Introduction to Probabilistic Machine Learning

Ralf Herbrich, Rainer Schlosser

Tutorial 3 – Recap Theory Unit 3

Overview

1. **Questions and Updates**
2. Recap: Main Concepts of Unit 3
3. Example: Conditional Independence & D-Separation
4. Simulating 1 vs 1 TrueSkill (discrete)

Course Overview

Week	Topic Lecture	Tutorial	Exercises
07.04. & 08.04.	1 Probability Theory	Intro Julia	
14.04. & 15.04.	2 Inference Methods and Decision-Making	no tutorial	**Exercise 1**
21.04. & 22.04.	**no lecture**	Theory Unit 1 & 2	(14.04. – 05.05.)
28.04. & 29.04.	**3 Graphical Models: Independence**	**Theory Unit 3**	
05.05. & 06.05.	4 Graphical Models: Exact Inference	Theory Unit 4	Exercise 2
12.05. & 13.05.	5 Graphical Models: Approximate Inference	Theory Unit 5	(05.05. – 19.05.)
19.05. & 20.05.	6 Bayesian Ranking	Theory Unit 6	Exercise 3
26.05. & 27.05.	7 Linear Basis Function Models	Theory Unit 7	(19.05. – 02.06.)
02.06. & 03.06.	8 Bayesian Regression	Theory Unit 8	Exercise 4
09.06. & 10.06.	**no lecture**	9 Bayesian Classification	(02.06. – 23.06.)
16.06. & 11.06.	10 Non-Bayesian Classification Learning	Theory Unit 9 & 10	
23.06. & 24.06.	11 Gaussian Processes	Theory Unit 11	Exercise 5
30.06. & 01.07.	12 Information Theory	Theory Unit 12	(23.06. – 07.07.)
07.07. & 08.07.	13 Real-World Applications		

Overview

1. Questions and Updates
2. **Recap: Main Concepts of Unit 3**
3. Example: Conditional Independence & D-Separation
4. Simulating 1 vs 1 TrueSkill (discrete)

Recap Unit 3: Overview of Concepts and Focus

a) Graphical Models & Bayesian Networks

b) Conditional Probabilities & Chain Rule

c) Conditional Independence

d) D-Separation

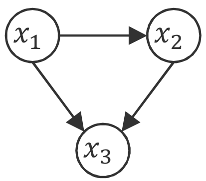
Recap: Conditional Probabilities in Bayesian Networks

- **Observation**. Any joint distribution $p(x_1, \dots, x_n)$ can be written as

$$p(x_1, \dots, x_n) = \prod_{i=1}^{n} p(x_i | x_1, \dots, x_{i-1})$$

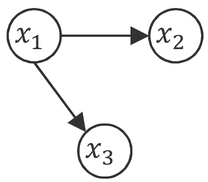
- **Bayesian Network**. *Given a joint distribution as a product of* ***conditional distributions***, $p(x_1, \dots, x_n) = \prod_{i=1}^{n} p(x_i | \mathrm{parents}_i)$, *a* Bayesian network *is a graph with a node for every variable* x_i, *and a* ***directed edge*** *from every variable* $x \in \mathrm{parent}_i$ to x_i. *If the variable is independent of all other variables, it has no incoming edges.*

- **Examples**: For 3 variables, we have these four generic Bayesian networks



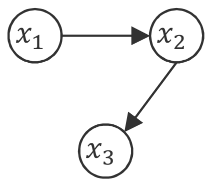
$p(x_1, x_2, x_3) = p(x_1) \cdot p(x_2|x_1) \cdot p(x_3|x_1, x_2)$

full mesh



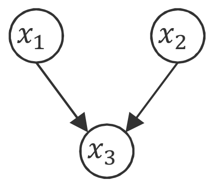
$p(x_1, x_2, x_3) = p(x_1) \cdot \prod_{i=2}^{3} p(x_i|x_1)$

star



$p(x_1, x_2, x_3) = p(x_1) \cdot \prod_{i=2}^{3} p(x_i|x_{i-1})$

chain



$p(x_1, x_2, x_3) = p(x_3|x_1, x_2) \cdot \prod_{i=1}^{2} p(x_i)$

sink

Joint Probabilities vs Conditional Probabilities

Joint Probabilities and short-hand notation:

$$p(x_1, x_2, x_3) = P(X_1 = x_1, X_2 = x_2, X_3 = x_3)$$

$$\sum_{x_1}\sum_{x_2}\sum_{x_3} p(x_1, x_2, x_3) = 1$$

Joint Probabilities vs Conditional Probabilities

Joint Probabilities and short-hand notation:

$$p(x_1, x_2, x_3) = P(X_1 = x_1, X_2 = x_2, X_3 = x_3) \qquad \sum_{x_1}\sum_{x_2}\sum_{x_3} p(x_1, x_2, x_3) = 1$$

Representation as **Conditional Probabilities** via Chain Rule in some „Order“

$$p(x_1, x_2, x_3) = p(x_2, x_3 \mid x_1) \cdot p(x_1) \qquad note: p(x_2, x_3) = p(x_3 \mid x_2) \cdot p(x_2)$$

$$= p(x_3 \mid x_1, x_2) \cdot p(x_2 \mid x_1) \cdot p(x_1)$$

Joint Probabilities vs Conditional Probabilities

Joint Probabilities and short-hand notation:

$$p(x_1, x_2, x_3) = P(X_1 = x_1, X_2 = x_2, X_3 = x_3) \qquad \sum_{x_1}\sum_{x_2}\sum_{x_3} p(x_1, x_2, x_3) = 1$$

Representation as **Conditional Probabilities** via Chain Rule in some „Order“

$$p(x_1, x_2, x_3) = p(x_2, x_3 \mid x_1) \cdot p(x_1) \qquad note: p(x_2, x_3) = p(x_3 \mid x_2) \cdot p(x_2)$$

$$= p(x_3 \mid x_1, x_2) \cdot p(x_2 \mid x_1) \cdot p(x_1)$$

Norm Identities:

$$\sum_{x_1}\sum_{x_2}\sum_{x_3} p(x_1, x_2, x_3) = \sum_{x_1}\sum_{x_2}\sum_{x_3} p(x_3 \mid x_1, x_2) \cdot p(x_2 \mid x_1) \cdot p(x_1)$$

$$= \sum_{x_1} p(x_1) \cdot \sum_{x_2} p(x_2 \mid x_1) \cdot \sum_{x_3} p(x_3 \mid x_1, x_2) = 1$$

Joint Probabilities vs Conditional Probabilities (Variants!)

Joint Probabilities and short-hand notation:

$$p(x_1, x_2, x_3) = P(X_1 = x_1, X_2 = x_2, X_3 = x_3) \qquad \sum_{x_1}\sum_{x_2}\sum_{x_3} p(x_1, x_2, x_3) = 1$$

Representation as **Conditional Probabilities** via Chain Rule in some „Order"

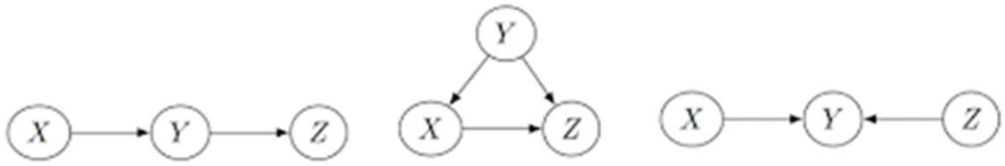
$$\begin{aligned} p(x_1, x_2, x_3) &= p(x_3 \mid x_1, x_2) \cdot p(x_2 \mid x_1) \cdot p(x_1) \\ &= p(x_2 \mid x_3, x_1) \cdot p(x_1 \mid x_3) \cdot p(x_3) \\ &= p(x_1 \mid x_2, x_3) \cdot p(x_3 \mid x_2) \cdot p(x_2) \end{aligned}$$

Norm Identities:

$$\begin{aligned} \sum_{x_1}\sum_{x_2}\sum_{x_3} p(x_1, x_2, x_3) &= \sum_{x_1} p(x_1) \cdot \sum_{x_2} p(x_2 \mid x_1) \cdot \sum_{x_3} p(x_3 \mid x_1, x_2) = 1 \\ &= \sum_{x_3} p(x_3) \cdot \sum_{x_1} p(x_1 \mid x_3) \cdot \sum_{x_2} p(x_2 \mid x_3, x_1) = 1 \\ &= \sum_{x_2} p(x_2) \cdot \sum_{x_3} p(x_3 \mid x_2) \cdot \sum_{x_1} p(x_1 \mid x_2, x_3) = 1 \end{aligned}$$

Recap: Conditional Independence

- In modelling specific data, domain experts often know whether or not two (latent) **measurements X & Z can affect each** other or not (i.e., are independent)



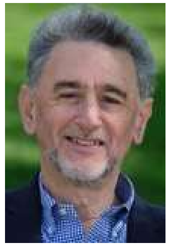
(i) Causal chain/mediator (ii) Confounder (iii) Collider

- Bayesian networks are useful to determine conditional independence.
- **Conditional Independence**. *A random variable* x_i *is conditionally independent of a random variable* x_j *given the variable* x_k *if for all values* a *of* x_k, b *of* x_j

$$p(x_i | x_j = b, x_k = a) = p(x_i | x_k = a)$$

  - Equivalent definition: $p(x_i, x_j | x_k = a) = p(x_i | x_k = a) \cdot p(x_j | x_k = a)$
  - Shorthand notation (Dawid, 1979): $x_i \perp x_j | x_k$

Hasso Plattner Institut
Digital Engineering · Universität Potsdam



Philip Dawid
(1946–)

Conditional Independence

Question: Does one variable X affect another variable Y?

Or: When does observing Y change the distribution of X?

Problem: Influence of third variables (e.g. Confounder/Collider)

Helping Concept: **Conditional Independence (CI)**

Involves: Compares the probabilities of two variables X and Y

conditioned on the observation of a third variable Z

Conditional Independence

Question: Does one variable X affect another variable Y?

Or: When does observing Y change the distribution of X?

Problem: Influence of third variables (e.g. Confounder/Collider)

Helping Concept: **Conditional Independence (CI)**

Involves: Compares the probabilities of two variables X and Y conditioned on the observation of a third variable Z

Question: Relation to usual Independence of 2 variables?

Conditional Independence

Question:	Does one variable X affect another variable Y? Or: When does observing Y change the distribution of X?
Problem:	Influence of third variables (e.g. Confounder/Collider)
Helping Concept:	**Conditional Independence (CI)**
Involves:	Compares the probabilities of two variables X and Y conditioned on the observation of a third variable Z
Different from:	**Independence (I)** of 2 Random Variables!

$$I \not\Rightarrow CI \;\; \& \;\; CI \not\Rightarrow I$$

Conditional Independence

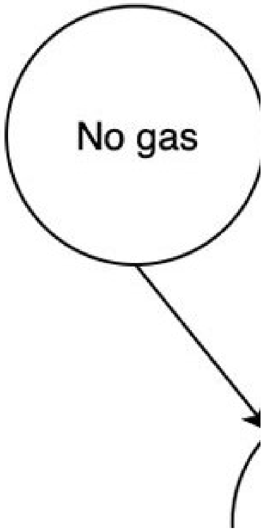
Question: Does one variable X affect another variable Y?

Or: When does observing Y change the distribution of X?

Problem: Influence of third variables (e.g. Confounder)

Helping Concept: **Conditional Independence (CI)**

Example: Are **Gas** (X) and **Battery** (Y) independent conditioned on observations of **Car Dead** (Z)?



Checking for Conditional Independence

Question: Are two variables X & Y **conditionally independent**?

(for a given set of other observed variables)!

Idea: Look for paths/connections between them

To confirm CI: (1)

(2)

(3)

Checking for Conditional Independence

Question: Are two variables X & Y **conditionally independent**?

(for a given set of other observed variables)!

Idea: Look for paths/connections between them

To confirm CI: (1) determine all „**paths**“

(2) check all paths whether they have a „**blocked node**“

(3)

Checking for Conditional Independence

Question: Are two variables X & Y **conditionally independent**?

(for a given set of other observed variables)!

Idea: Look for paths/connections between them

To confirm CI:

(1) determine all „**paths**"

(2) check all paths whether they have a „**blocked node**"

(3) Does **every** path have a blocked node?

If yes: then X & Y are CI   If no: then X & Y are not CI

Checking for Conditional Independence

Question: Are two variables X & Y **conditionally independent**?

(for a given set of other observed variables)!

Idea: Look for paths/connections between them

To confirm CI:

(1) determine all „**paths**“

(2) check all paths whether they have a „**blocked node**“

(3) Does **every** path have a blocked node?

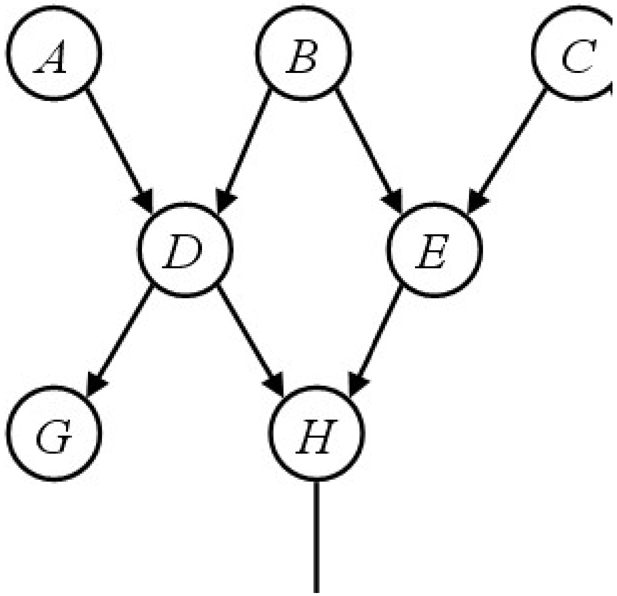
If yes: then X & Y are CI   If no: then X & Y are not CI

Left to understand:

a) What is a path?

b) What is a blocked node?

a) What is a Path?

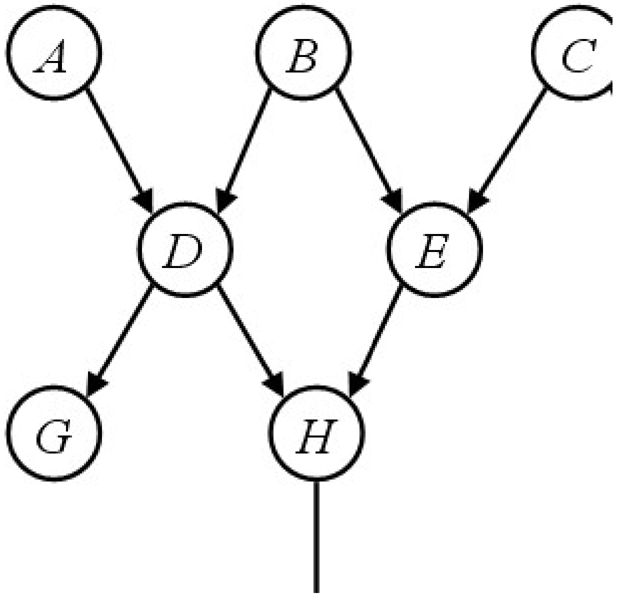


Look for: Connections from a node X to a node Y (*all directions allowed*)

Consider: Single routes *with* directed edges (without forks)

Example: Paths from G to E ??

a) What is a Path?



Look for: Connections from a node X to a node Y (*all directions allowed*)

Consider: Single routes *with* directed edges (without forks)

Example: From G to E (path 1: GDBE, path 2: GDHE)

b) What is a Blocked Node?

Blocked Node: A node Z that – e.g. when observed – is such that

a considered path in the graph is such that

the distribution of X **does not** change if Y is observed

b) What is a Blocked Node?

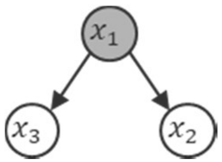
Blocked Node: A node Z that – e.g. when observed – is such that

a considered path in the graph is such that

the distribution of X **does not** change if Y is observed

When this is the case on a route of edges? Consider 3 nodes in a row (**triples**):

(1) Tail-to-Tail (Fork)



x1 (obs.) is a blocking node

(2) Head-to-Tail (Chain)

(3) Head-to-Head (Sink)

b) What is a Blocked Node?

Blocked Node: A node Z that – e.g. when observed – is such that

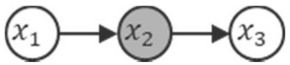
a considered path in the graph is such that

the distribution of X **does not** change if Y is observed

When this is the case on a route of edges? Consider 3 nodes in a row (**triples**):

(1) Tail-to-Tail (Fork)

(2) Head-to-Tail (Chain)



x2 (obs.) is a blocking node

(3) Head-to-Head (Sink)

b) What is a Blocked Node?

Blocked Node: A node Z that – e.g. when observed – is such that

a considered path in the graph is such that

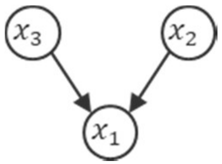
the distribution of X **does not** change if Y is observed

When this is the case on a route of edges? Consider 3 nodes in a row (**triples**):

(1) Tail-to-Tail (Fork)

(2) Head-to-Tail (Chain)

(3) Head-to-Head (Sink)

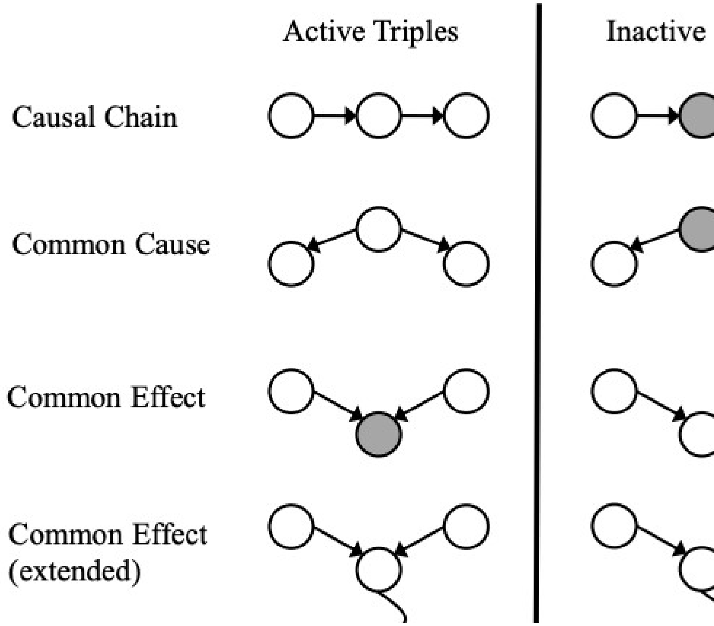


x1 (free) is a blocking node

Putting it together: D-Separation

Check: Are X & Y CI „for a given set of observed variables"?

X & Y are CI if: **All paths** contain **at least** one **inactive triple**



Putting it together: D-Separation

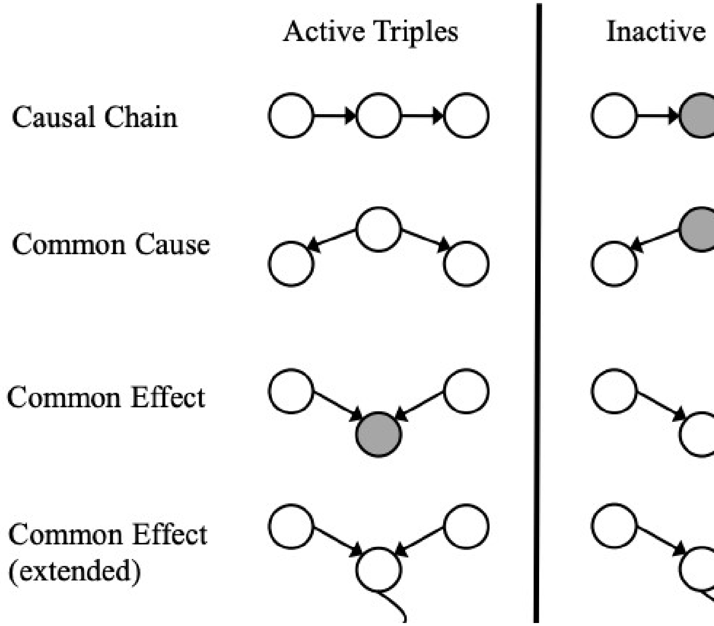
Check: Are X & Y CI „for a given set of observed variables“?

X & Y are CI if: **All paths** contain **at least** one **inactive triple**

(1) Tail-to-Tail (**Fork** with root observation)

(2) Head-to-Tail (**Chain** with middle observation)

(3) Head-to-Head (**Merge** with clean sink)



X & Y not CI if: If **there is** an **active path** between them (without any blocking node)

e.g., if they are neighbors (child/parent)

Overview

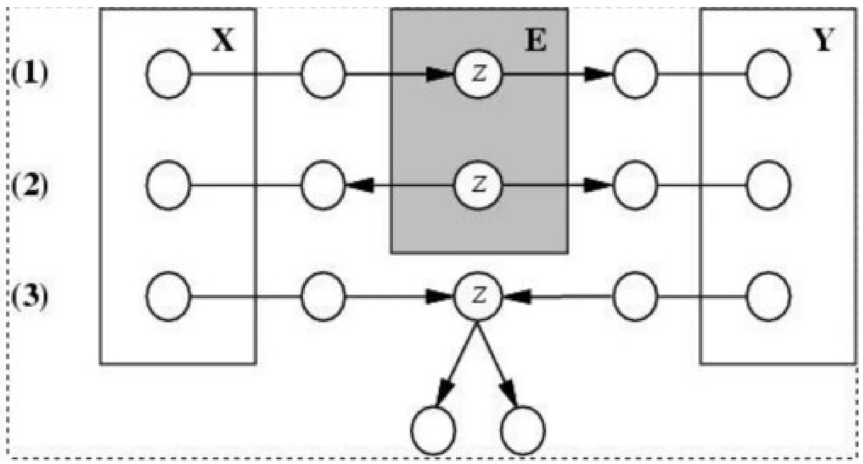
1. Questions and Updates
2. Recap: Main Concepts of Unit 3
3. **Example: Conditional Independence & D-Separation**
4. Recap: Main Concepts of Unit 4
5. Example: Message Passing in Factor Graphs
6. Hints for Exercise 2 (to be handed in Monday May 13, 7:00)

Conditional Independence: Examples & Generalizations

Look for: (1) Head-to-Tail (**Chain** with middle obs.)

(2) Tail-to-Tail (**Fork** with root observation)

(3) Head-to-Head (**Merge** with clean sink)

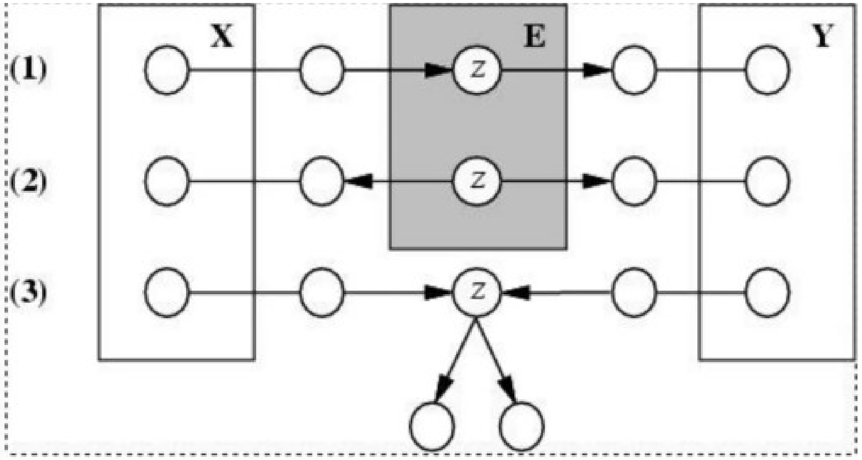


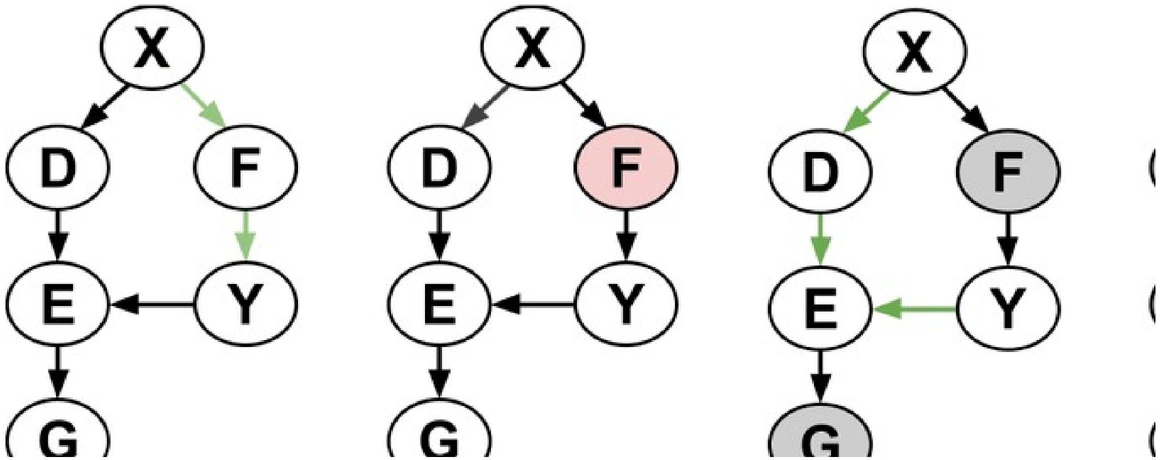
Conditional Independence: Examples & Generalizations

Look for: (1) Head-to-Tail (**Chain** with middle obs.)

(2) Tail-to-Tail (**Fork** with root observation)

(3) Head-to-Head (**Merge** with clean sink)





Recap: Proof for Case H2T (Unit 3, slide 12)

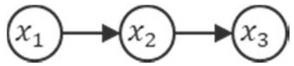
Proof for Chain Graph (Head-to-Tail): $p(x_1, x_3) = p(x_1) \cdot p(x_3 \mid x_1) \neq p(x_1) \cdot p(x_3)$

$$x_1 \rightarrow x_2 \rightarrow x_3$$

$$p(x_1, x_3) = p(x_1) \cdot \sum_{x_2} p(x_2|x_1) \cdot p(x_3|x_2) = p(x_1) \cdot p(x_3|x_1) \neq p(x_1) \cdot p(x_3)$$

Recap: Proof for Case H2T (Unit 3, slide 12)

Proof for Chain Graph (Head-to-Tail): $p(x_1, x_3) = p(x_1) \cdot p(x_3 \mid x_1) \neq p(x_1) \cdot p(x_3)$



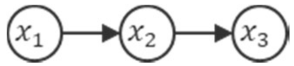
Note (1)

$$p(x_1, x_2, x_3) \overset{general}{=} p(x_1) \cdot p(x_2 \mid x_1) \cdot p(x_3 \mid x_1, x_2) \overset{chain}{=} p(x_1) \cdot p(x_2 \mid x_1) \cdot p(x_3 \mid x_2)$$

$$\Rightarrow p(x_3 \mid x_2) \overset{chain}{=} p(x_3 \mid x_1, x_2)$$

Recap: Proof for Case H2T (Unit 3, slide 12)

Proof for Chain Graph (Head-to-Tail): $p(x_1, x_3) = p(x_1) \cdot p(x_3 \mid x_1) \neq p(x_1) \cdot p(x_3)$



Note (1) $$p(x_1, x_2, x_3) \overset{general}{=} p(x_1) \cdot p(x_2 \mid x_1) \cdot p(x_3 \mid x_1, x_2) \overset{chain}{=} p(x_1) \cdot p(x_2 \mid x_1) \cdot p(x_3 \mid x_2)$$

$$\Rightarrow p(x_3 \mid x_2) \overset{chain}{=} p(x_3 \mid x_1, x_2)$$

Note (2) $$p(x_3 \mid x_2) \overset{general}{=} \frac{p(x_3, x_2)}{p(x_2)}$$

Does this also hold in the world „given x1"?

Recap: Proof for Case H2T (Unit 3, slide 12)

Proof for Chain Graph (Head-to-Tail): $p(x_1, x_3) = p(x_1) \cdot p(x_3 \mid x_1) \neq p(x_1) \cdot p(x_3)$

$$x_1 \rightarrow x_2 \rightarrow x_3$$

Note (1)

$$p(x_1, x_2, x_3) \overset{general}{=} p(x_1) \cdot p(x_2 \mid x_1) \cdot p(x_3 \mid x_1, x_2) \overset{chain}{=} p(x_1) \cdot p(x_2 \mid x_1) \cdot p(x_3 \mid x_2)$$

$$\Rightarrow p(x_3 \mid x_2) \overset{chain}{=} p(x_3 \mid x_1, x_2)$$

Note (2)

$$p(x_3 \mid x_2) \overset{general}{=} \frac{p(x_3, x_2)}{p(x_2)}$$

This also holds in the world „given x1"!

$$\frac{p(x_3, x_2 \mid x_1)}{p(x_2 \mid x_1)} \overset{general}{=} p(x_3 \mid x_2, x_1)$$

Recap: Proof for Case H2T (Unit 3, slide 12)

Proof for Chain Graph (Head-to-Tail): $p(x_1, x_3) = p(x_1)\cdot p(x_3 \mid x_1) \neq p(x_1)\cdot p(x_3)$

$$x_1 \to x_2 \to x_3$$

Note (1)

$$p(x_1,x_2,x_3) \overset{general}{=} p(x_1)\cdot p(x_2 \mid x_1)\cdot p(x_3 \mid x_1,x_2) \overset{chain}{=} p(x_1)\cdot p(x_2 \mid x_1)\cdot p(x_3 \mid x_2)$$

$$\Rightarrow p(x_3 \mid x_2) \overset{chain}{=} p(x_3 \mid x_1, x_2) \leftarrow$$

Note (2)

$$p(x_3 \mid x_2) \overset{general}{=} \frac{p(x_3,x_2)}{p(x_2)}$$

This also holds in the world „given x1"!

$$\frac{p(x_3,x_2 \mid x_1)}{p(x_2 \mid x_1)} \overset{general}{=} p(x_3 \mid x_2, x_1)$$

Hence:

$$p(x_1,x_3) = p(x_1)\cdot \sum_{x_2} p(x_2 \mid x_1)\cdot p(x_3 \mid x_2) \overset{(1)}{=} p(x_1)\cdot \sum_{x_2} p(x_2 \mid x_1)\cdot p(x_3 \mid x_1, x_2)$$

Recap: Proof for Case H2T (Unit 3, slide 12)

Proof for Chain Graph (Head-to-Tail): $p(x_1, x_3) = p(x_1) \cdot p(x_3 \mid x_1) \neq p(x_1) \cdot p(x_3)$

$$x_1 \rightarrow x_2 \rightarrow x_3$$

Note (1)

$$p(x_1, x_2, x_3) \overset{general}{=} p(x_1) \cdot p(x_2 \mid x_1) \cdot p(x_3 \mid x_1, x_2) \overset{chain}{=} p(x_1) \cdot p(x_2 \mid x_1) \cdot p(x_3 \mid x_2)$$

$$\Rightarrow p(x_3 \mid x_2) \overset{chain}{=} p(x_3 \mid x_1, x_2)$$

Note (2)

$$p(x_3 \mid x_2) \overset{general}{=} \frac{p(x_3, x_2)}{p(x_2)}$$

This also holds in the world „given x1"!

$$\frac{p(x_3, x_2 \mid x_1)}{p(x_2 \mid x_1)} \overset{general}{=} p(x_3 \mid x_2, x_1)$$

Hence:

$$p(x_1, x_3) = p(x_1) \cdot \sum_{x_2} p(x_2 \mid x_1) \cdot p(x_3 \mid x_2) \overset{(1)}{=} p(x_1) \cdot \sum_{x_2} p(x_2 \mid x_1) \cdot p(x_3 \mid x_1, x_2)$$

$$\overset{(2)}{=} p(x_1) \cdot \sum_{x_2} p(x_2 \mid x_1) \cdot \frac{p(x_3, x_2 \mid x_1)}{p(x_2 \mid x_1)} = p(x_1) \cdot \sum_{x_2} p(x_3, x_2 \mid x_1) = p(x_1) \cdot p(x_3 \mid x_1)$$

Overview

1. Questions and Updates
2. Recap: Main Concepts of Unit 3
3. Example: Conditional Independence & D-Separation
4. **Simulating 1 vs 1 TrueSkill (discrete)**

Simulating TrueSkill (cp. Unit 2, slide 11-12)

Consider TrueSkill 1 vs 1 with discrete variables!

Simulate Skills:

$$P(s_i) := 1/N, \qquad s_i = 1, \dots, N, \;\; i = 1, 2, \;\; N = 20$$

Simulate Performances:

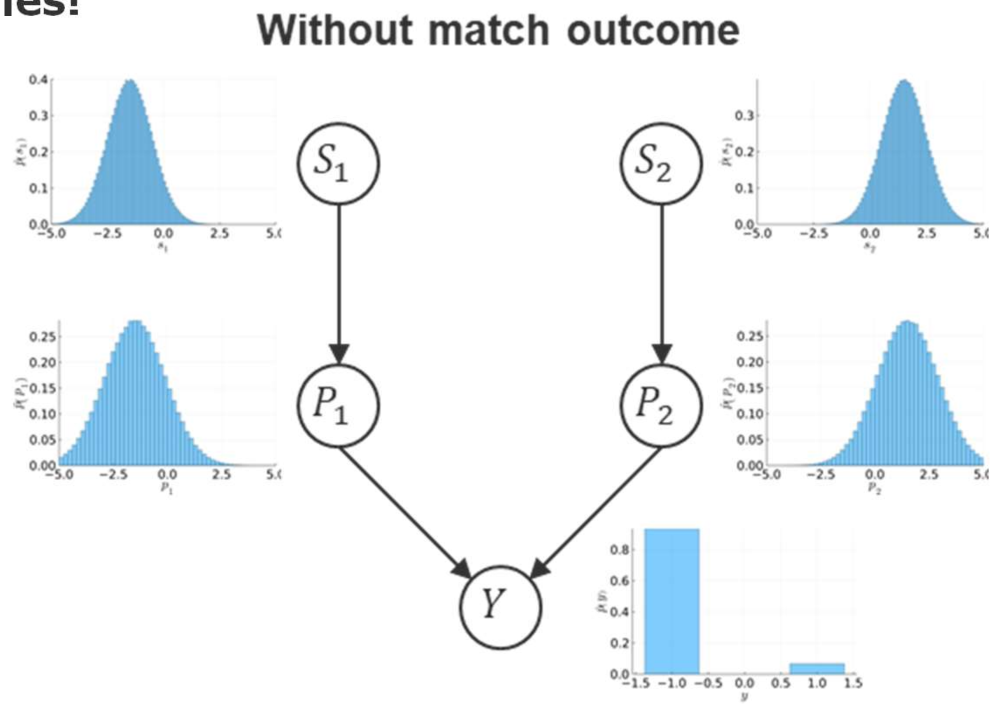
$$P(p_i \mid s_i) \propto N - |s_i - p_i|, \qquad s_i, p_i = 1, \dots, N, \;\; i = 1, 2$$

Evaluate Differences:

$$d = p_1 - p_2$$

Evaluate Outcomes:

$$y := 1_{\{d>0\}}$$



(a) Simulate e.g. 1000 vectors $(s_1^{(k)}, s_2^{(k)}, p_1^{(k)}, p_2^{(k)}, d^{(k)}, y^{(k)}), \qquad k = 1, \dots, 1000$

Simulating TrueSkill (cp. Unit 2, slide 11-12)

Consider TrueSkill 1 vs 1 with discrete variables!

Simulate Skills:

$$P(s_i) := 1/N, \qquad s_i = 1,\ldots,N,\ \ i = 1,2,\ \ N = 20$$

Simulate Performances:

$$P(p_i \mid s_i) \propto N - |s_i - p_i|, \qquad s_i, p_i = 1,\ldots,N,\ \ i = 1,2$$

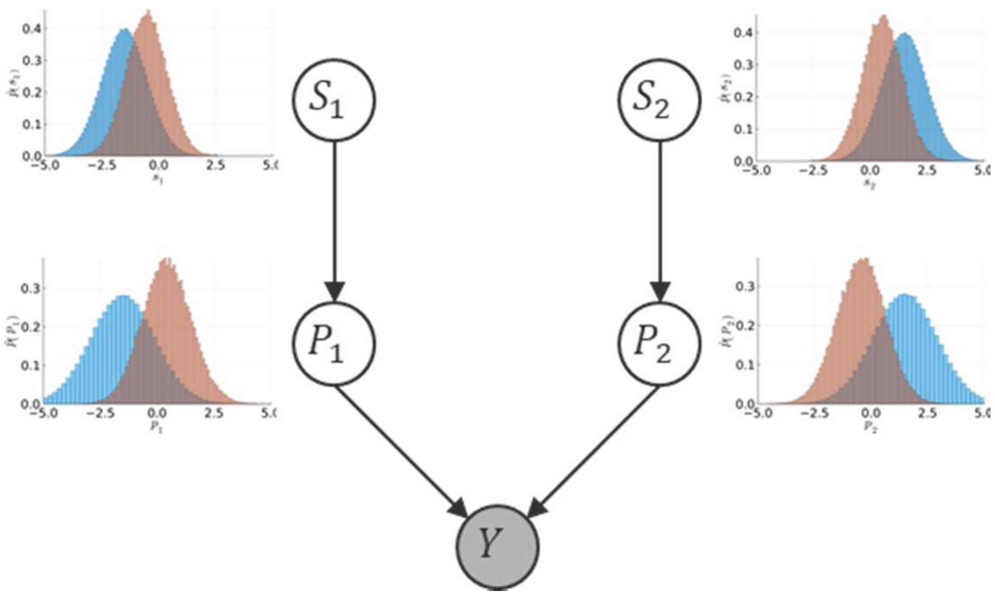
Evaluate Differences:

$$d = p_1 - p_2$$

Evaluate Outcomes:

$$y := 1_{\{d>0\}}$$

With match outcome ($y = 1$)



(b) Evaluate Skills & Performances conditioned on y=1! Doable?

Summary

- Recap I: Conditional Probabilities
- Recap II: Conditional Independence in Bayesian Networks
- Recap III: Simulation of Networks

See you next Week!