# Reasoning Under Uncertainty: Independence and Inference

Alan Mackworth

UBC CS 322 – Uncertainty 5

March 20, 2013

Textbook §6.3.1, 6.5, 6.5.1, 6.5.2

# Lecture Overview

- Recap: Bayesian Networks
- Inference in Special Types of Bayesian Networks
  - Markov Chains
  - Hidden Markov Models (HMMs)
- Inference in General Bayesian Networks
  - Observations and Inference
  - Time Permitting: Variable elimination

# Recap: Conditional Independence

**Definition (Conditional independence)**
Random variable X is (conditionally) independent of random variable Y given random variable Z, written $X \perp\!\!\!\perp Y \mid Z$ if, for all $x \in \mathrm{dom}(X)$, $y_j \in \mathrm{dom}(Y)$, $y_k \in \mathrm{dom}(Y)$ and $z \in \mathrm{dom}(Z)$ the following equation holds:

$$
\begin{aligned}
& P(X = x \mid Y = y_j, Z = z) \\
= & P(X = x \mid Y = y_k, Z = z) \\
= & P(X = x \mid Z = z)
\end{aligned}
$$

- Definition of $X \perp\!\!\!\perp Y \mid Z$ in distribution form: $P(X \mid Y, Z) = P(X \mid Z)$

# Recap: Bayesian Networks, Definition

**Definition (Bayesian Network)**

A Bayesian network consists of

- A directed acyclic graph $(V, E)$ whose nodes are labeled with random variables
- A domain for each random variable
- A conditional probability distribution for each variable $X$
  - Specifies $P(X|Parents(X))$
  - $Parents(X)$ is the set of variables $X'$ with $(X', X) \in E$
    - For nodes X without predecessors, $Parents(X) = \{\,\}$

- Chain rule: $P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|X_1,\ldots,X_{i-1})$
- Bayesian Network semantics:
  - A variable is conditionally independent of its non-descendants given its parents
  - $X_i \perp\!\!\!\perp \{X_1,\ldots,X_{i-1}\} \setminus Pa(X_i) \mid Pa(X_i)$
  - I.e., $P(X_i|X_1,\ldots,X_{i-1}) = P(X_i|pa(X_i))$

# Recap: Construction of Bayesian Networks

Encoding the joint over $\{X_1, \ldots, X_n\}$ as a Bayesian network:

- Totally order the variables: e.g., $X_1, \ldots, X_n$
- For every variable $X_i$, find the smallest set of parents $Pa(X_i) \subseteq \{X_1, \ldots, X_{i-1}\}$ such that $X_i \perp\!\!\!\perp \{X_1, \ldots, X_{i-1}\} \setminus Pa(X_i) \mid Pa(X_i)$
  - $X_i$ is conditionally independent from its other ancestors given its parents
- For every variable $X_i$, construct its conditional probability table
  - $P(X_i \mid Pa(X_i))$
  - This has to specify a conditional probability distribution $P(X_i \mid Pa(X_i) = pa(X_i))$ for every instantiation $pa(X_i)$ of $X_i$'s parents
  - If a variable has 3 parents each of which has a domain with 4 values, how many instantiations of its parents are there?

$3^4$ $4^3$ $3*4$ $4^3-1$

# Recap: Construction of Bayesian Networks

Encoding the JPD over $\{X_1, \ldots, X_n\}$ as a Bayesian network:

- Totally order the variables: e.g., $X_1, \ldots, X_n$
- For every variable $X_i$, find the smallest set of parents $Pa(X_i) \subseteq \{X_1, \ldots, X_{i-1}\}$ such that $X_i \perp\!\!\!\perp \{X_1, \ldots, X_{i-1}\} \setminus Pa(X_i) \mid Pa(X_i)$
  - $X_i$ is conditionally independent from its other ancestors given its parents
- For every variable $X_i$, construct its conditional probability table
  - $P(X_i \mid Pa(X_i))$
  - This has to specify a conditional probability distribution $P(X_i \mid Pa(X_i) = pa(X_i))$ for every instantiation $pa(X_i)$ of $X_i$'s parents
  - If a variable has 3 parents each of which has a domain with 4 values, how many instantiations of its parents are there?
    - $4 * 4 * 4 = 4^3$
    - For each of these $4^3$ values we need one probability distribution defined over the values of $X_i$
    - So need $[(|dom(X_i)| - 1) * 4^3]$ numbers in total for $X_i$'s CPT

# Recap of BN construction with a small example

- Two Boolean variables: Disease and Symptom

1. The causal ordering: Disease, Symptom
2. Chain rule:
   P(Disease, Symptom) = P(Disease) × P(Symptom |Disease)
3. Is Disease $\perp\!\!\!\perp$ Symptom | {} ?
   - I.e., are they marginally independent (conditioned on nothing)?

Yes No

| Disease *D* | Symptom S | *P(D,S)* |
|---|---|---|
| t | t | 0.0099 |
| t | f | 0.0001 |
| f | t | 0.0990 |
| f | f | 0.8910 |

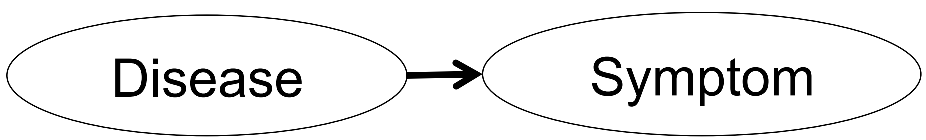
| Disease *D* | *P(D)* |
|---|---|
| t | 0.01 |
| f | 0.99 |

| Symptom S | *P(S)* |
|---|---|
| t | 0.1089 |
| f | 0.8911 |

# Recap of BN construction with a small example

- Two Boolean variables: Disease and Symptom

1. The causal ordering: Disease, Symptom
2. Chain rule:
   P(Disease, Symptom) = P(Disease) × P(Symptom |Disease)
3. Is Disease ⫫ Symptom | {} ?
   - I.e., are they marginally independent (conditioned on nothing)?
   - No! That would mean P(D,S) = P(D) × P(S), which is not true
   - We have to put an edge from the parent (Disease) to the child (Symptom)




| Disease *D* | Symptom S | *P(D,S)* |
|---|---|---|
| t | t | 0.0099 |
| t | f | 0.0001 |
| f | t | 0.0990 |
| f | f | 0.8910 |

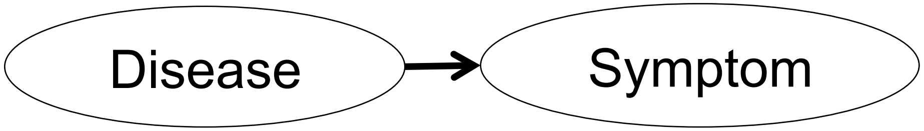
| Disease *D* | *P(D)* |
|---|---|
| t | 0.01 |
| f | 0.99 |

| Symptom S | *P(S)* |
|---|---|
| t | 0.1089 |
| f | 0.8911 |

# Recap of BN construction with a small example

- Which (conditional) probability tables do we need?

P(D) P(D|S) P(S|D) P(D,S)




| Disease *D* | Symptom S | *P(D,S)* |
|---|---|---|
| t | t | 0.0099 |
| t | f | 0.0001 |
| f | t | 0.0990 |
| f | f | 0.8910 |

| Disease *D* | *P(D)* |
|---|---|
| t | 0.01 |
| f | 0.99 |

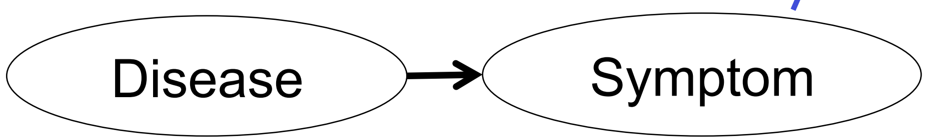
| Symptom S | *P(S)* |
|---|---|
| t | 0.1089 |
| f | 0.8911 |

# Recap of BN construction with a small example

- Which conditional probability tables do we need?
  - P(D) and P(S|D)
  - In general: for each variable X in the network: P( X|Pa(X) )

| *P(D=t)* |
|---|
| 0.01 |

| Disease *D* | $P(S=t \mid D)$ |
|---|---|
| t | 0.0099/(0.0099+0.0001)=0.99 |
| f | 0.099/(0.099+0.891)=0.1 |




| Disease *D* | Symptom S | *P(D,S)* |
|---|---|---|
| t | t | 0.0099 |
| t | f | 0.0001 |
| f | t | 0.0990 |
| f | f | 0.8910 |

| Disease *D* | *P(D)* |
|---|---|
| t | 0.01 |
| f | 0.99 |

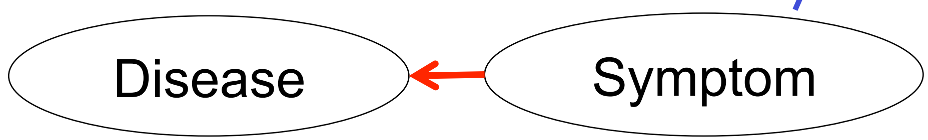
| Symptom S | *P(S)* |
|---|---|
| t | 0.1089 |
| f | 0.8911 |

# Recap of BN construction with a small example

- How about a different ordering? Symptom, Disease
  - We need distributions P(S) and P(D|S)
  - In general: for each variable X in the network: P( X|Pa(X) )

| Symptom S | *P(D=t\|S)* |
|---|---|
| t | 0.0099/(0.0099+0.099)=0.00909090 |
| f | 0.0001/(0.0001+0.891)=0.00011122 |

| *P(S=t)* |
|---|
| 0.1089 |




| Disease *D* | Symptom S | *P(D,S)* |
|---|---|---|
| t | t | 0.0099 |
| t | f | 0.0001 |
| f | t | 0.0990 |
| f | f | 0.8910 |

| Disease *D* | *P(D)* |
|---|---|
| t | 0.01 |
| f | 0.99 |

| Symptom S | *P(S)* |
|---|---|
| t | 0.1089 |
| f | 0.8911 |

# Remark: where do the conditional probabilities come from?

- The joint distribution is not normally the starting point
  - We would have to define exponentially many numbers

- First define the Bayesian network structure
  - Either by domain knowledge
  - Or by machine learning algorithms (see CPSC 540)
    - Typically based on local search

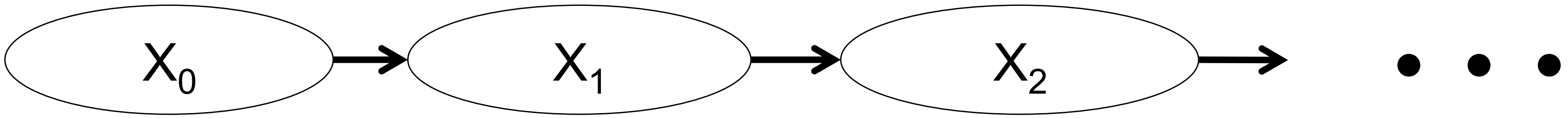
- Then fill in the conditional probability tables
  - Either by domain knowledge
  - Or by machine learning algorithms (see CPSC 340, CPSC 422)
    - Based on statistics over the observed data

# Lecture Overview

- Recap: Bayesian Networks
- Inference in Special Types of Bayesian Networks
  - Markov Chains
  - Hidden Markov Models (HMMs)
- Inference in General Bayesian Networks
  - Observations and Inference
  - Time Permitting: Factors and Variable Elimination

# Markov Chains

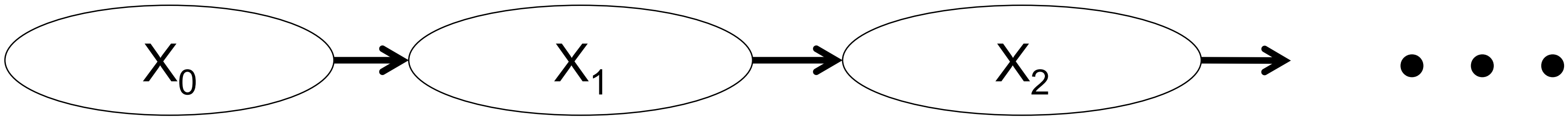
- A Markov chain is a special kind of belief network:




- $X_t$ represents a state at time t.
- Its dependence structure yields: $P(X_t|X_1, \ldots, X_{t-1}) = P(X_t|X_{t-1})$
  - This conditional probability distribution is called the state transition probability
  - Intuitively $X_t$ conveys all of the information about the history that can affect the future states:
    "The past is independent of the future given the present."

- JPD of a Markov Chain: $P(X_0,\ldots,X_T) = P(X_0) \times \prod_{t=1}^{T} P(X_t|X_{t-1})$

# Stationary Markov Chains



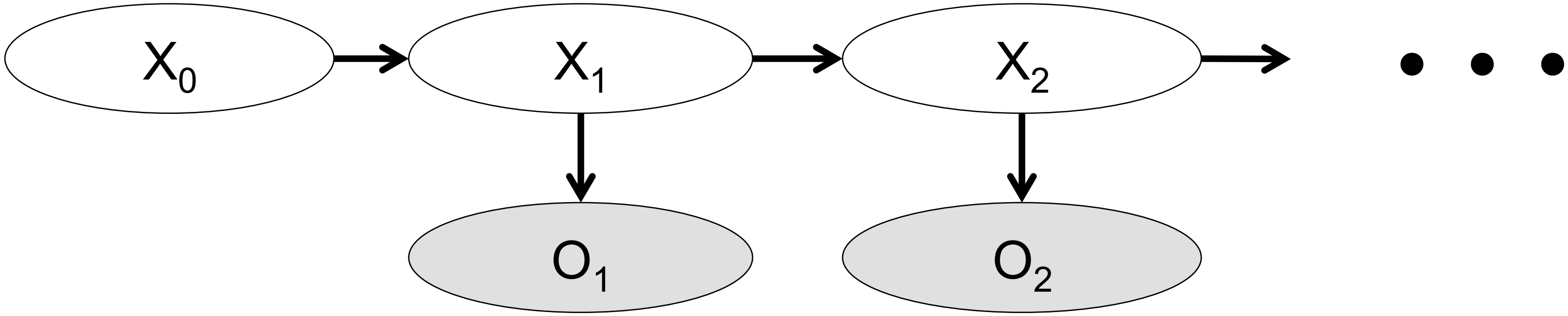

- A stationary Markov chain is when
  - All state transition probability tables are the same
  - I.e., for all t > 0, t' > 0: $P(X_t|X_{t-1}) = P(X_{t'}|X_{t'-1})$
- We only need to specify $P(X_0)$ and $P(X_t|X_{t-1})$.
  - Simple model, easy to specify
  - Often the natural model
  - The network can extend indefinitely in time
- Example: Drunkard's walk, robot random motion

# Lecture Overview

- Recap: Bayesian Networks
- Inference in Special Types of Bayesian Networks
  - Markov Chains
  - Hidden Markov Models (HMMs)
- Inference in General Bayesian Networks
  - Observations and Inference
  - Time Permitting: Factors and Variable Elimination

# Hidden Markov Models (HMMs)

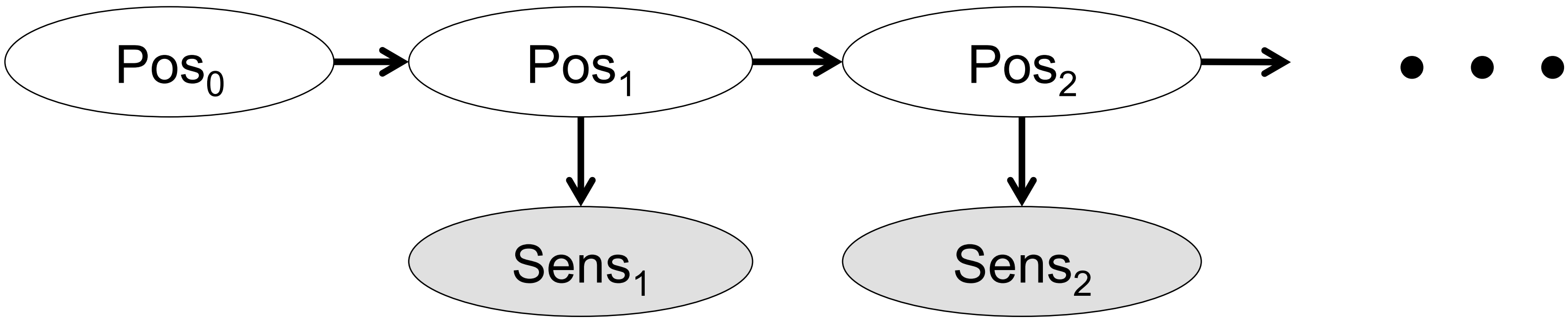
- A Hidden Markov Model (HMM) is a stationary Markov chain plus a noisy observation about the state at each time step:




- Same conditional probability tables at each time step
  - The state transition probability $P(X_t|X_{t-1})$
    - also called the system dynamics
  - The observation probability $P(O_t|X_t)$
    - also called the sensor model
- JPD of an HMM: $P(X_0, \ldots, X_T, O_1, \ldots, O_T)$
  $= P(X_0) \times \prod_{t=1}^{T} P(X_t|X_{t-1}) \times \prod_{t=1}^{T} P(O_t|X_t)$

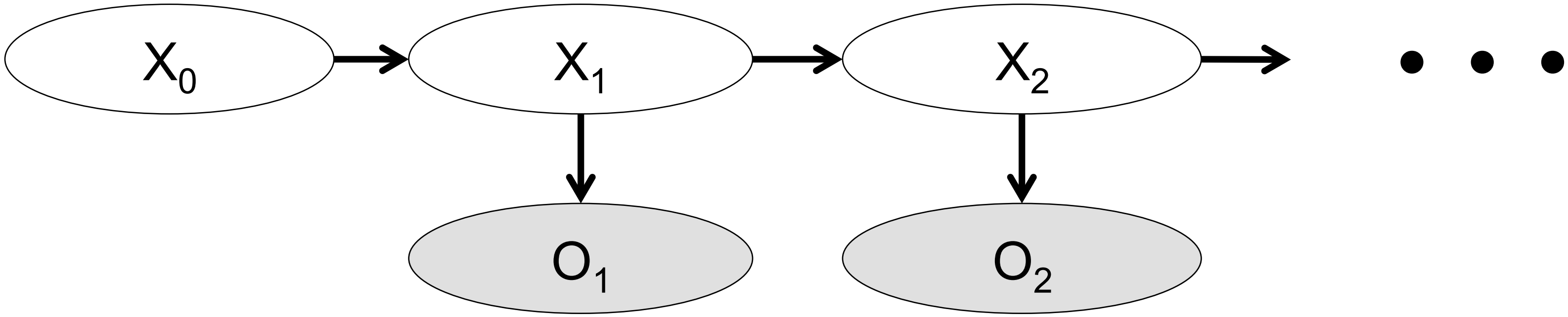
# Example HMM: Robot Tracking

- Robot tracking as an HMM:




- Robot is moving at random: $P(Pos_t|Pos_{t-1})$
- Sensor observations of the current state $P(Sens_t|Pos_t)$

# Filtering in Hidden Markov Models (HMMs)




- Filtering problem in HMMs:
  at time step t, we would like to know $P(X_t|o_1, \ldots, o_t)$
- We can derive simple update equations for this belief state:
  - We are given $P(X_0)$ (i.e., $P(X_0 \mid \{\})$)
  - We can compute $P(X_t|O_1, \ldots, O_t)$ if we know $P(X_{t-1}|o_1, \ldots, o_{t-1})$
  - Simple example of dynamic programming
  - See P&M text Section 6.5.3 (not responsible for this for exam!)

# Lecture Overview

- Recap: Bayesian Networks
- Inference in Special Types of Bayesian Networks
  - Markov Chains
  - Hidden Markov Models (HMMs)
- Inference in General Bayesian Networks
  - Observations and Inference
  - Time Permitting: Factors and Variable Elimination

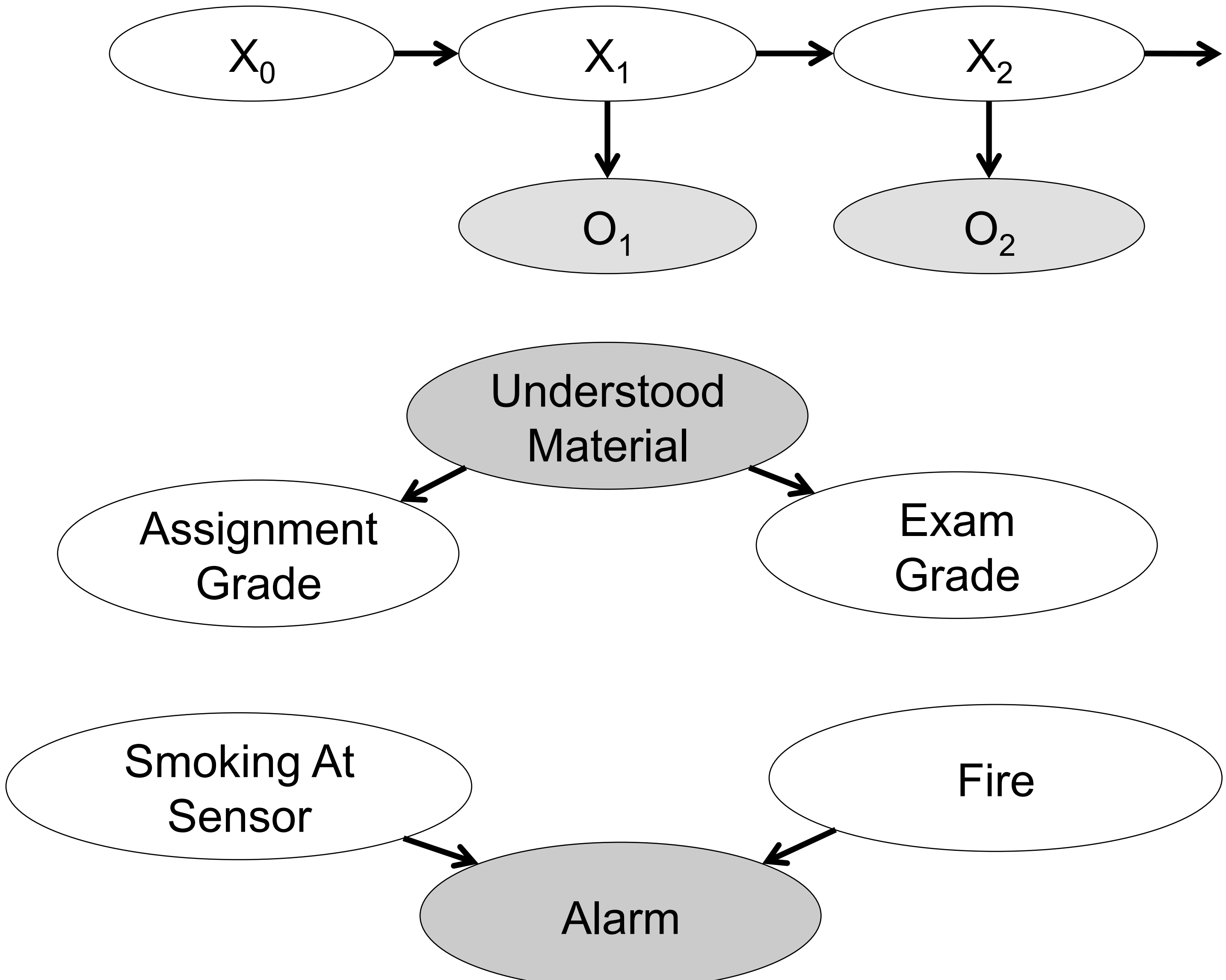
# Bayesian Networks: Incorporating Observations

- In the special case of Hidden Markov Models (HMMs):
  - we could easily incorporate observations
  - and do efficient inference (in particular: filtering)

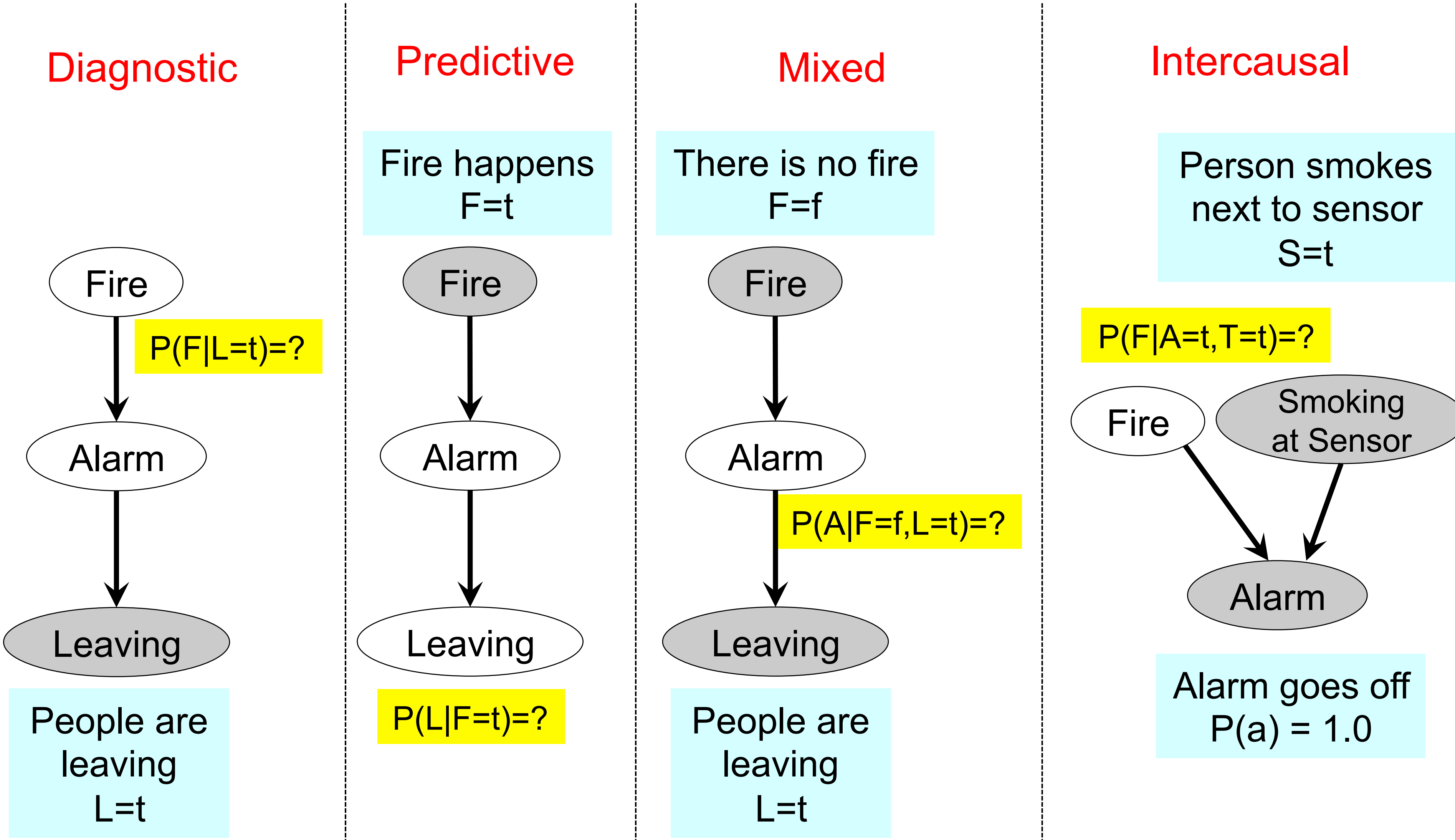
- Back to general Bayesian Networks
  - We can still incorporate observations
  - And we can still do (fairly) efficient inference

# Bayesian Networks: Incorporating Observations

We denote observed variables as shaded. Examples:



# Bayesian Networks: Types of Inference




We will use the same reasoning procedure for all of these types

# Inference in Bayesian Networks

Given:

- A Bayesian Network BN, and
- Observations of a subset of its variables E: E=e
- A subset of its variables Y that is queried

Compute: The conditional probability P(Y|E=e)

How: Run variable elimination algorithm

N.B. We can already do all this: See lecture “Uncertainty2” Inference by Enumeration topic.

The BN represents the JPD. Could just multiply out the BN to get full JPD and then do Inference by Enumeration BUT that’s extremely inefficient - does not scale.

# Lecture Overview

- Inference in General Bayesian Networks
  - Factors:
    - Assigning Variables
    - Summing out Variables
    - Multiplication of Factors
  - The variable elimination algorithm

# Factors

- A factor is a function from a tuple of random variables to the real numbers R
- We write a factor on variables $X_1, \dots, X_j$ as $f(X_1, \dots, X_j)$

- P(Z|X,Y) is a factor f (X,Y,Z)
  - Factors do not have to sum to one
  - P(Z|X,Y) is a set of probability distributions: one for each combination of values of X and Y

- P(Z=f|X,Y) is a factor f(X,Y)

$f(X, Y)_{Z=f}$

| X | Y | Z | val |
|---|---|---|---|
| t | t | t | 0.1 |
| t | t | f | 0.9 |
| t | f | t | 0.2 |
| t | f | f | 0.8 |
| f | t | t | 0.4 |
| f | t | f | 0.6 |
| f | f | t | 0.3 |
| f | f | f | 0.7 |

# Operation 1: assigning a variable

- We can make new factors out of an existing factor

- Our first operation:
  we can assign some or all of the variables of a factor.

f(X,Y,Z):

| X | Y | Z | val |
|---|---|---|---|
| t | t | t | 0.1 |
| t | t | f | 0.9 |
| t | f | t | 0.2 |
| t | f | f | 0.8 |
| ~~f~~ | ~~t~~ | ~~t~~ | ~~0.4~~ |
| ~~f~~ | ~~t~~ | ~~f~~ | ~~0.6~~ |
| ~~f~~ | ~~f~~ | ~~t~~ | ~~0.3~~ |
| ~~f~~ | ~~f~~ | ~~f~~ | ~~0.7~~ |

*What is the result of assigning X= t ?*

$f(X=t,Y,Z) = f(X, Y, Z)_{X = t}$

| Y | Z | val |
|---|---|---|
| t | t | 0.1 |
| t | f | 0.9 |
| f | t | 0.2 |
| f | f | 0.8 |

Factor of Y, Z

# More examples of assignment

f(X,Y,Z):

| X | Y | Z | val |
|---|---|---|---|
| t | t | t | 0.1 |
| t | t | f | 0.9 |
| t | f | t | 0.2 |
| t | f | f | 0.8 |
| f | t | t | 0.4 |
| f | t | f | 0.6 |
| f | f | t | 0.3 |
| f | f | f | 0.7 |

f(X=t,Y,Z)

Factor of Y,Z

| Y | Z | val |
|---|---|---|
| t | t | 0.1 |
| t | f | 0.9 |
| f | t | 0.2 |
| f | f | 0.8 |

f(X=t,Y,Z=f):

| Y | val |
|---|---|
| t | 0.9 |
| f | 0.8 |

Factor of Y

f(X=t,Y=f,Z=f): 0.8

Number

# Operation 2: Summing out a variable

- Our second operation on factors: we can marginalize out (or sum out) a variable
  - Exactly as before. Only difference: factors don't sum to 1
  - Marginalizing out a variable X from a factor f($X_1, \ldots, X_n$) yields a new factor defined on {$X_1, \ldots, X_n$} \ {X}

$$\left(\sum_{X_1} f\right)(X_2, \ldots, X_j) = \sum_{x \in dom(X_1)} f(X_1 = x, X_2, \ldots, X_j)$$

$f_3$=

| B | A | C | val |
|---|---|---|---|
| t | t | t | 0.03 |
| t | t | f | 0.07 |
| f | t | t | 0.54 |
| f | t | f | 0.36 |
| t | f | t | 0.06 |
| t | f | f | 0.14 |
| f | f | t | 0.48 |
| f | f | f | 0.32 |

$(\sum_B f_3)(A,C)$

| A | C | val |
|---|---|---|
| t | t | 0.57 |
| t | f | 0.43 |
| f | t | 0.54 |
| f | f | 0.46 |

# Operation 3: multiplying factors

$f_1(A,B)$:

| A | B | Val |
|---|---|---|
| t | t | 0.1 |
| t | f | 0.9 |
| f | t | 0.2 |
| f | f | 0.8 |

$f_2(B,C)$:

| B | C | Val |
|---|---|---|
| t | t | 0.3 |
| t | f | 0.7 |
| f | t | 0.6 |
| f | f | 0.4 |

$f_1(A,B) \times f_2(B,C)$:

| A | B | C | val |
|---|---|---|---|
| t | t | t | 0.03 |
| t | t | f | 0.1x0.7 |
| t | f | t | 0.9x0.6 |
| t | f | f | … |
| f | t | t | |
| f | t | f | |
| f | f | t | |
| f | f | f | |

# Operation 3: multiplying factors

- The product of factor $f_1(A, B)$ and $f_2(B, C)$, where $B$ is the variable in common, is the factor $(f_1 \times f_2)(A, B, C)$ defined by

$$(f_1 \times f_2)(A,B,C) = f_1(A,B) f_2(B,C)$$

- Note: A, B, and C can be sets of variables
  - The domain of $f_1 \times f_2$ is $A \cup B \cup C$

# Learning Goals For Today's Class

- Build a Bayesian Network for a given domain
- Understand basics of Markov Chains and Hidden Markov Models
- Classify the types of inference:
  - Diagnostic, Predictive, Mixed, Intercausal
- Understand factors

---

Assignment 4 available on Connect: Q1, Q2, Q3 and Q4 NOW. Q5: variable elimination (VE) next class.