Introduction to Probabilistic Graphical Models

Julius Pfrommer
julius.pfrommer@kit.edu

Karlsruher Institut of Technologie,
Vision and Fusion Laboratory (IES)

Fraunhofer IOSB

KIT Karlsruhe Institute of Technology

April 12, 2017

A ship has arrived in the harbor.
It has 4 masts and is carrying oranges.

How old is the captain?

A ship has arrived in the harbor.
It has 4 masts and is carrying oranges.

How old is the captain?

- Does the answer logically follow from the provided information?
- Can we make an educated guess?
- How sure are we about our response?

Discrete Probability Distributions

Random variable X with finite support $|\mathcal{X}| < \infty$.

$$P(X = x) \geq 0, \quad \sum\nolimits_x P(X = x) = 1$$

Chain Rule $$\begin{aligned} P(A, B) &= P(A \mid B) P(B) \\ &= P(B \mid A) P(A) \end{aligned}$$

Marginalization $$P(A) = \sum_{b \in \mathcal{B}} P(A, B = b)$$

Bayes' Theorem $$P(A \mid B) = \frac{P(B \mid A) P(A)}{P(B)}$$

The Aids Test

SPENDE BLUT
BEIM ROTEN KREUZ

- You are working at the Red Cross.
  Your job is to test blood donations for HIV.
- The probability of the test results T obviously depends on the infection of the sample I:

$P(T \mid I)$	$i_{\mathtt{t}}$	$i_{\mathtt{f}}$
$t_{\mathtt{t}}$	0.99	0.05
$t_{\mathtt{f}}$	0.01	0.95

- Around 0.2% of the general population are infected with HIV.

A blood donation comes up positive.
What are the chances it is actually infected?

The German Tank Problem

A sad part of German history…

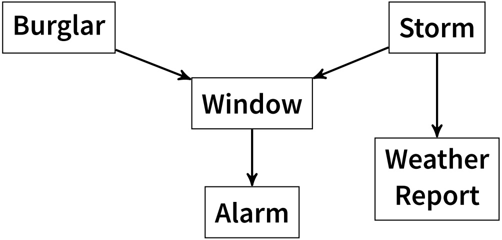
- Allied intelligence made estimates on the production capacity of the German wartime industry
- The Germans, always correct, labeled tanks with a serial number that included the date and the monthly build-count

Given the serial numbers of captured tanks built in June 1941, how many tanks did Germany actually build that month?

Inference from observations

Your vacation home in the French alps has an electronic alarm.

- The alarm notifies you when a window is opened
- Windows are opened by burglars or a storm
- Storm warnings are announced on the weather report



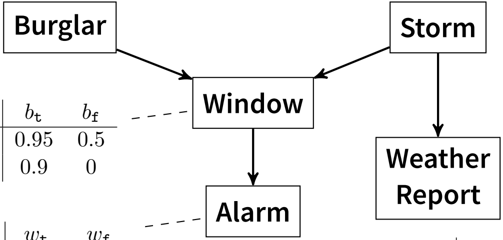
Inference from observations

Your vacation home in the French alps has an electronic alarm.

- The alarm notifies you when a window is opened
- Windows are opened by burglars or a storm
- Storm warnings are announced on the weather report

$P(B = b_{\mathtt{t}}) = 0.001$

$P(S = s_{\mathtt{t}}) = 0.01$



$P(W = w_{\mathtt{t}} \mid B, S)$	$b_{\mathtt{t}}$	$b_{\mathtt{f}}$
$s_{\mathtt{t}}$	0.95	0.5
$s_{\mathtt{f}}$	0.9	0

	$w_{\mathtt{t}}$	$w_{\mathtt{f}}$
$P(A = a_{\mathtt{t}} \mid W)$	0.95	0.01

	$s_{\mathtt{t}}$	$s_{\mathtt{f}}$
$P(R = r_{\mathtt{f}} \mid S)$	0.9	0.02

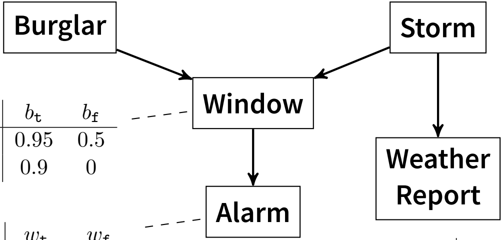
Inference from observations

Your vacation home in the French alps has an electronic alarm.

- The alarm notifies you when a window is opened
- Windows are opened by burglars or a storm
- Storm warnings are announced on the weather report

$P(B = b_{\mathtt{t}}) = 0.001$

$P(S = s_{\mathtt{t}}) = 0.01$



$P(W = w_{\mathtt{t}} \mid B, S)$	$b_{\mathtt{t}}$	$b_{\mathtt{f}}$
$s_{\mathtt{t}}$	0.95	0.5
$s_{\mathtt{f}}$	0.9	0

	$w_{\mathtt{t}}$	$w_{\mathtt{f}}$
$P(A = a_{\mathtt{t}} \mid W)$	0.95	0.01

	$s_{\mathtt{t}}$	$s_{\mathtt{f}}$
$P(R = r_{\mathtt{f}} \mid S)$	0.9	0.02

The alarm goes off with no storm warning. Was there a burglar?

Conditional Probability Tables (CPT)

Distributions with finite support can be represented as tables.

```
(report, storm) value
--------------- ------
(True, True)    0.009
(True, False)   0.0198
(False, True)   0.001
(False, False)  0.9702
```

3 ½ Operations on CPTs

Join Combine two CPT: $P(A \mid B), P(B) \Rightarrow P(A, B)$

Marginalize Reduce the domain of the CPT $P(A, B) \Rightarrow P(A)$

Eliminate Remove a variable with an observation
$P(A, B), P(B = b) = 1 \Rightarrow P(A \mid b)$

(Normalize) Ensure that $\sum_a P(A = a) = 1$

Computing the Posterior

The joint distribution over all variables is

$$P(B, S, W, A, R) = P(B)P(S)P(W \mid B, S)P(A \mid W)P(R \mid S)$$

Computing the Posterior

The joint distribution over all variables is

$$P(B, S, W, A, R) = P(B)P(S)P(W \mid B, S)P(A \mid W)P(R \mid S)$$

Compute the marginal distribution by "summing up"

$$\begin{aligned} P(b_{\mathtt{t}} \mid a_{\mathtt{t}}, r_{\mathtt{f}}) &= \frac{P(b_{\mathtt{t}}, a_{\mathtt{t}}, r_{\mathtt{f}})}{P(a_{\mathtt{t}}, r_{\mathtt{f}})} \\ &= \frac{\sum_{s \in \mathcal{S}, w \in \mathcal{W}} P(b_{\mathtt{t}}, s, w, a_{\mathtt{t}}, r_{\mathtt{f}})}{\sum_{b \in \mathcal{B}, s \in \mathcal{S}, w \in \mathcal{W}} P(b, s, w, a_{\mathtt{t}}, r_{\mathtt{f}})} \end{aligned}$$

Computing the Posterior

The joint distribution over all variables is

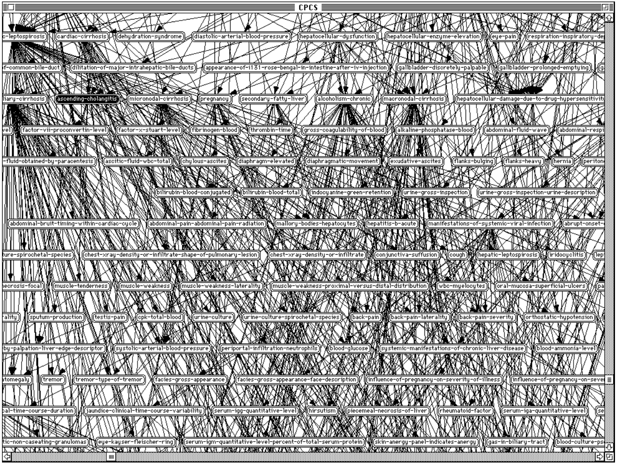
$$P(B, S, W, A, R) = P(B)P(S)P(W \mid B, S)P(A \mid W)P(R \mid S)$$

Compute the marginal distribution by "summing up"

$$\begin{aligned} P(b_{\mathtt{t}} \mid a_{\mathtt{t}}, r_{\mathtt{f}}) &= \frac{P(b_{\mathtt{t}}, a_{\mathtt{t}}, r_{\mathtt{f}})}{P(a_{\mathtt{t}}, r_{\mathtt{f}})} \\ &= \frac{\sum_{s \in \mathcal{S}, w \in \mathcal{W}} P(b_{\mathtt{t}}, s, w, a_{\mathtt{t}}, r_{\mathtt{f}})}{\sum_{b \in \mathcal{B}, s \in \mathcal{S}, w \in \mathcal{W}} P(b, s, w, a_{\mathtt{t}}, r_{\mathtt{f}})} \end{aligned}$$

Having a standard procedure is nice.
But this is way too inefficient!
Can we exploit the structure of the problem?

Realistic models are often very complex



Parker and Miller (1988): “Using Causal Knowledge to Create Simulated Patient Cases: CPCS Project as an Extension of INTERNIST-1”

Who ya gonna call?

- Pearl, J. (1984). **Heuristics: intelligent search strategies for computer problem solving**. Addison-Wesley.
- Pearl, J. (1988). **Probabilistic reasoning in intelligent systems: networks of plausible inference**. Morgan Kaufmann.
- Pearl, J. (2000). **Causality**. Cambridge university press.

Representing Conditional Independence

Conditional Independence

$$P(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z}) = P(\mathbf{X} \mid \mathbf{Z}) P(\mathbf{Y} \mid \mathbf{Z}) \Leftrightarrow (\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$$

Representing Conditional Independence

Conditional Independence

$$P(\mathbf{X}, \mathbf{Y} \,|\, \mathbf{Z}) = P(\mathbf{X} \,|\, \mathbf{Z}) P(\mathbf{Y} \,|\, \mathbf{Z}) \Leftrightarrow (\mathbf{X} \perp \mathbf{Y} \,|\, \mathbf{Z})$$

Inferring Conditional Independence
Let P a distribution and $G = (\mathbf{X}, \mathcal{E})$ a directed acyclic graph of

- Random variables $\mathbf{X}$
- Dependence relations $(i, j) \in \mathcal{E} \subseteq \mathbf{X}^2, i \neq j$.
  If $(i, j) \in \mathcal{E}$, then X_j and X_i are dependent variables on P.

Representing Conditional Independence

Conditional Independence

$$P(\mathbf{X}, \mathbf{Y} \,|\, \mathbf{Z}) = P(\mathbf{X} \,|\, \mathbf{Z}) P(\mathbf{Y} \,|\, \mathbf{Z}) \Leftrightarrow (\mathbf{X} \perp \mathbf{Y} \,|\, \mathbf{Z})$$

Inferring Conditional Independence

Let P a distribution and $G = (\mathbf{X}, \mathcal{E})$ a directed acyclic graph of

- Random variables $\mathbf{X}$
- Dependence relations $(i, j) \in \mathcal{E} \subseteq \mathbf{X}^2, i \neq j$.
  If $(i, j) \in \mathcal{E}$, then X_j and X_i are dependent variables on P.

With $\mathrm{Par}_G(X_i)$ the direct parent nodes of X_i and $\mathrm{Desc}_G(X_i)$ its descendants (including X_i itself), the graph G implies

$$(X_i \perp \mathbf{X} \setminus \mathrm{Desc}_G(X_i) \,|\, \mathrm{Par}_G(X_i)) \,.$$

Bayesian Networks

Definition of Bayesian Networks

A Bayesian Network $\mathcal{B} = (P, G)$ consists of

- a distribution P defined for a set of random variables $\mathbf{X}$,
- a graph $G = (\mathbf{X}, \mathcal{E})$,

where the conditional independences implied by G are a subset of the actual conditional independences on P.

Bayesian Networks

Definition of Bayesian Networks

A Bayesian Network $\mathcal{B} = (P, G)$ consists of

- a distribution P defined for a set of random variables $\mathbf{X}$,
- a graph $G = (\mathbf{X}, \mathcal{E})$,

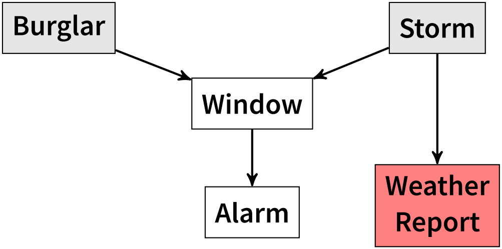
where the conditional independences implied by G are a subset of the actual conditional independences on P.

For any BN, the distribution P has a *factored representation*:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \mathrm{Par}_G(X_i))$$

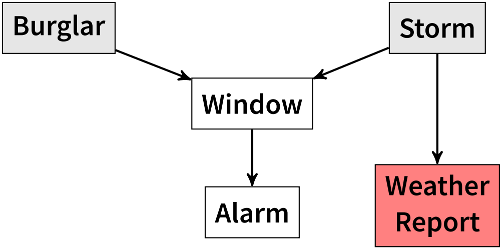
$$P(B, S, W, A, R) = P(B)P(S)P(W \mid B, S)P(A \mid W)P(R \mid S)$$

Conditional Independence Quiz I



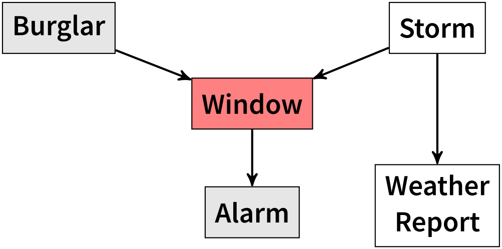
Weather Report is observed. Are *Burglar* and *Storm* conditionally independent?

Conditional Independence Quiz I



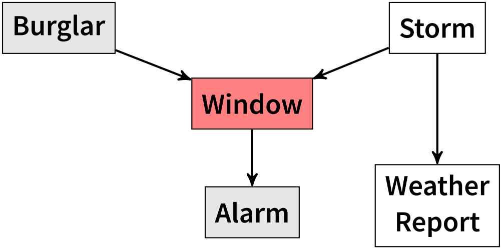
Weather Report is observed. Are *Burglar* and *Storm* conditionally independent?

Conditional Independence Quiz II



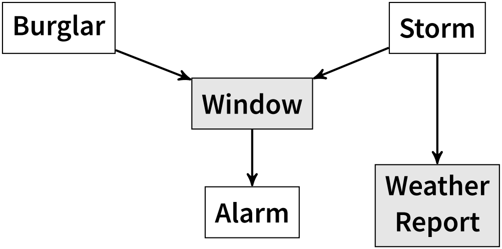
Window is observed. Are *Burglar* and *Alarm* conditionally independent?

Conditional Independence Quiz II



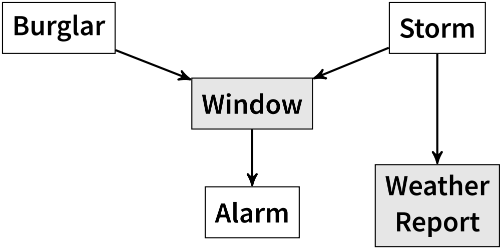
Window is observed. Are *Burglar* and *Alarm* conditionally independent?

Conditional Independence Quiz III



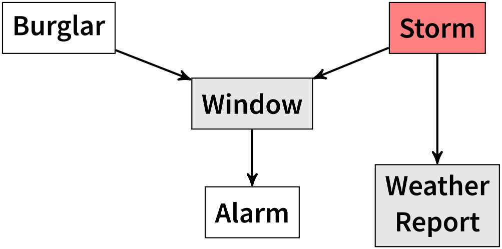
No observations are made. Are *Weather Report* and *Window* conditionally independent?

Conditional Independence Quiz III



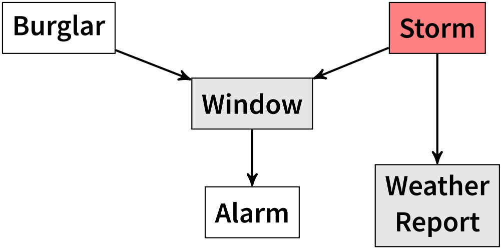
No observations are made. Are *Weather Report* and *Window* conditionally independent?

Conditional Independence Quiz IV



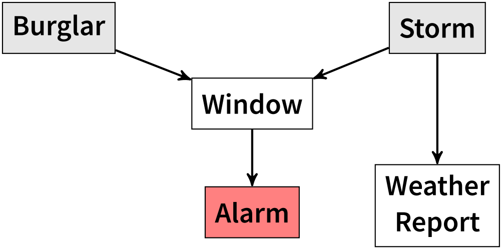
Storm is observed. Are *Weather Report* and *Window* conditionally independent?

Conditional Independence Quiz IV



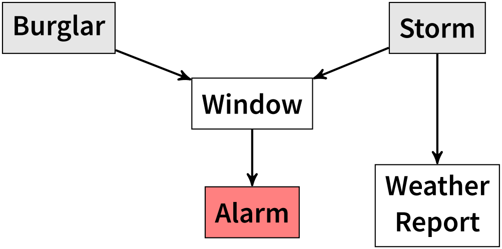
Storm is observed. Are *Weather Report* and *Window* conditionally independent?

Conditional Independence Quiz V



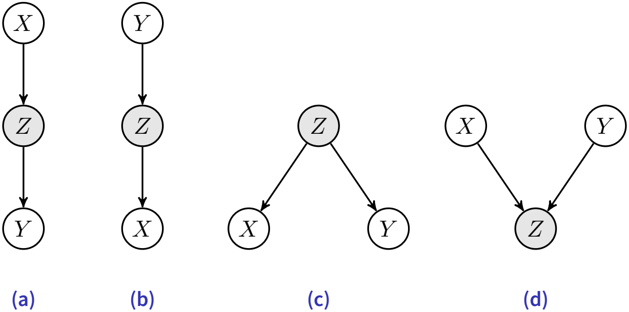
Alarm is observed. Are *Burglar* and *Storm* conditionally independent?

Conditional Independence Quiz V



Alarm is observed. Are *Burglar* and *Storm* conditionally independent?

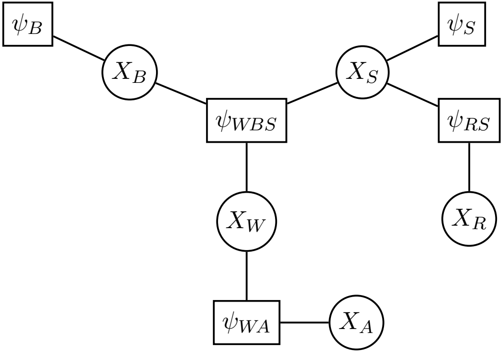
Condition Independence with Observations



Y is independent of X if …

a) **Causal Chain:** Z **is** observed
b) **Evidence Chain:** Z **is** observed
c) **Common Cause:** Z **is** observed
d) **Common Effect:** $\mathrm{Desc}_G(Z)$ **is not** observed

Factor Graphs



- Bipartite undirected **Factor Graph** $G = (V, F, \mathcal{E})$
- Variables $X_i, i \in V$ with Domain $\mathcal{X}_i$
- Factor-Functions $\psi_j : \mathcal{X}_j \to \mathbb{R}$ depend on variables $j \subseteq V$
- Undirected Edges $\mathcal{E} = \{(i, j) : i \in j\}$
- Neighbors are $j \in N(i)$ and $i \in j$

Belief Propagation

Messages are exchanged between variables i and factors j.
Messages are "ad-hoc" factors with domain $\mathcal{X}_i$.

$$m_{i \to j}(x_i) = \prod_{k \in N(i) \setminus j} m_{k \to i}(x_i)$$

$$m_{j \to i}(x_i) = \underbrace{\sum_{\substack{\mathbf{x}_j \in \mathcal{X}_j: \\ \mathbf{x}_{j_i} = x_i}}}_{\text{Marginalize}} \underbrace{\left[\psi_j(\mathbf{x}_j) \prod_{h \in j \setminus i} m_{h \to j}(x_{j_h}) \right]}_{\text{Join}}$$

Belief Propagation

Messages are exchanged between variables i and factors j.
Messages are “ad-hoc” factors with domain $\mathcal{X}_i$.

$$m_{i\to j}(x_i) = \prod_{k\in N(i)\setminus j} m_{k\to i}(x_i)$$

$$m_{j\to i}(x_i) = \underbrace{\sum_{\substack{\mathbf{x}_j\in\mathcal{X}_j:\\ \mathbf{x}_{j_i}=x_i}}}_{\text{Marginalize}} \underbrace{\left[\psi_j(\mathbf{x}_j) \prod_{h\in j\setminus i} m_{h\to j}(x_{j_h})\right]}_{\text{Join}}$$

Forward-Backward Schedule

- Message to neighbor j after receiving messages from $N(i) \setminus j$
- Converges after $2|\mathcal{E}|$ messages on **tree-structured** Factor Graphs
- The marginal distribution of variable X_i is

$$P(x_i) = \prod_{j\in N(i)} m_{j\to i}(x_i)$$

Where to go from here?

- **Dynamic Bayes Nets**: Observations over several time steps
  - Hidden Markov Models are the simplest Dynamic Bayes Nets
- **Continuous Domains**: Inference with Gaussians / Exponential Family Distributions (Wainwright, Jordan, et al., 2008)
- **Loopy Graphs** (Yedidia et al., 2005)
- **Undirected Models**: Markov Random Fields
- **Generalized Distributive Law**: Belief Propagation on General Algebraic Semirings
  - Turbo Codes (McEliece et al., 1998)
  - Fourier Transform (Kschischang et al., 2001)
  - SLAM (Paskin, 2002)
  - Queries in Relational Databases (Green et al., 2007)
  - Kalman Filter (Bickson et al., 2008)
  - SAT (Bacchus et al., 2009)
  - Sum-Product Networks (Friesen and Domingos, 2016)

Thanks for Listening.
Questions?

Karlsruhe Machine Learning,
Statistics and AI Meetup

References I

Bacchus, Fahiem, Shannon Dalmao, and Toniann Pitassi (2009). "Solving# SAT and Bayesian inference with backtracking search". In: *Journal of Artificial Intelligence Research* 34, pp. 391–442.

Bickson, Danny, Ori Shental, and Danny Dolev (2008). "Distributed Kalman filter via Gaussian belief propagation". In: *Communication, Control, and Computing, 2008 46th Annual Allerton Conference on*. IEEE, pp. 628–635.

Friesen, Abram L and Pedro Domingos (2016). "The sum-product theorem: A foundation for learning tractable models". In: *Proceedings of ICML*, pp. 1909–1918.

Green, Todd J, Grigoris Karvounarakis, and Val Tannen (2007). "Provenance semirings". In: *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, pp. 31–40.

References II

Kschischang, Frank R, Brendan J Frey, and H-A Loeliger (2001). “Factor graphs and the sum-product algorithm”. In: *Information Theory, IEEE Transactions on* 47.2, pp. 498–519.

McEliece, Robert J, David JC MacKay, and Jung-Fu Cheng (1998). “Turbo decoding as an instance of Pearl’s “belief propagation” algorithm”. In: *Selected Areas in Communications, IEEE Journal on* 16.2, pp. 140–152.

Parker, Ronnie C and Randolph A Miller (1988). “Using Causal Knowledge to Create Simulated Patient Cases: CPCS Project as an Extension of INTERNIST-1”. In: *Selected Topics in Medical Artificial Intelligence*. Springer, pp. 99–115.

Paskin, Mark A (2002). “Thin Junction Tree Filters for Simultaneous Localization and Mapping”. In: *Computer Science*.

References III

Wainwright, Martin J, Michael I Jordan, et al. (2008). "Graphical models, exponential families, and variational inference". In: *Foundations and Trends® in Machine Learning* 1.1–2, pp. 1–305.

Yedidia, Jonathan S, William T Freeman, and Yair Weiss (2005). "Constructing free-energy approximations and generalized belief propagation algorithms". In: *Information Theory, IEEE Transactions on* 51.7, pp. 2282–2312.