Structure Learning 2

Structure Scores

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

Structure Scores

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

   \[
   \text{Initial State} \\
   \begin{array}{c}
   A \\
   B \\
   C \\
   \end{array}
   \]

Structure Scores

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

   - Edge Addition
   - Edge Deletion
   - Edge Reversal

   \[
   \begin{array}{c}
   \text{Edge Addition} \\
   \begin{array}{c}
   A \\
   B \\
   C \\
   \end{array}
   \end{array}
   \begin{array}{c}
   \text{Edge Deletion} \\
   \begin{array}{c}
   A \\
   B \\
   C \\
   \end{array}
   \end{array}
   \begin{array}{c}
   \text{Edge Reversal} \\
   \begin{array}{c}
   A \\
   B \\
   C \\
   \end{array}
   \end{array}
   \]

   Can only perform an operator if it doesn’t lead to a cycle!
Structure Scores

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

4. Tricks for dealing with this:
   - Random restart, simulated annealing, tabu search and data perturbation

Likelihood Scores

\[
\max_{\theta} L(G, \theta) \quad \text{subject to} \quad L(G, \theta) \leq \max_{\theta} L(G, \theta)
\]

Graph structure that maximizes the likelihood

Maximum likelihood estimates of parameters

\[
\hat{\theta}_{\text{ML}}(G; D) = l(\hat{\theta}_{\text{ML}}; D)
\]

Log likelihood
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.
- $M[x, y] = M \cdot \hat{p}(x, y)
- M[y] = M \cdot \hat{p}(y)

Note that:
- $\hat{p}[y|x] = \hat{p}(y|x)
- \hat{p}[y] = \hat{p}(y)$

Likelihood Scores

Claim:

\[
\text{score}_L (G; \mathcal{D}) = M \sum_{i=1}^{n} I_p(X_i; \text{Parents}(X_i, G)) - M \sum_{i=1}^{n} H_p(X_i)
\]

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:

\[
l(\hat{\theta}_G; \mathcal{D}) = \sum_{i=1}^{n} \left[ \sum_{u \in \text{Val} (\text{Parents}(X_i, G))} \sum_{x_i} M[x_i, u_i] \log \hat{\theta}_x[u_i] \right]
\]

Take one of these terms and let $u_i = \text{Parents}(X_i, G)$

\[
\frac{1}{M} \sum_{u_i} \sum_{x_i} M[x_i, u_i] \log \hat{\theta}_x[u_i]
\]
Likelihood Scores

\[
\frac{1}{M} \sum_{u_i} \sum_{x_i} M[x_i, u_i] \log \hat{\theta}_{x_i | u_i} \\
= \sum_{u_i} \sum_{x_i} \hat{p}(x_i, u_i) \log \hat{p}(x_i | u_i) \\
= \sum_{u_i} \sum_{x_i} \hat{p}(x_i, u_i) \log \left( \frac{\hat{p}(x_i, u_i)}{\hat{p}(u_i)} \frac{\hat{p}(x_i)}{\hat{p}(x_i)} \right) \\
= \sum_{u_i} \sum_{x_i} \hat{p}(x_i, u_i) \log \left( \frac{\hat{p}(x_i, u_i)}{\hat{p}(u_i)} \frac{\hat{p}(x_i)}{\hat{p}(x_i)} \right) + \sum_{x_i} \hat{p}(x_i) \log \hat{p}(x_i) \\
= I_p(X; U) - \sum_{x_i} \hat{p}(x_i) \log \frac{1}{\hat{p}(x_i)} \\
= I_p(X; U) - H_p(X) \\
\]

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

Likelihood Scores

An alternate representation:

\[
\frac{1}{M} \text{score}_G(G; D) = H_p(X_1, \ldots, X_n) \\
= \sum_{i=1}^{n} I_p(X_i; X_{i-1}) - \text{Parents}(X_i, G) | \text{Parents}(X_i, G)) \\
\]

Measures to what extent the Markov properties of the graph are violated in the data (fewer violations ⇒ larger score)

Likelihood Scores

What are the implications of

\[
I_p(X_i; U_i) - H_p(X_i) \\
\]

Depends on network structure (because \( U_i = \text{Parents}(X_i, G) \)).

Does not depend on network structure

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

Problems with Likelihood Score

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

\[
\text{score}_G(G; D) \geq \text{score}_{G'}(G'; D) \\
\]

\[
G_1 \quad G_0 \\
X \quad X \\
Y \quad Y \\
\]
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 $P(\theta_g | G)$

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

Score $B(G; D) = \log P(D | G) + \log P(G)$

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

*“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 $\tilde{\theta}$ value
- Bayesian score works with a distribution of $\theta_G$ and averages $P(\mathcal{D}|\theta_G, G)$ over this distribution
  - Results in an expected likelihood

Marginal Likelihood (Single Variable case)

What about the marginal likelihood?

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

Shorthand: let $p_i = \frac{M[i]}{M}$ and $\alpha = \alpha_0 + \alpha_1$

Marginal Likelihood (Single Variable case)

Let the prior distribution over the parameters of $X$ be Dirichlet($\alpha_1, \alpha_0$)
Let the data $\mathcal{D} = \{x[1], ..., x[M]\}$ have $M[1]$ heads and $M[0]$ tails
Maximum likelihood value given $D$ is: $P(\mathcal{D}|\tilde{\theta}) = \left(\frac{M[1]}{M}\right)^{M[1]} \cdot \left(\frac{M[0]}{M}\right)^{M[0]}$

Note:

$$\int_{\theta_G} \text{Beta}(M[1]+\alpha_1, M[0]+\alpha_0)d\theta_G = 1$$

$$\int_{\theta_G} \frac{\Gamma(\alpha+M)}{\Gamma(\alpha_1+1, \alpha_0+1)} \cdot \prod_{i} \Gamma(\alpha_0+\alpha_1+1) = 1$$

$$\int_{\theta_G} p_i^{\alpha_0+\alpha_1+1} \cdot (1-p_i)^{\alpha_0+\alpha_1+1}d\theta_G = \frac{\Gamma(M[1]+\alpha_1, M[0]+\alpha_0)}{\Gamma(\alpha+M)}$$
Marginal Likelihood (Single Variable case)

\[
P(D|G) = \int_{\theta_0} p_1^{M(1)} p_0^{M(0)} \frac{\Gamma(\alpha_0 + \alpha_1)}{\Gamma(\alpha_0) \Gamma(\alpha_1)} p_1^{(\alpha_1-1)} p_0^{(\alpha_0-1)} d\theta_0
\]

\[
\frac{\Gamma(\alpha)}{\Gamma(\alpha_0) \Gamma(\alpha_1)} \int_{\theta_0} p_1^{M(1)+\alpha_1-1} p_0^{M(0)+\alpha_0-1} d\theta_0 = 1
\]

Note: The Gamma function is as follows:

\[
\Gamma(1) = 1
\]

\[
\Gamma(x + 1) = x \Gamma(x)
\]

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

\[
P(D|G) = \frac{\Gamma(\alpha)}{\Gamma(\alpha + M)} \cdot \prod_{i=1}^{k} \frac{\Gamma(\alpha_i + M[x^i])}{\Gamma(\alpha_i)}
\]

Bayesian Scoring

Global parameter independence:

Let \(G\) be a Bayesian network structure with parameters \(\theta = (\theta_{X_1|P(a(X_1))}, ..., \theta_{X_n|P(a(X_n))})\).

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

\[
P(\theta) = \prod_{i=1}^{n} P(\theta_{X_i|P(a(X_i))})
\]

Local parameter independence:

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

\[
P(\theta_{X|U}) = \prod_{u} P(\theta_{X|U})
\]
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: $P(\theta_X, \theta_Y|G_0) = P(\theta_X|G_0)P(\theta_Y|G_0)$

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:

```
 X ----> Y
```

| $X$ | $P(X)$ | $X$ | $Y$ | $P(Y|X)$ |
|-----|--------|-----|-----|----------|
| 0   | $\theta_x$ | 0   | 0   | $\theta_y^{00}$ |
| 1   | $\theta_x$   | 0   | 1   | $\theta_y^{01}$   |
|     |           | 1   | 0   |           |
|     |           | 1   | 1   |           |

$$P(D|G_{X \rightarrow Y}) = \left( \int_{x} \prod_{m=1}^{M} P(x[m]|\theta_x, G_{X \rightarrow Y})P(\theta_x|G_{X \rightarrow Y})d\theta_x \right) \cdot \left( \int_{y} \prod_{m=1}^{M} P(y[m]|\theta_y, G_{X \rightarrow Y})P(\theta_y|G_{X \rightarrow Y})d\theta_y \right)$$

Integral of a product of independent functions is the product of integrals:

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:

```
 X ----> Y
```

$$P(D|G_{X \rightarrow Y}) = \left( \int_{x} \prod_{m=1}^{M} P(x[m]|\theta_x, G_{X \rightarrow Y})P(\theta_x|G_{X \rightarrow Y})d\theta_x \right) \cdot \left( \int_{y} \prod_{m=1}^{M} P(y[m]|\theta_y, G_{X \rightarrow Y})P(\theta_y|G_{X \rightarrow Y})d\theta_y \right)$$

Integral of a product of independent functions is the product of integrals:

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:

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

\[
P(D | G_{X \rightarrow Y}) = \left( \prod_{s=1}^{M} \int_{\theta_{s \leftarrow X}} P(\theta_{s \leftarrow X}, G_{X \rightarrow Y}) \prod_{m=1}^{\bar{m}} \frac{\Gamma(\alpha_{s \leftarrow X} + M[x_{s \leftarrow X}, m])}{\Gamma(\alpha_{s \leftarrow X})} \right)
\]

Bayesian Scoring

If we have a Bayesian network with Dirichlet priors where \( P(\theta_{X | \theta_{X-\bar{X}} | G}) \) has hyperparameters \( \{ \alpha_{x_{j} | u_{i}} : j = 1, ..., |X_{i}| \} \) then

\[
P(D | G) = \prod_{i=1}^{n} \prod_{u_{i} \in \text{val}(X_{i})} \frac{\Gamma(\alpha_{x_{j} | u_{i}} + M[x_{j} | u_{i}])}{\Gamma(\alpha_{x_{j} | u_{i}})} \]

Where:

\[
\alpha_{x_{j} | u_{i}} = \sum_{j} \alpha_{x_{j} | u_{i}}
\]
Bayesian Scoring

If we use a Dirichlet parameter prior for all parameters in our network, then, because $M \to \infty$ (proof omitted), we have:

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

# of independent parameters in $G$

This is the Bayesian Information Criterion (BIC) score

Bayesian Scoring

This is the Bayesian Information Criterion (BIC) score:

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

Things to note:
- Entropy term $M \sum_{i=1}^{n} H_R(X_i)$ 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

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
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

Recall that

\[ \text{score}_{\text{B}}(G;D) = \log P(D | G) + \log P(G) \]

Grows linearly with the number of examples (dominates the score)

Structure prior (stays constant). Only matters for small sample sizes

• 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