Learning probabilistic Graphical Models

Péter Antal

Department of Measurement and Information Systems

Intelligent Data Analysis, November 25, 2016

Topics

- Assumptions:
  - Stability
  - Causal Markov Assumption
- Parameter learning
  - Complete data: ML/MAP
- Asymptotic learning
  - Hardness of learning: NP
  - IC
- Score-based methods
  - An information theoretic approach
  - Score equivalence

Stable distributions

Definition

The distribution P is said to stable (or faithfull), if there exists a DAG called perfect map exactly representing its (in)dependencies (i.e. $(X \perp\!\!\!\perp Y|Z)_G \Leftrightarrow (X \perp\!\!\!\perp Y|Z)_P \; \forall X, Y, Z \subseteq V$). The distribution P is stable w.r.t. a DAG G, if G perfectly represents its (in)dependencies.

Numerically encoded independencies cannot be represented structurally, i.e. by d-separation, thus cannot be learned with standard BN representation.

1. Consider $p(X, Y, Z)$ with binary X, Z and ternary Y. The conditionals $p(Y|X)$ and $p(Z|Y)$ can be selected such that $p(z|x) = p(z|\neg x)$. That is $(X \not\perp\!\!\!\perp Y)$ and $(Y \not\perp\!\!\!\perp Z)$, but $(X \perp\!\!\!\perp Z)$, demonstrating that the "naturally" expected transitivity of dependency can be destroyed numerically.
2. Consider $P(X, Y, Z)$ with binary variables, where $p(x) = p(y) = 0.5$ and $p(Z|X, Y) = 1(Z = XOR(X, Y))$. That is $(X \perp\!\!\!\perp Z)$ and $(Y \perp\!\!\!\perp Z)$, but $(\{X, Y\} \not\perp\!\!\!\perp Z)$, demonstrating that pairwise independence does not imply total independence.

The Causal Markov Condition

Definition

A DAG G is called a causal structure over variables V, if each node represents a variable and edges denote direct influences. A causal model is a causal structure extended with local models $p(X_i|pa(X_i, G))$ for each node describing the dependency of variable X_i on its parents $pa(X_i, G)$. As the conditionals are frequently from a parametric family, they are parameterized by $\boldsymbol{\theta_i}$, and $\boldsymbol{\theta}$ denotes the overall parameterezation, so a causal model is pair $(G, \boldsymbol{\theta}$.

Definition

A causal structure G and distribution P satisfies the Causal Markov Condition, if P obeys the local Markov condition w.r.t. G.

Note: Reichenbach's "common cause principle", i.e. hidden variables are allowed, only variables that influences two or more variables in V are necessary for causal sufficiency.
(The causal Markov condition implies sufficiency and stability implies necessity of G).

Constraint-based BN learning: IC

The Inductive Causation algorithm (assuming a stable distribution P):

1. *Skeleton:* Construct an undirected graph (skeleton), such that variables $X, Y \in \boldsymbol{V}$ are connected with an edge iff $\forall S (X \perp\!\!\!\perp Y | S)_P$, where $S \subseteq \boldsymbol{V} \setminus \{X, Y\}$.
2. *v-structures:* Orient $X \rightarrow Z \leftarrow Y$ iff X, Y are nonadjacent, Z is a common neighbour and $\neg \exists S$ that $(X \perp\!\!\!\perp Y | S)_P$, where $S \subseteq \boldsymbol{V} \setminus \{X, Y\}$ and $Z \in S$.
3. *propagation:* Orient undirected edges without creating new v-structures and directed cycle.

Theorem

The following four rules are necessary and sufficient.

R_1 *if* $(a \not\rightarrow c) \wedge (a \rightarrow b) \wedge (b - c)$, *then* $b \rightarrow c$

R_2 *if* $(a \rightarrow c \rightarrow b) \wedge (a - b)$, *then* $a \rightarrow b$

R_3 *if* $(a - b) \wedge (a - c \rightarrow b) \wedge (a - d \rightarrow b) \wedge (c \not\rightarrow d)$, *then* $a \rightarrow b$

R_4 *if* $(a - b) \wedge (a - c \rightarrow d) \wedge (c \rightarrow d \rightarrow b) \wedge (c \not\rightarrow b) \wedge (a - d)$, *then* $a \rightarrow b$

The complexity of BN learning

The NP-hardness of finding a Bayesian network for the observations (as minimal representation of the observed independencies, which is I-map).

Theorem

Let $\boldsymbol{V}$ be a set of variables with joint distribution $p(\boldsymbol{V})$. Assume that an oracle is available that reveals in $\mathcal{O}(1)$ time whether an independence statement holds in p. Let $0 < k \leq |\boldsymbol{V}|$ and $s = \frac{1}{2}n(n-1) - \frac{1}{2}k(k-1)$. Then, the problem of deciding whether or not there is a (non-minimal) Bayesian network that represents p with less or equal to s edges by consulting the oracle is NP-complete.

The NP-hardness of finding a best scoring Bayesian network (i.e. the NP-hardness of optimization over DAGs).

Theorem

Let $\boldsymbol{V}$ be a set of variables, D_N is a complete data set, $S(G, D_N)$ is a score function and a real value c. Then, the problem of deciding whether or not there exist a Bayesian network structure G_0 defined over the variables $\boldsymbol{V}$, where each node in G_0 has at most $1 < k$ parents, such that $p \leq S(G_0, D_N)$ is NP-complete.

Learning tree Bayesian networks: goal

Approximate the target distribution P with a tree-dependent distribution P^t using the Kullback-Leibler divergence (relative/cross-entropy measure).

Definition

For two discrete probability distributions P and Q with probabilities p_i and q_i the Kullback-Leibler divergence is

$$D_{KL}(P\|Q) = \mathrm{KL}(P\|Q) = \sum_i p_i \log(p_i/q_i) \tag{1}$$

Lemma

The KL divergence is nonnegative:

$$-\mathrm{KL}(P||Q) = \sum_i p_i \log(q_i/p_i) \leq \sum_i p_i((q_i/p_i) - 1) = 0 \tag{2}$$

using $\log(x) \leq x - 1$. *It is 0, iff P and Q are identical.*

Entropy, mutual information, KL divergence

⇒The KL divergence is not symmetric and it does not satisfy the triangle inequality, thus it is not a distance.
⇒The KL divergence dominates the L_1 distance, $L_1(P,Q) = \sum_i |p_i - q_i|$, and the L_2 distance, $L_2(P,Q) = (\sum_i (p_i - q_i)^2)^{1/2}$.
⇒The mutual information of X and Y with $P(X,Y)$ can be written as

$$I(X,Y) = \sum_{x,y} P(x,y)[\log \frac{P(x,y)}{P(x)P(y)} = KL(P(X,Y)||P(X)P(Y)), \tag{3}$$

which is 0, iff X and Y are independent.
⇒The joint entropy of X and Y with $P(X,Y)$ can be written as

$$H(X,Y) = H(X|Y) + I(X,Y) + H(Y|X), \tag{4}$$

where $H(Y|X)$ is the conditional entropy defined as

$$H(Y|X) = \sum_x P(x)H(Y|X=x) = \sum_x P(x) \sum_y P(y) \log P(y). \tag{5}$$

Learning tree Bayesian networks: parameter learning

If Q is a distribution defined by a tree Bayesian network t in learning P, then

$$\begin{aligned}
\mathrm{KL}(P\|Q) &= -\sum_x P(x)\sum_{i=1}^{n}\log Q(x_i|x_{j(i)}) + \sum_x P(x)\log P(x) \\
&= -\sum_{i=1}^{n}\sum_{x_i,x_{j(i)}} P(x_i,x_{j(i)})\log Q(x_i|x_{j(i)}) - H(X) \\
&= -\sum_{i=1}^{n}\sum_{x_{j(i)}} P(x_{j(i)})\sum_{x_i} P(x_i|x_{j(i)})\log Q(x_i|x_{j(i)}) - H(X)
\end{aligned}$$

which is maximal if $Q(x_i|x_{j(i)}) = P(x_i|x_{j(i)})$ for all $x_{j(i)}$.

Learning tree Bayesian networks: structure learning

Using the optimal parametrization in a tree Bayesian network t in learning P, we have

$$\begin{aligned} \mathrm{KL}(P\|Q) &= -\sum_{i=1}^{n}\sum_{x_i,x_{j(i)}} P(x_i,x_{j(i)})[\log\frac{P(x_i,x_{j(i)})}{P(x_i)P(x_{j(i)})} + \log P(x_i)] - H(X) \\ &= -\sum_{i=1}^{n} I(X_i,X_{j(i)}) + \sum_{i=1}^{n}\sum_{x_i} P(x_i)\log P(x_i) - H(X) \end{aligned}$$

which is maximized (optimal) if the tree t is a maximum weight spanning tree with weights $I(X_i,X_{j(i)})$ (mutual information).

Corollary

If the P target distribution is tree-based (tree-dependent), then the projected distribution in an optimal tree will be identical.

Learning tree Bayesian networks: pseudocode

Either using data or a prior knowledge base:

1. Compute $P(x_i, x_j)$ for all pairs of values.
2. Compute $I(X_i, X_j)$ for all pairs of variables.
3. Select largest branch and add it to the tree unless create a loop, otherwise discard it.
4. Repeat until $n - 1$ edges (or I() drops below a threshold$\Rightarrow$ forest....)

Chow&Liu (1968): Maximum Weight Spanning Tree (MWST) learning, Pearl(1988).

The ML learning: Optimality of relative frequencies

Theorem

Relative frequency is a ML estimator in multinomial sampling. Assume $i = 1, \ldots K$ outcomes assuming multinomial sampling with parameters $\theta = \{\theta_i\}$ and observed occurrencies $n = \{n_i\}$ ($N = \sum_i n_i$). Then

$$\log \frac{p(n|\hat{\theta})}{p(n|\theta)} = \log \frac{\prod_i (\hat{\theta}_i)^{n_i}}{\prod_i (\theta_i)^{n_i}} = \sum_i n_i \log \frac{\hat{\theta}_i}{\theta_i} = N \sum_i \hat{\theta}_i \log \frac{\hat{\theta}_i}{\theta_i} > 0.$$

where the last quantity is the KL divergence, which is always positive.

The ML learning I.

Using the optimal parameter selection of $\theta^*_{ijk} = N_{ijk}/N_{ij+}$ in structure G, where N_{ijk} are the occurrences of value x_k and parental configuration q_j for variable X_i and its parental set $Pa(X_i)$ (N_{ij+} is the appropriate sum), we get for the likelihood of structure G,

$$ML(G; D_N) = p(D_N|G, \boldsymbol{\theta}^*) = \prod_{l=1}^{N} \prod_{i=1}^{n} p(x_i^{(l)}|pa_i^{(l)}) \tag{6}$$

$$= \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \frac{N_{ijk}}{N_{ij+}}^{N_{ijk}} \tag{7}$$

by taking logarithm, rearranging and expanding with N

$$\log(ML(G; D_N)) = N \sum_{i=1}^{n} \sum_{j=1}^{q_i} \frac{N_{ij+}}{N} \sum_{k=1}^{r_i} \frac{N_{ijk}}{N_{ij+}} \log(N_{ijk}/N_{ij+}) \tag{8}$$

The ML Learning II

Using conditional entropy $H(Y|X) = \sum_x p(x) \sum_y p(y|x) \log(p(y|x))$, the chain rule $H(X,Y) = H(Y|X) + H(X)$ and the definition of mutual information $I(Y;X) = H(Y) - H(Y|X)$, it can be rewritten as

$$\log(ML(G;D_N)) = -N \sum_{i=1}^{n} H(X_i|Pa(X_i,G)) \quad (9)$$

$$= N \sum_{i=1}^{n} I(X_i;Pa(X_i,G)) - N \sum_{i=1}^{n} H(X_i) \quad (10)$$

$$(11)$$

This shows that the maximization of the ML score is equivalent with finding a BN parameterized with the observed frequencies that has minimum entropy or that we are finding a BN parameterized with the observed frequencies that has maximum mutual information between its children and their parents (10,

Complexity regularization

Because of the monotonicity of mutual information — if $Pa(X_i) \subset Pa(X_i)'$, then $I(X_i; Pa(X_i)) \leq I(X_i; Pa'(X_i))$ — so the complete network maximizes the maximum likelihood score. However score functions such as the MDL-score derived from the minimum description length (MDL) principle or the Bayesian information criterion (BIC)-score derived with a non-informative Bayesian approach contains various complexity penalty terms. We shall use only the BIC-score defined as follows

$$BIC(G; D_N) = \log(ML(G; D_N)) - 1/2 dim(G) \log(N) \tag{12}$$

where $dim(G)$ denotes the number of free parameters.

Score equivalence

Definition

A score function $S(G; D_N)$ is called score equivalent, if for each pair of observationally equivalent Bayesian network structure G_1, G_2 the scores are equal $S(G_1; D_N) = S(G_2; D_N)$ for all D_N.

Theorem

The $BIC(G; D_N)$ scoring metric is score equivalent.

The score equivalence of the *BIC* score is the direct consequence of the result that the number of free parameters (that is the term $dim(G)$) are equal in observationally equivalent Bayesian networks.

Asymptotic consistency

Theorem

Let $\boldsymbol{V}$ be a set of variables. Let the prior distribution $p(G)$ over Bayesian network structures be positive. Let $p(\boldsymbol{V})$ be a positive and stable distribution and G_0 is a corresponding perfect map (i.e. a Bayesian network representing exactly all the independencies in $p(\boldsymbol{V})$, see Def. **??***). Now, let D_N is an i.i.d. data set generated from $p(\boldsymbol{V})$. Then, for any network structure G over $\boldsymbol{V}$ that is not a perfect map of $p(\boldsymbol{V})$ we have that*

$$\lim_{N\to\infty} BD_e(G_0; D_N) - BD_e(G; D_N) = -\infty \text{ and also} \quad (13)$$

$$\lim_{N\to\infty} BIC_e(G_0; D_N) - BIC_e(G; D_N) = -\infty \quad (14)$$

Rate of convergence

Furthermore, a rate of convergence result is also derived and a corresponding sample complexity $N(\epsilon, \delta)$ to select an appropriate sample size for a given accuracy between the target distribution p_0 and the distribution p_{BN} represented by the learned Bayesian network with a given confidence

$$p(D_N : KL(p_0|p_{BN}) > \epsilon) < \delta \tag{15}$$

Thank you for your attention!

Questions?