Readings:
K&F: 16.3, 16.4, 17.3

Bayesian Param. Learning

Bayesian Structure Learning

Graphical Models – 10708
Carlos Guestrin
Carnegie Mellon University
October 6th, 2008

Decomposable score

- Log data likelihood

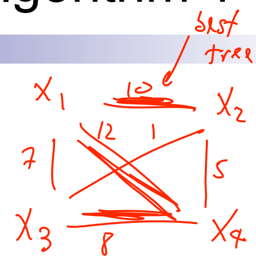
$$\log \hat{P}(\mathcal{D} \mid \theta, \mathcal{G}) = m \sum_i \hat{I}(X_i, \mathbf{Pa}_{X_i}) - m \sum_i \hat{H}(X_i)$$

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

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

Chow-Liu tree learning algorithm 1

- For each pair of variables X_i, X_j
  - Compute empirical distribution:
    $$\widehat{P}(x_i, x_j) \overset{\text{MLE}}{=} \frac{\text{Count}(x_i, x_j)}{m}$$
  - Compute mutual information:
    $$\widehat{I}(X_i, X_j) = \sum_{x_i, x_j} \widehat{P}(x_i, x_j) \log \frac{\widehat{P}(x_i, x_j)}{\widehat{P}(x_i)\widehat{P}(x_j)}$$
- Define a graph
  - Nodes $X_1, \ldots, X_n$
  - Edge (i,j) gets weight $\widehat{I}(X_i, X_j)$ w_{ij}



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

find maximum spanning tree

best tree BN

Chow-Liu tree learning algorithm 2

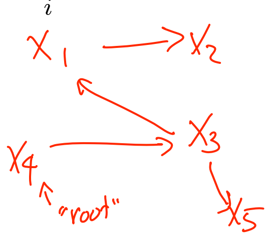
$$\log \widehat{P}(\mathcal{D} \mid \theta, \mathcal{G}) = M \sum_i \widehat{I}(x_i, \mathbf{Pa}_{x_i,\mathcal{G}}) - M \sum_i \widehat{H}(X_i)$$

- Optimal tree BN
  - Compute maximum weight spanning tree
  - Directions in BN: pick any node as root, breadth-first -search defines directions

all trees (BNs) with 1 parent & same skeleton have no immoralities

$\Rightarrow$ I equivalent

$\Rightarrow$ pick any



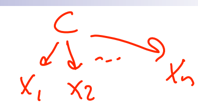
using Chow-Liu

OPTIMAL tree BN

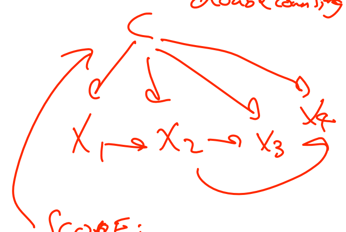
Can we extend Chow-Liu 1

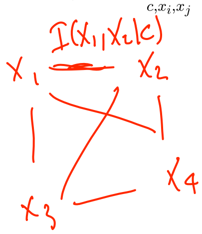
- Tree augmented naïve Bayes (TAN) [Friedman et al. '97]
  - Naïve Bayes model overcounts, because correlation between features not considered
  - Same as Chow-Liu, but score edges with:

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



TAN can reduce double counting





maximum spanning tree, ⇒ optimal TAN

SCORE:
$I(C,\emptyset) + I(C,X_1) + I(X_2, C X_1) + I(X_3, C X_2) + I(X_4, C X_2)$ + entropy

Can we extend Chow-Liu 2

- (Approximately learning) models with tree-width up to *k*
  - [Chechetka & Guestrin '07]
  - But, $O(n^{2k+6})$

What you need to know about learning BN structures so far

- Decomposable scores
  - Maximum likelihood
  - Information theoretic interpretation
- Best tree (Chow-Liu)
- Best TAN
- Nearly best k-treewidth (in $O(N^{2k+6})$)

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:
- Adding a parent always increases score!!!

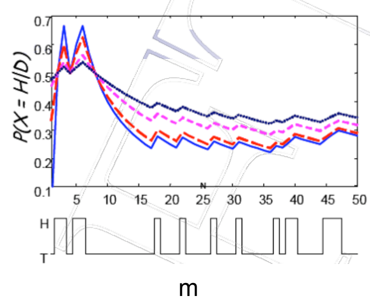
Bayesian score

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

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

Can we really trust MLE?

- What is better?
  - 3 heads, 2 tails
  - 30 heads, 20 tails
  - $3\text{x}10^{23}$ heads, $2\text{x}10^{23}$ tails



- Many possible answers, we need distributions over possible parameters

Bayesian Learning

- Use Bayes rule:

$$P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})}$$

- Or equivalently:

$$P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta)$$

Bayesian Learning for Thumbtack

$$P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta)$$

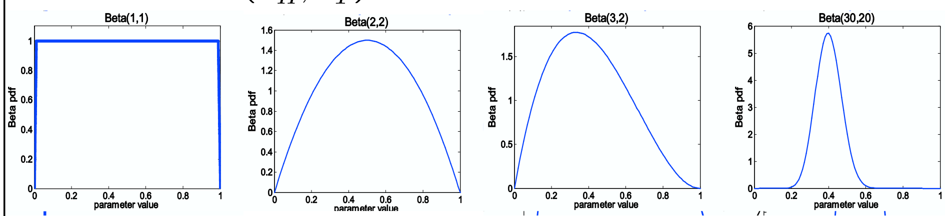
- Likelihood function is simply Binomial:

$$P(\mathcal{D} \mid \theta) = \theta^{m_H}(1-\theta)^{m_T}$$

- What about prior?
  - Represent expert knowledge
  - Simple posterior form
- Conjugate priors:
  - Closed-form representation of posterior (more details soon)
  - **For Binomial, conjugate prior is Beta distribution**

Beta prior distribution – P(θ)

$$P(\theta) = \frac{\theta^{\alpha_H - 1}(1-\theta)^{\alpha_T - 1}}{B(\alpha_H, \alpha_T)} \sim Beta(\alpha_H, \alpha_T)$$

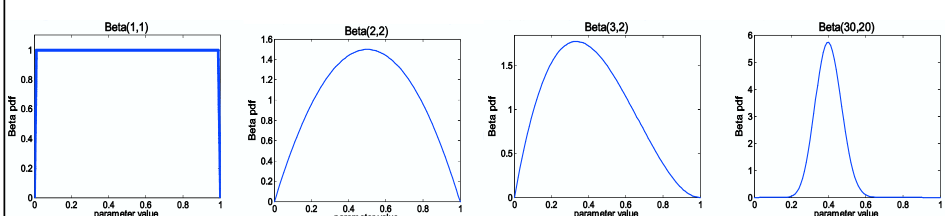


- Likelihood function: $P(\mathcal{D} \mid \theta) = \theta^{m_H}(1-\theta)^{m_T}$
- Posterior: $P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta) P(\theta)$

Posterior distribution

- Prior: $Beta(\alpha_H, \alpha_T)$
- Data: m_H heads and m_T tails
- Posterior distribution:

$$P(\theta \mid \mathcal{D}) \sim Beta(m_H + \alpha_H, m_T + \alpha_T)$$



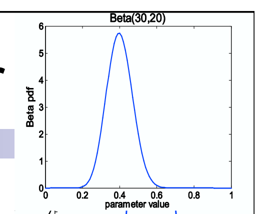
Conjugate prior

- Prior: $Beta(\alpha_H, \alpha_T)$
- Data: m_H heads and m_T tails (binomial likelihood)
- Posterior distribution:

$$P(\theta \mid \mathcal{D}) \sim Beta(m_H + \alpha_H, m_T + \alpha_T)$$

- Given likelihood function P(D|θ)
- (Parametric) prior of the form P(θ|α) is **conjugate** to likelihood function if posterior is of the same parametric family, and can be written as:
  - P(θ|α'), for some new set of parameters α'

Using Bayesian posterior



- Posterior distribution:

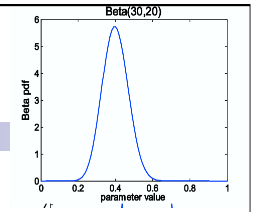
$$P(\theta \mid \mathcal{D}) \sim Beta(m_H + \alpha_H, m_T + \alpha_T)$$

- Bayesian inference:
  - No longer single parameter:

$$E[f(\theta)] = \int_0^1 f(\theta) P(\theta \mid \mathcal{D}) d\theta$$

  - Integral is often hard to compute

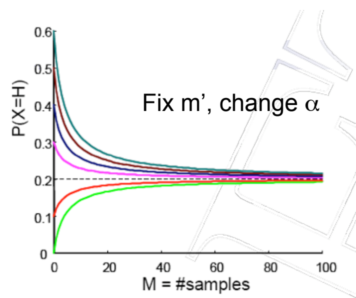
Bayesian prediction of a new coin flip

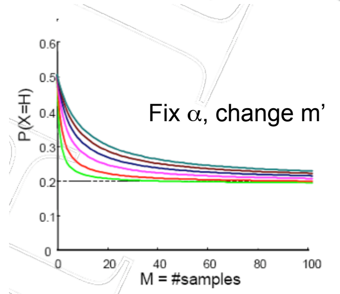


- Prior:
- Observed m_H heads, m_T tails, what is probability of m+1 flip is heads?

Asymptotic behavior and equivalent sample size

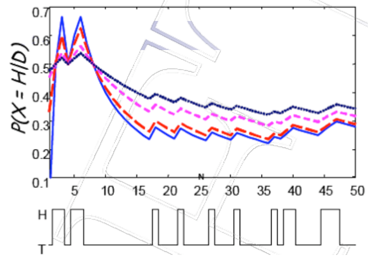
- Beta prior equivalent to extra thumbtack flips:
  - $E[\theta] = \frac{m_H + \alpha_H}{m_H + \alpha_H + m_T + \alpha_T}$
- As $m \to \infty$, prior is "forgotten"
- **But, for small sample size, prior is important!**
- **Equivalent sample size**:
  - Prior parameterized by α_H,α_T, or
  - m' (equivalent sample size) and α
  - $E[\theta] = \frac{m_H + \alpha m'}{m_H + m_T + m'}$





Bayesian learning corresponds to smoothing

$$E[\theta] = \frac{m_H + \alpha m'}{m_H + m_T + m'}$$



- m=0 $\Rightarrow$ prior parameter
- m$\rightarrow\infty$ $\Rightarrow$ MLE

Bayesian learning for multinomial

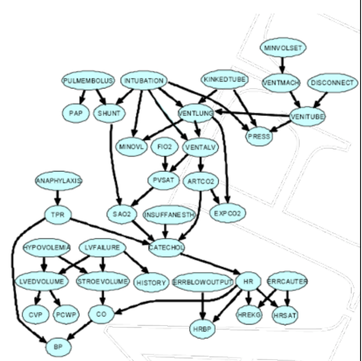
- What if you have a k sided coin???
- Likelihood function if **multinomial**:
  -
  -
- **Conjugate** prior for multinomial is **Dirichlet**:
  - $\theta \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k) \sim \prod_i \theta_i^{\alpha_i - 1}$
- **Observe** *m* data points, m_i from assignment i, **posterior**: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

Bayesian learning for two-node BN

- Parameters θ_X, $\theta_{Y|X}$
- Priors:
  - $P(\theta_X)$:
  - $P(\theta_{Y|X})$:

Very important assumption on prior: Global parameter independence

- **Global parameter independence:**
  - Prior over parameters is product of prior over CPTs

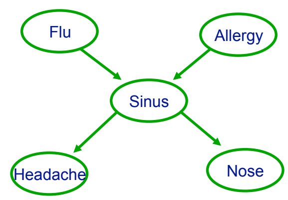


Global parameter independence, d-separation and local prediction

- Independencies in **meta BN**:

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

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



Within a CPT

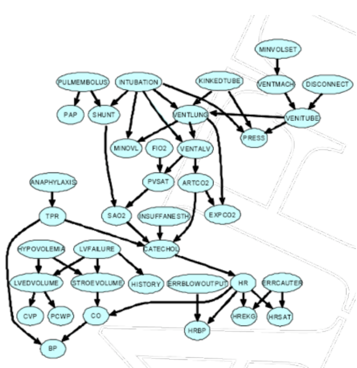
- Meta BN including CPT parameters:

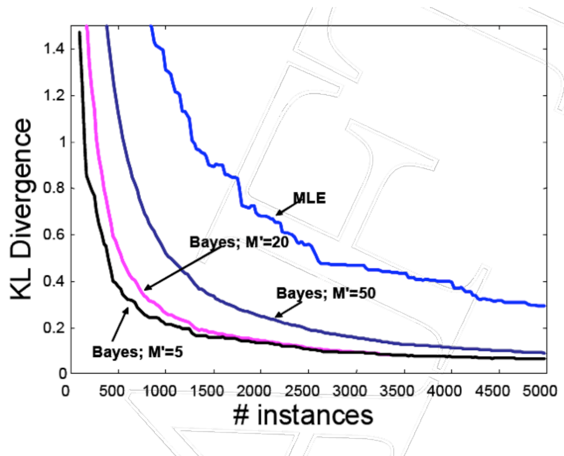
- Are $\theta_{Y|X=t}$ and $\theta_{Y|X=f}$ d-separated given *D*?
- Are $\theta_{Y|X=t}$ and $\theta_{Y|X=f}$ independent given *D*?
  - Context-specific independence!!!
- Posterior decomposes:

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

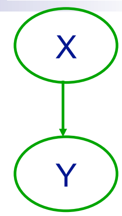
Announcements

- **Project description is out on class website**:
  - Individual or groups of two only
  - Suggested projects on the class website, or do something related to your research (preferable)
    - Must be something you started this semester
    - The semester goes really quickly, so be realistic (and ambitious ☺)
  - Must be related to Graphical Models! ☺
- **Project deliverables**:
  - one page proposal due Wednesday (10/8)
  - 5-page milestone report Nov 3rd in class
  - Poster presentation on Dec. 1st, 3-6pm in NSH Atrium
  - Write up, 8-pages, due Dec 3rd by 3pm by email to instructors (no late days)
  - All write ups in NIPS format (see class website), page limits are strict
- **Objective**:
  - Explore and apply concepts in **probabilistic graphical models**
  - Doing a fun project!

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 $\rightarrow$ 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:
  - Local and global 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*) $\propto$ $c^{|G|}$
- 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→∞:

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

BIC approximation, a decomposable score

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

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

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?

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, $P(\theta_{X_i|\mathbf{Pa}_{X_i}})$ = Dirichlet($\alpha,\ldots,\alpha$)
  - K2 is "inconsistent"

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)
- **BDe prior**:
- Has "consistency property":