Structure Learning 2

Structure Scores

• Searching for highest-scoring network structure is intractable
• Need to resort to heuristic search (e.g., hillclimbing)
• Need:
  1. Search space
  2. Scoring function
  3. Search procedure

1. Search space
  • Start with initial state (e.g., disconnected graph or randomly generated one)

Initial State

A

B

C

Structure Scores

1. Search space
  • Move to a neighboring state by applying an operator:

Edge Addition

A

B

C

Edge Deletion

A

B

C

Edge Reversal

A

B

C

Can only perform an operator if it doesn't lead to a cycle!
2. Scoring function:
   - Two general classes of scoring functions:
     1. Likelihood scoring functions
     2. Bayesian scoring functions
   - More about this in a bit…assume we have a scoring function for now

3. Search procedure
   - Greedy search: pick the best scoring neighboring state to move to
   - Repeat until convergence
   - Converges to a local optimum

Likelihood Scores

\[
\max_{G, \theta} L(G, \theta; D) = \max_{G} \left[ \max_{\theta} L(G, \theta; D) \right] = \max_{G} L(G, \theta; D)
\]

Graph structure that maximizes the likelihood

Maximum likelihood estimates of parameters

\[
\text{score}_L(G; D) = l(\theta_G; D)
\]

Log likelihood

Tricks for dealing with this:
- random restart, simulated annealing, tabu search and data perturbation
Likelihood Scores

Let $M$ be the number of samples. We use the notation $M[x]$ to be the count of $x$ in the data.

Let $\hat{p}$ be the empirical distribution observed in the data. Eg.
1. $M[x, y] = M \cdot \hat{p}(x, y)$
2. $M[y] = M \cdot \hat{p}(y)$

Note that:
1. $\hat{p}_{y|x} = \hat{p}(y|x)$
2. $\hat{p}_y = \hat{p}(y)$

Likelihood Scores

Claim:

$$score_L(G; D) = M \sum_{i=1}^{n} I_p(X_i; Parents(X_i, G)) - M \sum_{i=1}^{n} H_p(X_i)$$

$$= M \sum_{i=1}^{n} [I_p(X_i; Parents(X_i, G)) - H_p(X_i)]$$

Likelihood Scores

Mutual Information

$$I_p(X; Y) = \sum_{x,y} \hat{p}(x,y) \log \frac{\hat{p}(x,y)}{\hat{p}(x)\hat{p}(y)}$$

$$= \frac{1}{M} \sum_{x,y} M[x,y] \log \left( \frac{M[x,y]}{M[x]M[y]} \right)$$

Likelihood Scores

Proof:

$$l(\theta; D) = \sum_{i=1}^{n} \left[ \sum_{u_i \in Val(Parents(X_i, G))} \sum_{x_i} M[x_i, u_i] \log \hat{\theta}_{x_i|u_i} \right]$$

$$= M \sum_{i=1}^{n} \left[ \frac{1}{M} \sum_{u_i} \sum_{x_i} M[x_i, u_i] \log \hat{\theta}_{x_i|u_i} \right]$$
Likelihood Scores

What are the implications of 
$I_p(X_i; U_i) - H_p(X_i)$

Depends on network structure (because $U_i = \text{Parents}(X_i, G)$). Only need to maximize this.

The likelihood of a network measures how informative $\text{Parents}(X_i)$ are about $X_i$

Note that if $\text{Parents}(X_i, G) = \emptyset$, then $I_p(X_i; \text{Parents}(X_i, G)) = 0$

Likelihood Scores

An alternate representation:

$\frac{1}{M} \sum_{u_i} \sum_{x_i} M[x_i, u_i] \log \hat{\theta}_{x_i|u_i}$

$= \sum_{u_i} \sum_{x_i} \hat{P}(x_i, u_i) \log \hat{P}(x_i|u_i)$

$= \sum_{u_i} \sum_{x_i} \hat{P}(x_i, u_i) \log \left( \frac{\hat{P}(x_i, u_i)}{\hat{P}(u_i) \cdot \hat{P}(x_i)} \right)$

$= \sum_{u_i} \sum_{x_i} \hat{P}(x_i, u_i) \log \left( \frac{\hat{P}(x_i, u_i)}{\hat{P}(u_i) \cdot \hat{P}(x_i)} \right) + \sum_{x_i} \left( \sum_{u_i} \hat{P}(x_i, u_i) \right) \log \hat{P}(x_i)$

$= I_p(X_i; U_i) - \sum_{x_i} \hat{P}(x_i) \log \frac{1}{\hat{P}(x_i)}$

$= I_p(X_i; U_i) - H_p(X_i)$

Note that if $\text{Parents}(X_i, G) = \emptyset$, then $I_p(X_i; \text{Parents}(X_i, G)) = 0$

Problems with Likelihood Score

Never prefers a simpler network over a more complex one eg.

$G_1$

$X \rightarrow Y$

$G_0$

$X$

$Y$

Measures to what extent the Markov properties of the graph are violated in the data (fewer violations $\Rightarrow$ larger score)
Problems with Likelihood Score

- Exhibits a conditional independence only if it holds exactly in the empirical distribution.
  - Due to noise, this almost never happens.
- Learns a fully connected graph.
  - Overfits the training data and does not generalize well to unseen cases.
- Needs a penalty for learning overly complex structures.

Bayesian Scoring

Bayesian Score

- Bayesian philosophy: if you are uncertainty about something, put a distribution over it.
- In structure learning, we have uncertainty over the structure and the parameters.
- We will have two prior distributions:
  - Structure prior $P(G)$
  - Parameter prior $P(\theta_G | G)$

Bayesian Score

Recall: $P(G|D) = \frac{P(D|G)P(G)}{P(D)} = \alpha P(G|D)P(G)$

score$_B(G; D) = \log P(D | G) + \log P(G)$

Marginal Likelihood (dominates the score)

Structure prior

$P(D|G) = \int_{\Theta_G} P(D|\theta_G, G)P(\theta_G | G)d\theta_G$

"Averages" out $P(D|\theta_G, G)$ over the distribution of $\theta_G$. Contrast this with maximum likelihood which finds the $\theta_G$ that maximizes the likelihood of the data.
Bayesian score

- How does the Bayesian score improve over the likelihood score?
  - By avoiding overfitting
- Likelihood score commits to a single \( \hat{\theta} \) value
- Bayesian score works with a distribution of \( \theta \) and averages \( P(D|\theta, G) \) over this distribution
  - Results in an expected likelihood

Marginal Likelihood (Single Variable case)

- Suppose we have a single binary random variable \( X \)
- Let the prior distribution over the parameters of \( X \) be \( \text{Dirichlet}(\alpha_1, \alpha_0) \)
- Let the data \( D = \{x[1], ..., x[M]\} \) have \( M[1] \) heads and \( M[0] \) tails
- Maximum likelihood value given \( D \) is:
  \[
P(D|\hat{\theta}) = \left( \frac{M[1]}{M} \right)^{M[1]} \cdot \left( \frac{M[0]}{M} \right)^{M[0]}
  \]

Shorthand: let \( p_i = \frac{M[i]}{M} \) and \( \alpha = \alpha_0 + \alpha_1 \)
Bayesian Scoring

Global parameter independence:
Let $\mathcal{G}$ be a Bayesian network structure with parameters $\theta = (\theta_{X_1 | P(alpha(X_1))}, \ldots, \theta_{X_n | P(alpha(X_n))})$.

The distribution $P(\theta)$ satisfies global parameter independence if it has the form:

$$P(\theta) = \prod_{i=1}^{n} P(\theta_{X_i | P(alpha(X_i))})$$

Bayesian Scoring

Local parameter independence:
Let $X$ be a variable with parents $U$. We say that distribution $P(\theta_{X|U})$ satisfies local parameter independence if:

$$P(\theta_{X|U}) = \prod_{u} P(\theta_{X|u})$$

Example:

| X | Y | P(Y|X) |
|---|---|-------|
| 0 | 0 | $\theta_{0,0}$ |
| 0 | 1 | $\theta_{0,1}$ |
| 1 | 0 | $\theta_{1,0}$ |
| 1 | 1 | $\theta_{1,1}$ |

Note that the Gamma function is as follows:

$$\Gamma(\alpha) = \int_{0}^{\infty} e^{-x} x^{\alpha-1} dx$$

ie. it is a continuous generalization of the factorial:

$$\Gamma(n+1) = n!$$

We can easily generalize to a multinomial distribution over the space of values $x^1, \ldots, x^k$ with a prior $\text{Dirichlet}(\alpha_1, \ldots, \alpha_k)$:

$$P(D|\mathcal{G}) = \frac{\Gamma(\alpha)}{\Gamma(\alpha + M)} \cdot \frac{\Gamma(\alpha_1 + 1)}{\Gamma(\alpha_1)} \cdot \cdots \cdot \frac{\Gamma(\alpha_k + M[x^1])}{\Gamma(\alpha_k)}$$
Bayesian Scoring

Now suppose there are two binary random variables X and Y. Let $g_0$ be a graph with X and Y and no edges.

$$P(D|g_0) = \int_{\Theta_X \times \Theta_Y} P(D|\theta_X, \theta_Y, g_0) P(\theta_X, \theta_Y|g_0) d[\theta_X, \theta_Y]$$

1. Decompose likelihood in terms of each variable

$$P(D|\theta_X, \theta_Y, g_0) = \prod_{i=1}^M P(x|m)[\theta_X, g_0] P(y|m)[\theta_Y, g_0]$$

2. Global Parameter Independence: $P(\theta_X, \theta_Y|g_0) = P(\theta_X|g_0)P(\theta_Y|g_0)$

Bayesian Scoring

Now suppose there are two binary random variables X and Y and let $g_{X \rightarrow Y}$ be the graph below:

$$P(D|g_{X \rightarrow Y}) = \left( \int_{\Theta_X} \prod_{m=1}^M P(x|m)\theta_{X \rightarrow Y} P(\theta_{X \rightarrow Y}|g_{X \rightarrow Y}) d\theta_X \right) \cdot \left( \int_{\Theta_Y} \prod_{m=1}^M P(y|m)|\theta_{X \rightarrow Y} P(\theta_{Y|x}|g_{X \rightarrow Y}) d\theta_Y \right)$$

Note: decomposes into one term for each random variable.
Bayesian Scoring

Now suppose there are two binary random variables $X$ and $Y$ and let $G_{X \rightarrow Y}$ be the graph below:

![Bayesian Scoring Graph]

One term for each parameter family. Each term has a closed form solution.

$$P(D|G_{X \rightarrow Y}) = \left( \int_{\theta} \prod_{m=1}^{M} P(y[m]|\theta, G_{X \rightarrow Y}) \right) \prod_{k} \Gamma(\alpha + M \mu[k]) \Gamma(\alpha)$$

The general case: let $G$ be a network structure, and let $P(\theta|G)$ be a parameter prior satisfying global parameter independence. Then:

$$P(D|G) = \prod_{i=1}^{n} \prod_{\theta_{(k)}} \int_{\theta_{(k)}} P(x_i[m]|\theta_{(k)}) \prod_{\theta_{(k)}} P(\theta_{(k)}) d\theta_{(k)}$$

If $P(\theta)$ also satisfies local parameter independence, then

$$P(D|G) = \prod_{i=1}^{n} \prod_{\theta_{(s)}} \int_{\theta_{(s)}} P(x_i[m]|\theta_{(s)}) \prod_{\theta_{(s)}} P(\theta_{(s)}) d\theta_{(s)}$$

Bayesian Scoring

If we have a Bayesian network with Dirichlet priors where $P(\theta|pa(X_i)|G)$ has hyperparameters $\left\{ \alpha^j_{X_i|u_i} \right\}$ then

$$P(D|G) = \prod_{i=1}^{n} \prod_{u_i \in pa(X_i) \cap \theta} \Gamma\left( \alpha^j_{X_i|u_i} + M[x_i/u_i] \right) \Gamma\left( \alpha^j_{X_i|u_i} \right)$$

Where:

$$\alpha^j_{X_i|u_i} = \sum_{j} \alpha^j_{X_i|u_i}$$

Bayesian Scoring

If we have a Bayesian network with Dirichlet priors where $P(\theta|pa(X_i)|G)$ has hyperparameters $\left\{ \alpha^j_{X_i|u_i} \right\}$ then

$$P(D|G) = \prod_{i=1}^{n} \prod_{u_i \in pa(X_i) \cap \theta} \Gamma\left( \alpha^j_{X_i|u_i} + M[x_i/u_i] \right) \Gamma\left( \alpha^j_{X_i|u_i} \right)$$

Where:

- Iterates over # of random variables
- Iterates over # of instantiations of parents of $X_i$
- Iterates over # of values of $X_i$
If we use a Dirichlet parameter prior for all parameters in our network, then, because $M \to \infty$ (proof omitted), we have:

$$\log P(D|G) = l(\hat{\theta}_G; D) - \frac{\log M}{2} \text{Dim}[G] + o(1)$$

# of independent parameters in $G$

This is the Bayesian Information Criterion (BIC) score.

This is the Bayesian Information Criterion (BIC) score:

$$\text{score}_{\text{BIC}}(G; D) = l(\hat{\theta}_G; D) - \frac{\log M}{2} \text{Dim}[G] + O(1)$$

Can also interpret this as the # of bits to encode the model and the data given the model (minimum description length).

Assume that our data are generated by some distribution $P^*$ for which the network $G^*$ is a perfect map.

We say that a scoring function is consistent if the following properties hold as the amount of data $M \to \infty$, with probability that approaches 1 (over all possible choices of data set $D$):

- The structure $G^*$ will maximize the score
- All structures $G$ that are not I-equivalent to $G^*$ will have strictly lower score

Things to note:

- Entropy term $M \sum_{i=1}^{n} H_p(X_i)$ can be ignored (doesn’t depend on graph structure)
- Trades off fit to data and model complexity
  - The stronger the dependence of a variable on its parents, the higher the score (grows linearly)
  - The more complex the network, the lower the score (grows logarithmically)
- As $M$ grows, the score pays more attention to the data fit
**Bayesian Scoring**

- The BIC score (and the Bayesian score) is consistent [proof omitted]

- In practice though, the BIC score tends to have a very strong preference for simpler structures

**Structure Priors**

- Typically assign uniform priors over structures
- If you can provide an informed structure prior, you could penalize edges in the graph:
  - $P(G) \propto c^{|G|}$ (where $c < 1$ and $|G|$ is the number of edges)

- Mathematically convenient to have structure prior with structure modularity:
  - $P(G) \propto \prod_i P(Pa(X_i) = Pa^G(X_i))$

  *Uses local properties not global properties of the graph*

Recall that

$$\text{score}_B(G; D) = \log P(D | G) + \log P(G)$$

- Grows linearly with the number of examples (dominates the score)
- Structure prior (stays constant). Only matters for small sample sizes