Initialization Heuristics for Greedy Bayesian Network Structure Learning

Walter Perez Urcia Denis Deratani Mauá

Universidade de São Paulo
Instituto de Matemática e Estadística
Departamento de Ciências da Computação

KDMiLE 2015

Contents

Introduction

- Variables $X_1, \ldots, X_n$ takes values in $\Omega_1, \ldots, \Omega_n$
  - $X_1 : Gender, \Omega_1 = \{Male, Female\}$
  - $X_2 : City\ size, \Omega_2 = \{Big, Small\}$
- Factored possibility space $\Omega = \Omega_1 \times \ldots \times \Omega_n$
- Event is a subset of Ω
- Probability function maps events α and β into real values such that
  - $0 \leq \mathbb{P}(\alpha) \leq 1$
  - $\mathbb{P}(\Omega) = 1$
  - $\mathbb{P}(\alpha \cup \beta) = \mathbb{P}(\alpha) + \mathbb{P}(\beta) - \mathbb{P}(\alpha \cap \beta)$

- Every assignment of value to a variable correspond to an event:

$$Gender = M \leftrightarrow \alpha = \{(M, s), (M, b)\}$$

- The probability distribution of a variable X maps assignments of the variable to the respective probabilities:

$$\mathbb{P}(X = x) = \mathbb{P}(\{\omega : \omega \text{ consistent with } x\})$$

- We denote the probability distribution of X as $\mathbb{P}(X)$ and the probability of an arbitrary event $\{X = x\}$ as $\mathbb{P}(x)$
- It follows from the properties of probability function that

$$\sum_X \mathbb{P}(X) = \sum_{x \in \Omega_X} \mathbb{P}(X = x) = 1$$

- A joint assignment to a set of variables is an event

$$\textit{Gender} = M \text{ and } \textit{City size} = b \leftrightarrow \alpha = \{(M, b)\}$$

- The joint probability distribution of a set of variables is a function that maps joint assignments to their event probabilities:

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(\{\omega : \omega \text{ consistent with } x, y\})$$

- We denote the probability distribution of X and Y as $\mathbb{P}(X)$ and the probability of an arbitrary joint event as $\mathbb{P}(x, y)$
- It follows from the properties of probability function that

$$\sum_{X,Y} \mathbb{P}(X, Y) = \sum_{x \in \Omega_X} \sum_{x \in \Omega_Y} \mathbb{P}(X = x, Y = y) = 1$$

- Maps assignments of two variables to conditional probabilities:

$$\mathbb{P}(X = x \mid Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)}$$

- Represented as $\mathbb{P}(X \mid Y)$
- Analogously, we can define join conditional probability distribution $\mathbb{P}(X, Y \mid Z, W)$

By definition of conditional probability:

$$\mathbb{P}(\alpha \mid \beta)\mathbb{P}(\beta) = \mathbb{P}(\alpha \cap \beta)$$

For events $\alpha_1, \ldots, \alpha_n$ it follows that

$$\mathbb{P}(\alpha_1 \cap \ldots \cap \alpha_n) = \mathbb{P}(\alpha_1) \prod_{i=2}^{n} \mathbb{P}(\alpha_i \mid \alpha_1 \cap \ldots \cap \alpha_{i-1})$$

In terms of variables:

$$\mathbb{P}(A, B, C) = \mathbb{P}(A)\mathbb{P}(B \mid A)\mathbb{P}(C \mid B, A)$$

$$\mathbb{P}(\beta \mid \alpha) = \frac{\mathbb{P}(\alpha \mid \beta)}{\mathbb{P}(\alpha)}\mathbb{P}(\beta)$$

- Prior probability: $\mathbb{P}(\beta)$
- Posterior probability: $\mathbb{P}(\beta \mid \alpha)$
- Data Likelihood: $\mathbb{P}(\alpha \mid \beta)$
- Evidence probability: $\mathbb{P}(\alpha)$

- Bayes' rule can be seen as a way of revising beliefs in light of new information/knowledge: start with $\mathbb{P}(\beta)$, observe α then set $\mathbb{P}(\beta)' = \mathbb{P}(\beta \mid \alpha)$
- This way of thinking is known as Bayesian Reasoning

Events α and β are independent if:

$$\mathbb{P}(\alpha \cap \beta) = \mathbb{P}(\alpha)\mathbb{P}(\beta)$$

- The following are equivalent definitions:
  - Either $\mathbb{P}(\alpha \mid \beta) = \mathbb{P}(\alpha)$ or $\mathbb{P}(\beta) = 0$
  - Either $\mathbb{P}(\beta \mid \alpha) = \mathbb{P}(\beta)$ or $\mathbb{P}(\alpha) = 0$
- Knowing β is irrelevant to determining the value of α
- Knowing α is irrelevant to determining the value of β

Variables A and B are independent if:

$$\mathbb{P}(A = a, B = b) = \mathbb{P}(A = a)\mathbb{P}(B = b)$$

for all values of a and b.
Another way to write this is:

$$\mathbb{P}(A, B) = \mathbb{P}(A)\mathbb{P}(B)$$

Events α and β are independent conditional on event γ if:

$$\mathbb{P}(\alpha \cap \beta \mid \gamma) = \mathbb{P}(\alpha \mid \gamma)\mathbb{P}(\beta \mid \gamma)$$

The following are equivalent definitions:

- Either $\mathbb{P}(\alpha \mid \beta, \gamma) = \mathbb{P}(\alpha \mid \gamma)$ or $\mathbb{P}(\beta \mid \gamma) = 0$
- Either $\mathbb{P}(\beta \mid \alpha, \gamma) = \mathbb{P}(\beta \mid \gamma)$ or $\mathbb{P}(\alpha \mid \gamma) = 0$

Analogously, variables A and B are conditionally independent given C if

$$\mathbb{P}(A, B \mid C) = \mathbb{P}(A \mid C)\mathbb{P}(B \mid C)$$

for every assignment to A, B and C

The presence of independences reduces the number of probability values to specify:

- No independences: $\mathbb{P}(A, B, C)$, k^3 values
- A, B and C are dependent, and A and B are conditionally independent given C: $\mathbb{P}(A, B \mid C) = \mathbb{P}(A \mid C)\mathbb{P}(B \mid C)$, $k + 2k^2$ values
- A, B and C are independent: $\mathbb{P}(A, B, C) = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$, $3k$ values

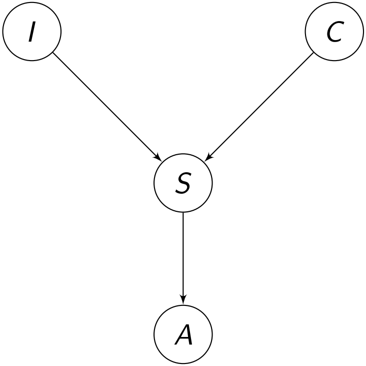
Bayesian Network

Markov property

Given its parents, every variable is conditionally independent from its non-descendants non-parents

Factorization property

$$\mathbb{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \mid Pa(X_i))$$



The directed acyclic graph (DAG) above has joint probability distribution:

$$\mathbb{P}(I, C, S, A) = \mathbb{P}(I)\mathbb{P}(C \mid I)\mathbb{P}(S \mid C, I)\mathbb{P}(A \mid S, C, I)$$

$$= \mathbb{P}(I)\mathbb{P}(C)\mathbb{P}(S \mid C, I)\mathbb{P}(A \mid S)$$

A Bayesian Network consists of

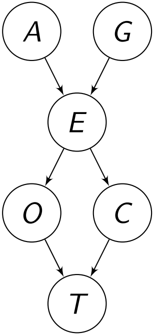
- A DAG G over a set of variables $X_1, \ldots, X_n$
- Probability constraints: $\mathbb{P}(X_i = k \mid Pa(X_i) = j) = \theta_{ijk}$

Joint Probability Distribution

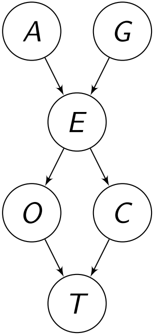
There is a unique probability function consistent with a BN:

$$\mathbb{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \mid Pa(X_i)) = \prod_{i=1}^{n} \theta_{ijk}$$

- Age (A): young, adult, old
- Gender (G): male, female
- Education (E): primary, high school, university
- Occupation (O): employee, self-employed
- City size (C): big, small
- Transport (T): private (car), public (bus, train, etc)

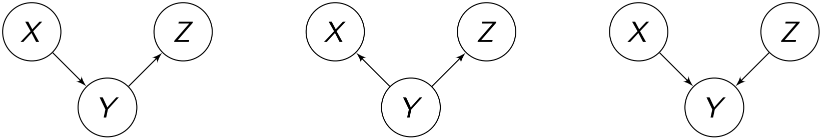


- Education rates have been increasing over years; young people are more likely to have university degrees than old people
- Women are more likely to invest in their education than men; women outnumber men in the vast majority of university-level courses
- High education levels is key to getting prestigious professions; jobs requiring university degrees are more easily available in big cities
- Preferred means of transport depends on occupation and city size



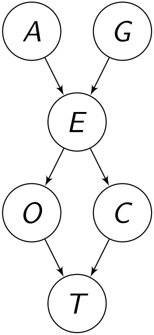
$$\mathbb{P}(A = young) = 0.3$$
$$\mathbb{P}(A = adult) = 0.5$$
$$\mathbb{P}(A = old) = 0.2$$
$$\mathbb{P}(E = high \mid A = young, G = F) = 0.7$$
$$\vdots$$
$$\mathbb{P}(C = small \mid E = high) = 0.25$$
$$\vdots$$

An arc $X \rightarrow Y$ can be interpreted as "X causes Y"



causal chain   common cause   common effect

Defining and verifying causality is difficult and controversial: We can loosely define X causes Y if X temporarily precedes and direct influences Y



We can query a Bayesian Network about unspecified probabilities

- Are women more likely to prefer public transport over men: $\mathbb{P}(T = public \mid G = F) > \mathbb{P}(T = public \mid G)$?
- What is the distribution of ages for people who use private means of transport: $\mathbb{P}(A \mid T = private)$?

Learning BN

Constraint-based approaches

Perform multiple conditional independence hypothesis testing in order to build a DAG

Score-based approaches

Associate every DAG with a polynomial-time computable score value and search for structure with high score values

Learning as optimization

Given dataset D, select G that maximizes decomposable score function $sc(G, D)$

- Score $sc(G, D)$ is usually a mix of data fitness F and model complexity P:

$$sc(G, D) = F(G) + \psi(N) \times P(G)$$

  with $\psi(N) \geq 0$ is a function of data size $|D| = N$
- We usually ommit dependence on D: $sc(G)$

Learning as optimization

Select G that maximizes decomposable score function

$$G^* = arg \max_{G:G \text{ is a DAG}} sc(G)$$

$$G^* = arg \max_{G} \sum_{i} sc(X_i, Pa(X_i))$$

Greedy Search is a popular approach to find an approximate solution. It relies on the definition of a neighborhood space among solutions and on local moves that search for improving solution in the neighborhood of an incumbent solution

```
GreedySearch( Dataset D ) : return a BN G
  G = Initial_Solution(X_1, ..., X_n)
  For a number of iterations K
    best_neighbor = find_best_neighbor(G)
    if score(best_neighbor) > score(G) then
      G = best_neighbor
  Return G
```

Different neighborhoods and local moves rise to different methods such as:

- Structure-based
- Equivalence-based
- Order-based

Based on the observation that the problem of learning a Bayesian network can be written as

$$G^* = \arg\max_{<} \max_{G \text{ consistent with } <} \sum_{i=1}^{n} sc(X_i, Pa(X_i))$$

$$G^* = \arg\max_{<} \sum_{i=1}^{n} \max_{P \subseteq \{X_j < X_i\}} sc(X_i, P)$$

which means that if an optimal ordering over the variables is known, an optimal DAG can be found by maximizing the local scores independently

```
OrderBasedGreedySearch( Dataset D ) : return a BN
  L = Get_Order(X_1, ..., X_n)
  For a number of iterations K
    current_sol = L
    For each i = 1 to n-1 do
      L_i = swap(L, i, i+1)
      if score(L_i) > score(current_sol)
        current_sol = L_i
    if score(current_sol) > score(L) then
      L = current_sol
  Return network(L)
```

where $swap(L, i, i+1)$ swaps the values $L[i]$ and $L[i+1]$

Usually *Get_Order* generates a random order, which means that it has $n!$ possibilities.

Initializing Heuristics

Conclusions

We....

Future Work

We....

Thanks!