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) = $\sum_i$ 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

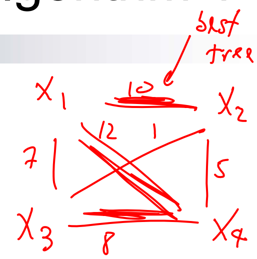
- For each pair of variables X_i, X_j
  - Compute empirical distribution:

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

  - 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

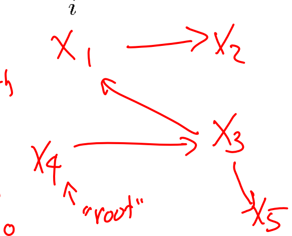
best tree BN

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
⇒ I equivalent
⇒ pick any

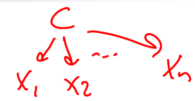


using Chow-Liu OPTIMAL tree BN

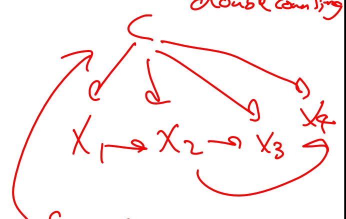
Can we extend Chow-Liu 1

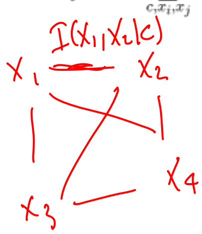
- Tree augmented naïve Bayes (TAN)
[Friedman et al. '97]
  - Naïve Bayes model overcounts, because correlation between features not considered
  - Same as Chow-Liu, but score edges with:

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



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!

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

- Information never hurts:

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

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

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

- Adding a parent always increases score!!!

MLE $\Rightarrow$ Complete Graph

Bayesian score

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

notes LD

$P(D \mid G, \theta_G)$

prior over graphs, eg $P(G) \propto e^{-c\,|\text{number of edges}|}$

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

prior over CPT parameters

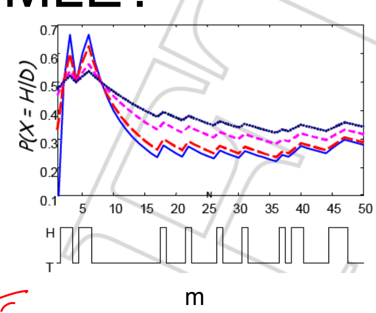
$$= \frac{\int_{\theta_G} P(D \mid G, \theta_G)\, P(\theta_G \mid G)\, P(G)\, d\theta_G}{P(D)}$$

$$\underbrace{\log P(\mathcal{G} \mid D)}_{\text{posteriors}} = \overbrace{\log P(\mathcal{G})}^{\text{prior over graphs}} + \log \int_{\theta_{\mathcal{G}}} P(D \mid \mathcal{G}, \theta_{\mathcal{G}}) P(\theta_{\mathcal{G}} \mid \mathcal{G})\, d\theta_{\mathcal{G}}$$

+constant ← log P(D)

Can we really trust MLE?

- What is better?
  - 3 heads, 2 tails $\quad \theta_{MLE} = \frac{3}{5}$
  - 30 heads, 20 tails $\quad \theta_{MLE} = \frac{3}{5}$
  - $3\text{x}10^{23}$ heads, $2\text{x}10^{23}$ tails $\quad \theta_{MLE} = \frac{3}{5}$



- Many possible answers, we need distributions over possible parameters

Bayesian Learning

- Use Bayes rule: likelihood prior

posterior

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

- Or equivalently:

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

Bayesian Learning for Thumbtack

posterior likelihood prior

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

- Likelihood function is simply Binomial:

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

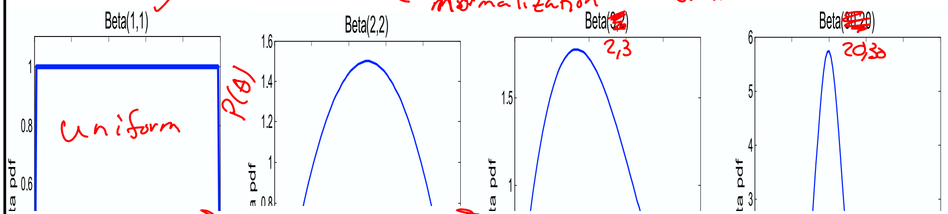
- What about prior?
  - Represent expert knowledge
  - Simple posterior form
- Conjugate priors:
  - Closed-form representation of posterior (more details soon)
  - **For Binomial, conjugate prior is Beta distribution**

Beta prior distribution – P(θ)

prior

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

normalization constant (drop)



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

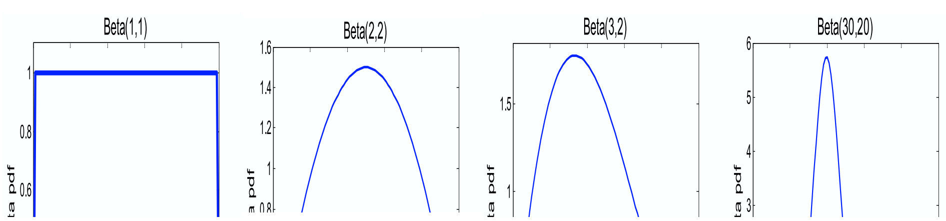
$$= \theta^{m_H}(1-\theta)^{m_T}\,\theta^{\alpha_H-1}(1-\theta)^{\alpha_T-1} = \theta^{m_H+\alpha_H-1}(1-\theta)^{m_T+\alpha_T-1}$$

$$\sim Beta(m_H+\alpha_H,\ m_T+\alpha_T)$$

Posterior distribution

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

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



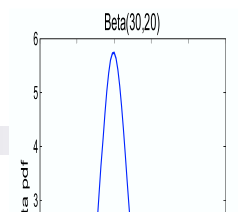
Conjugate prior

- 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:

posterior

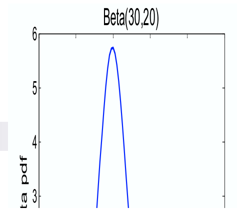
$$E[f(\theta)] = \int_0^1 f(\theta) P(\theta \mid \mathcal{D}) d\theta$$

utility

  - Integral is often hard to compute

often → mean parameter
mode parameter

Bayesian prediction of a new coin flip



- Prior: Beta(α_H, α_T)
- Observed m_H heads, m_T tails, what is probability of m+1 flip is heads?

Posterior Beta$(\alpha_H + m_H, \alpha_T + m_T)$

$$P(m{+}1\ \text{flip} = \text{heads} \mid m_H, m_T)$$

$$= \int_\theta \underbrace{P(m{+}1\ \text{flip}{=}H \mid \theta)}_{\theta} P(\theta \mid m_H, m_T)\, d\theta$$

$$= \int_\theta \theta \underbrace{P(\theta \mid m_H, m_T)}_{\text{Beta}(\alpha_H + m_H, \alpha_T + m_T)}\, d\theta \equiv \text{mean} = \frac{\alpha_H + m_H}{\alpha_H + m_H + \alpha_T + m_T}$$

Asymptotic behavior and equivalent sample size

Prior Beta$(\alpha m', (1-\alpha) m')$

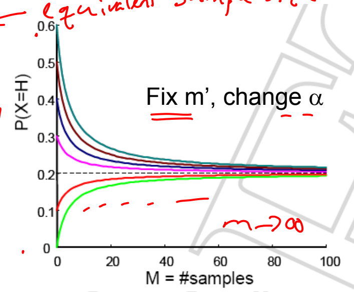
$m' \leftarrow$ equivalent sample size

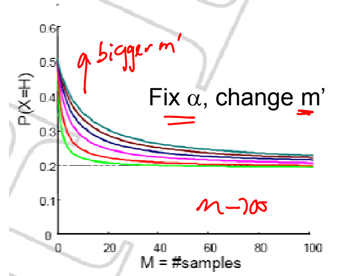
- Beta prior equivalent to extra thumbtack flips:
  - $$E[\theta] = \frac{m_H + \alpha_H}{m_H + \alpha_H + m_T + \alpha_T}$$

$\alpha_H = \alpha m'$

$\alpha_T = (1-\alpha) m'$

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



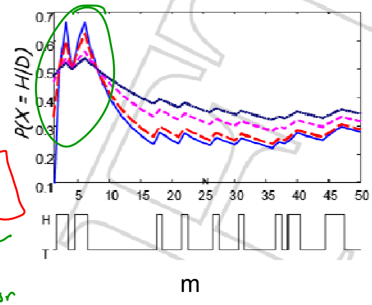


Bayesian learning corresponds to smoothing

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

$$= \frac{m}{m+m'} \underbrace{\left[\frac{m_H}{m}\right]}_{\text{MLE estimate}} + \frac{m'}{m+m'} \left[\underbrace{\frac{\alpha m'}{m'}}_{\alpha}\right]$$

prior "mean"



- m=0 ⇒ prior parameter
- m→∞ ⇒ MLE

$\text{Beta}(\alpha_H, \alpha_T) \leftarrow$ mode $\dfrac{\alpha_H - 1}{\alpha_H + \alpha_T - 2}$

Bayesian learning for multinomial

$m_i \leftarrow$ # observations of class, value or side i

- What if you have a k sided coin???
- Likelihood function if **multinomial**:
  - $P(D \mid \theta_1, \dots \theta_k) = \theta_1^{m_1} \theta_2^{m_2} \dots \theta_k^{m_k}$
  - $\sum \theta_i = 1, \quad \theta_i \geq 0$
- **Conjugate** prior for multinomial is **Dirichlet**:
  - $\theta \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_k) \sim \prod_i \theta_i^{\alpha_i - 1}$

    $\alpha_i \geq 0$
- **Observe** *m* data points, m_i from assignment i, **posterior**:

$$P(\theta_1 \dots \theta_k \mid m_1 \dots m_k) \propto P(m_1 \dots m_k \mid \theta_1 \dots \theta_k)\ P(\theta)$$

$$= \text{Dirichlet}(\alpha_1 + m_1, \alpha_2 + m_2, \dots, \alpha_k + m_k)$$

- **Prediction**:

$$E[\theta_i] = \frac{m_i + \alpha_i}{\sum_j (m_j + \alpha_j)}$$

Bayesian learning for two-node BN

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

$X \to Y$

$P(X)$ $\quad P(Y|X)$

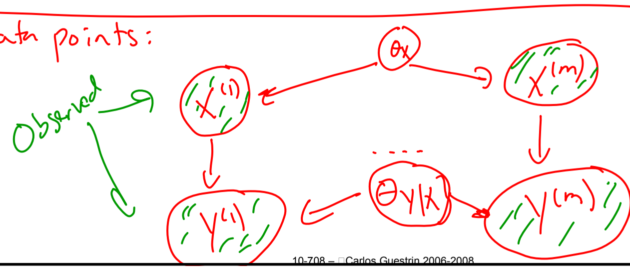
$\text{Dirichlet}(\alpha_{X=1}, \alpha_{X=2}, \ldots, \alpha_{X=x})$

for each value $X=x$

a set of parameters $\theta_{Y|X=x}$

$P(\theta_{Y|X=x}) = \text{Dirichlet}(\alpha_{Y=1|X=x}, \ldots, \alpha_{Y=k|X=x})$

m data points:



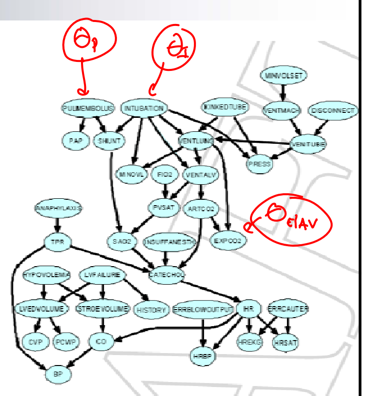
$\theta_X \perp \theta_{Y|X} \mid D$

x y

Very important assumption on prior: Global parameter independence

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

$$P(\theta) = \prod_i P(\theta_{X_i} \mid Pa_{X_i})$$



Global parameter independence, d-separation and local prediction

- Independencies in **meta BN**:

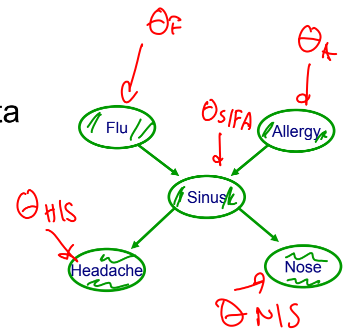
add prior vars to the BN

$P(\theta) = P(\theta_F)\, P(\theta_A)\, P(\theta_{S|FA})\, P(\theta_{N|S})\, P(\theta_{H|S})$

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

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

params indep. given data



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

$$E[\theta_{X=x|U=u}] = \frac{Count(X=x, U=u) + Count'(X=x, U=u)}{Count(U=u) + Count'(U=u)}$$