Readings:
K&F: 3.3, 3.4, 16.1, 16.2, 16.3, 16.4

Learning P-maps
Param. Learning

Graphical Models – 10708
Carlos Guestrin
Carnegie Mellon University
September 24th, 2008

Perfect maps (P-maps)

- I-maps are not unique and often not simple enough
- Define "simplest" *G* that is I-map for *P*
  - A BN structure *G* is a **perfect map** for a distribution *P* if I(*P*) = I(*G*)
- Our goal:
  - Find a perfect map!
  - Must address equivalent BNs

Inexistence of P-maps 1

- XOR (this is a hint for the homework)

$A = B \text{ XOR } C$

$A \perp B$ | $\neg\, A \perp B \mid C$
$B \perp C$ | $\neg\, A \perp C \mid B$
$C \perp A$ | $\neg\, B \perp C \mid A$

I map : Minimal

$A \perp B$

$A \rightarrow C \leftarrow B$

not P-map
can't read
$B \perp C$
$A \perp C$

P-MAP ?
extra credit

Obtaining a P-map

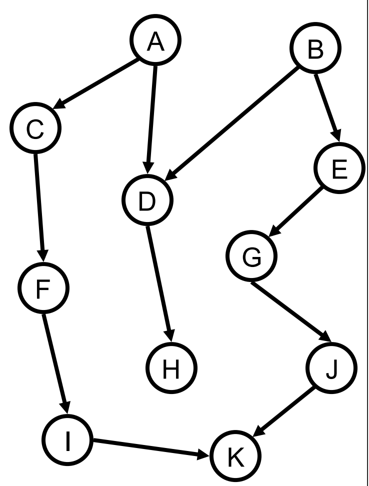
- Given the independence assertions that are true for *P*
- Assume that there exists a perfect map G^*
  - Want to find G^*
- Many structures may encode same independencies as G^*, when are we done?
  - Find all equivalent structures simultaneously!

I-Equivalence

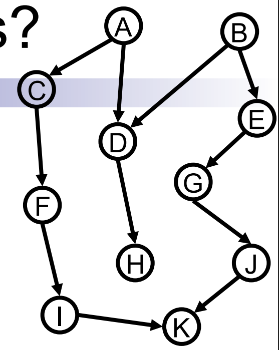
- Two graphs G_1 and G_2 are **I-equivalent** if $I(G_1) = I(G_2)$
- **Equivalence class** of BN structures
  - Mutually-exclusive and exhaustive partition of graphs
- How do we characterize these equivalence classes?

Skeleton of a BN

- **Skeleton** of a BN structure G is an **undirected graph** over the same variables that has an edge X–Y for every X→Y or Y→X in G
- (Little) **Lemma:** Two I-equivalent BN structures must have the same skeleton



What about V-structures?



- **V-structures are key property of BN structure**

- **Theorem:** If G_1 and G_2 have the same skeleton and V-structures, then G_1 and G_2 are I-equivalent

Same V-structures not necessary

- **Theorem:** If G_1 and G_2 have the same skeleton and V-structures, then G_1 and G_2 are I-equivalent

- Though sufficient, same V-structures not necessary

Immoralities & I-Equivalence

- Key concept not V-structures, but "immoralities" (unmarried parents ☺)
  - X → Z ← Y, with no arrow between X and Y
  - Important pattern: X and Y independent given their parents, but not given Z
  - (If edge exists between X and Y, we have *covered* the V-structure)
- **Theorem:** G_1 and G_2 have the same skeleton and immoralities if and only if G_1 and G_2 are I-equivalent

Obtaining a P-map

- Given the independence assertions that are true for *P*
  - Obtain skeleton
  - Obtain immoralities
- From skeleton and immoralities, obtain every (and any) BN structure from the equivalence class

Identifying the skeleton 1

- When is there an edge between X and Y?

- When is there no edge between X and Y?

Identifying the skeleton 2

- Assume d is max number of parents (d could be n)
- For each X_i and X_j
  - $E_{ij} \leftarrow$ true
  - For each $\mathbf{U} \subseteq \mathbf{X} - \{X_i, X_j\}$, $|\mathbf{U}| \leq d$
    - Is $(X_i \perp X_j \mid \mathbf{U})$?
      - $E_{ij} \leftarrow$ false
  - If E_{ij} is true
    - Add edge X – Y to skeleton

Identifying immoralities

- Consider X – Z – Y in skeleton, when should it be an immorality?

- Must be $X \rightarrow Z \leftarrow Y$ (immorality):
  - When X and Y are **never independent** given **U,** if $Z \in \mathbf{U}$

- Must **not** be $X \rightarrow Z \leftarrow Y$ (not immorality):
  - When there exists **U** with $Z \in \mathbf{U}$, such that X and Y are **independent** given **U**

From immoralities and skeleton to BN structures

- Representing BN equivalence class as a **partially-directed acyclic graph** (PDAG)

- Immoralities force direction on some other BN edges
- Full (polynomial-time) procedure described in reading

What you need to know

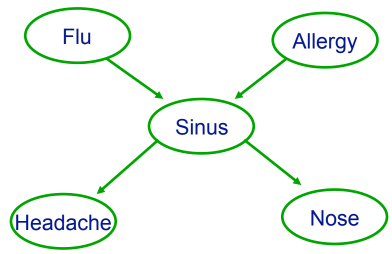
- Minimal I-map
  - every *P* has one, but usually many
- Perfect map
  - better choice for BN structure
  - not every *P* has one
  - can find one (if it exists) by considering I-equivalence
  - Two structures are I-equivalent if they have same skeleton and immoralities

Announcements

- Recitation tomorrow
  - Don't miss it!
- No class on Monday ☹

Review

- Bayesian Networks
  - Compact representation for probability distributions
  - Exponential reduction in number of parameters
  - Exploits independencies



- Next – Learn BNs
  - parameters
  - structure

Thumbtack – Binomial Distribution

- P(Heads) = θ, P(Tails) = 1-θ

- Flips are i.i.d.:
  - Independent events
  - Identically distributed according to Binomial distribution
- Sequence *D* of α_H Heads and α_T Tails

$$P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

Maximum Likelihood Estimation

- **Data:** Observed set D of α_H Heads and α_T Tails
- **Hypothesis:** Binomial distribution
- Learning θ is an optimization problem
  - What's the objective function?
- MLE: Choose θ that maximizes the probability of observed data:

$$\begin{aligned}\widehat{\theta} &= \arg\max_{\theta} \; P(\mathcal{D} \mid \theta) \\ &= \arg\max_{\theta} \; \ln P(\mathcal{D} \mid \theta)\end{aligned}$$

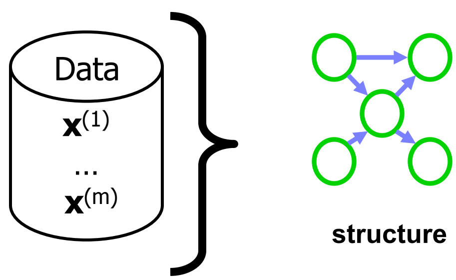
Your first learning algorithm

$$\begin{aligned}\widehat{\theta} &= \arg\max_{\theta} \; \ln P(\mathcal{D} \mid \theta) \\ &= \arg\max_{\theta} \; \ln \theta^{\alpha_H}(1-\theta)^{\alpha_T}\end{aligned}$$

- Set derivative to zero: $\boxed{\frac{d}{d\theta} \ln P(\mathcal{D} \mid \theta) = 0}$

Learning Bayes nets

	Known structure	Unknown structure
Fully observable data		
Missing data		

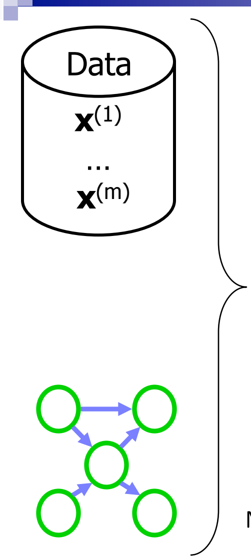


+

CPTs – $P(X_i \mid \mathbf{Pa}_{X_i})$

parameters

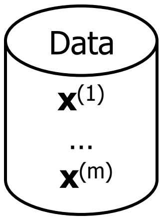
Learning the CPTs

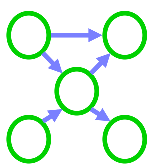


For each discrete variable X_i

MLE: $$P(X_i = x_i \mid X_j = x_j) = \frac{\text{Count}(X_i = x_i, X_j = x_j)}{\text{Count}(X_j = x_j)}$$

Learning the CPTs



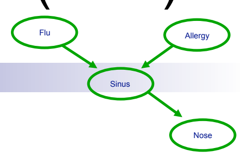


For each discrete variable X_i

MLE: $$P(X_i = x_i \mid X_j = x_j) = \frac{\text{Count}(X_i = x_i, X_j = x_j)}{\text{Count}(X_j = x_j)}$$

WHY??????????

Maximum likelihood estimation (MLE) of BN parameters – example



- Given structure, log likelihood of data:

$$\log P(\mathcal{D} \mid \theta_{\mathcal{G}}, \mathcal{G})$$

Maximum likelihood estimation (MLE) of BN parameters – General case

- Data: $\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(m)}$
- Restriction: $\mathbf{x}^{(j)}[\mathbf{Pa}_{Xi}] \to$ assignment to $\mathbf{Pa}_{Xi}$ in $\mathbf{x}^{(j)}$
- Given structure, log likelihood of data:

$$\log P(\mathcal{D} \mid \theta_{\mathcal{G}}, \mathcal{G})$$

Taking derivatives of MLE of BN parameters – General case

$$\log P(\mathcal{D} \mid \theta_{\mathcal{G}}, \mathcal{G}) = \sum_{j=1}^{m} \sum_{i=1}^{n} \log P\left(X_i = x_i^{(j)} \mid \mathbf{Pa}_{X_i} = \mathbf{x}^{(j)}\left[\mathbf{Pa}_{X_i}\right]\right)$$

General MLE for a CPT

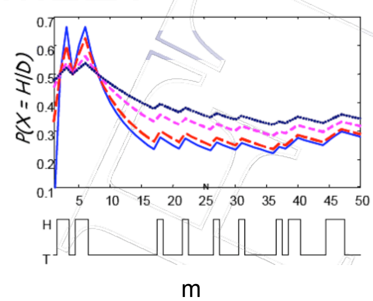
- Take a CPT: P(X|**U**)
- Log likelihood term for this CPT

- Parameter $\theta_{X=x|\mathbf{U}=\mathbf{u}}$:

MLE: $P(X = x \mid \mathbf{U} = \mathbf{u}) = \theta_{X=x|\mathbf{U}=\mathbf{u}} = \frac{\text{Count}(X = x, \mathbf{U} = \mathbf{u})}{\text{Count}(\mathbf{U} = \mathbf{u})}$

Can we really trust MLE?

- What is better?
  - 3 heads, 2 tails
  - 30 heads, 20 tails
  - 3×10^{23} heads, 2×10^{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

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

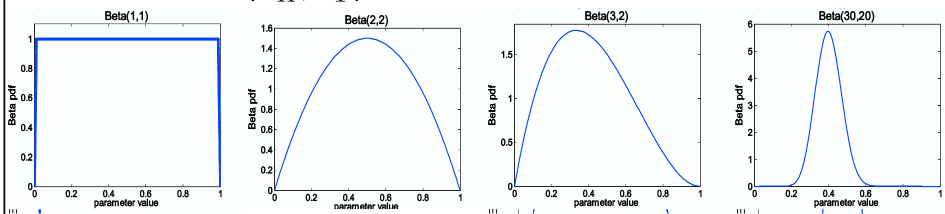
- Likelihood function is simply Binomial:

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

- 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)

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



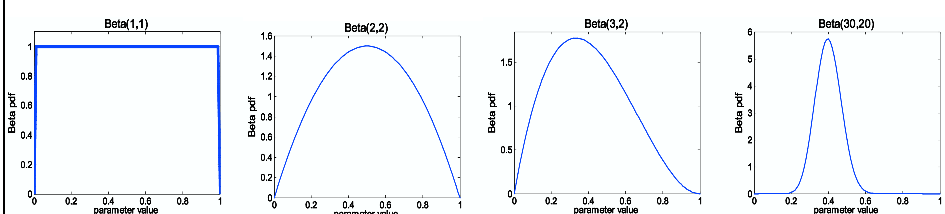
- Likelihood function: $P(\mathcal{D} \mid \theta) = \theta^{m_H}(1-\theta)^{m_T}$
- Posterior: $P(\theta \mid \mathcal{D}) \;\; \propto \;\; P(\mathcal{D} \mid \theta)P(\theta)$

10-708 – ©Carlos Guestrin 2006-2008 31

Posterior distribution

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

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



10-708 – ©Carlos Guestrin 2006-2008 32

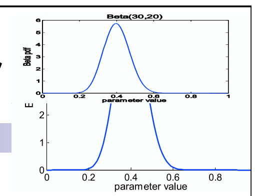
Conjugate prior

- 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:

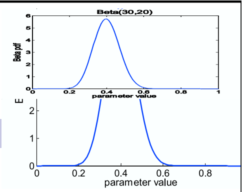
$$P(\theta \mid \mathcal{D}) \sim Beta(m_H + \alpha_H, m_T + \alpha_T)$$

- 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

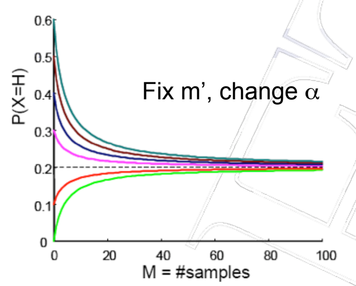
Bayesian prediction of a new coin flip

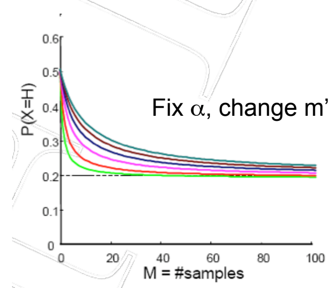


- Prior:
- Observed m_H heads, m_T tails, what is probability of m+1 flip is heads?

Asymptotic behavior and equivalent sample size

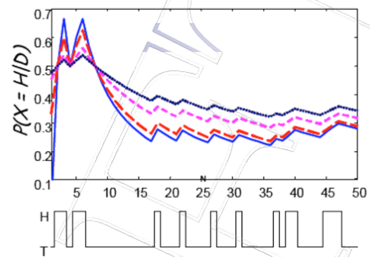
- Beta prior equivalent to extra thumbtack flips:
  - $E[\theta] = \frac{m_H + \alpha_H}{m_H + \alpha_H + m_T + \alpha_T}$
- As $m \rightarrow \infty$, prior is "forgotten"
- **But, for small sample size, prior is important!**
- **Equivalent sample size**:
  - Prior parameterized by α_H,α_T, or
  - m'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

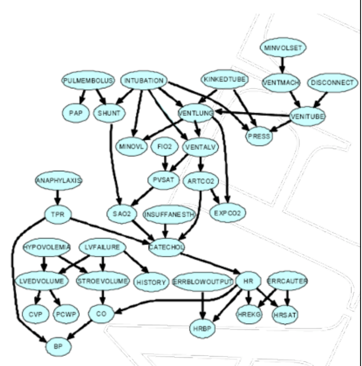
- What if you have a k sided coin???
- Likelihood function if **multinomial**:
  -
  -
- **Conjugate** prior for multinomial is **Dirichlet**:
  - $\theta \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k) \sim \prod_i \theta_i^{\alpha_i - 1}$
- **Observe** *m* data points, m_i from assignment i, **posterior**: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

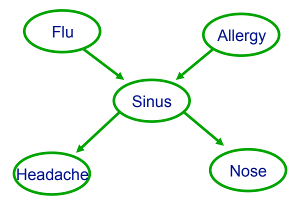


Global parameter independence, d-separation and local prediction

- 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

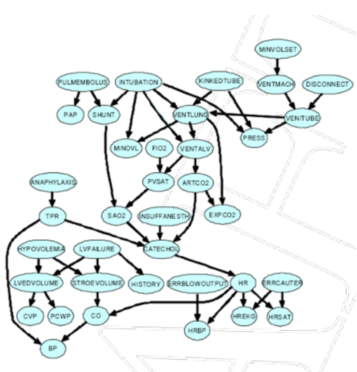
- Meta BN including CPT parameters:

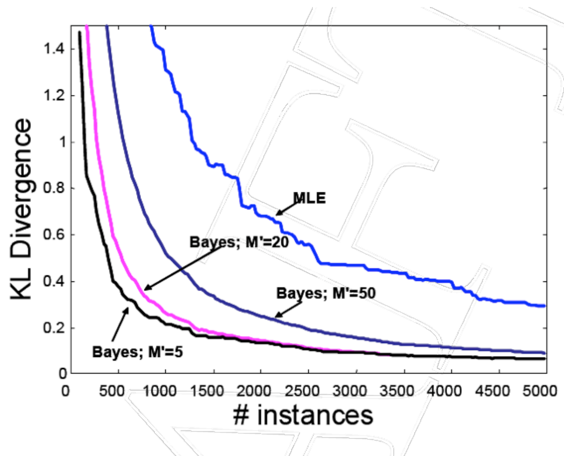
- Are $\theta_{Y|X=t}$ and $\theta_{Y|X=f}$ d-separated given *D*?
- Are $\theta_{Y|X=t}$ and $\theta_{Y|X=f}$ independent given *D*?
  - Context-specific independence!!!
- Posterior decomposes:

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

- MLE:
  - score decomposes according to CPTs
  - optimize each CPT separately
- Bayesian parameter learning:
  - motivation for Bayesian approach
  - Bayesian prediction
  - conjugate priors, equivalent sample size
  - Bayesian learning $\Rightarrow$ smoothing
- Bayesian learning for BN parameters
  - Global parameter independence
  - Decomposition of prediction according to CPTs
  - Decomposition within a CPT