Knowledge Engineering

Sargur Srihari
srihari@cedar.buffalo.edu

Topics

- Picking Variables
- Determining Structure
- Determining Probabilities

Knowledge Engineering

- Going from given distribution to Bayesian network is more complex
- We have a vague model of the world
  - Need to crystallize it into network structure and parameters
- Task has several components
  - Each is subtle
  - Mistakes have consequences in quality of answers

Three tasks in model building

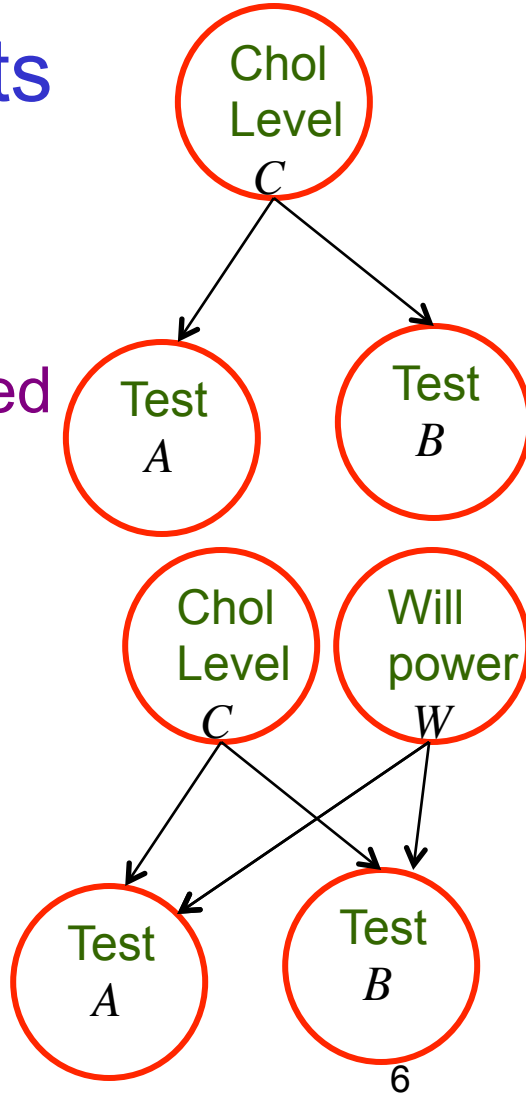
- All three tasks are hard:
  1. Picking variables
     - Many ways to pick entities and attributes
  2. Determining structure
     - Many structures hold
  3. Determining probabilities
     - Eliciting probabilities from people is hard

1. Picking Variables

- Model should contain variables
  - we can observe or that we will query
- Choosing variables is one of the hardest tasks
  - There are implications throughout the model
- Common problem: ill-defined variables
  - In medical domain: variable "Fever"
    - Temperature at time of admission?
    - Over prolonged period?
    - Thermometer or internal temperature?
  - Interaction of fever with other variables depend on specific interpretation

Need for Hidden Variables

- There are several Cholestorol Tests
- For accurate answers:
  - Nothing to eat after 10:00pm
  - If person eats, all tests become correlated
- Hidden variable: willpower
  - Including it will render:
    - cholestorol tests conditionally independent given true cholestorol level and willpower
- Hidden variables: to avoid all variables being correlated



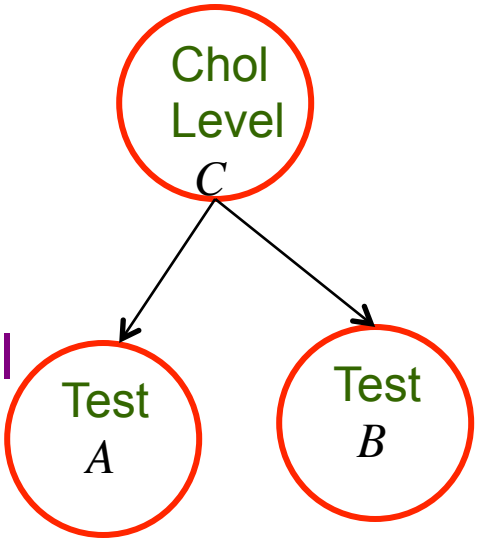
$A \perp B | C, W$

Some variables not needed

- Not necessary to include every variable
  - SAT score may depend on partying previous night
  - Probability already accounts for poor score despite intelligence

Picking Domain for Variables

- Reasonable domain of values to be chosen
- If partitions not fine enough conditional independence assumptions may be false
- Task of determining cholestorol level (C)
  - Two tests A and B
  - $(A \perp B|C)$
    - C: Normal if $\leq$ 200, High if > 200
    - Both tests fail if chol level has a marginal value(say 210)
      - Conditional independence assump. is false!
    - Introduce marginal value

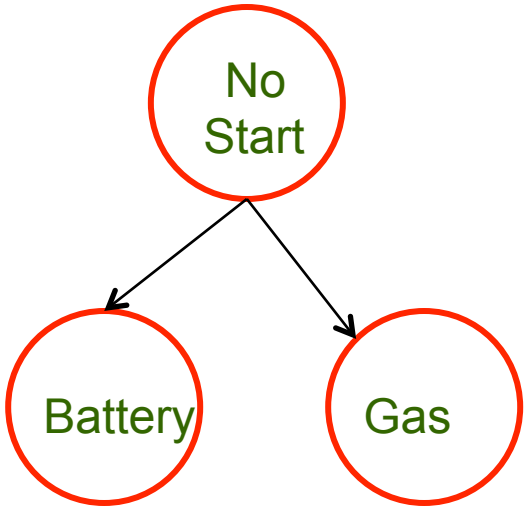


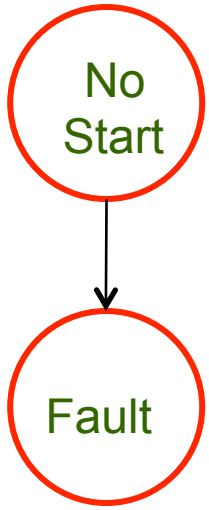
2. Picking Structure

- Many structures are consistent if we pick same set of independences
- Choose structure that reflects causal order and dependencies
  - Causes are parents of the effect
  - Causal graphs tend to be sparser
- Backward Construction Process
  - Lung cancer should have smoking as a parent
  - Smoking should have gender as a parent

Modeling weak influences

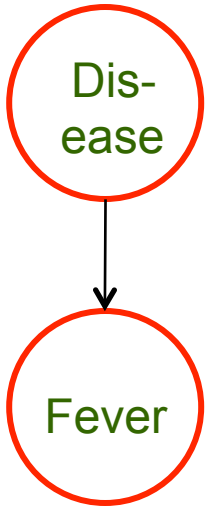
- Reasoning in a Bayesian network strongly depends on connectivity
- Adding edges can make it expensive to use
- Make approximations to decrease complexity





3. Picking Probabilities

- Zero Probabilities
  - Common mistake
    - Event extremely unlikely but not impossible
    - Can never condition away: irrecoverable errors
- Orders of Magnitude
  - Small diffs in low probs can make large differences in conclusions
    - 10^{-4} is very different from 10^{-5}
- Relative Values
  - Probability of fever higher with pneumonia than with flu



Disease \Fever	High	Lo
Pneum	0.9	0.1
Flu	0.6	0.4

Sensitivity Analysis

- Useful tool for estimating network parameters
- Determine extent to which a given probability parameter affects outcome
- Allows us to determine whether it is important to get a particular CPD entry right
- Helps figure out which CPD entries are responsible for an answer that does not match our intuition