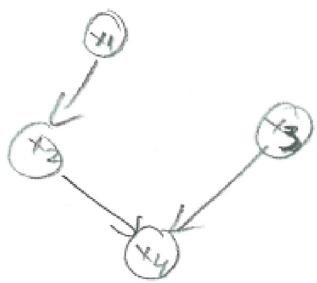
Alin Tomescu, http://people.csail.mit.edu/~alinush
6.867 Machine learning | Prof. Tommi Jaakkola | Week 14, Tuesday, December 3rd, 2013| Lecture 24

Lecture 24: More Bayesian networks

Today we'll focus on learning Bayesian networks from data.

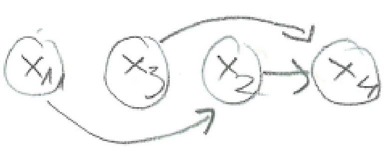
Graph:

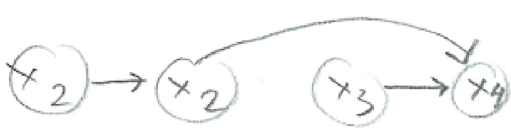
- Nodes are associated with random variables $X_1 \dots X_n$
- Graph is acyclic, as a representation of dependencies between the variables it must be acyclic
- Graph comes from specifying independence relations between variables
    - D-separation criterion gives us these independence relations
- Graph implies a partial ordering on the variables
    - Any variable coming later in the ordering must…
    - (See figure below)



$x_1 \perp\!\!\!\perp x_3$

$x_1 \perp\!\!\!\perp x_4 \mid x_2$





Distribution:

Distribution reflects the graph (consistency), and this implies:

$$P(x_1, \dots, x_n) = \prod_{i=1}^{n} P_i(x_i | x_{pa_i})$$

Whatever the graph states the distribution must hold (see figure below).

Graph		Distr
D-sep $\Leftrightarrow$ indep	$\Rightarrow$	P

The reverse might not be true.

Alin Tomescu, http://people.csail.mit.edu/~alinush

Learning Bayesian networks from data

Assumptions: complete data (means we have a value assignment for each observation), discrete variables

$$x_1, \dots, x_n, \text{where } x_i \in \{1, \dots, r_i\}$$

Complete data:

	x_1	x_2	x_3	x_4
Obs. 1	10	8	3	15
Obs. 2	7	13	9	10

Assume these are i.i.d. samples from some $p^*(x_1, \dots, x_n)$

$$D = \{(x_1^t, \dots, x_n^t), t = 1, \dots, T\}$$

When we learn we have 3 problems to solve

1. Parameter estimation, for a given graph G
   a. ML, MAP, Bayesian
2. Model selection problem (score)
   a. Bayesian information criterion, Bayesian score
3. Graph search problem (must find highest scoring graph)

Parameter estimation

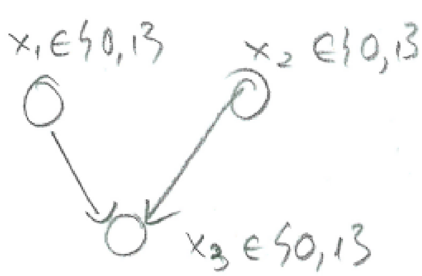
Given a graph G, we know:

$$P(x_1, \dots, x_n) = \prod_{i=1}^{n} P_i(x_i | x_{pa_i})$$

Now we parameterize this distribution:

$$P(x_1, \dots, x_n; \theta) = \prod_{i=1}^{n} P_i(x_i | x_{pa_i}; \theta)$$

We will assume that the model is **fully parameterized,** which means I am fully exploring the freedom to choose this distribution and the P_i conditional probabilities. (see figure below).



$P_3(x_3|x_1,x_2;\theta_3)$

x_1, x_2	$x_3 = 0$	$x_3 = 1$
0 0	.9	.1
1 0	.5	.5
0 1	.7	.3
1 1	.1	.9

θ_3

$$l(\theta;D) = \sum_{t=1}^{T} \log P(x_1^t, \dots, x_n^t; \theta) = \sum_{t=1}^{T}\sum_{i=1}^{n} \log P_i\left(x_i^t | x_{pa_i}^t; \theta_i\right) = \sum_{i=1}^{n}\left[\sum_{t=1}^{T} \log P_i\left(x_i^t | x_{pa_i}^t; \theta_i\right)\right]$$
$$= \sum_{i=1}^{n} \sum_{x_i, x_{pa_i}} n_i\left(x_i, x_{pa_i}\right) \log P_i\left(x_i^t | x_{pa_i}^t; \theta_i\right)$$

$$n_i\left(x_i, x_{pa_i}\right) = \sum_{t=1}^{T} \mathbf{1}\left(x_i^t = x_i\right) \prod_{j \in pa_i} \mathbf{1}\left(x_j^t = x_j\right)$$

How do we solve for the parameters? What is the ML parameter estimate? Since the model is fully parameterized I know each conditional probability can be chosen independently (because the parameters are not tied across different conditional probability tables).

For the i^{th} variable:

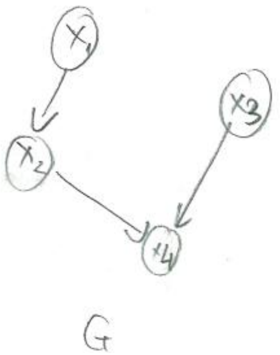
$$\sum_{x_{pa_i}}\left[\sum_{x_i} n_i\left(x_i, x_{pa_i}\right) \log P\left(x_i | x_{pa_i}; \theta_i\right)\right]$$

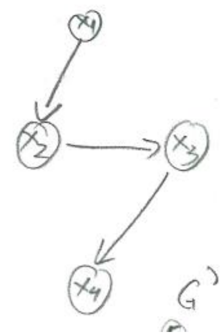
This equation corresponds to a particular row in the conditional probability table that we drew for $P_3(x_3|x_1,x_2;\theta_3)$.

Fix x_{pa_i}, then

$$P\left(x_i | x_{pa_i}; \theta_i\right) = \frac{n_i\left(x_i, x_{pa_i}\right)}{\sum_{x_i'} n_i\left(x_i', x_{pa_i}\right)}$$

Model selection





$$l(\hat{\theta}, D, G) = \sum_{i=1}^{n} \sum_{x_i, x_{pa_i}} n_i(x_i, x_{pa_i}) \log P(x_i | x_{pa_i}; \hat{\theta}_i) = \sum_{i=1}^{n} l(i|pa_i, D)$$

$$BIC(G) = l(\hat{\theta}, D, G) - \frac{\#\, param}{2} \log T$$

$$G \Leftrightarrow pa_1, \dots, pa_n, x_i \in \{1, \dots, r_i\}$$

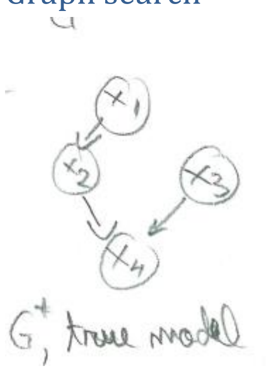
$$score(i|pa_i, D) = l(i|pa_i, D) - \frac{\#\, param}{2} \log T = l(i|pa_i, D) - \frac{\left(\prod_{j \in pa_i} r_j\right)(r_i - 1)}{2} \log T$$

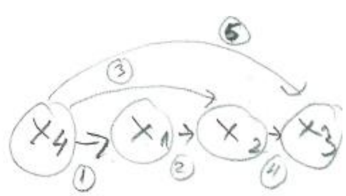
Heavily penalizes models with large number of parents.

Now we get a **decomposable score:**

$$BIC(G) = \sum_{i=1}^{n} score(i|pa_i, D)$$

Graph search





Step 1: Evaluate $score(i|pa_i)$ for each $i = 1, \dots, n$, for each $pa_i \subseteq \{1, \dots, n\} - \{i\}$

Step 2: Find the highest scoring acyclic graph that maximizes $\sum_{i=1}^{n} score(i|pa_i, D)$

$$O(n2^{n-1})$$