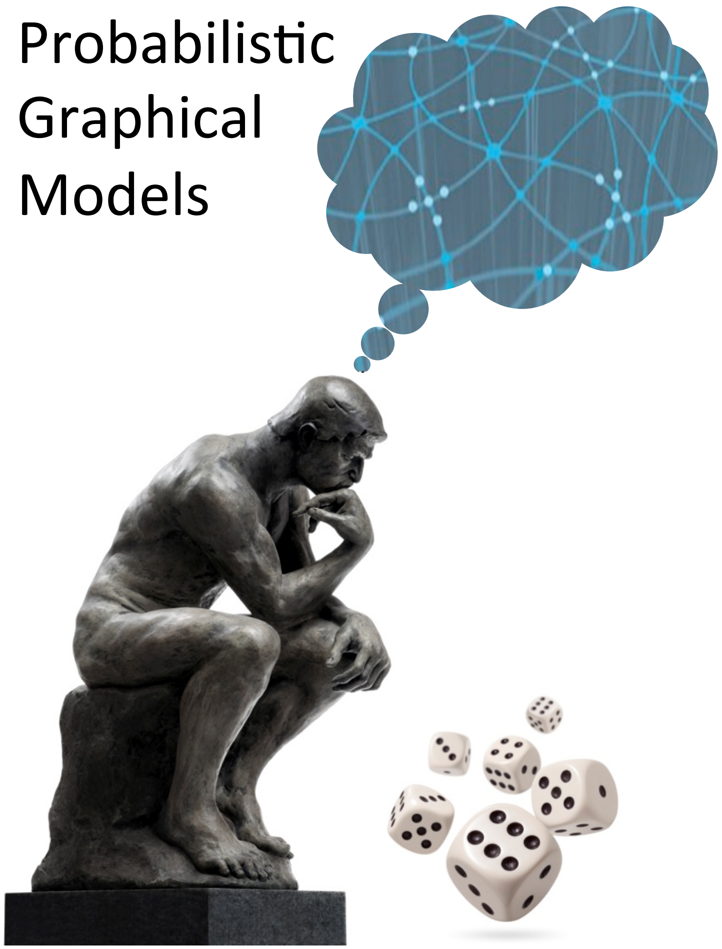
Probabilistic Graphical Models



Learning

BN Structure

Bayesian Score

Bayesian Score

Marginal likelihood

Prior over structures

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

Marginal probability of Data

independent of $\mathcal{G}$

$$\mathrm{score}_B(\mathcal{G} : \mathcal{D}) = \log P(\mathcal{D} \mid \mathcal{G}) + \log P(\mathcal{G})$$

Marginal Likelihood of Data Given G

$$\mathrm{score}_B(\mathcal{G} : \mathcal{D}) = \log P(\mathcal{D} \mid \mathcal{G}) + \log P(\mathcal{G})$$

Likelihood

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

Prior over parameters

$$P(\mathcal{D} \mid \mathcal{G}) = \int P(\mathcal{D} \mid \mathcal{G}, \boldsymbol{\theta}_{\mathcal{G}}) P(\boldsymbol{\theta}_{\mathcal{G}} \mid \mathcal{G}) d\boldsymbol{\theta}_{\mathcal{G}}$$

Marginal Likelihood Intuition

$$P(\mathcal{D} \mid \mathcal{G}) = P(\boldsymbol{x}[1], \ldots, \boldsymbol{x}[M] \mid \mathcal{G})$$

$$P(\boldsymbol{x}[1] \mid \mathcal{G})$$
$$P(\boldsymbol{x}[2] \mid \boldsymbol{x}[1], \mathcal{G})$$
$$\ldots$$
$$P(\boldsymbol{x}[M] \mid \boldsymbol{x}[1], \ldots, \boldsymbol{x}[M-1], \mathcal{G})$$

$\hat{\theta}_{\mathcal{G}}$ depends on all of D

Daphne Koller

Marginal Likelihood: BayesNets

variables

prior

suff stats

$$P(\mathcal{D} \mid \mathcal{G}) = \prod_i \left(\prod_{\boldsymbol{u}_i \in Val(\mathbf{Pa}^{\mathcal{G}}_{X_i})} \boxed{\frac{\Gamma(\alpha_{X_i|\boldsymbol{u}_i})}{\Gamma(\alpha_{X_i|\boldsymbol{u}_i} + M[\boldsymbol{u}_i])}} \prod_{x_i^j \in Val(X_i)} \left[\frac{\Gamma(\alpha_{x_i^j|\boldsymbol{u}_i} + M[x_i^j, \boldsymbol{u}_i])}{\Gamma(\alpha_{x_i^j|\boldsymbol{u}_i})} \right] \right).$$

suff stats

prior

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$$

$$\Gamma(x) = x \cdot \Gamma(x-1)$$

Daphne Koller

Marginal Likelihood Decomposition

$$P(\mathcal{D} \mid \mathcal{G}) = \prod_i \prod_{\boldsymbol{u}_i \in Val(\mathbf{Pa}^{\mathcal{G}}_{X_i})} \frac{\Gamma(\alpha_{X_i|\boldsymbol{u}_i})}{\Gamma(\alpha_{X_i|\boldsymbol{u}_i} + M[\boldsymbol{u}_i])} \prod_{x_i^j \in Val(X_i)} \left[\frac{\Gamma(\alpha_{x_i^j|\boldsymbol{u}_i} + M[x_i^j, \boldsymbol{u}_i])}{\Gamma(\alpha_{x_i^j|\boldsymbol{u}_i})} \right],$$

$$\log P(\mathcal{D} \mid \mathcal{G}) = \sum_i \mathrm{FamScore}_B(X_i \mid \mathbf{Pa}^{\mathcal{G}}_{X_i} \ : \ \mathcal{D})$$

Structure Priors

$$\text{score}_B(\mathcal{G} : \mathcal{D}) = \underline{\log P(\mathcal{D} \mid \mathcal{G})} + \underline{\log P(\mathcal{G})}$$

- Structure prior P(G)
  - Uniform prior: P(G) ∝ constant
  - Prior penalizing # of edges: P(G) ∝ $c^{|G|}$ (0<c<1)
  - Prior penalizing # of parameters
- Normalizing constant across networks is similar and can thus be ignored P(G)

Parameter Priors

- Parameter prior $P(\theta|G)$ is usually the BDe prior
  - α: equivalent sample size
  - B_0: network representing prior probability of events
  - Set $\alpha(x_i, pa_i^G) = \alpha\, P(x_i, pa_i^G \mid B_0)$
    - Note: pa_i^G are not the same as parents of X_i in B_0

      B_0
- A single network provides priors for all candidate networks
- Unique prior with the property that I-equivalent networks have the same Bayesian score

Daphne Koller

BDe and BIC

- As $M \rightarrow \infty$, a network G with Dirichlet priors satisfies

#instances

#independent parameters

$$\underbrace{\log P(\mathcal{D} \mid \mathcal{G})}_{\text{likelihood score}} = \underbrace{\ell(\hat{\boldsymbol{\theta}}_{\mathcal{G}} : \mathcal{D})}_{\text{log likelihood of } \mathcal{D} \text{ given MLE } \hat{\theta}_{\mathcal{G}}} - \frac{\log M}{2} \mathrm{Dim}[\mathcal{G}] + \underbrace{O(1)}_{\text{constant relative to } M \text{ — doesn't grow with } M}$$

$$\underbrace{\mathrm{score}_{BIC}(\mathcal{G} : \mathcal{D}) = \ell(\hat{\boldsymbol{\theta}}_{\mathcal{G}} : \mathcal{D}) - \frac{\log M}{2} \mathrm{Dim}[\mathcal{G}]}_{\text{as } M \rightarrow \infty \text{ score is consistent}}$$

Summary

- Bayesian score averages over parameters to avoid overfitting
- Most often instantiated as BDe
  - BDe requires assessing prior network
  - Can naturally incorporate prior knowledge
  - I-equivalent networks have same score
- Bayesian score
  - Asymptotically equivalent to BIC (as $m \to \infty$)
  - Asymptotically consistent learns correct network as $m \to \infty$
  - But for small M, BIC tends to underfit