Readings:
K&F: 17.3, 17.4, 17.5.1, 8.1, 12.1

Structure Learning
(The Good), The Bad, The Ugly

A little inference too…

Graphical Models – 10708
Carlos Guestrin
Carnegie Mellon University
October 8th, 2008

Decomposable score

- Log data likelihood

$$\log \hat{P}(\mathcal{D} \mid \theta, \mathcal{G}) = m \sum_i \hat{I}(X_i, \mathbf{Pa}_{X_i}) - m \sum_i \hat{H}(X_i)$$

- Decomposable score:
  - Decomposes over families in BN (node and its parents)
  - Will lead to significant computational efficiency!!!
  - Score(G : D) = $\sum_i$ FamScore(X_i|$\mathbf{Pa}_{X_i}$: D)

for MLE: $FamScore(X_i \mid Pa_{X_i} : D) = m \hat{I}(X_i; Pa_{X_i}) - m \hat{H}(X_i)$

Chow-Liu tree learning algorithm 1

- For each pair of variables X_i, X_j
  - Compute empirical distribution:

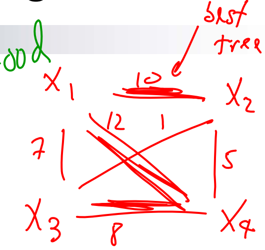
$$\hat{P}(x_i, x_j) \overset{\text{MLE}}{=} \frac{\text{Count}(x_i, x_j)}{m}$$

  - Compute mutual information:

$$\hat{I}(X_i, X_j) = \sum_{x_i, x_j} \hat{P}(x_i, x_j) \log \frac{\hat{P}(x_i, x_j)}{\hat{P}(x_i)\hat{P}(x_j)}$$

- Define a graph
  - Nodes $X_1, \ldots, X_n$
  - Edge (i,j) gets weight $\hat{I}(X_i, X_j)$ (w_{ij})

The Good



max trees ↑ score(tree) $= \sum_{ij} I(X_i, X_j) = \sum_{ij} w_{ij}$

find maximum spanning tree → best tree BN

Maximum likelihood score overfits!

↑ $\log \hat{P}(\mathcal{D} \mid \theta, \mathcal{G}) = m \sum_i \hat{I}(X_i, \mathbf{Pa}_{X_i,\mathcal{G}}) - m \sum_i \hat{H}(X_i)$

- Information never hurts:

↑ $I(X_i, Pa_{X_i}) = H(X_i) - H(X_i \mid Pa_{X_i})$

$H(A \mid B) \leq H(A \mid C) \quad C \subseteq B$

the more parents the higher $I(X_i, Pa_{X_i})$

- Adding a parent always increases score!!!

MLE ⇒ Complete Graph

Bayesian score

- Prior distributions:
  - Over structures
  - Over parameters of a structure
- Posterior over structures given data:

$$\log P(\mathcal{G} \mid D) \propto \log P(\mathcal{G}) + \log \int_{\theta_{\mathcal{G}}} P(D \mid \mathcal{G}, \theta_{\mathcal{G}}) P(\theta_{\mathcal{G}} \mid \mathcal{G}) d\theta_{\mathcal{G}}$$

Bayesian learning for multinomial

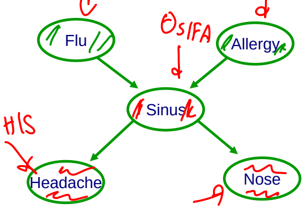
- What if you have a k sided coin???
- Likelihood function if **multinomial**:
- **Conjugate** prior for multinomial is **Dirichlet**:
  - $\theta \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k) \sim \prod_i \theta_i^{\alpha_i - 1}$
- **Observe** *m* data points, m_i from assignment i, **posterior**:
- **Prediction**:

Global parameter independence, d-separation and local prediction

- Independencies in **meta BN**:

- **Proposition**: For fully observable data *D*, if prior satisfies global parameter independence, then

$$P(\theta \mid \mathcal{D}) = \prod_i P(\theta_{X_i \mid \mathbf{Pa}_{X_i}} \mid \mathcal{D})$$

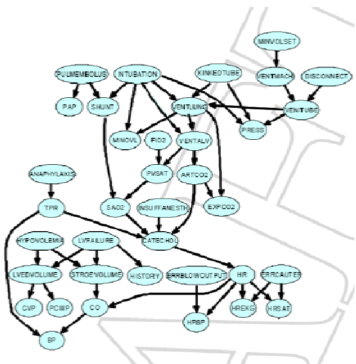


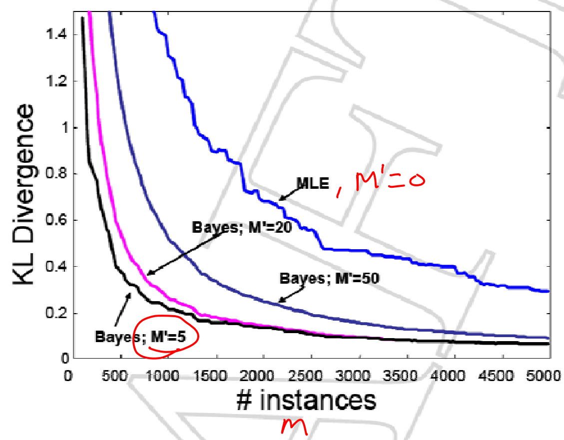
Priors for BN CPTs

(more when we talk about structure learning)

- Consider each CPT: P(X|**U**=**u**)
- Conjugate prior:
  - Dirichlet($\alpha_{X=1|\mathbf{U}=\mathbf{u}}, \ldots, \alpha_{X=k|\mathbf{U}=\mathbf{u}}$)
- More intuitive:
  - "prior data set" *D*' with m' equivalent sample size
  - "prior counts":
  - prediction:

An example





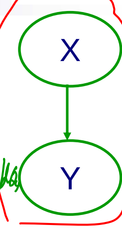
What you need to know about parameter learning

- Bayesian parameter learning:
  - motivation for Bayesian approach
  - Bayesian prediction
  - conjugate priors, equivalent sample size
  - Bayesian learning ⇒ smoothing
- Bayesian learning for BN parameters
  - Global parameter independence
  - Decomposition of prediction according to CPTs
  - Decomposition within a CPT

Bayesian score and model complexity

$$\log P(D \mid \mathcal{G}) = \log \int_{\theta_\mathcal{G}} P(D \mid \mathcal{G}, \theta_\mathcal{G}) P(\theta_\mathcal{G} | \mathcal{G}) d\theta_\mathcal{G}$$

True model:



P(Y=t|X=t) = 0.5 + α

P(Y=t|X=f) = 0.5 - α

- Structure 1: X and Y independent
  - Score doesn't depend on alpha
- Structure 2: X → Y
  - Data points split between P(Y=t|X=t) and P(Y=t|X=f)
  - For fixed M, only worth it for large α
    - Because posterior over parameter will be more diffuse with less data

Bayesian, a decomposable score

$$\log P(D \mid \mathcal{G}) = \log \int_{\theta_\mathcal{G}} P(D \mid \mathcal{G}, \theta_\mathcal{G}) P(\theta_\mathcal{G} | \mathcal{G}) d\theta_\mathcal{G}$$

- As with last lecture, assume:
  - Parameter independence
- Also, prior satisfies **parameter modularity**:
  - If X_i has same parents in *G* and *G'*, then parameters have same prior
- Finally, structure prior P(*G*) satisfies **structure modularity**
  - Product of terms over families
  - E.g., P(*G*) ∝ $c^{|G|}$
- Bayesian score decomposes along families!

BIC approximation of Bayesian score

- Bayesian has difficult integrals
- For Dirichlet prior, can use simple Bayes information criterion (BIC) approximation
  - In the limit, we can forget prior!
  - **Theorem**: for Dirichlet prior, and a BN with Dim(G) independent parameters, as m→∞:

$$\log P(D \mid \mathcal{G}) = \log P(D \mid \mathcal{G}, \theta_{\mathcal{G}}) - \frac{\log m}{2}\text{Dim}(\mathcal{G}) + O(1)$$

BIC approximation, a decomposable score

- BIC: $\text{Score}_{\text{BIC}}(\mathcal{G} : D) = \log P(D \mid \mathcal{G}, \theta_{\mathcal{G}}) - \frac{\log m}{2}\text{Dim}(\mathcal{G})$
- Using information theoretic formulation:

$$\text{Score}_{\text{BIC}}(\mathcal{G} : D) = m\sum_i \hat{I}(X_i, \mathbf{Pa}_{X_i,\mathcal{G}}) - m\sum_i \hat{H}(X_i) - \frac{\log m}{2}\sum_i \text{Dim}(P(X_i \mid \mathbf{Pa}_{X_i,\mathcal{G}}))$$

Consistency of BIC and Bayesian scores

Consistency is limiting behavior, says nothing about finite sample size!!!

- A scoring function is **consistent** if, for true model G^*, as $m \to \infty$, with probability 1
  - G^* maximizes the score
  - All structures **not I-equivalent** to G^* have strictly lower score
- **Theorem**: BIC score is consistent
- **Corollary**: the Bayesian score is consistent
- What about maximum likelihood score? NO

MLE: not Consistent

$Score_{MLE}(\text{Complete Graph}) = Score_{MLE}(G^*)$

BIC:

$Penalty(\text{Complete Graph}) > Penalty(G^*)$

Priors for general graphs

- For finite datasets, prior is important!
- Prior over structure satisfying prior modularity
- What about prior over parameters, how do we represent it?
  - *K2 prior*: fix an α, $P(\theta_{X_i|\mathbf{Pa}_{X_i}}) = \text{Dirichlet}(\alpha,\ldots,\alpha)$ $\forall u$
  - K2 is "inconsistent"

all vars have k values

$\lvert Pa_{X_i}\rvert$	"equivalent sample size"
0	$k\alpha$
1	for each assign. parents $P(\theta_{X_i \mid Pa_{X_i}=u}) \sim Dir(\alpha,\alpha,\ldots)$ $k^2\alpha$
2	$k^3\alpha$

BDe prior

- Remember that Dirichlet parameters analogous to "fictitious samples"
- Pick a fictitious sample size m'
- For each possible family, define a prior distribution $P(X_i,\mathbf{Pa}_{X_i})$
  - Represent with a BN
  - Usually independent (product of marginals) $P'(X_i, Pa_{X_i}) = P'(X_i)\prod_{j \in Pa_{X_i}} P'(X_j)$
- **BDe prior**: most common uniform

$$P(\theta_{X_i} \mid Pa_{X_i}=u) \sim \text{Dirichlet}(m' P'(X_i=1, Pa_{X_i}=u), \ldots, m' P'(X_i=K, Pa_{X_i}=u))$$

- Has "consistency property":

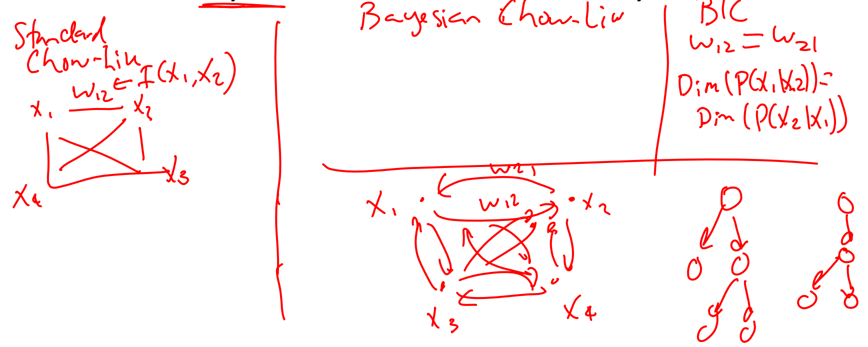
$Score_{BDe}(G:D)$ is consistent

Score equivalence

- If *G* and *G'* are I-equivalent then they have same score
- **Theorem 1**: Maximum likelihood score and BIC score satisfy score equivalence
- **Theorem 2**:
  - If P(*G*) assigns same prior to I-equivalent structures (e.g., edge counting)
  - and parameter prior is dirichlet
  - then **Bayesian score satisfies score equivalence** *if and only if* prior over parameters represented as a **BDe prior**!!!!!!

Chow-Liu for Bayesian score

- Edge weight $w_{X_j \to X_i}$ is advantage of adding X_j as parent for X_i



- Now have a directed graph, need directed spanning forest
  - Note that adding an edge can hurt Bayesian score – choose forest not tree
  - Maximum spanning forest algorithm works