Bayesian Statistics

Instructor: GP Saggese, PhD - `gsaggese@umd.edu`

References:

- AIMA (Artificial Intelligence: a Modern Approach)
  - Chap 12, Quantifying uncertainty
  - Chap 13: Probabilistic reasoning
  - Chap 14: Probabilistic reasoning over time

Quantifying uncertainty

- **Quantifying uncertainty**
- Probabilistic reasoning

Logic-based AI Acting Under Uncertainty

- Real-world agents face **uncertainty** from:
  - Partial observability (agent can't see the full state)
  - Non-determinism (actions don't always have predictable outcomes)
  - Adversarial conditions (other agents may interfere)
- In logic-based AI systems:
  - Actions are represented using **rules** like:
    - "If preconditions P hold, then action A causes effect E"
  - Example:
    - "If I turn the car key, the engine starts"
    - But: the battery might be dead, there's no fuel, the starter is broken, etc.
- Logical agents approach
  - Use a **belief state**: set of all possible current world states
  - Construct **contingent plans** that handle every possible sensor report
    - Must consider all possible explanations, even unlikely ones
    - Plans become large and complex
    - No guaranteed plan may exist, yet action is required

Causal and exhaustive augmentation

- To use propositional logic, augment the left-side of $X \implies Y$ to make it:
    1. **Causal**: identify true causal-effect relationships
    2. **Exhaustive**: identify all possible conditions leading to the outcome
- **Logical qualification problem**: trying to enumerate all the preconditions necessary for an action to succeed
- **Problems**
    1. **Laziness**: too much work to create all possible rules
    2. **Theoretical ignorance**: lack of understanding
        - Science doesn't always have a complete theory of the domain
        - E.g., medical science doesn't know all the rules
    3. **Practical ignorance**: lack of facts
        - Even if we knew all the rules, we might not have all the information needed
        - E.g., not all necessary tests can be run for a particular patient
- This led to expert systems failure and AI winter (mid 1980s, 1990s)
    - The real world is complex and open-ended
    - Logical rules can't capture all necessary and sufficient conditions

Failure of logic-based AI: wet grass example

- Consider the propositions:
  - $Rain$ = "it rains"
  - $WetGrass$ = "the grass is wet"
  - $Cover$ = "there is a protective cover over the grass"
  - $Evaporate$ = "the water evaporates quickly"
  - $Sprinkler$ = "the sprinkler system is on"
  - Dew = "there is morning dew"
- $Rain \implies WetGrass$ is not true in general
  - If it rains but there is a cover over the grass, the grass will not be wet
  - If it rains but there is high temperature, the wet grass might dry quickly
- $WetGrass \implies Rain$ is not true in general
  - The grass could be wet because of a sprinkler system
  - The grass could be wet because of morning dew
- Identify all exceptions, alternative explanations, and dependencies
  1. Causal
     - $Rain \implies (WetGrass \vee (Cover \vee Evaporate \ldots)$: "if it rains and there is no other source of water, the grass will be wet"
  2. Exhaustive
     - $WetGrass \iff (Rain \vee (Sprinkler \vee Dew \ldots)$: "if it rains and there is no protective cover, the grass will be wet"

Acting Under Uncertainty: solution

- We can't use propositional logic under uncertainty
  - Need approaches (like probabilistic reasoning) that handle uncertainty and partial knowledge
- Acting under uncertainty requires combining:
  - **Probabilities**: for possible outcomes
  - **Utilities**: for evaluating desirability of each outcome
- **Key idea:**
  - Rational choice = plan that maximizes expected utility
  - Evaluate plans based on performance on average, given known information
  - Even if success is not guaranteed
- Rational decision depends on:
  - **Performance measure**: combines goals like punctuality, comfort, legal compliance
  - **Belief**: agent's internal estimate of outcome likelihoods

Probability and knowledge

- The confusing part is that there no uncertainty in the actual world
  - E.g., the grass is wet, but either it has rained or not
- Probabilities relate to a knowledge state, not the real world
  - Updating knowledge can change probability statements
- E.g., updating belief about wet grass and rain:
  - Initially, we observe wet grass, and from past data we know that:
    - $\Pr(Rain|WetGrass) = 0.8$: 80% chance it rained if grass is wet
  - Learn new information:
    - Sprinkler was on
    - Wet grass could be due to the sprinkler, not rain
    - Belief changes: $\Pr(Rain|WetGrass) = 0.4$
  - Further observe:
    - Weather report says there was no rain
    - Certain it did not rain, despite wet grass
    - Overrides prior evidence: $\Pr(Rain|WetGrass) = 0$

Probabilistic reasoning

- Quantifying uncertainty
- **Probabilistic reasoning**

Full joint probability distribution

- Consider a set of random variables $X_1, X_2, \ldots, X_n$
- The **full joint probability distribution** assigns a probability to every possible world:

$$\Pr(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

  - A possible world = a particular assignment of values to all variables
  - Can answer any probabilistic query about the domain
- **Cons**
  - Size grows exponentially k^n with the number of variables n and size k
  - Impractical for real-world problems with many variables
  - Manually specifying each entry is tedious
- **Independence** and **conditional independence** simplify modeling
  - In the real world, many variables are not fully dependent on all others
  - Reduces the number of parameters needed
  - Makes compact and structured representations possible
    - E.g., factorized probabilistic models, Bayesian networks

Independence of Random Variables: Definition

- Two random variables X and Y are **independent** iff:

$$\Pr(X, Y) = \Pr(X) \cdot \Pr(Y)$$

- Equivalently, knowing Y tells us nothing about X, $\Pr(X|Y) = \Pr(X)$
- E.g.,
  - The events "coin flip result" and "weather" are independent
  - $\Pr(\text{Coin=Heads}|\text{Weather=Rainy}) = \Pr(\text{Coin=Heads})$
- Independence reduces the number of parameters needed to model a system
  - Allows factorization of joint distribution, if all variables are mutually independent, e.g.,

$$\Pr(X_1, X_2, X_3) = \Pr(X_1) \cdot \Pr(X_2) \cdot \Pr(X_3)$$

Conditional Independence: Definition

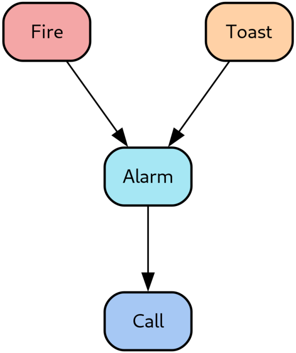
- Two random variables X and Y are **conditionally independent** given a random variable Z iff knowing Z makes X and Y independent:

$$\Pr(X, Y|Z) = \Pr(X|Z)\Pr(Y|Z)$$

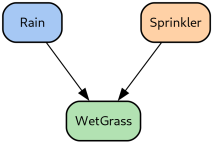
- E.g.,
  - X = "it is raining today"
  - Y = "if a person is carrying an umbrella"
  - Z = "the weather forecast"
  - Without Z, there is a relationship between X and Y (X and Y are not independent)
  - Given Z, rain X may not directly influence whether a person carries an umbrella Y
  - Thus, X and Y can be conditionally independent given Z
- True independence is rare; conditional independence is more common and useful
- Conditional independence simplifies probabilistic models
  - It reduces the joint conditional distribution to the product of individual conditional distributions

Conditional Independence: Example

- Two events can become independent once we know a third event
- **Example:**
  - *Fire*: "there is a fire"
  - *Toast*: "someone burned toast"
  - *Alarm*: "the alarm rings"
  - *Call*: "a friend calls to check on you"
- **Dependencies**:
  - *Alarm* depends on *Fire* or *Toast*
  - *Call* depends on whether *Alarm* rings
- **Conditional independence**:
  - $\Pr(\mathit{Call} \mid \mathit{Alarm}, \mathit{Fire}) = \Pr(\mathit{Call} \mid \mathit{Alarm})$
  - Once we know the alarm rang, the specific cause doesn't affect whether the friend calls
- **Interpretation**:
  - *Call* is conditionally independent of *Fire* given *Alarm*
  - Knowing the alarm rang "blocks" the path of influence from *Fire* to *Call*



Conditional Independence: Garden Example



- Garden world with *Rain*, *Sprinkler*, and *WetGrass*
- Is $\Pr(Rain|Sprinkler) = \Pr(Rain)$?
  - **No**: if the sprinkler is on, it's less likely it rained
  - *Rain* and *Sprinkler* are not independent
- Is $\Pr(Rain|Sprinkler, WetGrass) = \Pr(Rain|WetGrass)$?
  - **Yes**: knowing the grass is wet, whether the sprinkler was on tells us nothing more about the rain
  - *Rain* and *Sprinkler* are conditionally independent given *WetGrass*
- **Interpretation**:
  - Without *WetGrass*: *Rain* and *Sprinkler* affect each other because they both explain *WetGrass*
  - With *WetGrass*: once *WetGrass* is observed, the "explaining away" effect occurs
- **"Explaining away" occurs when**
  - Two variables (causes) independently influence a third variable (effect)
  - Observing the effect creates a dependence between the causes

Bayesian Networks: Definition

- Aka:
  - "Bayes net"
  - "Belief networks"
  - "Probabilistic networks"
  - "Graphical models" (somehow a broader class of statistical models)
  - "Causal networks" (arrows have constraints that have special meaning)

- **Formal definition (syntax)**
  - A Bayesian network is a Directed Acyclic Graph (DAG)

1. **Nodes** X_i correspond to random variables (discrete or continuous)
2. **Edges** connect nodes $X \rightarrow Y$ representing direct dependencies among variables
   - We say that $X = Parent(Y)$
   - The edges form a direct acyclic graph (DAG)
3. Each node X_i is associated with a **conditional probability**:

$$\Pr(X_i | Parents(X_i))$$

   quantifying the effect of the parents on the node

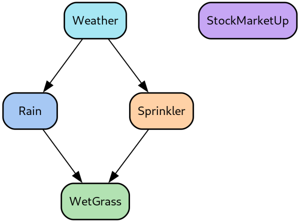
- CPD specifies the probability of the node given its parents
- If a node has no parents, it has a **prior probability**

Bayesian network: intuition

- Bayesian networks can represent:
  - **Any full joint** distribution
  - Often **very concisely**, representing dependencies among variables
- The topology of the network (nodes and edges) specifies conditional independence relationships
  - E.g., $X \to Y$ means "X has a direct influence on Y", i.e., "X relates to Y" (not necessarily "causes")
  - Domain experts can decide what relationships exist among domain variables, determining the topology
- In the Bayesian network graph:
  - Nodes are directly influenced by their parents
  - Nodes are indirectly influenced by all their ancestors
- Conditional probabilities can be specified/estimated

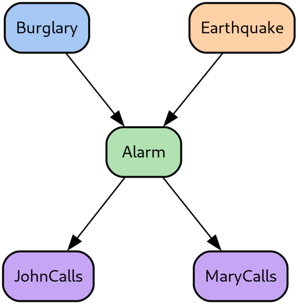
Bayesian Networks: Wet Grass Example

- Consider a world with 5 variables
  - *Rain*, *Sprinkler*, *WetGrass*, *StockMarketUp*, *Weather*
  - *Weather* affects both *Rain* and *Sprinkler*
  - *WetGrass* is affected by both *Rain* and *Sprinkler*
  - *StockMarketUp* is independent of all the other variables
- Independence assumptions:
  - *Rain* and *Sprinkler* are **conditionally dependent** given *Weather*
  - *Rain* and *Sprinkler* are **conditionally independent** given *WetGrass*, but only if *Weather* is not observed
  - *StockMarketUp* is completely independent of all other variables



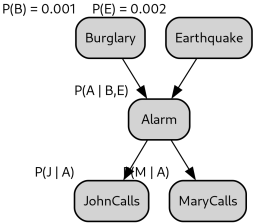
Bayesian Networks: Burglar Example

- (Famous example from Judea Pearl)
- There is an *Alarm* system installed at a home in LA
  - Fairly reliable at detecting *Burglary*
  - Also responds to minor *Earthquakes* (false positive)
- You have two neighbors, *John* and *Mary*, who will *Call* you when they hear the *Alarm*
  - *John*:
    - Almost always *Call*s when he hears the alarm
    - Sometimes confuses telephone ringing with the *Alarm* and *Call*s (false positive)
  - *Mary*:
    - Misses the alarm 30% of the cases (false negative)



Bayesian networks: burglar example (2/3)

- The structure of the graph shows that:
  - *Burglary* and *Earthquake* affects the event *Alarm*
  - *JohnCalls* and *MaryCalls* depend only on the *Alarm*, and not on *Burglary* and *Earthquake*



Bayesian networks: burglar example (3/3)

- The probability of *Burglary* is 0.001
- The probability of *Earthquake* is 0.002
- Compute $\Pr(Alarm) = f(Burglary, Earthquake)$ since events are independent

Burglary	**Earthquake**	**P(Alarm| B,E)**
True	True	0.70
True	False	0.01
False	True	0.70
False	False	0.01

- *JohnCalls* and *MaryCalls* are represented by:

Alarm (A)	**P(JohnCalls| .)**
True	0.90
False	0.05

Alarm (A)	**P(MaryCalls| .)**
True	0.70
False	0.01

Conditional Probability Table

- Aka CPT
- Each row contains the conditional probability of the node under a conditioning case (i.e., a possible combination of the values for the parent nodes)
  - Natural for discrete variables, but extendable to continuous variables
- **Note**: a conditional probability table summarizes an infinite set of circumstances in the table
  - E.g., *MaryCalls* could depend on her being at work, asleep, passing of a helicopter, . . .

Conditional probability table: examples

Alarm (A)	P(JohnCalls| .)	P(-JohnCalls | .)
True	0.90	0.10
False	0.05	0.95

Alarm (A)	P(JohnCalls| .)
True	0.90
False	0.05

P(Burglary)
.001

Burglary	Earthquake	P(Alarm | .)
T	T	.95
T	F	.94
...	...	...

- The sum of probabilities of the actions must be 1
- Removing the redundancy
- A node without parents has an unconditional probability
- A node with k parents has 2^k possible rows in the table

Bayesian Networks: Semantics

- There are two equivalent semantic interpretations:

1. **Joint Distribution View**
    - The network encodes the **joint probability distribution** over all variables
    - Computed as the product of local conditional probabilities:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \text{Parents}(X_i))$$

    - Helps in constructing models and understanding overall behavior
2. **Conditional Independence View**
    - The structure encodes **conditional independency** between variables
    - A variable is conditionally independent of its non-descendants given its parents
    - Useful for efficient inference and reasoning

Chain rule for a joint distribution

- A joint distribution can always be expressed using the chain rule for any:
  - Set of RVs
  - Ordering of the RVs
- We express one variable conditionally to the remaining ones

$$\Pr(x_1, ..., x_{n-1}, x_n) = \Pr(x_n | x_{n-1}, ..., x_1) \Pr(x_{n-1}, ..., x_1)$$

- Then we apply the same formula recursively, until we get an unconditional probability

$$\begin{aligned}
&\Pr(x_1, ..., x_n) \\
&= \Pr(x_n | x_{n-1}, ..., x_1) \Pr(x_{n-1}, ..., x_1) \\
&= \Pr(x_n | x_{n-1}, ..., x_1) \Pr(x_{n-1} | x_{n-2}, ..., x_1) \Pr(x_{n-2}, ..., x_1) \\
&... \\
&= \Pr(x_n | x_{n-1}, ..., x_1) \Pr(x_{n-1} | x_{n-2}, ..., x_1) \Pr(x_{n-2} | x_{n-3}, ..., x_1) ... \Pr(x_2 | x_1) \Pr(x_1) \\
&= \prod_{i=1}^{n} \Pr(x_i | x_{i-1}, ..., x_1)
\end{aligned}$$

Probability of a statement from a Bayesian network

- The full joint distribution represents the probability of an assignment to each variable $X_i = x_i$: $\Pr(x_1, ..., x_n) = \Pr(X_1 = x_1 \wedge ... \wedge X_n = x_n)$
- To evaluate a Bayesian network
  - Sort the nodes in topological order (there are several orderings consistent with the directed graph structure)
  - Use the chain rule with the topological ordering:

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

  - Since the probability of each node is conditionally independent of its predecessors (all nodes) given its parents

$$\Pr(X_i | X_{i-1}, ..., X_1) = \Pr(X_i | Parents(X_i))$$

  - Express the joint probability in terms of the CPTs:

$$\Pr(X_1, ..., X_n) = \prod_{i=1}^{n} \Pr(X_i | Parents(X_i))$$

Probability of a statement from a Bayesian network: example

- Given Pearl LA example, we want to compute the probability that:
  - The alarm has sounded: Alarm
  - Neither a burglary nor an earthquake has occurred: $\neg \mathit{Burglary} \wedge \neg \mathit{Earthquake}$
  - Both John and Mary call: $\mathit{JohnCalls}, \mathit{MaryCalls}$
- The solution is to compute:

$$\begin{aligned}&\Pr(\mathit{JohnCalls}, \mathit{MaryCalls}, \mathit{Alarm}, \neg \mathit{Burglary}, \neg \mathit{Earthquake})\\&= \Pr(\mathit{JohnCalls}|\mathit{Alarm}) \Pr(\mathit{MaryCalls}|\mathit{Alarm}) \Pr(\mathit{Alarm}|\neg \mathit{Burglary} \wedge \neg \mathit{Earthquake})\end{aligned}$$

Constructing a Bayesian network

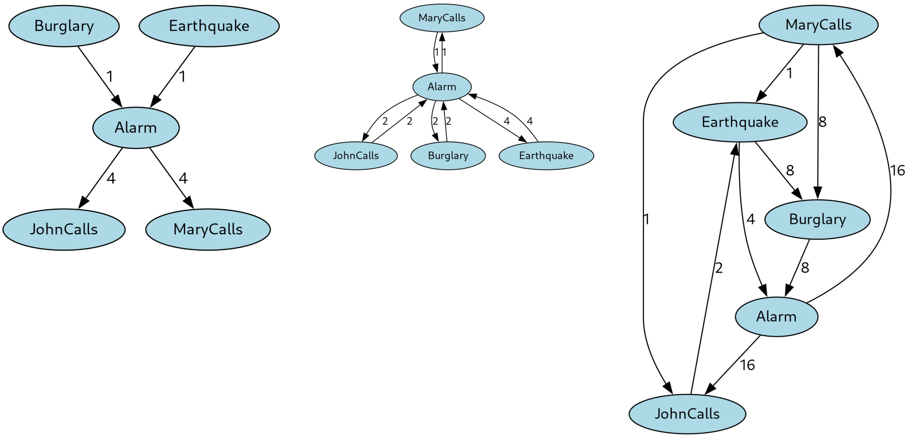
- Gather domain knowledge
  - Identify key variables and their potential interactions
  - Understand the problem context and objectives
- Determine the random variables required to model the problem X_i
  - List all relevant random variables necessary to describe the system
- Order the nodes according to the dependencies implied by cause-effects
  - Determine causal relationships between variables
  - The Bayesian network is minimal when nodes are ordered by cause-effect
- For each node, pick the minimum set of parents $Parents(X_i)$
  - Select parents that directly influence the node X_i
  - Avoid redundant connections, ensuring the network remains minimal
  - Add edges to represent the dependencies
- Estimate the conditional probability CPTs $\Pr(X_i|Parents(X_i))$ for each node
  - Gather data or expert opinion to estimate probabilities
  - Use statistical techniques for parameter estimation if necessary
- Validate the network structure with domain experts
  - Ensure that the network is a Directed Acyclic Graph (DAG)
  - E.g., test the network by predicting known outcomes and comparing with actual data

Bayesian networks

- Bayesian networks are a representation with several interesting properties
  - **Complete**
    - Encode all information in a joint probability
  - **Consistent** (non-redundant)
    - In a Bayesian network, there are no redundant probability values
    - One (e.g., a domain expert) can't create a Bayesian network violating probability axioms
  - **Compact** (locally structured, sparse)
    - Each subcomponent interacts directly with a limited number of other components
    - Typically yields linear (not exponential) growth in complexity
    - Sometimes we ignore real-world dependency to keep the graph simple
- **Fully connected systems**
  - Domains where each variable is influenced by all others
  - The Bayesian network is fully connected, with complexity like the joint probability

Ordering of nodes

- The complexity of the Bayesian network depends on the choice in ordering the nodes

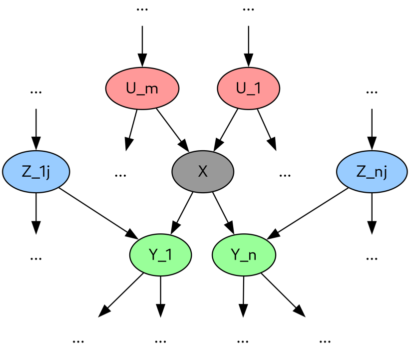


Causal vs diagnostic models

- A **causal model** goes from causes to symptoms
  - Often simpler (i.e., fewer dependencies) and "easier" to estimate
- A **diagnostic model** goes from symptoms to causes
  - E.g., $\mathit{MaryCalls} \rightarrow \mathit{Alarm}$, or $\mathit{Alarm} \rightarrow \mathit{Burglary}$
  - These relationship are:
    - Tenuous
    - Difficult to estimate (or unnatural)

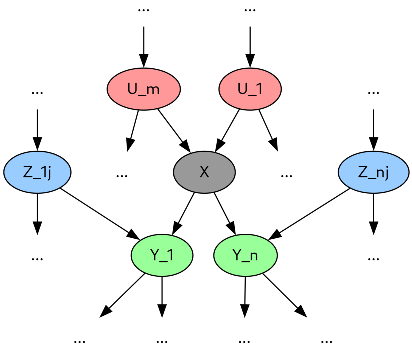
Markov blanket of a node

- The Markov blanket of a node X_i consists of:
  - The parents of X_i (red nodes), i.e., the nodes that influence X_i
  - The children of X_i (green nodes), i.e., the nodes that are directly influenced by X_i
  - The spouses of X_i (blue nodes), i.e., the nodes that are parents of the children nodes, i.e., that share a child with the node in question



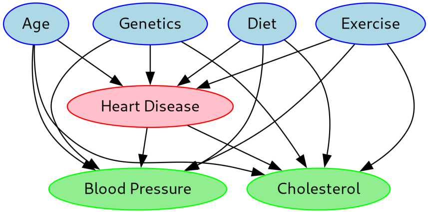
Conditional independence on the Markov blanket

- By construction, each variable is conditionally independent of its predecessors, given its parents
- In a Bayesian network, a variable is conditionally independent of *all other nodes* in the network given its Markov blanket (its parents, its children, and its spouses)
- The Markov blanket of a node X_i contains all the nodes necessary to predict the state of the node X_i, making the network irrelevant
    - This enables efficient and localized inference



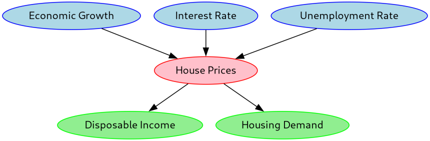
Markov blanket: medical example

- Consider risk factors and outcomes for heart disease
- Target node
  - H: Heart disease
- Parent nodes (direct influence of H, risk factors)
  - A: Age
  - G: Genetic predisposition
  - D: Diet
  - E: Exercise level
- Child (direct influenced by H, outcomes)
  - BP: Blood pressure
  - C: Cholesterol level
- Note that A, G, D, E also influence BP and C so they are spouse nodes of H
- Knowing the state of A, G, D, E, BP allows to compute H, without any other information



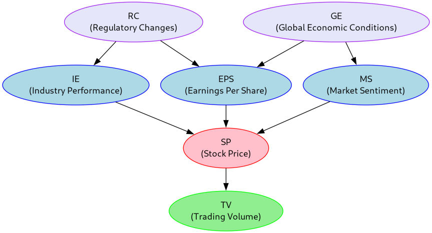
Markov blanket: economic example

- Consider factors affecting house prices in a particular region
- Target node
  - HP: House prices
- Parent
  - E: Economic growth
  - IR: Interest rate
  - UE: Unemployment rate
- Child
  - DI: Disposable income
    - The house price affects how much money people have left after housing costs
  - D: Demand for houses
    - Higher prices can reduce demand



Markov blanket: finance example

- Consider factors affecting an individual company's stock price
- Target node
  - *SP*: Stock Price
- Parent
  - *EPS*: Earnings per share
  - *IE*: Industry performance
  - *MS*: Market sentiment
- Child
  - *TV*: Trading volume
    - Changes in stock price influence how much stock is being traded
- Spouse
  - *RC*: Regulatory changes in the technology sector
    - Influences *IE* and *EPS*, but not directly *TV*
  - *GE*: Global economic conditions
    - Influences *MS* and *EPS*, but not directly *TV*



Specifying a Conditional Probability Table

- Even with a small number of parents k, the Conditional Probability Table (CPT) for a node requires $O(2^k)$ values in the worst case
- Often, the relationship is not completely arbitrary
- **Deterministic nodes** have values specified by their parents, without uncertainty, e.g.,
  - A logical relationship:
    - $\mathit{IsNorthAmerican} = \mathit{IsCanadian} \vee \mathit{IsUS} \vee \mathit{IsMexican}$
  - A numerical relationship:
    - $\mathit{BestPrice} = \min(\mathit{Price}_i)$

Noisy logical relationships

- Noisy logical relationships are a probabilistic version of a logical relationship
  - E.g., noisy-OR, noisy-MAX distribution
  - Noisy nodes can be simpler to describe given the k parents

Example

- A "noisy-OR" is a probabilistic version of a logical $\vee$
  - E.g., in propositional logic $Fever \iff Cold \vee Flu \vee Malaria$
- The assumptions are:
  1. All the possible causes are listed (one can use a leak node for "misc causes")
  2. There is uncertainty about the ability of the parents to be the cause of the child node, i.e., a probability that a cause is inhibited
  3. The probabilities of inhibition are independent
- Under these assumptions:

$$
\begin{aligned}
&\Pr(fever|parents(Fever)) \\
&1cm = 1 - \Pr(\neg fever|cold, \neg flu, \neg malaria)\cdot \\
&1cm \Pr(\neg fever|\neg cold, flu, \neg malaria)\cdot \\
&1cm \Pr(\neg fever|\neg cold, \neg flu, malaria)
\end{aligned}
$$

Context-specific independence

- A variable exhibits **context-specific independence** if it is conditionally independent of its parents given certain values of others, e.g.,
- *Damage* occurs during a period of time depending on the *Ruggedness* of your car and whether an *Accident* occurred in that period:

  $\Pr(Damage|Ruggedness, Accident) = d1$ else $d2(Ruggedness)$ if Accident

  where $d1$ and $d2$ are distributions

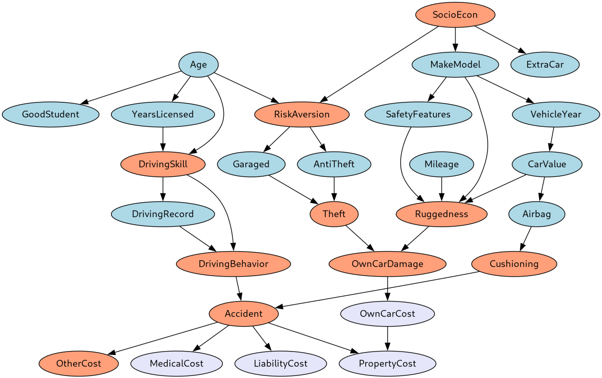
Bayesian networks with continuous variables

- Many real world problems involve continuous quantities
  - E.g., height, mass, temperature, money
- We can't specify the Conditional Probability Table (CPT) for continuous RVs, but we can use:
  1. Discretization (i.e., use intervals)
     - Cons: loss of accuracy and large CPTs
  2. Continuous variables
     - Families of probability density functions (e.g., Gaussian distribution)
     - Non-parametric PDFs
- **Hybrid Bayesian** networks mix discrete and continuous variables in a Bayesian network
  - E.g., a customer buys some fruit depending on its cost

Bayesian network: car insurance company (1/2)

- A car insurance company:
  - Receives an application from an individual to insure a specific vehicle
  - Decides on appropriate annual premium to charge (based on the claims and pay out)
- Build a Bayes network that captures the causal structure of the domain
- There are 3 kind of claims
  - *MedicalCost*: injuries sustained by the applicant
  - *LiabilityCost*: lawsuits filed by other parties against applicant
  - *PropertyCost*: vehicle damage to either party and theft of the vehicle
- Input information
  - About the applicant: *Age*, *YearsWithLicense*, *DrivingRecord*, *GoodStudent*
  - About the vehicle: *MakeModel*, *VehicleYear*, *Airbag*, *SafetyFeatures*
  - About the driving situation: *Mileage*, *HasGarage*

Bayesian network: car insurance company (2/2)



- Blue nodes: information provided by the applicants
- Brown nodes: hidden variables (i.e., not input nor output)
- Lavender nodes: target variables