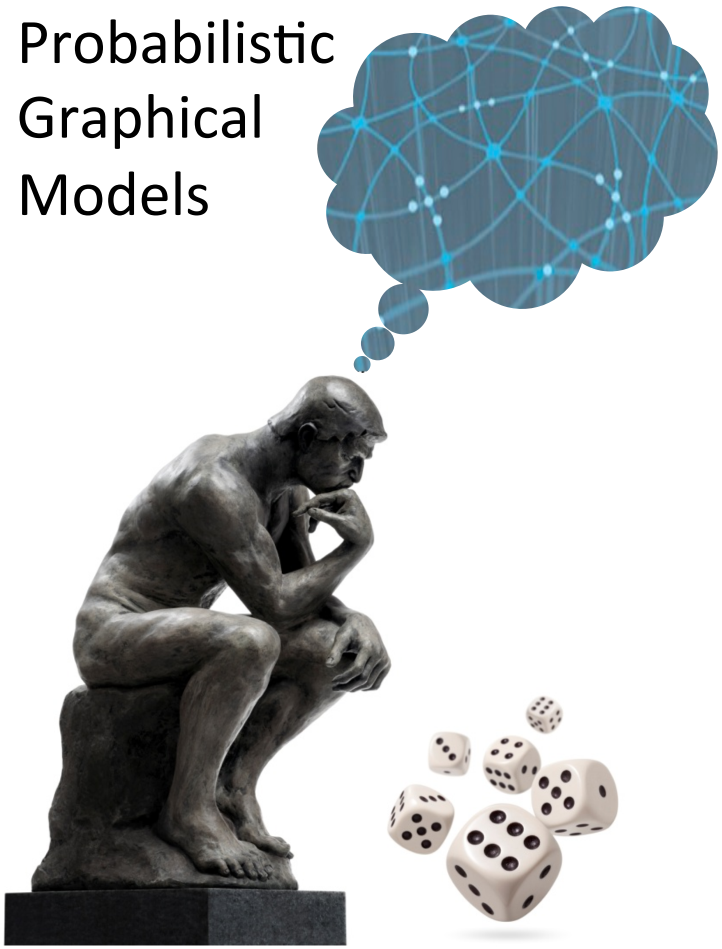
Probabilistic
Graphical
Models



Learning

BN Structurds

Likelihood Structure Score

Likelihood Score

- Find $\underline{(G,\theta)}$ that maximize the likelihood

$$\underline{\mathrm{score}_L}(\mathcal{G} \ : \ \mathcal{D}) = \underbrace{\ell((\hat{\boldsymbol{\theta}}, \mathcal{G}) : \mathcal{D})}$$

$\hat{\theta}$ = MLE of params. given $\mathcal{G}$ and $\mathcal{D}$

Example

$\mathcal{G}_0$: X Y

$\mathcal{G}_1$: X → Y

$$\text{score}_L(\mathcal{G}_0 \ : \ \mathcal{D}) = \sum_m (\log \hat\theta_{x[m]} + \log \hat\theta_{y[m]})$$

$$\text{score}_L(\mathcal{G}_1 \ : \ \mathcal{D}) = \sum_m (\log \hat\theta_{x[m]} + \log \hat\theta_{y[m]|x[m]})$$

$$\text{score}_L(\mathcal{G}_1 \ : \ \mathcal{D}) - \text{score}_L(\mathcal{G}_0 \ : \ \mathcal{D}) = \sum_m (\log \hat\theta_{y[m]|x[m]} - \log \hat\theta_{y[m]})$$

$$= \sum_{x,y} M[x,y] \log \hat\theta_{y|x} - \sum_y M[y] \log \hat\theta_y$$

$$= M \sum_{x,y} \hat P(x,y) \log \hat P(y \mid x) - M \sum_y \hat P(y) \log \hat P(y)$$

$$= \frac{\hat P(y,x)}{\hat P(x)}$$

$$= M \left(\sum_{x,y} \hat P(x,y) \log \hat P(y \mid x) - \sum_{x,y} \hat P(x,y) \log \hat P(y) \right)$$

$$= M \sum_{x,y} \hat P(x,y) \log \frac{\hat P(x,y)}{\hat P(x) \hat P(y)} = M \cdot \mathbf{I}_{\hat P}(X;Y)$$

$\hat P$ = empirical distribution

$M[x,y] = M \hat P[x,y]$

$\sum_x \hat P(x,y) = \hat P(y)$

mutual information

Daphne Koller

General Decomposition

- The Likelihood score decomposes as:

mutual information — independent of G

$$\underbrace{\text{score}_L(\mathcal{G} \; : \; \mathcal{D})} = M \underbrace{\sum_{i=1}^{n} \mathbf{I}_{\hat{P}}(X_i; \mathbf{Pa}^G_{X_i})} - M \underbrace{\sum_{i} \boldsymbol{H}_{\hat{P}}(X_i)}$$

Score is higher if X_i is correlated with parents

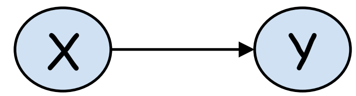
$$\mathbf{I}_P(\boldsymbol{X};\boldsymbol{Y}) = \sum_{\boldsymbol{x},\boldsymbol{y}} P(\boldsymbol{x},\boldsymbol{y}) \log \frac{P(\boldsymbol{x},\boldsymbol{y})}{P(\boldsymbol{x})P(\boldsymbol{y})}$$

$$\boldsymbol{H}_P(\boldsymbol{X}) = -\sum_{\boldsymbol{x}} P(\boldsymbol{x}) \log P(\boldsymbol{x})$$

Limitations of Likelihood Score

$\mathcal{G}_0$

$\mathcal{G}_1$



$$\text{score}_L(\mathcal{G}_1 \ : \ \mathcal{D}) - \text{score}_L(\mathcal{G}_0 \ : \ \mathcal{D}) = M\mathbf{I}_{\hat{P}}(X;Y)$$

- Mutual information is always ≥ 0
- Equals 0 iff X, Y are independent
  - In empirical distribution $\hat{P}$ $\quad$ $I_{\hat{P}}(X;Y) > 0$ $\quad$ almost always
- Adding edges can't hurt, and almost always helps
- Score maximized for fully connected network

Daphne Koller

Avoiding Overfitting

- Restricting the hypothesis space
  - restrict # of parents or # of parameters
- Scores that penalize complexity:
  - Explicitly
  - Bayesian score averages over all possible parameter values

Summary

- Likelihood score computes log-likelihood of D relative to G, using MLE parameters for G
  - Parameters optimized for D
- Nice information-theoretic interpretation in terms of (in)dependencies in G
- Guaranteed to overfit the training data (if we don't impose constraints)