

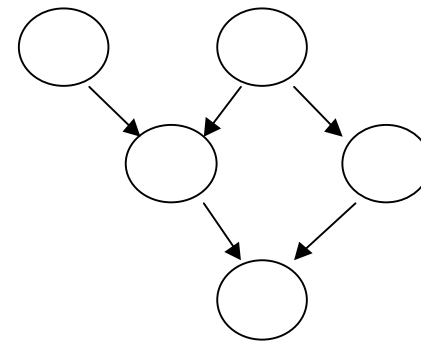
Embedded Bayesian Network Classifiers  
David Heckerman and Christopher Meek  
Microsoft Research

EE 390L, Data Mining  
Michael Anderson  
mande@mail.utexas.edu

# Modeling with Bayesian Networks

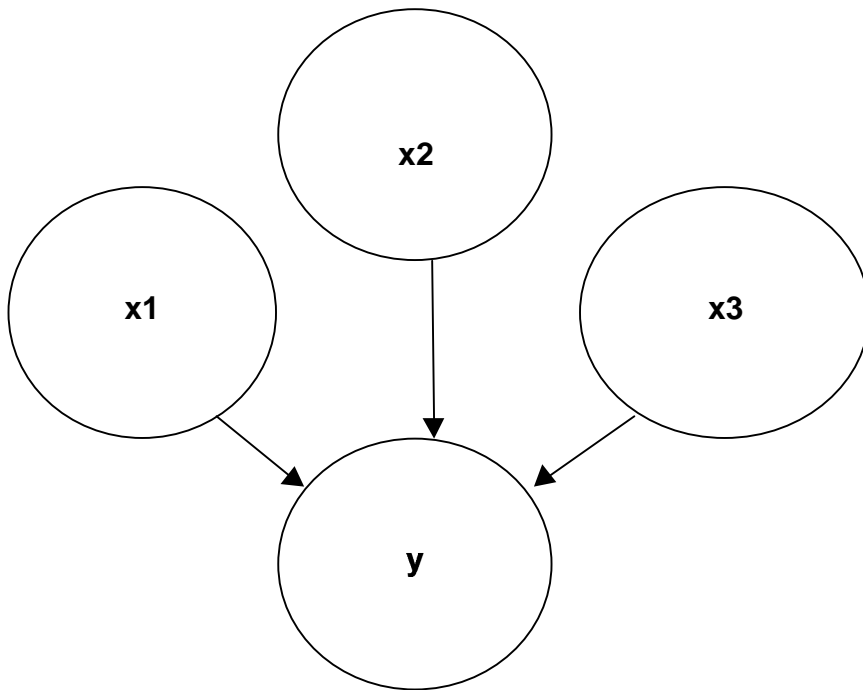
- goal is to map a high-dimension data space to a low-dimension function and parameter space

$$(\mathbf{y}, \mathbf{x}) \rightarrow (\mathbf{m}(\cdot), \theta_m)$$



- Bayesian networks have advantages
  - explanatory power (cause and effect)
  - distribution functions depend only on immediate neighbors

# but....



- The “curse of dimensionality” extends to parameter spaces
- for a network of dichotomous variables, parameters multiply
  - $x_1, x_2, x_3$  each have 1
  - $y$  has 8
- **What if we just....**

## ...reverse the arrows!

- take advantage of symmetry in Bayes Theorem

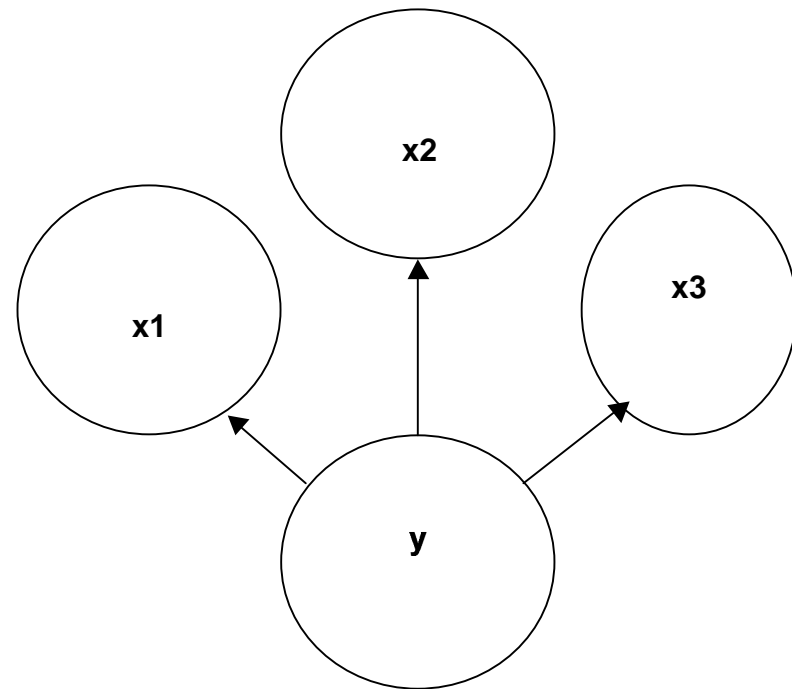
$$P(Y | X) = \frac{P(X | Y)P(Y)}{P(X)}$$

- map into new space

$$(m(\cdot), \theta_m) \leftrightarrow (m'(\cdot), \theta'_m)$$

- an embedded Bayesian network classifier obeys

$$P(Y | \mathbf{X}, \mathbf{m}, \theta_m) = P(Y | \mathbf{X}, \mathbf{m}', \theta'_m)$$



# New Parameters for Old

- H&M outline transformations to a lower-dimensional parameter space
- space is non-linear in parameters
- first step uses  $\lambda$  parameters of *softmax regression*

$$P(Y = k | \mathbf{x}, \mathbf{m}(\cdot), \theta_m) = \frac{e^{\lambda_{kX}}}{1 + \sum_{j=1}^n e^{\lambda_{jX}}}$$

$$\lambda_{kX} = \ln \frac{P(Y = k | \mathbf{x}, \mathbf{m}(\cdot), \theta_m)}{P(Y = 0 | \mathbf{x}, \mathbf{m}(\cdot), \theta_m)} = \ln \frac{\theta(Y = k | \mathbf{pa}_y, \mathbf{m}(\cdot), \theta_m)}{\theta(Y = 0 | \mathbf{pa}_y, \mathbf{m}(\cdot), \theta_m)} + \sum_{y \rightarrow x_i} \ln \frac{\theta(X_i = k | \mathbf{pa}_i^{y=k}, m(\cdot), \theta_m)}{\theta(X_i = 0 | \mathbf{pa}_i^{y=k}, m(\cdot), \theta_m)}$$

- statisticians call this *logistic regression*

# Using the Marginal Likelihood

- EBNCs can be used to more efficiently calculate *marginal likelihood* of a database, given a model

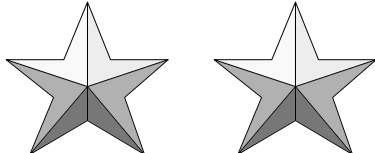
$$\ln P(D | \mathbf{m}(\cdot)) \approx \sum_{i=1}^{|\mathbf{x}|} \left[ L_i(\hat{\phi}_i) - \frac{d_i}{2} \ln N \right] \quad L_i(\theta_i) = \sum_{l=1}^{|D|} \ln P(x_{i,l} | \mathbf{pa}_{i,l}, \phi_i, \mathbf{m}(\cdot))$$

$\theta \rightarrow \lambda \rightarrow \phi$

- Form is similar to Schwartz' *Bayesian Information Criterion*, a measure of model fit
- By selecting model with maximum likelihood, can perform *feature selection*

# Comments

- flipping the arrows is a neat trick -- easy to think of, hard to do
- discussion of mapping to lower-D parameter space was opaque
- use of non-standard jargon (softmax regression, etc.) obscured relationship to existing models and techniques
- BIC and feature selection are not specific to EBNCs

- rating:  (out of 4)