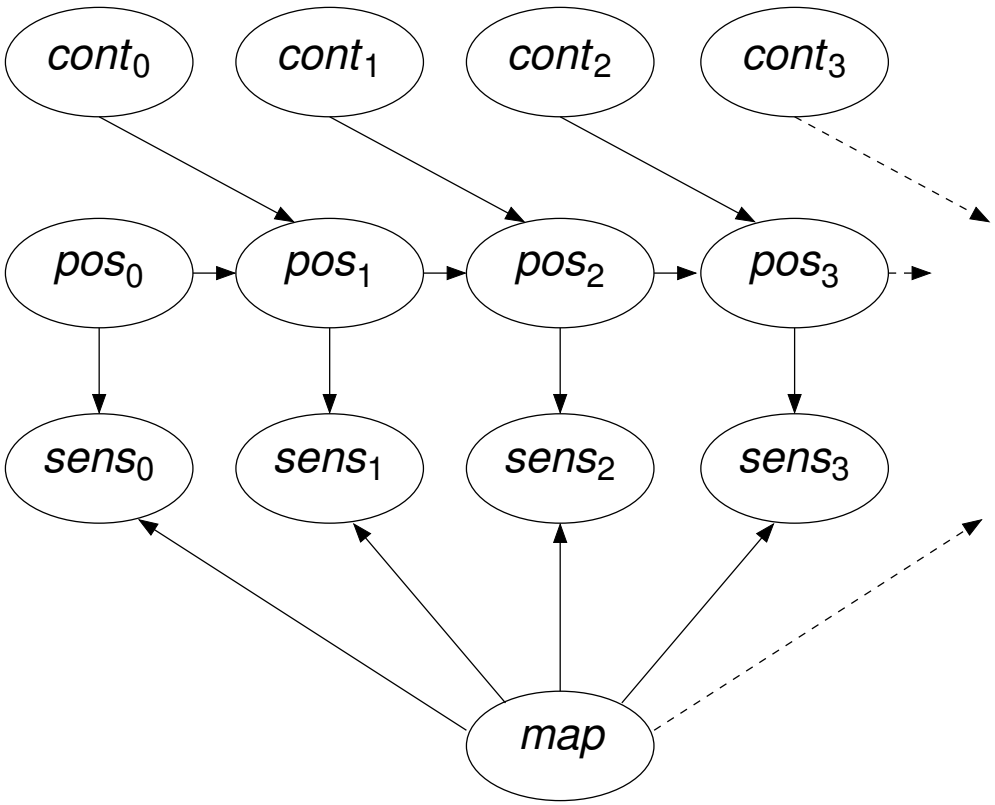
- For $\mathcal{N} = (\mathcal{G}, \boldsymbol{\theta})$: $\mathcal{G}$ is an *I*-map for $P_{\mathcal{N}}$.
- If $\mathcal{G}$ is an *I*-map for P, then there exist parameters $\boldsymbol{\theta}$, such that $P_{(\mathcal{G},\boldsymbol{\theta})} = P$
- For all $\mathcal{G}$ exists a parameterization $\boldsymbol{\theta}$, such that $\mathcal{G}$ is a perfect map for $P_{(\mathcal{G},\boldsymbol{\theta})}$

But:

- $\mathcal{G}$ is not always a perfect map for $P_{(\mathcal{G},\boldsymbol{\theta})}$ (one can have parameterizations $\boldsymbol{\theta}$ that lead to independencies in $P_{(\mathcal{G},\boldsymbol{\theta})}$ which cannot be read off the structure $\mathcal{G}$).
- There exist distributions P, for which no $\mathcal{G}$ is a perfect map.

Why is D-separation important?

- Gaining insight: given a (correct) Bayesian network model, can derive insight into the dependencies among the variables
- Debugging a model: given a Bayesian network model, check whether entailed independence relations are plausible
- Correctness of algorithms: certain computational procedures depend on validity of special independence relations



$cont_t$: control input at time t

pos_t: position at time t

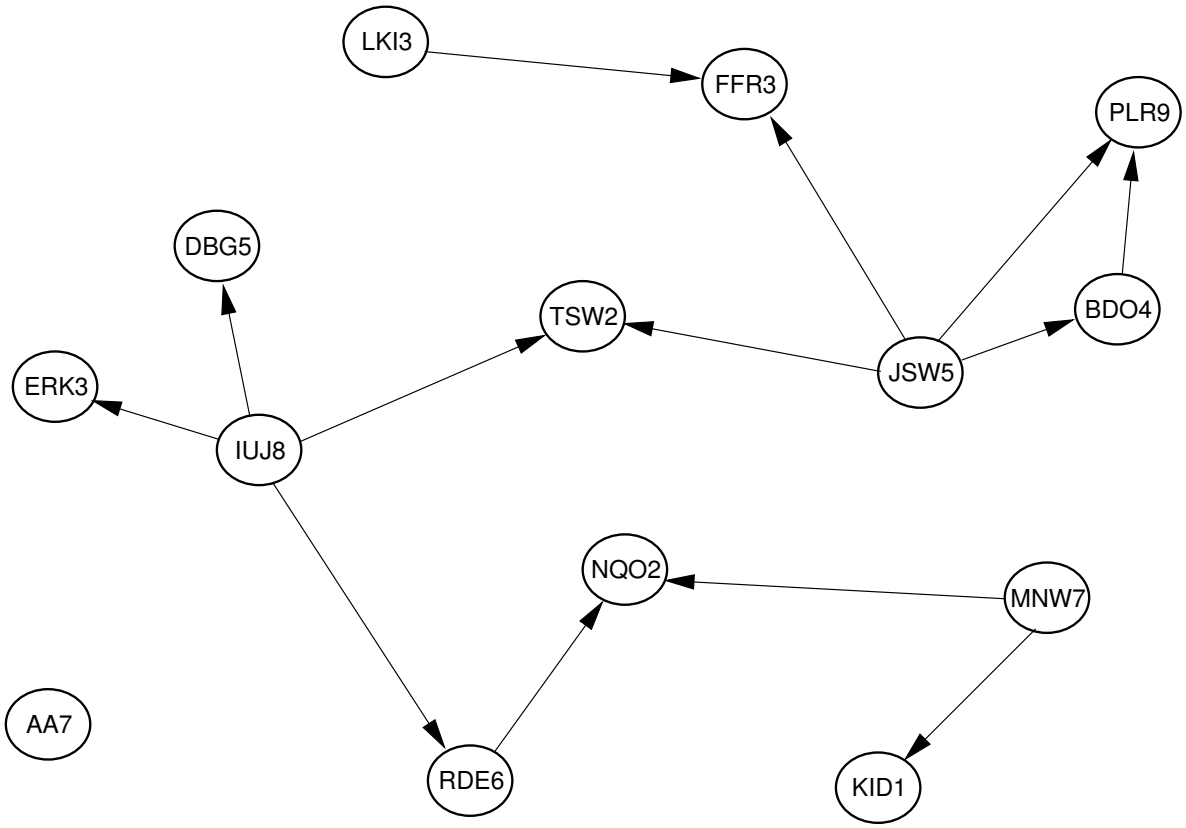
$sens_t$: sensor reading at time t

map: map of the environment

Independence properties extracted from the designed model:

pos_t is independent of $cont_{t-2}, sens_{t-1}$ given pos_{t-1}

pos_t is *not* independent of $sens_{t-1}$ given pos_{t-1} *and* $sens_t$



Independence relations discovered from the learned network:

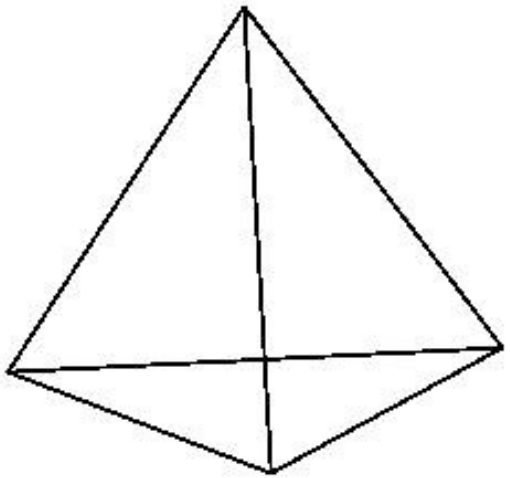
AA7 is independent of everything else

NQO2 is independent of *everything except* MNW7 *and* KID1 given RDE6

Example: two binary RV's A, B. Space of all joint distributions of A, B is

$$\Delta^4 = \{(p_1, p_2, p_3, p_4) \in [0,1]^4 \mid \sum_i p_i = 1\}$$

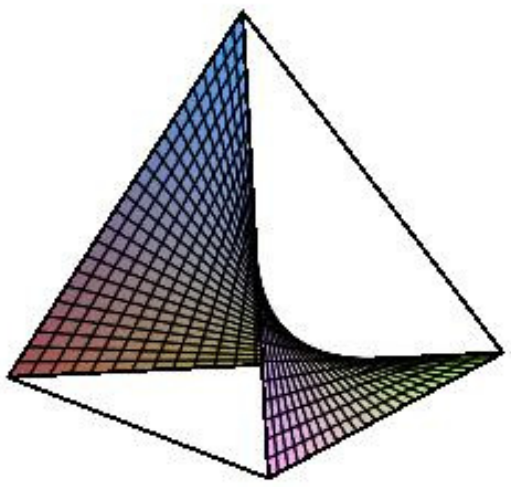
Δ^4 is a 3-dimensional tetrahedron in 4-dimensional space:



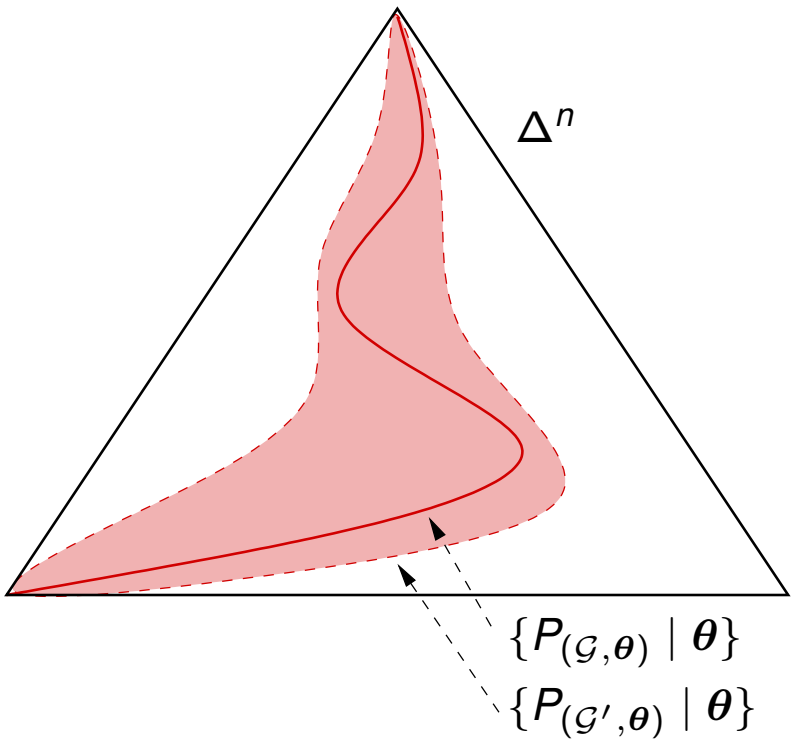
Example: two binary RV's A, B. Space of all joint distributions of A, B is

$$\Delta^4 = \{(p_1, p_2, p_3, p_4) \in [0,1]^4 \mid \sum_i p_i = 1\}$$

Δ^4 is a 3-dimensional tetrahedron in 4-dimensional space:



Distributions in which A, B are independent form a 2-dimensional manifold in Δ^4



- The distributions for which $\mathcal{G}$ is an *I*-map form a low-dimensional manifold in the probability polytope Δ^n.
- These are the distributions that can be represented using $\mathcal{G}$ with a suitable parameterization $\boldsymbol{\theta}$
- If $\mathcal{G}'$ is obtained from $\mathcal{G}$ by adding edges, then the distributions representable with $\mathcal{G}'$ are a superset of the distributions representable with $\mathcal{G}$.
- For any maximal DAG $\mathcal{G}^{max}$: $\{P_{(\mathcal{G}^{max},\boldsymbol{\theta})} \mid \boldsymbol{\theta}\} = \Delta^n$