CS559 Machine Learning
Probabilistic Graphical Model, Bayesian Network

Tian Han

Department of Computer Science
Stevens Institute of Technology

Week 11

Outline

- Introduction to Graphical Model
- Bayesian Network

Introduction to Graphical Model

Probabilistic Graphical Models

- A simple way to visualize the structure of a probabilistic model

Probabilistic Graphical Models

- A simple way to visualize the structure of a probabilistic model
- Insight into the properities of the model

Probabilistic Graphical Models

- A simple way to visualize the structure of a probabilistic model
- Insight into the properities of the model
- Complex computations can be expressed in terms of graphical manipulations

Probabilistic Graphical Models

- A simple way to visualize the structure of a probabilistic model
- Insight into the properities of the model
- Complex computations can be expressed in terms of graphical manipulations
- Two types:
  - **Directed** graphical models: Bayesian Networks, latent Variable Model, etc
  - **Undirected** graphical models: Markov network (Markov Random Field), Energy-based Model etc

Graphical Models

- Graphical models are graph-based representations of various factorization assumptions of distributions

Graphical Models

- Graphical models are graph-based representations of various factorization assumptions of distributions
- These factorizations are typically equivalent to *conditional independence* statement amongst (sets of) variables in the distributions.

Directed graphical model for joint probability

- The product rule of probability over three variables a, b, c: (different ordering has different factorization)

$$\begin{aligned} p(a,b,c) &= p(c|a,b)p(a,b) \\ &= p(c|a,b)p(b|a)p(a) \end{aligned}$$

Directed graphical model for joint probability

- The product rule of probability over three variables a, b, c: (different ordering has different factorization)

$$
\begin{aligned}
p(a,b,c) &= p(c|a,b)p(a,b) \\
&= p(c|a,b)p(b|a)p(a)
\end{aligned}
$$

- A directed graphical model representing the joint probability distribution over a, b, c:

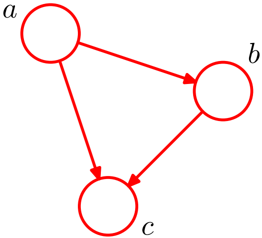


Figure: [C. Bishop, PRML]

Definitions

- A graph G contains:
  - Nodes, also known as vertices. Represents a random variable (or group of random variables).
  - Edges, also known as links between nodes. Express probabilistic relationships between these variables.

Definitions

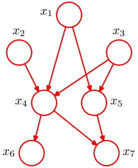
- A graph G contains:
  - Nodes, also known as vertices. Represents a random variable (or group of random variables).
  - Edges, also known as links between nodes. Express probabilistic relationships between these variables.
- Edges may be directed or undirected (may also have associated weights)

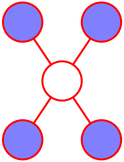
Definitions

- A graph G contains:
  - Nodes, also known as vertices. Represents a random variable (or group of random variables).
  - Edges, also known as links between nodes. Express probabilistic relationships between these variables.
- Edges may be directed or undirected (may also have associated weights)
- Directed Graphs (all edges are directed)

Definitions

- A graph G contains:
    - Nodes, also known as vertices. Represents a random variable (or group of random variables).
    - Edges, also known as links between nodes. Express probabilistic relationships between these variables.
- Edges may be directed or undirected (may also have associated weights)
- Directed Graphs (all edges are directed)
- Undirected Graphs (all edges are undirected)





More Definitions

- Path: The Path $A \to B$ from node A to node B is a sequence of nodes that connects A to B.

More Definitions

- Path: The Path $A \to B$ from node A to node B is a sequence of nodes that connects A to B.
- Cycle: A cycle is a directed path that starts and returns to the same node

More Definitions

- Path: The Path $A \to B$ from node A to node B is a sequence of nodes that connects A to B.
- Cycle: A cycle is a directed path that starts and returns to the same node
- Directed Acyclic Graph (DAG): A DAG is a graph G with directed edges (arrows on each link) between the nodes such that by following a path of nodes from one node to another along the direction of each edge, no path will revisit the same node.

More Definitions

- Path: The Path $A \rightarrow B$ from node A to node B is a sequence of nodes that connects A to B.
- Cycle: A cycle is a directed path that starts and returns to the same node
- Directed Acyclic Graph (DAG): A DAG is a graph G with directed edges (arrows on each link) between the nodes such that by following a path of nodes from one node to another along the direction of each edge, no path will revisit the same node.
- Parent and Children: if a link going from node a to node b, then a is the parent of b.

More Definitions

- Path: The Path $A \to B$ from node A to node B is a sequence of nodes that connects A to B.
- Cycle: A cycle is a directed path that starts and returns to the same node
- Directed Acyclic Graph (DAG): A DAG is a graph G with directed edges (arrows on each link) between the nodes such that by following a path of nodes from one node to another along the direction of each edge, no path will revisit the same node.
- Parent and Children: if a link going from node a to node b, then a is the parent of b.
- Graphs can be represented using: the edge list, the adjacency matrix.

The Fundamental Questions

- Representation
  - How to capture/model uncertainties in possible worlds?
  - How to encode our domain knowledge/assumptions/constraints?

The Fundamental Questions

- Representation
  - How to capture/model uncertainties in possible worlds?
  - How to encode our domain knowledge/assumptions/constraints?
- Inference
  - How do I answer questions/predictions according to my model and/or based given data?

$$P(X_i|\mathcal{M})$$

The Fundamental Questions

- Representation
  - How to capture/model uncertainties in possible worlds?
  - How to encode our domain knowledge/assumptions/constraints?
- Inference
  - How do I answer questions/predictions according to my model and/or based given data?

$$P(X_i|\mathcal{M})$$

- Learning
  - What model is "right" for my data?

$$\mathcal{M} = \text{arg max}_{\mathcal{M} \in M} F(D; \mathcal{M})$$

Basic Probability Concepts

- Representation: what is the joint probability distribution on multiple variables?
  - How many state configurations?
  - Are they all needed to be represented?
  - Do we get any scientific insight?

Basic Probability Concepts

- Representation: what is the joint probability distribution on multiple variables?
  - How many state configurations?
  - Are they all needed to be represented?
  - Do we get any scientific insight?
- Inference: If not all variables are observable, how to compute the conditional distribution of latent variables given evidence?

Basic Probability Concepts

- Representation: what is the joint probability distribution on multiple variables?
  - How many state configurations?
  - Are they all needed to be represented?
  - Do we get any scientific insight?
- Inference: If not all variables are observable, how to compute the conditional distribution of latent variables given evidence?
- Learning: where do we get all this probabilities?
  - Maximal-likelihood estimation? but how many data do we need?
  - Are there other est. principles?
  - Where do we put domain knowledge in terms of plausible relationships between variables, and plausible values of the probabilities?

Two Important Rules

1. Chain rule or Product rule: Let $S_1, S_2, ..., S_n$ be events, and $p(S_i) > 0$, then:

$$p(S_1, S_2, ..., S_n) = p(S_1)p(S_2|S_1) \cdots p(S_n|S_{n-1}, ..., S_1)$$

   Let X, Y be two variables, then

$$p(X, Y) = p(X|Y)p(Y)$$

2. Sum rule: Let X, Y be two variables, then:

$$p(X) = \sum_{Y} p(X, Y)$$

Marginalization

- Suppose X and Y are random variables with distribution $p(X, Y)$
  X: Intelligence, Val(X) = {"Very High", "High"}
  Y : Grade, Val(Y) = {"a", "b"}

Marginalization

- Suppose X and Y are random variables with distribution $p(X, Y)$
  X: Intelligence, Val(X) = {"Very High", "High"}
  Y : Grade, Val(Y) = {"a", "b"}
- Joint distribution specified by:

	vh	h
a	0.7	0.15
b	0.1	0.05

Marginalization

- Suppose X and Y are random variables with distribution $p(X, Y)$
  X: Intelligence, Val(X) = {“Very High”, “High”}
  Y : Grade, Val(Y) = {“a”, “b”}
- Joint distribution specified by:

	vh	h
a	0.7	0.15
b	0.1	0.05

- $p(Y = a) =?$

Marginalization

- Suppose X and Y are random variables with distribution $p(X, Y)$
  X: Intelligence, Val(X) = {"Very High", "High"}
  Y : Grade, Val(Y) = {"a", "b"}
- Joint distribution specified by:

	vh	h
a	0.7	0.15
b	0.1	0.05

- $p(Y = a) =? \ 0.85$

Marginalization

- Suppose X and Y are random variables with distribution $p(X, Y)$
  X: Intelligence, Val(X) = {"Very High", "High"}
  Y : Grade, Val(Y) = {"a", "b"}
- Joint distribution specified by:

	vh	h
a	0.7	0.15
b	0.1	0.05

- $p(Y = a) =? \ 0.85$
- More generally, suppose we have a joint distribution $p(X_1, ..., X_n)$. Then,

$$p(X_i = x_i) = \sum_{x_1} \sum_{x_2} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_n} p(x_1, ..., x_n)$$

Conditioning

- Suppose X and Y are random variables with distribution $p(X,Y)$
  X: Intelligence, Val(X) $= \{$“Very High”, “High”$\}$
  Y : Grade, Val(Y) $= \{$“a”, “b”$\}$

Conditioning

- Suppose X and Y are random variables with distribution $p(X, Y)$
  X: Intelligence, Val(X) = {"Very High", "High"}
  Y : Grade, Val(Y) = {"a", "b"}
- Can compute the conditional probability:

$$
\begin{aligned}
p(Y = a|X = vh) &= \frac{p(Y = a, X = vh)}{p(X = vh)} \\
&= \frac{p(Y = a, X = vh)}{p(Y = a, X = vh) + p(Y = b, X = vh)} \\
&= \frac{0.7}{0.7 + 0.1} = 0.875
\end{aligned}
$$

Example: Medical diagnosis

- Variable of each **symptom** (e.g. "fever", "cough", "fast breathing", "shaking", "nausea", "vomiting")

Example: Medical diagnosis

- Variable of each **symptom** (e.g. "fever", "cough", "fast breathing", "shaking", "nausea", "vomiting")
- Variable for each **disease** (e.g. "flu", "pneumonia", "common cold", "bronchitis", "tuberculosis")

Example: Medical diagnosis

- Variable of each **symptom** (e.g. "fever", "cough", "fast breathing", "shaking", "nausea", "vomiting")
- Variable for each **disease** (e.g. "flu", "pneumonia", "common cold", "bronchitis", "tuberculosis")
- **Diagnosis** is performed by inference in the model:

$$p(\text{common cold} = 1 | \text{cough} = 1, \text{fever} = 1, \text{vomiting} = 0)$$

Example: Medical diagnosis

- Variable of each **symptom** (e.g. "fever", "cough", "fast breathing", "shaking", "nausea", "vomiting")
- Variable for each **disease** (e.g. "flu", "pneumonia", "common cold", "bronchitis", "tuberculosis")
- **Diagnosis** is performed by inference in the model:

$$p(\text{common cold} = 1 | \text{cough} = 1, \text{fever} = 1, \text{vomiting} = 0)$$

- One famous model, Quick Medical Reference (QMR-DT), has 600 diseases and 4000 findings.

Representing the distribution

- Naively, could represent multivariate distributions with table of probabilities for each outcome (assignment)

Representing the distribution

- Naively, could represent multivariate distributions with table of probabilities for each outcome (assignment)
- How many outcomes are there in QMR-DT? 2^{4600}

Representing the distribution

- Naively, could represent multivariate distributions with table of probabilities for each outcome (assignment)
- How many outcomes are there in QMR-DT? 2^{4600}
- Estimation of joint distribution would require a huge amount of data

Representing the distribution

- Naively, could represent multivariate distributions with table of probabilities for each outcome (assignment)
- How many outcomes are there in QMR-DT? 2^{4600}
- Estimation of joint distribution would require a huge amount of data
- Inference of conditional probabilities, e.g. $p(\text{common cold} = 1 | \text{cough} = 1, \text{fever} = 1, \text{vomiting} = 0)$ would require summing over exponentially many variables values

Structure through independence

- If $X_1, ..., X_n$ are independent, then:

$$p(X_1, ..., X_n) = p(X_1), ..., p(X_n)$$

Structure through independence

- If $X_1, ..., X_n$ are independent, then:

$$p(X_1, ..., X_n) = p(X_1), ..., p(X_n)$$

- Right hand side entries can be described by just n numbers (if $|Val(X_i)| = 2$)!

Structure through independence

- If $X_1, ..., X_n$ are independent, then:

$$p(X_1, ..., X_n) = p(X_1), ..., p(X_n)$$

- Right hand side entries can be described by just n numbers (if $|Val(X_i)| = 2$)!
- However, this is not a very useful model–observing a variable X_i cannot influence our predictions of X_j

Structure through independence

- If $X_1, ..., X_n$ are independent, then:

$$p(X_1, ..., X_n) = p(X_1), ..., p(X_n)$$

- Right hand side entries can be described by just n numbers (if $|Val(X_i)| = 2$)!
- However, this is not a very useful model–observing a variable X_i cannot influence our predictions of X_j
- If $X_1, ..., X_n$ are conditionally independent given Y, denoted as $X_i \perp X_{-i} | Y$ then:

$$\begin{aligned} p(y, x_1, ..., x_n) &= p(y)p(x_1|y)\prod_{i=2}^{n} p(x_i|x_1, ..., x_{i-1}, y) \\ &= p(y)p(x_1|y)\prod_{i=2}^{n} p(x_i|y) \end{aligned}$$

Structure through independence

- If $X_1, ..., X_n$ are independent, then:

$$p(X_1, ..., X_n) = p(X_1), ..., p(X_n)$$

- Right hand side entries can be described by just n numbers (if $|Val(X_i)| = 2$)!
- However, this is not a very useful model–observing a variable X_i cannot influence our predictions of X_j
- If $X_1, ..., X_n$ are conditionally independent given Y, denoted as $X_i \perp X_{-i} | Y$ then:

$$\begin{aligned} p(y, x_1, ..., x_n) &= p(y)p(x_1|y) \prod_{i=2}^{n} p(x_i | x_1, ..., x_{i-1}, y) \\ &= p(y)p(x_1|y) \prod_{i=2}^{n} p(x_i|y) \end{aligned}$$

- Simple, yet powerful

Example: naive Bayes

- Suppose that the features are conditionally independent given class label Y. Then

$$p(y, x_1, ..., x_n) = p(y) \prod_{i=1}^{n} p(x_i|y)$$

Example: naive Bayes

- Suppose that the features are conditionally independent given class label Y. Then

$$p(y, x_1, ..., x_n) = p(y) \prod_{i=1}^{n} p(x_i|y)$$

- **Estimate** the model with maximum likelihood. **Predict** with:

$$p(Y = 1|x_1, ..., x_n) = \frac{p(Y = 1) \prod_{i=1}^{n} p(x_i|Y = 1)}{\sum_{y=\{0,1\}} p(Y = y) \prod_{i=1}^{n} p(x_i|Y = y)}$$

Example: naive Bayes

- Suppose that the features are conditionally independent given class label Y. Then

$$p(y, x_1, ..., x_n) = p(y) \prod_{i=1}^{n} p(x_i|y)$$

- **Estimate** the model with maximum likelihood. **Predict** with:

$$p(Y = 1|x_1, ..., x_n) = \frac{p(Y = 1) \prod_{i=1}^{n} p(x_i|Y = 1)}{\sum_{y=\{0,1\}} p(Y = y) \prod_{i=1}^{n} p(x_i|Y = y)}$$

- Are the independence assumptions made here reasonable?

Example: naive Bayes

- Suppose that the features are conditionally independent given class label Y. Then

$$p(y, x_1, ..., x_n) = p(y) \prod_{i=1}^{n} p(x_i|y)$$

- **Estimate** the model with maximum likelihood. **Predict** with:

$$p(Y=1|x_1, ..., x_n) = \frac{p(Y=1) \prod_{i=1}^{n} p(x_i|Y=1)}{\sum_{y=\{0,1\}} p(Y=y) \prod_{i=1}^{n} p(x_i|Y=y)}$$

- Are the independence assumptions made here reasonable?
- Philosophy: Nearly all probabilistic models are wrong, but many are nonetheless useful

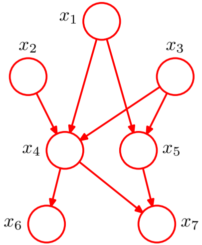
Bayesian Network

Bayesian Networks (BN)

- A bayesian network is a directed acyclic graph (i.e., DAG) in which each node has associated with the conditional probability of the node given its parents.

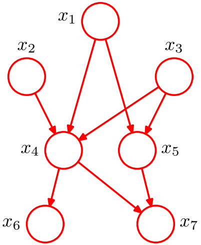
Bayesian Networks (BN)

- A bayesian network is a directed acyclic graph (i.e., DAG) in which each node has associated with the conditional probability of the node given its parents.
- The joint distribution is obtained by taking the product of the conditional probabilities:



Bayesian Networks (BN)

- A bayesian network is a directed acyclic graph (i.e., DAG) in which each node has associated with the conditional probability of the node given its parents.
- The joint distribution is obtained by taking the product of the conditional probabilities:



- $p(x_1, x_2, x_3, x_4, x_5, x_6, x_7) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3)p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5)$

Bayesian Networks

- A graph with K nodes, the joint distribution is given by:

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k | pa_k)$$

  where pa_k denotes the set of parents of x_k and $\mathbf{x} = \{x_1, ..., x_K\}$

- This key equation expresses the *factorization* properties of the joint distribution for a directed graphical model.

Alarm Example

- Sally's burglar Alarm(A) is sounding

Alarm Example

- Sally's burglar Alarm(A) is sounding
- Has she been Burgled (B), or was the alarm triggered by an Earthquake (E)?

Alarm Example

- Sally's burglar Alarm(A) is sounding
- Has she been Burgled (B), or was the alarm triggered by an Earthquake (E)?
- She turns the car Radio (R) on for news of earthequakes

Alarm Example

- Sally's burglar Alarm(A) is sounding
- Has she been Burgled (B), or was the alarm triggered by an Earthquake (E)?
- She turns the car Radio (R) on for news of earthequakes
- Without loss of generality, we can write:

$$\begin{aligned} p(A,R,E,B) &= p(A|R,E,B)P(R,E,B) \\ &= p(A|R,E,B)p(R|E,B)p(E,B) \\ &= p(A|R,E,B)p(R|E,B)p(E|B)p(B) \end{aligned}$$

Alarm Example (cont)

Assumptions

- The alarm is not directly influenced by any report on the radio $p(A|R, E, B) = p(A|E, B)$

Alarm Example (cont)

Assumptions

- The alarm is not directly influenced by any report on the radio $p(A|R, E, B) = p(A|E, B)$
- The radio broadcast is not directly influenced by the burglar variable $p(R|E, B) = p(R|E)$

Alarm Example (cont)

Assumptions

- The alarm is not directly influenced by any report on the radio $p(A|R, E, B) = p(A|E, B)$
- The radio broadcast is not directly influenced by the burglar variable $p(R|E, B) = p(R|E)$
- Burglaries don't directly "cause" earthquakes $p(E|B) = P(E)$

Alarm Example (cont)

Assumptions

- The alarm is not directly influenced by any report on the radio $p(A|R, E, B) = p(A|E, B)$
- The radio broadcast is not directly influenced by the burglar variable $p(R|E, B) = p(R|E)$
- Burglaries don't directly "cause" earthquakes $p(E|B) = P(E)$
- Therefore, we have:

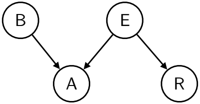
$$p(A, R, E, B) = p(A|E, B)p(R|E)p(E)p(B)$$

Alarm Example - probability table

- DAG for $p(A, R, E, B) = p(A|E, B)p(R|E)p(E)p(B)$

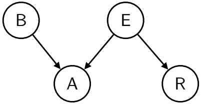
Alarm Example - probability table

- DAG for $p(A, R, E, B) = p(A|E, B)p(R|E)p(E)p(B)$



Alarm Example - probability table

- DAG for $p(A, R, E, B) = p(A|E, B)p(R|E)p(E)p(B)$

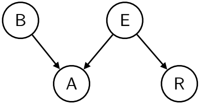


- Probability table for $p(A|B, E)$:

Alarm $= 1$	Burglar	Earthquake
0.9999	1	1
0.99	1	0
0.99	0	1
0.0001	0	0

Alarm Example - probability table

- DAG for $p(A, R, E, B) = p(A|E, B)p(R|E)p(E)p(B)$



- Probability table for $p(A|B, E)$:

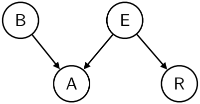
Alarm $= 1$	Burglar	Earthquake
0.9999	1	1
0.99	1	0
0.99	0	1
0.0001	0	0

- Probability table for $p(R|E)$:

Radio $= 1$	Earthquake=1
1	1
0	0

Alarm Example - probability table

- DAG for $p(A, R, E, B) = p(A|E, B)p(R|E)p(E)p(B)$



- Probability table for $p(A|B, E)$:

Alarm = 1	Burglar	Earthquake
0.9999	1	1
0.99	1	0
0.99	0	1
0.0001	0	0

- Probability table for $p(R|E)$:

Radio = 1	Earthquake=1
1	1
0	0

- $p(B = 1) = 0.01$ and $p(E = 1) = 0.000001$

Alarm Example - probability table

- Initial evidence: the alarm is sounding

$$
\begin{aligned}
p(B=1|A=1) &= \frac{\sum_{E,R} p(B=1,E,A=1,R)}{\sum_{B,E,R} P(B,E,A=1,R)} \\
&= \frac{\sum_{E,R} p(A=1|B=1,E)p(B=1)p(R|E)p(E)}{\sum_{B,E,R} p(A=1|B,E)p(B)p(E)p(R|E)} \\
&\approx 0.99
\end{aligned}
$$

Alarm Example - inference

- Additional evidence: the radio broadcasts an earthquake warning

Alarm Example - inference

- Additional evidence: the radio broadcasts an earthquake warning
- A similar calculation gives: $p(B = 1|A = 1, R = 1) \approx 0.01$

Alarm Example - inference

- Additional evidence: the radio broadcasts an earthquake warning
- A similar calculation gives: $p(B = 1|A = 1, R = 1) \approx 0.01$
- Initially, because the alarm sounds, Sally thinks that she's been burgled. However, this probability drops dramatically when she hears that there has been an earthquake.

Alarm Example - inference

- Additional evidence: the radio broadcasts an earthquake warning
- A similar calculation gives: $p(B = 1|A = 1, R = 1) \approx 0.01$
- Initially, because the alarm sounds, Sally thinks that she's been burgled. However, this probability drops dramatically when she hears that there has been an earthquake.
- The earthquake "explains away" to an extent the fact that the alarm is ringing

Wet grass example

- One morning Tracey leaves her house and realizes that her grass is wet (T).

Wet grass example

- One morning Tracey leaves her house and realizes that her grass is wet (T).
- Is it due to overnight Rain (R) or did she forget to turn off the sprinkler (S) last night?

Wet grass example

- One morning Tracey leaves her house and realizes that her grass is wet (T).
- Is it due to overnight Rain (R) or did she forget to turn off the sprinkler (S) last night?
- Next she notices that the grass of her neighbor, Jack, is also wet (J). This explains away to some extent the possibility that her sprinkler was left on, and she concludes therefore that it has probably been raining.

Wet grass example (cont)

- $R \in \{0, 1\}, R = 1$ means that it has been raining, and 0 otherwise

Wet grass example (cont)

- $R \in \{0, 1\}, R = 1$ means that it has been raining, and 0 otherwise
- $S \in \{0, 1\}, S = 1$ means that Tracey has forgotten to turn off the sprinkler, and 0 otherwise

Wet grass example (cont)

- $R \in \{0, 1\}, R = 1$ means that it has been raining, and 0 otherwise
- $S \in \{0, 1\}, S = 1$ means that Tracey has forgotten to turn off the sprinkler, and 0 otherwise
- $J \in \{0, 1\}, J = 1$ means that Jack's grass is wet, and 0 otherwise

Wet grass example (cont)

- $R \in \{0, 1\}, R = 1$ means that it has been raining, and 0 otherwise
- $S \in \{0, 1\}, S = 1$ means that Tracey has forgotten to turn off the sprinkler, and 0 otherwise
- $J \in \{0, 1\}, J = 1$ means that Jack's grass is wet, and 0 otherwise
- Joint probability

$$\begin{aligned} p(T, J, R, S) &= p(T|J, R, S)p(J, R, S) \\ &= p(T|J, R, S)p(J|R, S)p(R, S) \\ &= p(T|J, R, S)p(J|R, S)p(R|S)p(S) \end{aligned}$$

Wet grass example (cont)

- The number of values that need to be specified in general scales exponentially with the number of variables in the model. This is impractical in general and motivates simplifications

Wet grass example (cont)

- The number of values that need to be specified in general scales exponentially with the number of variables in the model. This is impractical in general and motivates simplifications
- Conditional independence:
  - $p(T|J,R,S) = p(T|R,S)$

Wet grass example (cont)

- The number of values that need to be specified in general scales exponentially with the number of variables in the model. This is impractical in general and motivates simplifications
- Conditional independence:
  - $p(T|J,R,S) = p(T|R,S)$
  - $p(J|R,S) = p(J|R)$

Wet grass example (cont)

- The number of values that need to be specified in general scales exponentially with the number of variables in the model. This is impractical in general and motivates simplifications
- Conditional independence:
  - $p(T|J, R, S) = p(T|R, S)$
  - $p(J|R, S) = p(J|R)$
  - $p(R|S) = p(R)$

Wet grass example (cont)

- Original equation:

$$\begin{aligned} p(T, J, R, S) &= p(T|J, R, S)p(J, R, S) \\ &= p(T|J, R, S)p(J|R, S)p(R, S) \\ &= p(T|J, R, S)p(J|R, S)p(R|S)p(S) \end{aligned}$$

Wet grass example (cont)

- Original equation:

$$\begin{aligned} p(T, J, R, S) &= p(T|J, R, S)p(J, R, S) \\ &= p(T|J, R, S)p(J|R, S)p(R, S) \\ &= p(T|J, R, S)p(J|R, S)p(R|S)p(S) \end{aligned}$$

- Now it becomes:

$$p(T, J, R, S) = p(T|R, S)p(J|R)p(R)p(S)$$

Wet grass example - probability table

- $p(R = 1) = 0.2$ and $p(S = 1) = 0.1$

Wet grass example - probability table

- $p(R = 1) = 0.2$ and $p(S = 1) = 0.1$
- $p(J = 1|R = 1) = 1$, $p(J = 1|R = 0) = 0.2$

Wet grass example - probability table

- $p(R=1)=0.2$ and $p(S=1)=0.1$
- $p(J=1|R=1)=1$, $p(J=1|R=0)=0.2$
- Probability table for $p(T|R,S)$:

T = 1	R	S
1	1	0
1	1	1
0.9	0	1
0	0	0

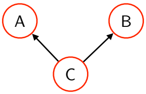
Wet grass example - inference

$$
\begin{aligned}
p(S=1|T=1) &= \frac{p(S=1,T=1)}{p(T=1)} = \frac{\sum_{J,R} p(T=1,J,R,S=1)}{\sum_{J,R,S} p(T=1,J,R,S)} \\
&= \frac{\sum_{J,R} p(T=1|R,S=1)p(J|R)p(R)p(S=1)}{\sum_{J,R,S} p(T=1|R,S)p(J|R)p(R)p(S)} \\
&= \frac{\sum_{R} p(T=1|R,S=1)p(R)p(S=1)}{\sum_{R,S} p(T=1|R,S)p(R)p(S)} \\
&= \frac{0.9\cdot 0.8\cdot 0.1+0.2\cdot 0.1}{0.9\cdot 0.8\cdot 0.1+1\cdot 0.2\cdot 0.1+0\cdot 0.8\cdot 0.9+1\cdot 0.2\cdot 0.9} \approx 0.3382
\end{aligned}
$$

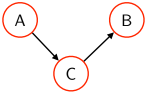
Wet grass example - inference

$$
\begin{aligned}
p(S=1|T=1,J=1) &= \frac{p(S=1,T=1,J=1)}{p(T=1,J=1)} \\
&= \frac{\sum_R p(T=1,J=1,R,S=1)}{\sum_{R,S} p(T=1,J=1,R,S)} \\
&= \frac{\sum_R p(J=1|R)p(T=1|R,S=1)p(R)p(S=1)}{\sum_{R,S} p(J=1|R)p(T=1|R,S)p(R)p(S)} \\
&= \frac{0.0344}{0.2144} \approx 0.1604
\end{aligned}
$$

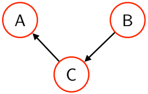
Independence in Bayesian Network - Marginal independence



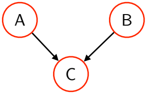
A,B are marginally dependent: $A \not\perp B$
$p(A,B) = \sum_C p(A|C)P(B|C)P(C) \neq p(A)p(B)$



A,B are marginally dependent: $A \not\perp B$
$p(A,B) = \sum_C p(A)P(C|A)P(B|C) \neq p(A)p(B)$

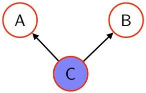


A,B are marginally dependent: $A \not\perp B$



A,B are marginally independent: $A \perp B$
$p(A,B,C) = p(A)p(B)p(C|A,B) \rightarrow p(A,B) = p(A)p(B)$

Independence in Bayesian Network - Conditional independence

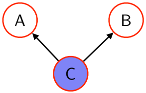


A,B are conditionally independent given C

$p(A, B|C) = \frac{p(A,B,C)}{p(C)} = \frac{p(A|C)p(B|C)p(C)}{p(C)} = p(A|C)p(B|C)$

Node C is said to be *tail-to-tail*.

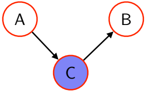
Independence in Bayesian Network - Conditional independence



A,B are conditionally independent given C

$p(A,B|C) = \frac{p(A,B,C)}{p(C)} = \frac{p(A|C)p(B|C)p(C)}{p(C)} = p(A|C)p(B|C)$

Node C is said to be *tail-to-tail*.

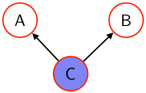


A,B are conditionally independent given C

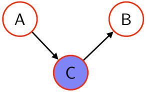
$p(A,B|C) = \frac{p(A)p(C|A)p(B|C)}{p(C)} = \frac{p(A,C)p(B|C)}{p(C)} = p(A|C)p(B|C)$

Node C is said to be *head-to-tail*.

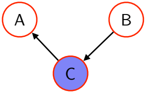
Independence in Bayesian Network - Conditional independence



A,B are conditionally independent given C
$p(A, B|C) = \frac{p(A,B,C)}{p(C)} = \frac{p(A|C)p(B|C)p(C)}{p(C)} = p(A|C)p(B|C)$
Node C is said to be *tail-to-tail*.

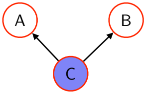


A,B are conditionally independent given C
$p(A, B|C) = \frac{p(A)p(C|A)p(B|C)}{p(C)} = \frac{p(A,C)p(B|C)}{p(C)} = p(A|C)p(B|C)$
Node C is said to be *head-to-tail*.

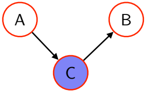


A,B are conditionally independent given C
$p(A, B|C) = \frac{p(A|C)p(C|B)p(B)}{p(C)} = \frac{p(A|C)p(B,C)}{p(C)} = p(A|C)p(B|C)$

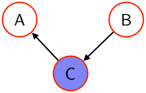
Independence in Bayesian Network - Conditional independence



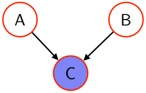
A,B are conditionally independent given C
$p(A,B|C) = \frac{p(A,B,C)}{p(C)} = \frac{p(A|C)p(B|C)p(C)}{p(C)} = p(A|C)p(B|C)$
Node C is said to be *tail-to-tail*.



A,B are conditionally independent given C
$p(A,B|C) = \frac{p(A)p(C|A)p(B|C)}{p(C)} = \frac{p(A,C)p(B|C)}{p(C)} = p(A|C)p(B|C)$
Node C is said to be *head-to-tail*.



A,B are conditionally independent given C
$p(A,B|C) = \frac{p(A|C)p(C|B)p(B)}{p(C)} = \frac{p(A|C)p(B,C)}{p(C)} = p(A|C)p(B|C)$



A,B are conditionally dependent given C
$p(A,B|C) \propto p(C|A,B)p(A)p(B) \neq p(A|C)p(B|C)$
Node C is said to be *head-to-head*.

Example: Fuel system on the car

- $B \in \{0, 1\}$: representing the state of a battery that is either charged $(B = 1)$ or flat $(B = 0)$.
- $F \in \{0, 1\}$: representing the state of the fuel tank that is either full of fuel $(F = 1)$ or empty $(F = 0)$.
- $G \in \{0, 1\}$: representing the state of an electric fuel gauge and which indicates either full $(G = 1)$ or empty $(G = 0)$.

Example: Fuel system on the car

- $B \in \{0, 1\}$: representing the state of a battery that is either charged $(B = 1)$ or flat $(B = 0)$.
- $F \in \{0, 1\}$: representing the state of the fuel tank that is either full of fuel $(F = 1)$ or empty $(F = 0)$.
- $G \in \{0, 1\}$: representing the state of an electric fuel gauge and which indicates either full $(G = 1)$ or empty $(G = 0)$.

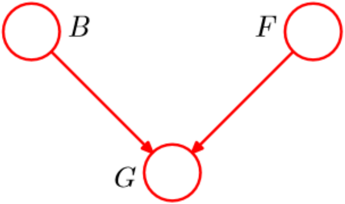


Figure: [C. Bishop, PRML]

Example: Fuel system on the car

Probability table: $p(G, B, F) = p(B)p(F)p(G|B, F)$

$$\begin{aligned} p(B = 1) &= 0.9 \\ p(F = 1) &= 0.9 \end{aligned}$$

G = 1	B	F
0.8	1	1
0.2	1	0
0.2	0	1
0.1	0	0

Example: Fuel system on the car

- $p(F = 0) = 0.1$: before observe any data, the prior probability of the fuel tank being empty.

Example: Fuel system on the car

- $p(F = 0) = 0.1$: before observe any data, the prior probability of the fuel tank being empty.
- $p(F = 0|G = 0)$: suppose we observe the fuel gauge and discover it reads empty, i.e., $G = 0$.

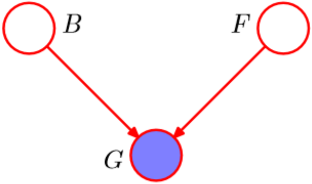


Figure: [C. Bishop, PRML]

Example: Fuel system on the car

- $p(F=0)=0.1$: before observe any data, the prior probability of the fuel tank being empty.
- $p(F=0|G=0)$: suppose we observe the fuel gauge and discover it reads empty, i.e., $G=0$.

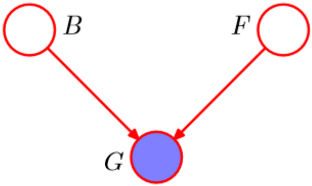


Figure: [C. Bishop, PRML]

- $p(F=0|G=0) > p(F=0)$ observing that the gauge reads empty makes it more likely that the tank is indeed empty.

$$p(F=0|G=0)=\frac{p(G=0|F=0)p(F=0)}{p(G=0)}\approx 0.257$$

Example: Fuel system on the car

- Now observed the states of both the fuel gauge and the battery:

$$p(F = 0 | G = 0, B = 0) \approx 0.111$$

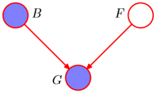


Figure: [C. Bishop, PRML]

Example: Fuel system on the car

- Now observed the states of both the fuel gauge and the battery:

$$p(F = 0|G = 0, B = 0) \approx 0.111$$

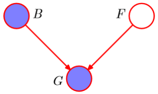


Figure: [C. Bishop, PRML]

- The probability that the tank is empty has *decreased*(from 0.257 to 0.111) as a result of the observation of the state of the battery. Finding out that the battery is flat *explains away* the observation that the fuel gauge reads empty.

Example: Fuel system on the car

- Now observed the states of both the fuel gauge and the battery:

$$p(F = 0|G = 0, B = 0) \approx 0.111$$

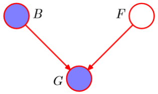


Figure: [C. Bishop, PRML]

- The probability that the tank is empty has *decreased*(from 0.257 to 0.111) as a result of the observation of the state of the battery. Finding out that the battery is flat *explains away* the observation that the fuel gauge reads empty.
- The state of the fuel tank (F) and that of the battery (B) have indeed become dependent on each other as a result of observing the reading on the fuel gauge (G).

D-separation ("directed separated") in Bayesian networks

- Algorithm to find out whether $A \perp B | \mathbf{C}$ is implied by a given directed acyclic graph.

D-separation ("directed separated") in Bayesian networks

- Algorithm to find out whether $A \perp B | \mathbf{C}$ is implied by a given directed acyclic graph.
- Look at all possible paths from A to B, any such path is said to be **blocked** if it includes a node such that either
    1. the arrows on the path meet either head-to-tail or tail-to-tail at the node, and the node is in the set C
    2. the arrows meet head-to-head at the node, and neither the node, nor any of its descendants, is in the set C.

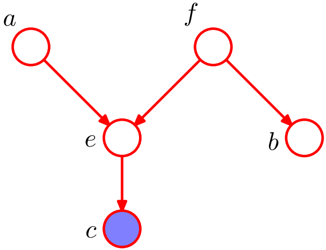
D-separation ("directed separated") in Bayesian networks

- Algorithm to find out whether $A \perp B | \mathbf{C}$ is implied by a given directed acyclic graph.
- Look at all possible paths from A to B, any such path is said to be **blocked** if it includes a node such that either
  1. the arrows on the path meet either head-to-tail or tail-to-tail at the node, and the node is in the set C
  2. the arrows meet head-to-head at the node, and neither the node, nor any of its descendants, is in the set C.
- If all such paths are blocked, then A is said to be *d-separated* from B by C. And the joint distribution for all variables in the graph will satisfy $A \perp B | \mathbf{C}$.

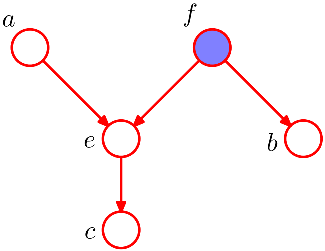
D-separation ("directed separated") in Bayesian networks

- Algorithm to find out whether $A \perp B | \mathbf{C}$ is implied by a given directed acyclic graph.
- Look at all possible paths from A to B, any such path is said to be **blocked** if it includes a node such that either
  1. the arrows on the path meet either head-to-tail or tail-to-tail at the node, and the node is in the set C
  2. the arrows meet head-to-head at the node, and neither the node, nor any of its descendants, is in the set C.
- If all such paths are blocked, then A is said to be *d-separated* from B by C. And the joint distribution for all variables in the graph will satisfy $A \perp B | \mathbf{C}$.
- d-separation reduces statistical independencies (hard) to connectivity in graphs (easy)
- Important because it allows us to quickly prune the Bayesian network, finding just the relevant variables for answering a query

D-separation examples



(a) $a \not\perp b | c$

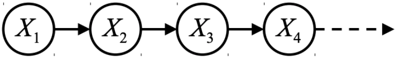


(b) $a \perp b | f$

Some frequently used graphical models

Markov Models

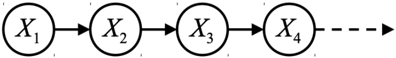
- A Markov model is a chain-structured BN
  - Each node is identically distributed (stationarity)
  - Value of X at a given time is called the state
  - As a BN:



$$P(X_1) \qquad P(X_2|X_1) \qquad P(X_T|X_{T-1})$$

  - Parameters: called transition probabilities or dynamics, specify how the state evolves over time (also, initial probs)

Markov Models - Conditional Independence



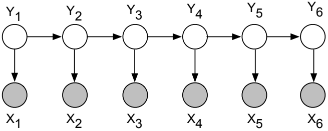
$P(X_1)$ $\quad P(X_2|X_1)$ $\quad P(X_T|X_{T-1})$

- Basic conditional independence
  - Past and future independent of the present
  - Each time step only depends on the previous
  - This is called the (first order) Markov property
- Note that the chain is just a (growing) BN
  - We can always use generic BN reasoning on it (if we truncate the chain)

Hidden Markov models

- Markov chains not so useful for most agents
  - Eventually you dont know anything anymore
  - Prediction needs previous observations.

- Hidden Markov Models



  - Frequently used for speech recognition and video modelling
  - Underlying Markov chain over states S
  - Joint distribution factors as:

$$p(\mathbf{y}, \mathbf{x}) = p(y_1)p(x_1|y_1)\prod_{t=2}^{T} p(y_t|y_{t-1})p(x_t|y_t)$$

L.R. Rabiner.A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE. Volume: 77 , Issue: 2 , Feb 1989.

Hidden Markov models

- Joint distribution factors as:

$$p(\mathbf{y}, \mathbf{x}) = p(y_1)p(x_1|y_1) \prod_{t=2}^{T} p(y_t|y_{t-1})p(x_t|y_t)$$

  - $p(y_1)$ is the distribution for the starting state
  - $p(y_t|y_{t-1})$ is the transition probability between any two states
  - $p(x_t|y_t)$ is the emission probability
- What are the conditional independencies here? e.g. $Y_1 \perp \{Y_3, ..., Y_6\} | Y_2$
- Markov assumptions:
  - The current state is conditionally independent of all the past states given the states in the previous time step.
  - The current evidence is only dependent on the current state.

Acknowledgement and Further Reading

Slides are adapted from Dr. Y. Ning's Spring 19 offering of CS-559.

Further Reading:
Chapter 8.1 8.2 of *Pattern Recognition and Machine Learning* by C. Bishop.