6.825 Techniques in Artificial Intelligence

Where do Bayesian Networks Come From?

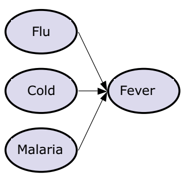
- Human experts
- Learning from data
- A combination of both

Human Experts

- Encoding rules obtained from experts, e.g. physicians for PathFinder
- Extracting these rules are very difficult, especially getting reliable probability estimates
- Some rules have a simple deterministic form:

- But, more commonly, we have many potential causes for a symptom and any one of these causes are sufficient for a symptom to be true

Multiple Independent Causes

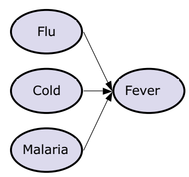


P(Fever | Flu) = 0.6
P(Fever | Cold) = 0.4
P(Fever | Malaria) = 0.9

In general, the table in the Fever node gives prob of fever given all combination of values of Flu, Cold and Malaria P(Fev | Flu, Col, Mal)

Big, and hard to assess

Noisy Or Example



P(Fever | Flu) = 0.6
P(Fever | Cold) = 0.4
P(Fever | Malaria) = 0.9

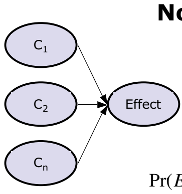
We are assuming that the causes act independently, which reduces the set of numbers that we need to acquire

Look only at the causes that are true:

$$P(Fev \mid Flu, \neg Col, Mal) = 1 - P(\neg Fev \mid Flu, Mal)$$

$$P(\neg Fev \mid Flu, Mal) = P(\neg Fev \mid Flu) P(\neg Fev, Mal) = (0.4)(0.1) = 0.04$$

Noisy Or



- Store $P(E|C_i)$ for all C_i
- Given a set, C_T, of true causes

$$\begin{aligned}\Pr(E|C) &= 1 - \Pr(\neg E|C) \\ &= 1 - \Pr(\neg E|C_T) \\ &= 1 - \prod_{C_i \in C_T} \Pr(\neg E \mid C_i) \\ &= 1 - \prod_{C_i \in C_T} (1 - \Pr(E \mid C_i))\end{aligned}$$

Recitation Problem

- Compute the conditional probability table for P(Fever | Flu, Cold, Malaria), for all assignments to the variables Flu, Cold, and Malaria.

Learning Bayesian Networks

- Instance of the general problem of probability density estimation
  - discrete space
  - interesting structure
- Four cases
  - structure known or unknown
  - all variables observable or some unobservable

This lecture: all variables observable, structure known or unknown

Known Structure

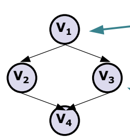
- Given nodes and arcs of a Bayesian network with m nodes
- Given a data set $D = \{\langle v_1^1,\ldots,v_m^1\rangle,\ldots,\{\langle v_1^k,\ldots,v_m^k\rangle\}$

values of nodes in sample 1   values of nodes in sample k

- Elements of D are assumed to be independent given M
- Find the model M (in this case, CPTs) that maximizes Pr(D|M)
- Known as the maximum likelihood model
- Humans are good at providing structure, data is good at providing numbers

Estimating Conditional Probabilities

- Use counts and definition of conditional probability
- Initializing all counters to 1 avoids 0 probabilities and converges on the maximum likelihood estimate



$$P(V_1) \approx \frac{\#(V_1=true)+1}{k+2}$$

$$P(V_3|V_1) \approx \frac{\#(V_3=true \wedge V_1=true)+1}{\#(V_1=true)+2}$$

$$P(V_3|\neg V_1) \approx \frac{\#(V_3=true \wedge V_1=false)+1}{\#(V_1=false)+2}$$

generally, the number of possible values of the variable on the left of the bar

Goodness of Fit

- Given data set D and model M, measure goodness of fit using log likelihood
- Assume each data sample generated independently

$$\Pr(D|M) = \prod_j \Pr(v^j|M)$$
$$= \prod_j \prod_i \Pr(N_i = v_i^j | Parents(N_i), M)$$

- Easier to compute the log; monotonic

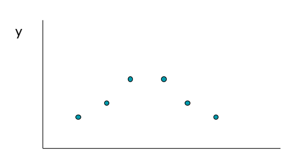
$$\log \Pr(D|M) = \log \prod_j \prod_i \Pr(N_i = v_i^j | Parents(N_i), M)$$
$$= \sum_j \sum_i \log \Pr(N_i = v_i^j | Parents(N_i), M)$$

Learning the Structure

- For a fixed structure, our counting estimates of the CPT converge to the maximum likelihood model
- What if we get to pick the structure as well?
- In general, the best model will have no conditional independence relationships
- Undesirable, for reasons of overfitting

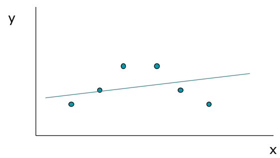
Overfitting

- Given a set of data points



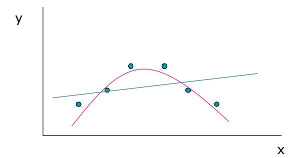
Overfitting

- Given a set of data points, you could
  - fit them with a line, with a lot of error



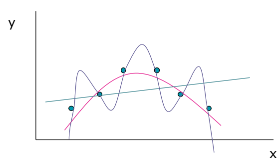
Overfitting

- Given a set of data points, you could
  - fit them with a line, with a lot of error
  - fit with a parabola, with a little error



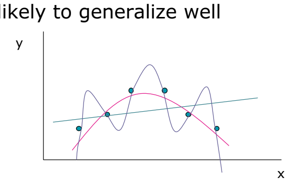
Overfitting

- Given a set of data points, you could
  - fit them with a line, with a lot of error
  - fit with a parabola, with a little error
  - fit with 10th order polynomial, with no error



Overfitting

- Given a set of data points, you could
  - fit them with a line, with a lot of error
  - fit with a parabola, with a little error
  - fit with 10th order polynomial, with no error
- 10th order polynomial over fits
  - less robust to variations in data
  - less likely to generalize well



Scoring Metric

- What if we want to vary the structure?

- We want a network that has conflicting properties
  - good fit to data: log likelihood
  - low complexity: total number of parameters

- Try to maximize scoring metric, by varying M (structure and parameters) given D

$$\log \Pr(D|M) - \alpha \# M$$

- Parameter α controls the tradeoff between fit and complexity

Search in Structure Space

- No direct way to find the best structure
- Too many to enumerate them all

- Start with some initial structure
- Do local search in structure space
  - neighborhood: add, delete, or reverse an arc
  - maintain no directed cycles
  - once you pick a structure, compute maximum-likelihood parameters, and then calculate the score of the model
  - increase score (or decrease sometimes, as in walkSAT or simulated annealing)

Initialization

Lots of choices!

- no arcs
- choose random ordering $V_1 \dots V_n$
  - variable V_i has all parents $V_1 \dots V_{n-1}$
  - variable V_i has parents randomly chosen from $V_1 \dots V_{n-1}$
- best tree network (can be computed in polynomial time)
  - compute pairwise mutual information between every pair of variables
  - find maximum-weight spanning tree

Recitation problem

Consider a domain with three binary nodes: A, B, and C

1. How many possible network structures are there over three nodes?

Data set: {<0,1,1>, <0, 1, 1>, <1,0,0>}

2. What parameter estimates would you get for the CPTs in each of the network structures on the following slide?
3. What is the log likelihood of the data given each of the models (given the estimates from the previous part)?
4. Do parts 2 and 3 again without the Bayesian correction (or with it, if you didn't use it the first time)
5. How many parameters are there in each of the models? (Don't count p and 1-p as separate parameters)

There are too many network structures for everyone to do every problem. So, if the day of your birthday is 0 mod 3, then do structures s1 and s2. If it's 1 mod 3, then do structures s3 and s4. And if it's 2 mod 3, then do structures s5 and s6.

Recitation Problem