Bayesian Networks

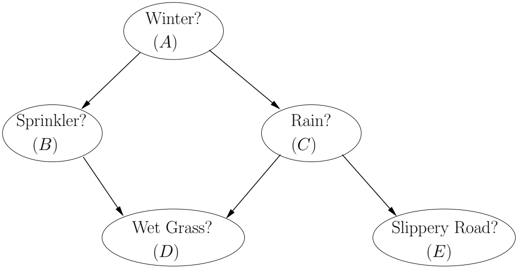
Brandon Malone

Much of this material is adapted from Chapter 4 of Darwiche's book

January 23, 2014

Graph concepts and terminology

We have a **directed acyclic graph** in which the set of **nodes** represent **random variables**, $\mathcal{X}$.



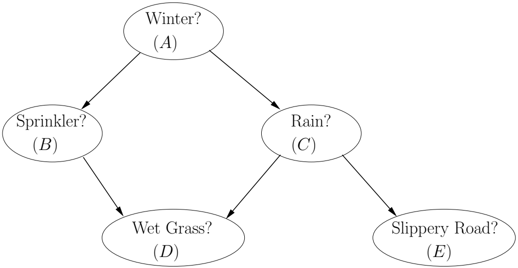
Pa_X: the **parents** of variable/node X

$Desc_X$: the **descendents** of X

$NonDesc_X$: the **non-descendents** of X, $\mathcal{X} \setminus \{X\} \setminus Pa_X \setminus Desc_X$

Graph concepts and terminology

We have a **directed acyclic graph** in which the set of **nodes** represent **random variables**, $\mathcal{X}$.



Trail or **pipe**. Any sequence of **edges** which connects two variables
Example: *Sprinkler* $\rightarrow$ *Wet Grass* $\leftarrow$ *Rain* $\rightarrow$ *Slippery Road*
N.B. The *direction* of the edge is not considered.

Valve. A variable in a trail

Probability terminology and notation

We have a **conditional probability distribution** represented as a table, called a **conditional probability table**.

A	B	$\Theta_{B\|A}$
T	T	0.20
T	F	0.80
F	T	0.75
F	F	0.25

Family. The variable X and its parents Pa_X, B and $\{A\}$ here

Parameters. The conditional probability distributions, $Pr(X = x|Pa_X = pa)$, often denoted $\theta_{x|pa}$

Each instantiation of Pa_X gives a different conditional distribution for X, so $\sum_x \theta_{x|pa} = 1$ for each pa.

Probability terminology and notation

We have a **conditional probability distribution** represented as a table, called a **conditional probability table**.

A	B	$\Theta_{B\|A}$
T	T	0.20
T	F	0.80
F	T	0.75
F	F	0.25

Compatability. A parameter $\theta_{x|pa}$ is compatible with a (partial) instantiation $\mathbf{z}$ if they assign the same value to common variables. We use $\theta_{x|pa} \sim \mathbf{z}$ to indicate compatibility.

Conditional independence. $I(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ means that $\mathbf{X}$ is independent of $\mathbf{Y}$ given $\mathbf{Z}$.

Factorized distributions

How can we use chain rule to write $Pr(A, B, C, D, E)$?

Factorized distributions

How can we use chain rule to write $Pr(A, B, C, D, E)$?

$$Pr(A, B, C, D, E) = Pr(A)Pr(B|A)Pr(C|A, B)Pr(D|A, B, C)Pr(E|A, B, C, D)$$

How many parameters does this require?

Factorized distributions

How can we use chain rule to write $Pr(A, B, C, D, E)$?

$$Pr(A, B, C, D, E) = Pr(A)Pr(B|A)Pr(C|A, B)Pr(D|A, B, C)Pr(E|A, B, C, D)$$

How many parameters does this require?

What if $I(E, \{C\}, \{A, B, D\})$?

Factorized distributions

How can we use chain rule to write $Pr(A, B, C, D, E)$?

$$Pr(A, B, C, D, E) = Pr(A)Pr(B|A)Pr(C|A, B)Pr(D|A, B, C)Pr(E|A, B, C, D)$$

How many parameters does this require?

What if $I(E, \{C\}, \{A, B, D\})$?

$$Pr(A, B, C, D, E) = Pr(A)Pr(B|A)Pr(C|A, B)Pr(D|A, B, C)Pr(E|C)$$

How many parameters does this require?

Factorized distributions

How can we use chain rule to write $Pr(A, B, C, D, E)$?

$$Pr(A, B, C, D, E) = Pr(A)Pr(B|A)Pr(C|A, B)Pr(D|A, B, C)Pr(E|A, B, C, D)$$

How many parameters does this require?

What if $I(E, \{C\}, \{A, B, D\})$?

$$Pr(A, B, C, D, E) = Pr(A)Pr(B|A)Pr(C|A, B)Pr(D|A, B, C)Pr(E|C)$$

How many parameters does this require?

What if (additionally) $I(C, \{A\}, \{B\})$ and $I(D, \{B, C\}, \{A\})$?

Factorized distributions

How can we use chain rule to write $Pr(A, B, C, D, E)$?

$$Pr(A, B, C, D, E) = Pr(A)Pr(B|A)Pr(C|A, B)Pr(D|A, B, C)Pr(E|A, B, C, D)$$

How many parameters does this require?

What if $I(E, \{C\}, \{A, B, D\})$?

$$Pr(A, B, C, D, E) = Pr(A)Pr(B|A)Pr(C|A, B)Pr(D|A, B, C)Pr(E|C)$$

How many parameters does this require?

What if (additionally) $I(C, \{A\}, \{B\})$ and $I(D, \{B, C\}, \{A\})$?

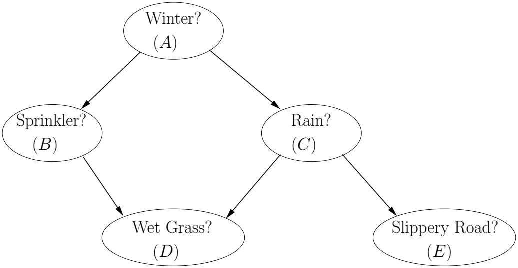
$$Pr(A, B, C, D, E) = Pr(A)Pr(B|A)Pr(C|A)Pr(D|B, C)Pr(E|C)$$

How many parameters does this require?

Graphical structures as factorized distributions

What is the relationship between the factorized distribution and the graphical structure?

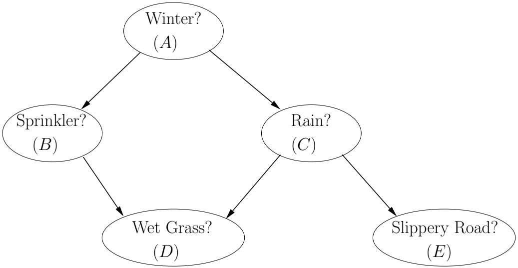
$$Pr(A, B, C, D, E) = Pr(A)Pr(B|A)Pr(C|A)Pr(D|B, C)Pr(E|C)$$



Graphical structures as factorized distributions

What is the relationship between the factorized distribution and the graphical structure?

$$Pr(A, B, C, D, E) = Pr(A)Pr(B|A)Pr(C|A)Pr(D|B, C)Pr(E|C)$$



The graph structure encodes the conditional independencies.

Local Markov property

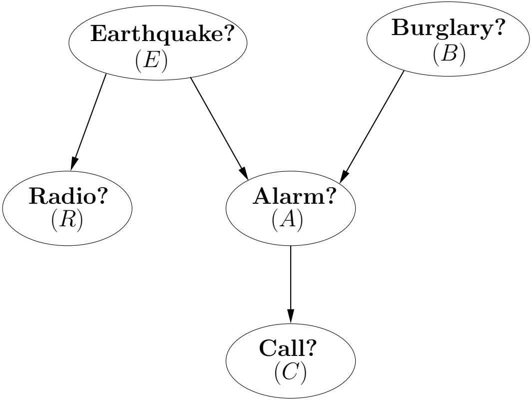
Local Markov property. Given a DAG structure G which encodes conditional independencies, we interpret G to compactly represent the following independence statements:

$$I(X, Pa_X, NonDesc_X) \quad \text{for all variables } X \text{ in DAG } G.$$

These conditional independencies are denoted as $Markov(G)$.

Local Markov property - Simple example

What conditional independencies are implied by the local Markov property for this DAG?



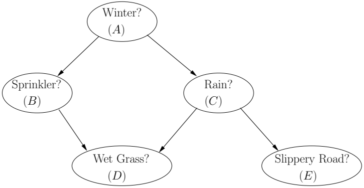
Bayesian network definition

A **Bayesian network** for a set of random variables $\mathcal{X}$ is a pair (G, Θ) where

- G is a DAG over $\mathcal{X}$, called the *structure*
- Θ is a set of CPDs (always CPTs in this course), one for each $X \in \mathcal{X}$, called the *parameterization*

Sample Bayesian network

Structure	Parameterization



A	Θ_A
T	.6
F	.4

A	B	$\Theta_{B\|A}$
T	T	.2
T	F	.8
F	T	.75
F	F	.25

C	E	$\Theta_{E\|C}$
T	T	.7
T	F	.3
F	T	0
F	F	1

A	C	$\Theta_{C\|A}$
T	T	.8
T	F	.2
F	T	.1
F	F	.9

B	C	D	$\Theta_{D\|B,C}$
T	T	T	.95
T	T	F	.05
T	F	T	.9
T	F	F	.1
F	T	T	.8
F	T	F	.2
F	F	T	0
F	F	F	1

Chain rule of Bayesian networks

Given a Bayesian network $\mathcal{B}$ and an instantiation $\mathbf{z}$, then

$$Pr(\mathbf{z}|\mathcal{B}) = \prod_{\theta_{x|pa} \sim \mathbf{z}} \theta_{x|pa}$$

That is, the probability of $\mathbf{z}$ is the probability of each variable given its parents.

In the future, we will typically omit $\mathcal{B}$ unless we need to distinguish between networks.

Class work

Suppose a Bayesian network has n variables, and each variable can take up to d values. Additionally, no variable has more than k parents.

How many parameters does an explicit distribution require?

How many parameters does a Bayesian network require? Use $O(\cdot)$.

Using the network on the handout, compute the following probabilities. Remember marginalization and Bayes' rule.

- $Pr(A = \mathrm{T}, B = \mathrm{T}, C = \mathrm{F}, D = \mathrm{T}, E = \mathrm{F})$
- $Pr(A = \mathrm{T}, B = \mathrm{T}, C = \mathrm{F})$
- $Pr(A = \mathrm{T}, B = \mathrm{T} | C = \mathrm{F})$
- $Pr(A = \mathrm{T}, B = \mathrm{T} | C = \mathrm{F}, D = \mathrm{T}, E = \mathrm{F})$

Graphoid axioms

The local Markov property tells us that

$$I(X, Pa_X, NonDesc_X) \quad \text{for all variables } X \text{ in DAG } G.$$

The **graphoid axioms** allow us to derive *global* independencies based on the graph structure.

- Symmetry
- Decomposition
- Weak union
- Contraction
- Intersection

Symmetry

If learning something about **Y** tells us nothing about **X**, then learning something about **X** tells us nothing about **Y**.

$$I(\mathbf{X}, \mathbf{Z}, \mathbf{Y}) \Leftrightarrow I(\mathbf{Y}, \mathbf{Z}, \mathbf{X})$$

Note that conditional independence is always w.r.t. some set of variables **Z** as evidence.

Decomposition

If learning something about $\mathbf{Y} \cup \mathbf{W}$ tells us nothing about $\mathbf{X}$, then learning something about $\mathbf{Y}$ or $\mathbf{W}$ individually tells us nothing about $\mathbf{X}$.

$$I(\mathbf{X}, \mathbf{Z}, \mathbf{Y} \cup \mathbf{W}) \Rightarrow I(\mathbf{X}, \mathbf{Z}, \mathbf{Y}) \text{ and } I(\mathbf{X}, \mathbf{Z}, \mathbf{W})$$

This allows us to reason about subsets. In particular,

$$I(X, Pa_X, \mathbf{W}) \quad \text{for all } \mathbf{W} \in NonDesc_X.$$

Given the topological ordering on the variables, this axiom proves the chain rule for Bayesian networks.

Weak union

If learning something about $\mathbf{Y} \cup \mathbf{W}$ tells us nothing about $\mathbf{X}$, then $\mathbf{Y}$ will not make $\mathbf{W}$ relevant.

$$I(\mathbf{X}, \mathbf{Z}, \mathbf{Y} \cup \mathbf{W}) \Rightarrow I(\mathbf{X}, \mathbf{Z} \cup \mathbf{Y}, \mathbf{W})$$

In particular, X is independent of $\mathbf{W} \in NonDesc_X$ given Pa_X and the other non-descendents.

Contraction

If learning something about $\mathbf{W}$ after learning $\mathbf{Y}$ tells us nothing about $\mathbf{X}$, then the combined information $\mathbf{Y} \cup \mathbf{W}$ was irrelevant to begin with.

$$I(\mathbf{X}, \mathbf{Z}, \mathbf{Y}) \text{ and } I(\mathbf{X}, \mathbf{Z} \cup \mathbf{Y}, \mathbf{W}) \Rightarrow I(\mathbf{X}, \mathbf{Z}, \mathbf{Y} \cup \mathbf{W})$$

Intersection

If **W** is irrelevant given **Y** and **Y** is irrelevant given **W**, then both **Y** and **W** were irrelevant to begin with*.

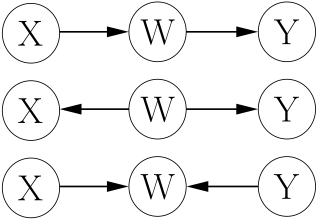
$$I(\mathbf{X}, \mathbf{Z} \cup \mathbf{W}, \mathbf{Y}) \text{ and } I(\mathbf{X}, \mathbf{Z} \cup \mathbf{Y}, \mathbf{W}) \Rightarrow I(\mathbf{X}, \mathbf{Z}, \mathbf{Y} \cup \mathbf{W})$$

* This holds only when the distribution is strictly positive.

Paths and valves

A **pipe** is a path from one variable to another.

Three types of **valves** compose a pipe.



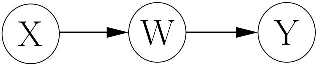
- Chain
- Common cause
- Common effect

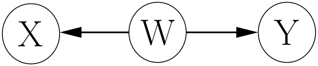
The "common effect" is often referred to as a "v-structure" when X and Y are not connected.

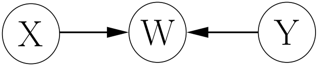
Open and closed valves

We can consider independence as “flow” through a pipe.

In particular, **X** and **Y** are independent given **Z** if all pipes between them are closed. A pipe is closed if any of its valves are closed.

- 

Closed iff $W \in \mathbf{Z}$
- 

Closed iff $W \in \mathbf{Z}$
- 

Closed iff $W, Desc_W \notin \mathbf{Z}$

Formally, this is called **d-separation** and is written $dsep(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$.

Complexity of d-separation

How many paths are there between nodes in **X** and **Y**?

So is d-separation practically useful?

Complexity of d-separation

How many paths are there between nodes in $\mathbf{X}$ and $\mathbf{Y}$?

So is d-separation practically useful?

Testing $dsep(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ is equivalent to testing if $\mathbf{X}$ and $\mathbf{Y}$ are connected in a new graph.

- Delete outgoing edges from nodes in $\mathbf{Z}$
- (Recursively) Delete any leaf which does not belong to $\mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z}$

So we can determine d-separation in linear time and space.

Markov blanket

The **Markov blanket** for a variable X is a set of variables $\mathbf{B}$ such that $X \notin \mathbf{B}$ and $I(X, \mathbf{B}, \mathbf{V} \setminus \mathbf{B} \setminus \{X\})$.

Which variables $\mathbf{B}$ d-separate a variable X from all of the other variables $(\mathbf{V} \setminus \mathbf{B} \setminus \{X\})$?

Markov blanket

The **Markov blanket** for a variable X is a set of variables $\mathbf{B}$ such that $X \notin \mathbf{B}$ and $I(X, \mathbf{B}, \mathbf{V} \setminus \mathbf{B} \setminus \{X\})$.

Which variables $\mathbf{B}$ d-separate a variable X from all of the other variables $(\mathbf{V} \setminus \mathbf{B} \setminus \{X\})$?

The parents, children and spouses.

Soundness, completeness and equivalence

Every independence found by d-separation is true for any distribution which factorizes according to the BN.

There *could* be independencies that d-separation cannot find (because it only uses the structure).

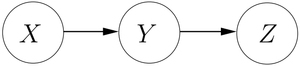
What are the independencies given by these networks?

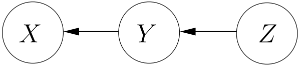
Soundness, completeness and equivalence

Every independence found by d-separation is true for any distribution which factorizes according to the BN.

There *could* be independencies that d-separation cannot find (because it only uses the structure).

What are the independencies given by these networks?





Different network structures can result in the same independencies. These networks are **Markov equivalent**.

Terminology for d-separation

A BN is an **independence map** (I-MAP) of *Pr* if every independence declared by d-separation holds in *Pr*.

An I-MAP is **minimal** if it ceases to be an I-MAP when any edge is deleted.

A BN is a **dependency map** (D-MAP) of *Pr* if the lack of d-separation implies a dependence in *Pr*.

Class work

Use the network on the handout (the Asian network) to answer the following independence questions.

- List the Markov blanket of all variables.
- $dsep(P, \{A, T, C, S, B, D\}, X)$
- $dsep(P, \{T, C\}, \{A, S\})$
- $dsep(P, \{C, D\}, B)$
- $dsep(B, S, P)$
- $dsep(\{B, C\}, S, P)$
- $dsep(\{B, C\}, P, \{A, T, X\})$

Class work

Use the network on the handout (the Asian network) to answer the following independence questions.

- List the Markov blanket of all variables.
- $dsep(P, \{A, T, C, S, B, D\}, X)$ No
- $dsep(P, \{T, C\}, \{A, S\})$ Yes
- $dsep(P, \{C, D\}, B)$ No
- $dsep(B, S, P)$ Yes
- $dsep(\{B, C\}, S, P)$ No
- $dsep(\{B, C\}, P, \{A, T, X\})$ No

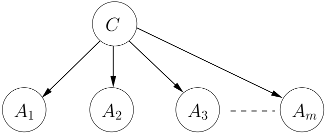
Recap

During this class, we discussed

- Basic terminology and notation for probability and graphs
- Bayesian networks as a parameterized model
- BNs as a factorization of a joint probability distribution
- BNs as a concise representation of conditional independencies based on d-separation
- Equivalence among BNs based on induced independencies

Next time in probabilistic models

- Discriminitive vs. generative learning
- Multinomial naive Bayes for document classification



- Hidden Markov models for gene prediction