Structure learning in Bayesian networks

40-956 Advanced Topics in AI: Probabilistic Graphical Models
Sharif University of Technology

Soleymani
Spring 2015
Recall: Learning problem

- Target: true distribution $P^*$ that maybe correspond to $M^* = (K^*, \theta^*)$
  - Hypothesis space: specified probabilistic graphical models

- Data: set of instances sampled from $P^*$

- Learning goal: selecting a model $\tilde{M}$ to construct the best approximation to $M^*$ according to a performance metric
  - We may use Bayesian model averaging instead of selecting a model $\tilde{M}$
Structure learning of Bayesian networks

For a fixed set of variables (i.e. nodes), we intend to learn the directed graph structure $\mathcal{G}^*$:

- Which edges are included in the model?
  - Fewer edges miss dependencies
  - More edges cause spurious dependencies
- We often prefer sparser structures that show better generalization
Learning the structure of DGMs

- Approaches
  - Constraint-based
    - Find conditional independencies in data using statistical tests
    - Find an I-map graph for the obtained conditional independencies
  - Score-based
    - View a BN structure and parameter learning as a density estimation task and address structure learning as a model selection problem
    - Score function measures how well the model fits the observed data
      - Maximum likelihood
      - Bayesian score
Score-based approach

- **Input:**
  - Hypothesis space: set of possible structures
  - Training data

- **Output:** A network that maximizes the score function

- Thus, we need a **score function** (including priors, if needed) and a **search strategy** (i.e. optimization procedure) to find a structure in the hypothesis space
Structural search

- Hypothesis space
  - General graphs: Number of graphs on $M$ nodes: $O(2^{M^2})$
  - Trees: Number of trees on $M$ nodes $O(M!)$

- Combinatorial optimization problem
  - Many heuristics to solve this problem but no guarantee on attaining optimality
Find \( \langle \mathcal{G}, \theta_{\mathcal{G}} \rangle \) that maximize the likelihood:

\[
\max_{\mathcal{G}, \theta_{\mathcal{G}}} P(\mathcal{D}|\mathcal{G}, \theta_{\mathcal{G}}) = \max_{\mathcal{G}} \max_{\theta_{\mathcal{G}}} P(\mathcal{D}|\mathcal{G}, \theta_{\mathcal{G}})
\]

\[
= \max_{\mathcal{G}} P(\mathcal{D}|\mathcal{G}, \hat{\theta}_{\mathcal{G}})
\]

Likelihood Score

\[
\text{Score}_L(\mathcal{G}, \mathcal{D}) = P(\mathcal{D}|\mathcal{G}, \hat{\theta}_{\mathcal{G}})
\]
Likelihood score: Information-theoretic interpretation

- Given structure, log likelihood of data:

\[
\ln P(\mathcal{D}|\theta_{\mathcal{G}}, \mathcal{G}) = \ln \prod_{n=1}^{N} P(x^{(n)}|\theta_{\mathcal{G}}, \mathcal{G}) = \sum_{n=1}^{N} \ln \prod_{i} P(x_{i}^{(n)}|P_{aX_{i}}, \theta_{X_{i}P_{aX_{i}}}, \mathcal{G})
\]

\[
= \sum_{i} \sum_{n=1}^{N} \ln P(x_{i}^{(n)}|P_{aX_{i}}, \theta_{X_{i}P_{aX_{i}}}, \mathcal{G})
\]

\[
= N \sum_{i} \sum_{x_{i}\in Val(X_{i})} \sum_{u_{i}\in Val(P_{aX_{i}})} \frac{\text{Count}(x_{i}, u_{i})}{N} \ln P(x_{i}|u_{i}, \theta_{X_{i}P_{aX_{i}}}, \mathcal{G})
\]

\[
\ln P(\mathcal{D}|\hat{\theta}_{\mathcal{G}}, \mathcal{G}) = N \sum_{i} \sum_{x_{i}\in Val(X_{i})} \sum_{u_{i}\in Val(P_{aX_{i}})} \hat{P}(x_{i}, u_{i}) \ln \left( \hat{P}(x_{i}|u_{i}) \frac{\hat{P}(x_{i})}{\hat{P}(x_{i})} \right)
\]

\[
= N \sum_{i} \left( I_{\hat{P}}(X_{i}; P_{aX_{i}}) - H_{\hat{P}}(X_{i}) \right)
\]

This score measures the strength of the dependencies between variables and their parents.
Limitations of likelihood score

- We have $I_P(X; Y) \geq 0$ and $I_P(X; Y) = 0$ iff $P(X, Y) = P(X)P(Y)$
- In the training data, almost always $I_\hat{P}(X; Y) > 0$
  - Due to statistical noise in the training data, exact independence almost never occurs in the empirical distribution

- Thus, ML score never prefers the simpler network
  - Score is maximized for fully connected network
  - Therefore likelihood score fails to generalize well

\[
I(X; Y \cup Z) \geq I(X; Y) \\
I(X; Y \cup Z) = I(X; Y) \text{ iff } X \perp Z | Y
\]
Avoiding overfitting

- Restricting the hypothesis space explicitly:
  - restrict number of parents or number of parameters

- Impose penalty on the complexity of the structure:
  - Explicitly
  - Bayesian score
Tree structure: Chow-Liu algorithm

- Compute $w_{ij} = I(X_i, X_j)$ for each pair of nodes $X_i$ and $X_j$

- Find a maximum weight spanning tree (MWST)
  - tree with the greatest total weight
    
    $\max_T \sum_{(i,j) \in T} w_{ij}$

- Give directions to edges in MST
  - pick an arbitrary node as the root and draw arrows going away from it (e.g., using a BFS algorithm)

we can find an optimal tree in polynomial time
BIC score

\[ \text{Score}_{\text{BIC}}(\mathcal{G}) = \log P(\mathcal{D}|\widehat{\theta}_\mathcal{G}) - \frac{\log N}{2} \text{Dim}[\mathcal{G}] \]

\[ = N \sum_i I_\hat{\rho}(X_i; P a_{X_i}) - N \sum_i H_\hat{\rho}(X_i) - \frac{\log N}{2} \text{Dim}[\mathcal{G}] \]

- Its negation is also known as Minimum Description Length (MDL)
- For the larger \( N \), the more emphasis is given to the fit to data
  - Mutual information grows linearly with \( N \) while complexity grows logarithmically with \( N \)
Asymptotic consistency of BIC score

Definition: A scoring function is called consistent if as \( N \to \infty \) with probability \( \to 1 \):
- \( G^* \) (or any I-equivalent structure) maximizes the score
- All non-I-equivalent structures have strictly lower score

Theorem: As \( N \to \infty \), the structure that maximizes the BIC score is the true structure \( G^* \) (or any I-equivalent structure)
- Asymptotically, spurious dependencies will not contribute to likelihood and will be penalized
- Required dependencies will be added due to linear growth of likelihood term compared to logarithmic growth of model complexity
Bayesian score

- Uncertainty over parameters and structure
  - Try to average over all possible parameter values and also consider the structure prior

- Bayesian score:
  \[ \text{Score}_B(G) = \log P(\mathcal{D}|G) + \log P(G) \]

- We consider the whole distribution over parameters \( \theta_G \) for the structure \( G \) to find the marginal likelihood:
  \[ P(\mathcal{D}|G) = \int_{\theta_G} P(\mathcal{D}|G, \theta_G)P(\theta_G|G)d\theta_G \]
Marginal likelihood

\[ P(\mathcal{D}|\mathcal{G}) = P(x^{(1)}, ..., x^{(N)}|\mathcal{G}) = P(x^{(1)}|\mathcal{G}) \times P(x^{(2)}|x^{(1)}, \mathcal{G}) \times \cdots \times P(x^{(N)}|x^{(1)}, ..., x^{(N-1)}, \mathcal{G}) \]

- Dirichlet priors and single variable:

\[ P(x^{(1)}, ..., x^{(N)}|\mathcal{G}) = \frac{\prod_{k=1}^{K} \alpha_k (\alpha_k + 1) \cdots (\alpha_k + M[k] - 1)}{\alpha (\alpha + 1) \cdots (\alpha + N - 1)} \]

\[ = \frac{\Gamma(\alpha)}{\Gamma(\alpha + N)} \times \prod_{k=1}^{K} \frac{\Gamma(\alpha_k + M[k])}{\Gamma(\alpha_k)} \]

\[ \alpha (\alpha + 1) \cdots (\alpha + N - 1) = \frac{\Gamma(\alpha + N)}{\Gamma(\alpha)} \]

\[ M[k] = \sum_{n=1}^{N} I(x^{(n)} = k) \]

\[ \alpha = \sum_{k=1}^{K} \alpha_k \]
Marginal likelihood in Bayesian networks: global & local parameter independence

- Let $P(\theta_G|G)$ satisfy global parameter independence assumptions:
  
  $$P(D|G) = \prod_i \int_{\theta_{X_i}|Pa_{X_i}} \prod_{n=1}^N P(x_i^{(n)}|P_{a_{X_i}^G}^{(n)}, \theta_{X_i|Pa_{X_i}^G}, G) P(\theta_{X_i|Pa_{X_i}^G}|G) d\theta_{Pa_{X_i}^G}$$

- Let $P(\theta_G|G)$ satisfy global & local parameter independence assumptions:

  $$P(D|G) = \prod_i \prod_{u_i \in Val(Pa_{X_i}^G)} \int_{\theta_{X_i|u_i}} \prod_{n=1}^N P(x_i^{(n)}|u_i, \theta_{X_i|u_i}, G) P(\theta_{X_i|u_i}|G) d\theta_{X_i|u_i}$$

- If Dirichlet priors $P(\theta_{X_i|u_i}|G)$ has hyperparameters $\{\alpha_{x_i|u_i}^G | x \in Val(X_i)\}$:

  $$P(D|G) = \prod_i \prod_{u_i \in Val(Pa_{X_i}^G)} \frac{\Gamma \left( \sum_{x_i \in Val(X_i)} \alpha_{x_i|u_i}^G \right)}{\Gamma \left( \sum_{x_i \in Val(X_i)} (\alpha_{x_i|u_i}^G + M_i[x_i,u_i]) \right)} \prod_{x_i \in Val(X_i)} \frac{\Gamma \left( \alpha_{x_i|u_i}^G + M_i[x_i,u_i] \right)}{\Gamma \left( \alpha_{x_i|u_i}^G \right)}$$
Priors

- **Structure prior:** $P(\mathcal{G})$
  - Uniform prior: $P(\mathcal{G}) \propto \text{constant}$
  - Prior penalizing number of edges: $P(\mathcal{G}) \propto c^{\lvert E(\mathcal{G}) \rvert} \ (0 < c < 1)$
  - Prior penalizing number of parameters

- **Parameter priors:** $P(\theta|\mathcal{G})$ is usually instantiated as **BDe prior**
  \[
  \alpha \left( X_i, Pa_{X_i}^G \right) = \alpha P' \left( X_i, Pa_{X_i}^G \right)
  \]
  $\alpha$: equivalent sample size
  $B_0$: prior network representing $P'$
  $Pa_{X_i}^G$ are not the same as parents of $X_i$ in $B_0$

- **BDe prior**
  - A single network $B_0$ provides priors for all candidate networks
  - Unique prior with the property that l-equivalent networks have the same Bayesian score
Asymptotic behavior of Bayesian score

- As $N \to \infty$, a network $\mathcal{G}$ with Dirichlet priors satisfies:

$$\log P(\mathcal{D}|\mathcal{G}) = \log P(\mathcal{D}|\theta_{\mathcal{G}}) - \frac{\log N}{2} \text{Dim}[\mathcal{G}] + O(1)$$

- Thus, Bayesian score is asymptotically equivalent to BIC and asymptotically consistent

$O(1)$ term does not grow with $M$
General graph structure

- **Theorem:**
  - The problem of learning a BN structure with at most $d \ (d \geq 2)$ parents is NP-hard.

- Search in the space of graph structures using local search algorithms
  - Algorithms like hill-climbing and simulated Annealing
  - Exploit score decomposition during the search to alleviate computational overhead
Optimization by local searches

- Graphs as states
- Score function is used as the objective function
- Neighbors of a state are found using these modifications:
  - Edge addition
  - Edge deletion
  - Edge reversal

We only consider legal neighbors that result in DAGs
Decomposable score

- Decomposable structure score:

\[
\text{Score}(\mathcal{G}; \mathcal{D}) = \sum_i \text{FamScore}(X_i|Pa_{X_i}^G; \mathcal{D})
\]

\(\text{FamScore}(X|U; \mathcal{D})\) measures how well the set of variables \(U\) serves as parent of \(X\) in the data set \(\mathcal{D}\).

- Decomposability is useful to reduce the computational overhead of evaluating different structures during search.
  - if we have a decomposable score, then a local change in the structure does not change the score of the remaining parts of the structure.
Decomposability of scores

- Likelihood score is decomposable

- Bayesian score is decomposable when:
  - parameter priors satisfy **global parameter independence** and **parameter modularity**
    - Parameter modularity:
      $$Pa_{X_i}^G = Pa_{X_i}^{G'} = U \Rightarrow P(\theta_{X_i|U}|G) = P(\theta_{X_i|U}|G')$$
  - and structure prior satisfies **structure modularity**
    $$P(G) \propto \prod_i P(Pa_{X_i} = Pa_{X_i}^G)$$
Summary

- Score functions
  - Likelihood score
  - BIC score
  - Bayesian score

- Optimal structure for trees is found using a greedy algorithm

- When hypothesis space includes general graphs, the model selection problem will be NP-hard and we use local search strategies to optimize the structure
Reference

- Koller and Friedman, Chapter 18
  - 18.1-18.3, 18.4.1, 18.4.3