Bayesian Parameter Estimation in Bayesian Networks

Sargur Srihari srihari@cedar.buffalo.edu

Topics

 Bayesian network where parameters are variables

- Global parameter independence
 - Leads to global decomposition
- How to choose priors for Bayesian learning
 - K2 Prior
 - BDe Prior
- Comparison of Bayesian and MLE in ICU example

Fully Bayesian Approach

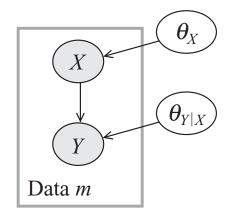
- In the full Bayesian approach to BN learning:
 - Parameters are considered to be random variables
- Need a joint distribution over unknown parameters θ and data instances D
- This joint distribution itself can be represented as a Bayesian network
 - instances and parameters of variables

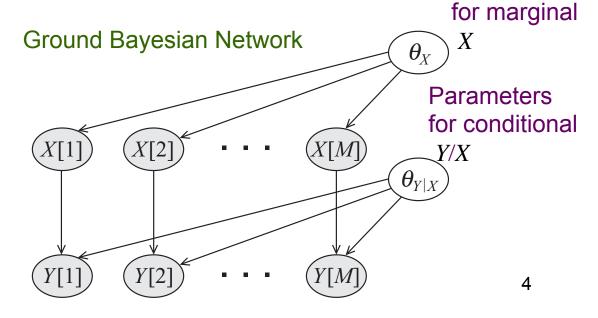
A simple example with 2 variables

- Consider the network $X \rightarrow Y$
- Training data $D=\{X[m],Y[m]\}$ for m=1,...M
- Unknown Parameter vectors θ_X and $\theta_{Y|X}$
- Meta network (BN of BN) for describing learning set-up

 Parameters

Plate Model





Global Parameter Independence

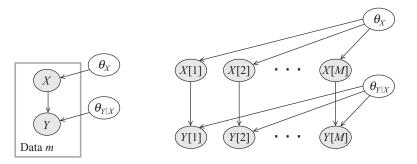
If G is a Bayesian network with parameters

$$\theta = (\theta_{X_1 \mid Pa_{X_1}}, ..., \theta_{X_n \mid Pa_{X_n}})$$

- Global Parameter independence assumption:
 - Priors for individual parameters tells us nothing about another
 - A prior $P(\theta)$ satisfies global independence if

$$P(\theta) = \prod_{i} P(\theta_{X_i \mid PaX_i})$$

Network with global parameter independence:



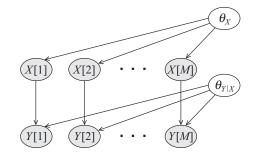
Caution on Parameter Independence

- Although we assume global parameter independence it should be considered with care
- Extension of student example
 - Student takes multiple classes
 - We want to learn distribution of grade given intelligence and difficulty
 - Same instructor for two courses may make parameters dependent

Global Decomposition

- Important conclusion if we assume global parameter independence
- E.g., if x[m] and y[m] are observed for all m, then θ_X and $\theta_{Y|X}$ are d-separated

$$P(\theta_{X}, \theta_{YX} \mid \mathcal{D}) = P(\theta_{X} \mid \mathcal{D}) P(\theta_{YX} \mid \mathcal{D})$$



- Practical ramification
 - Given data set D, we can determine the posterior over θ_X independently of the posterior over $\theta_{Y|X}$

General Networks

- We considered simple network $X \rightarrow Y$
 - Concluded that each problem (of determining posterior) can be handled separately and combined
- Now we generalize conclusion
 - For general networks
- ${\cal G}$ is a Bayesian network with parameters heta
- Need to specify prior $P(\theta)$ over all possible parameterizations of network

Prediction Problem

- Decomposition of posterior allows simplification of prediction
 - In $X \rightarrow Y$ example we want to compute probability of instance x[M+1], y[M+1] given samples x[1], y[1], ..., x[M], y[M]
 - We need to integrate over the parameter $P(x[M+1],y[M+1] \mid D) = \int P(x[M+1],y[M+1] \mid D,\theta) P(\theta \mid D) d\theta$
 - Since posterior probability decomposes into a product we can show that

$$P(x[M+1], y[M+1] | D) = \left(\int P(x[M+1] | \theta_X) P(\theta_X | D) d\theta_X \right) \left(\int P(y[M+1] | \theta_{Y|X}) P(\theta_{Y|X} | D) d\theta_{Y|X} \right)$$

We have that the prediction problems for X
and Y can be solved separately

Priors for BN Learning

- Each node X in a BN has a set of multinomial distributions $\theta_{Xi|paXi}$, one for each instantiation pa_{Xi} of X_i 's parents Pa_{Xi}
- Each of these parameters has a separate
 Dirichlet prior governed by hyperparameters

$$\alpha_{\mathbf{X}_i|pa_{\mathbf{X}_i}} = \left(\alpha_{\mathbf{x}_i^1|pa_{\mathbf{X}_i}},..,\alpha_{\mathbf{x}_i^{K_i}|pa_{\mathbf{X}_i}}\right) \qquad \qquad \text{α are pseudocounts}$$

– Where K_i is the number of values of X_i

$$\theta \sim Dirichlet(\alpha_1,...,\alpha_k) \text{ if } P(\theta) \propto \prod_k \theta_k^{\alpha_k - 1}$$

$$\alpha = \sum_j \alpha_j \quad \text{Note that } \alpha \text{ is the total no of imaginary samples}$$
which does not sum to I

- Prediction: $P(x[M+1] = x^k \mid D) = \frac{M[k] + \alpha_k}{M + \alpha}$
- How to determine α and α_j s?

K2 Prior

- In a software system called K2
- Use a fixed prior

$$lpha_{x_i^j|pa_{X_i}}=1$$
 Add 1 to every cell in CPD

- For all hyper-parameters in the network
- Has consequences that are conceptually unsatisfying

Inconsistency of K2 Prior

- For every multinomial parameter
 - a fixed Dirichlet $D(\alpha,...,\alpha)$
 - Typically $\alpha=1$, called K2 prior
- K2 prior is inconsistent

- y^0 y^I +1 +1
- Ex 1: Y is binary valued y^0, y^1 with no parents



- If we take Dirichlet(1,1) for θ_Y our imaginary sample size for Y is 2
- Ex 2: Introduce parent Xwith four values x^0 - x^3 to get $X \rightarrow Y$ Y is still binary valued y^0, y^1

- If Dirichlet(1,1) is prior for all parameters $\theta_{Y|x^i}$
- Imaginary sample size for Y is δ , two for each context x^i
- No. of imaginary samples we have seen for a event should not depend on structure chosen



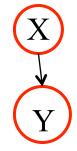
Common Approach

- Hyperparameter α_{x^k} is an imaginary count of prior experience
- Have a data set D' of prior examples
- We set $\alpha_{X_i \mid pa_{X_i}} = \alpha[x_i, pa_{X_i}]$
 - where α [x_i , pa_{Xi}] is the no. of times $X_i = x_i$ and $Pa_{Xi} = pa_{Xi}$ in D'
 - Count the parent-child combinations in imaginary data
- Equivalent to MLE with combined D and D'
- Requires storing a large set of pseudoinstances
 - A sample for every parent-child combination

Machine Learning

Common Approach Example

- Generalizes K2
- Example:
 - Y is binary valued y^0 , y^1
 - Parent X with four values x^0 - x^3 to get $X \rightarrow Y$



- $\alpha [y_i, pa_{Yi}]$
- Samples:

y^0	y^{I}
2	2

	x^0	x^{I}	x^2	x^3
\mathcal{Y}^0	2	2	2	2
y^{I}	2	2	2	2

BDe (Bayesian Dirichlet equivalent) prior

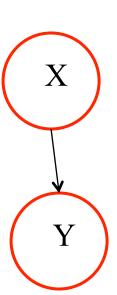
- Store size of imaginary data set α and
 P'[X₁,..., X_n] of frequencies of events
 - probability of every possible ξ
- Set the parameters as follows:

$$\alpha_{\mathcal{X}_{i}\mid pa_{X_{i}}} = \alpha \bullet P'(x_{i}, pa_{X_{i}})$$

- This will divide data set proportionately
- Apply this to $X \rightarrow Y$ problem

$$\alpha_{y} = \alpha \bullet P'(y) = \sum_{x^{i}} \alpha \bullet P'(y, x^{i}) = \sum_{x^{i}} \alpha_{y|x^{i}}$$





How do we represent P' for BDe?

- Use a Bayesian network so we can efficiently infer the quantities $P'(x_i, pa_{Xi})$
- When we have no prior knowledge we can set P' to be uniform
 - Implies an empty BN with uniform distributions for each variable
- Network structure is used only to provide parameters' priors not to guide structure search

BDe satisfies Score Equivalence

- BDe has a property useful in structure learning
 - Bayesian score

$$P(G \mid D) = \frac{P(D \mid G)P(G)}{P(D)}$$

$$score_{B}(G : D) = \log P(D \mid G) + \log P(G)$$

Score of a graph with Dirichlet Prior

$$score_{BIC}(G:D) = \ell(\theta_G:D) - \frac{\log M}{2}Dim[G]$$

Score decomposability

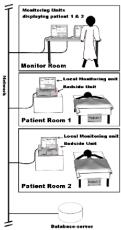
$$score(G) = \sum_{i=1}^{n} score(X_i, Pa_G(X_i)).$$

- Score equivalence
 - Networks with same Bde score are I-equivalent 17

Application to ICU monitoring

- Decision support system (medical diagnostics)
- Monitoring severely ill patients in ICU
- Goal is to calculate probabilities for a differential diagnosis based on available evidence





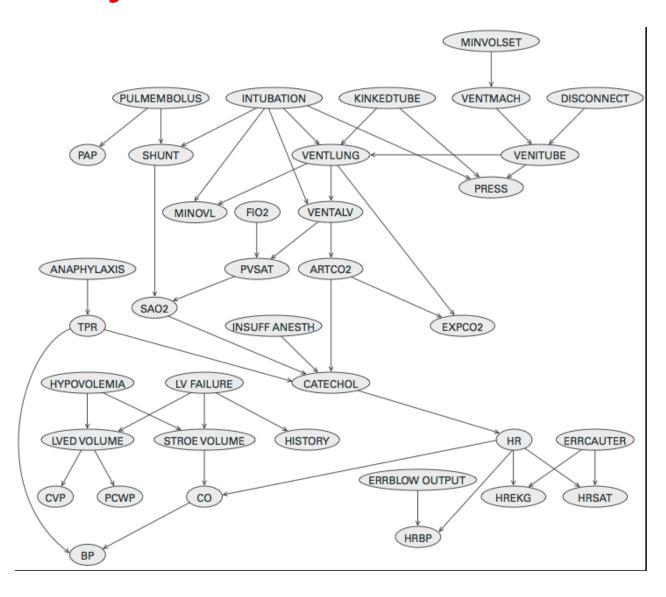
In differential diagnosis there are multiple alternatives which are systematically eliminated

18

Variables in ICU Alarm

- Medical knowledge encoded as 37 variables:
 - 8 diagnoses (at top level, no parents)
 - 16 findings
 - 13 intermediate variables
- Some top-level variables:
 - Anaphylaxis (severe allergic reaction)
 - Hypovolemia (shock of decreased blood volume)
 - Kinked Tube
 - Disconnect
 - Left Ventricle Failure

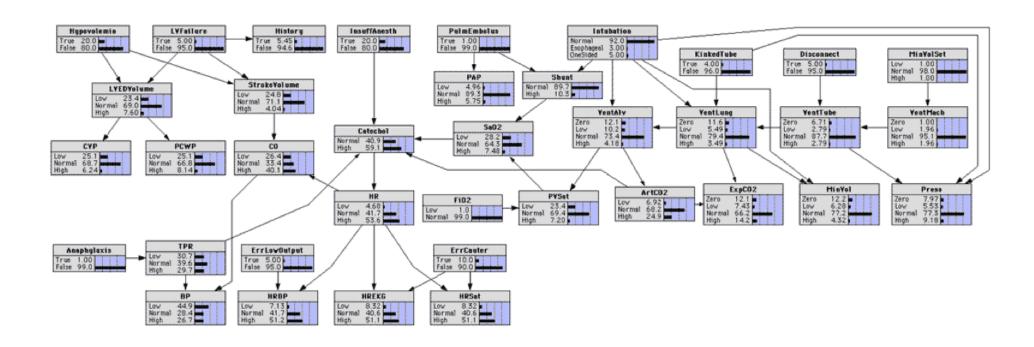
Bayesian Network for ICU-Alarm



Hand-constructed by experts for monitoring patients in an Intensive Care Unit

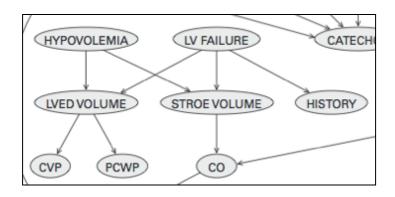
Benchmark for Bayesian Network Learning (Parameter Estimation)

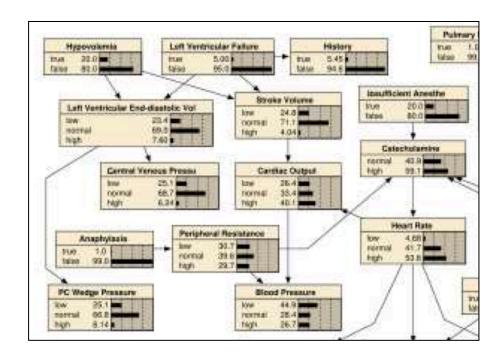
ICU alarm network



ALARM network: A Logical Alarm Reduction Mechanism

Contingency Tables between Dependent Variables





A contingency tables usually shows dependency between a pair of variables These show marginal probabilities of each variable

Parameters for ICU Alarm

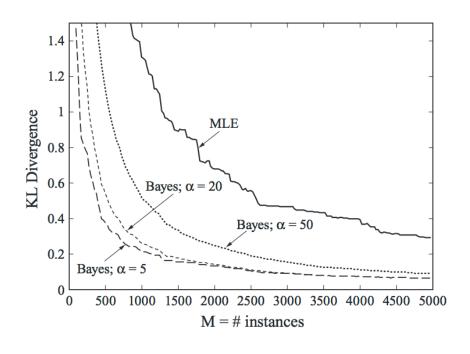
- Network Parameters
 - 37 nodes with table CPDs
 - 504 parameters (instead of 2⁵⁴)
- Evaluate ability of parameter estimation method
 - To reconstruct network parameters from data
- Inputs for parameter estimation
 - Training samples
 - Obtained by sampling from specified network
 - Network structure

Learning curve for ICU alarm

Measuring the quality of the learned network

Relative entropy between generating distribution P^* and learned distribution P Provides a global measure of extent to which learned distribution resembles true distribution

Relative entropy to true model



Lowest error achieved Very weak prior $\alpha=5$ Enough to provide smoothing

Higher values introduce a bias Effect of high bias disappears as samples increase

A Static Bayesian Network

