Tutorial on Learning Bayesian Networks for Relational Data: Definitions

Oliver Schulte and Ted Kirkpatrick

School of Computing Science
Simon Fraser University
Vancouver-Burnaby, Canada

September 22, 2016

Abstract

The mathematical concepts that the tutorial introduces and illustrates.

1 General Notation

We use boldface to denote sets and lists of objects; for instance, $\{a_1, a_2, \ldots, a_n\} \equiv \boldsymbol{a}$. The notation $|S|$ denotes the cardinality of a set S. Fix a set of random variables $\boldsymbol{X} = \{X_1, \ldots, X_n\}$. The notation $P(X_i = x) \equiv P(x)$ denotes the probability of random variable X_i taking on value x. We also use the set notation $P(\boldsymbol{X} = \boldsymbol{x}) \equiv P(\boldsymbol{x})$ to denote the joint probability that each random variable X_i takes on value $\boldsymbol{x}_i$.

2 Possible Worlds

A **world** is a triple $\langle \mathcal{I}, \mathcal{V}, \mathcal{F} \rangle$ where

1. $\mathcal{I}$ is a set of individuals,
2. $\mathcal{V}$ is a set of values, including $\mathrm{T}, \mathrm{F}, na$ and
3. $\mathcal{F}$ is a set of functions, called **functors**, that map one or more individuals to a value.

For simplicity we consider worlds with a finite number of functors and values only, but that is not essential for the results presented. Individuals are denoted by lower case constants. A functor

$$f : \mathcal{I}^k \rightarrow \mathcal{V}$$

with k arguments has **arity** k. A functor that returns a Boolean value $\{\mathrm{T}, \mathrm{F}\}$ is called a **predicate**, usually written with uppercase letters like P, R. Other functors are usually writen with lowercase letters.

A **sample** from world $\langle \mathcal{I}, \mathcal{V}, \mathcal{F} \rangle$ is a subworld $\mathcal{D} = \langle \mathcal{I}' \subset \mathcal{I}, \mathcal{V}, \mathcal{F}|\mathcal{I}' \rangle$ that specifies the values of functors for arguments drawn from a finite subset $\mathcal{I}'$ of individuals.

2.1 Possible Worlds

A **possible world** is a world where each functor belongs to exactly one of the following groups.

Classes A set of unary predicates.

Relationships A set of predicates with arity > 1.

Class Attributes For each class, a set of functors associated with the class. A class attribute is defined only for instances of the class. Thus if f is an attribute of $Class$, then $f(a) = na$ whenever $Class(a) = \mathrm{F}$.

Relationship Attributes For each relationship, a set of functors associated with the relationship. A relationship attribute is defined only for instances of the relationship. Thus if f is an attribute of $Relation$, then $f(\boldsymbol{a}) = na$ whenever $Relation(\boldsymbol{a}) = \mathrm{F}$.

In a possible world, each individual belongs to at least one class. The set of individuals that belong to a class is given by $\{a \in \mathcal{I} : Class(a) = \mathrm{T}\}$. The term class may refer both to the functor and to the associated set of indviduals. Relational functors are typed as well, meaning that each argument position is associated with a class. If any argument is from the wrong class, the functor returns na.

3 Relational Random Variables

A relational random variable is a logical term with a probabilistic semantics. Terms in logic are analogues of random variables in probability theory because both are built out of functions: Applying a function to a set of terms produces another term. Likewise, applying a function to a set of random variables produces another random variable.

Formally, a term is (i) an individual constant, (ii) a population variable, or (iii) **a functional term** of the form $f(\tau_1, \ldots, \tau_k)$ where the type of each term τ_i matches the argument type of f. A **relational random variable** (RRV) is a functional term. [maybe use "first-order variable" until we present a probabilistic interpretation]

When discussing general statistical concepts, the special syntactic structure of an RRV is often not important. In such cases, we denote them using the traditional random variable notation like X, Y.[1] A term/random variable is **ground** if it contains only constants; otherwise it is **first-order** (FORV). If we wish to emphasize that a RRV is ground, we indicate this using notation like X^*, Y^*.

[1] Unfortunately the tradition in statistics clashes with the equally strong tradition in logic of using X, Y to denote population variables.

3.1 Instantiations and Groundings

An **instantiation** $\mathbb{A}\backslash a$ replaces a population variable by an individual instance from the same class, denoted by a constant [cite Cussens UAI paper]. An instantiation for a set of population variables specifies an instance for each. Using boldface vector notation, $\mathbb{A}\backslash \boldsymbol{a} = \{\mathbb{A}_1\backslash a_1, \ldots, \mathbb{A}_k\backslash a_k\}$ denotes an instantiation for a set of population variables.

Applying an instantiation to a term/random variable replaces every occurrence of a population variable in the term/random variable by the instance specified. We write $X_{\mathbb{A}\backslash \boldsymbol{a}}$ for applying the instantiation $\mathbb{A}\backslash \boldsymbol{a}$ to X. An instantiation may also be applied to a set of random variables. An instantiation may replace all, some, or none of the population variables in a list of random variables. If an instantiation replaces all population variables, it is a **grounding**. Since a grounding produces a ground object, we indicate this by adding a * marker to the result of applying the grounding. For example, the notation $\boldsymbol{X}^*_{\mathbb{A}\backslash \boldsymbol{a}}$ indicates that the grounding $\mathbb{A}\backslash \boldsymbol{a}$ specifies a value for each population variable in the list of random variables $\boldsymbol{X}$. Without an * marker, an instantiation may replace by constants any number of population variables, including none or all.

3.2 Logical Formulas and Joint Assignments

A probabilistic semantics assigns a probability to each logical formula. Logical formulas are built out of basic equations of the form $\tau_1 = \tau_2$. These basic statements can be combined using the standard Boolean operations (and, or, not), and extended with quantifiers (for all, exists). The statistical analogue to adding quantifiers is extending the Boolean algebra over the basic events $\tau_1 = \tau_2$ to a σ-algebra. We describe the random selection semantics for a restricted language that is sufficient to represent the queries that are usually modelled with Bayesian networks: Conjunctions of basic **assignments**. A basic assignment specifies a value for a random variable, and a conjunction of them represents a joint assignment. Using random variable notation, a joint assignment is written as $(X_1 = x_1, \ldots, X_n = x_n) \equiv \boldsymbol{X} = \boldsymbol{x}$. The notation

$$P_w(\boldsymbol{X} = \boldsymbol{x}) = p$$

denotes that in world w, the frequency of the joint assignment $\boldsymbol{X} = \boldsymbol{x}$ is p. In Halpern's logical framework, this is equivalent to the statement

$$w \models P(\boldsymbol{X} = \boldsymbol{x}) = p$$

which can be read as "in world w, the probability assertion $P(\boldsymbol{X} = \boldsymbol{x}) = p$ is true". We next describe Halpern's probabilistic semantics for relational random variables.

4 The Random Selection Semantics for First-Order Relational Random Variables

The key idea of the random selection semantics is to view a first-order variable $\mathbb{A}$ as a random variable ranging over instances of the class associated with $\mathbb{A}$. In the following we assume for a population variable a uniform distribution over its class. The random

selection semantics treats different population variables as independent of each other, so their joint distribution is the product of the individual distributions.

Given that population variables are random variables, first-order terms represent functions of random variables. A function of random variables themselves represents a random variable, as usual in propability theory: The probability that function f takes on value y is the probability mass of all arguments $\boldsymbol{x}$ such that $f(\boldsymbol{x}) = y$.

For a ground term, an assignment $X^* = x$ of a value is either true or false in a possible world. The probability for such an assignment in a given world is therefore the extreme values 0 and 1. While this does not represent a frequency, it is technically convenient to include this case in the definition of $P_w(\cdot)$. These ideas lead to the following definitions.

$$P_w(\mathbb{A} = a) \equiv \frac{1}{|\mathcal{I}_{\mathbb{A}}|} \tag{1}$$

$$P_w(\mathbb{A}_1 = a_1, \ldots, \mathbb{A}_k = a_k) \equiv \frac{1}{|\mathcal{I}_{\mathbb{A}_1}| \times \cdots \times |\mathcal{I}_{\mathbb{A}_k}|} \tag{2}$$

$$P_w(\boldsymbol{X}^* = \boldsymbol{x}) \equiv \begin{cases} 1 & \text{if } w \text{ assigns value } x_i \text{ to each } X_i^* \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$P_w(\boldsymbol{X} = \boldsymbol{x} | \mathbb{A} = \boldsymbol{a}) \equiv P_w(\boldsymbol{X}^*_{\mathbb{A}\backslash\boldsymbol{a}} = \boldsymbol{x}) \tag{4}$$

$$P_w(\boldsymbol{X} = \boldsymbol{x}) = \sum_{\boldsymbol{a}} P_w(\boldsymbol{X} = \boldsymbol{x} | \mathbb{A} = \boldsymbol{a}) \times P_w(\mathbb{A} = \boldsymbol{a}) \tag{5}$$

where in Equation (5) the index $\boldsymbol{a}$ ranges over the set of groundings of the first-order random variables in $\boldsymbol{X}$. Equation (4) shows that we can view a grounding of first-order terms in two ways: in a logical view, it is a textual replacement of all first-order variables by constants. In a probabilistic view, it represents conditioning on the values of population variables. The next proposition shows that this equivalence holds for instantiations in general, even if they replace only some of the population variables.

Proposition 4.1 *For any possible world and instantiation $\mathbb{A}\backslash\boldsymbol{a}$, conditioning = grounding:*

$$P_w(\boldsymbol{X} = \boldsymbol{x} | \mathbb{A} = \boldsymbol{a}) = P_w(\boldsymbol{X}_{\mathbb{A}\backslash\boldsymbol{a}} = \boldsymbol{x}).$$

4.1 Proportion Interpretation

The next proposition shows that the joint probability $P_w(\cdot)$ of first-order random variables can be interpreted as a proportion, namely the number of groundings that match the joint assignment, divided by the total number of possible groundings. The **grounding count** is given by

$$\mathrm{n}\left[\boldsymbol{X} = \boldsymbol{x}; w\right] \equiv \sum_{\boldsymbol{a}} P_w(\boldsymbol{X} = \boldsymbol{x} | \mathbb{A} = \boldsymbol{a}) \tag{6}$$

where $\boldsymbol{a}$ ranges over the set of groundings of the first-order random variables in $\boldsymbol{X}$.

Therefore

$$P_w(\boldsymbol{X} = \boldsymbol{x}) = \frac{\mathrm{n}\left[\boldsymbol{X} = \boldsymbol{x}; w\right]}{|\mathcal{I}_{\mathbb{A}_1}| \times \cdots \times |\mathcal{I}_{\mathbb{A}_k}|}. \tag{7}$$

That is, the relational frequency of a first-order formula is the number of groundings that satisfy the formula, divided by the number of possible groundings given the type constraints. Following proposition 4.1, we define the **conditional grounding count** of a joint assignment by

$$\mathrm{n}\left[\boldsymbol{X} = \boldsymbol{x} | \mathbb{A} = \boldsymbol{a}; w\right] \equiv \mathrm{n}\left[\boldsymbol{X}_{\mathbb{A}\backslash\boldsymbol{a}} = \boldsymbol{x}; w\right]. \tag{8}$$

4.2 Data Table Visualization

We can visualize the frequency distribution $P_w()$ in terms of a **groundings table** as follows [**?**]. A possible world w assigns a value to each ground random variable X^*. Let $\boldsymbol{X}^*_{\mathbb{A}\backslash\boldsymbol{a}}$ be a grounding for n relational random variables. Then $\boldsymbol{x}_w^{\mathbb{A}\backslash\boldsymbol{a}}$ denotes an n-dimensional vector whose entries list one value for each of the n ground random variables $\boldsymbol{X}^*_{\mathbb{A}\backslash\boldsymbol{a}}$, as determined by w. We can visualize the vectors $\boldsymbol{x}_w^{\mathbb{A}\backslash\boldsymbol{a}}$ in a table with n columns whose rows are indexed by the possible groundings $\boldsymbol{a}$. This groundings table is the counterpart to a data table for a set of random variables $\boldsymbol{X}$ in the case of i.i.d. data. Equation (7) entails that the frequency $P_w(\boldsymbol{X} = \boldsymbol{x})$ is the number of rows in the groundings table that match the joint assignment $\boldsymbol{X} = \boldsymbol{x}$, divided by the total number of rows in the groundings table.

4.3 Ground Relational Random Variables

The random selection semantics cannot be used to define a distribution over the values of a ground term. Halpern's type 2 semantics instead requires the specification of a distribution μ over worlds. The probability $P^*(f(\boldsymbol{a}) = x)$ is then the probability sum of all worlds that satisfy $f(\boldsymbol{a}) = x$.

While in principle, probabilities over first-order formulas need not be related to probabilities over completely ground formulas, the **Halpern instantiation principle** connects the two via the equation:

$$P_w(\boldsymbol{X} = \boldsymbol{x}) = P^*(\boldsymbol{X}^* = \boldsymbol{x}) \tag{9}$$

for any grounding $\boldsymbol{X}^*$ of the random variables $\boldsymbol{X}^*$. Halpern's original principle is restricted to a set of terms that contain a single population variable only. We use the more general version of the instantiation principle that extends it to formulas with multiple population variables (for discussion see [**?**]).

5 Bayesian Networks for Relational Data

5.1 Bayesian Networks

A **Bayesian Network (BN)** is a directed acyclic graph (DAG) whose nodes comprise a set of random variables [**?**]. Depending on context, we interchangeably refer to the

nodes and (random) variables of a BN. The conditional probability parameters of a Bayesian network specify the distribution of a child node given an assignment of values to its parent node. For an assignment of values to its nodes, a BN defines the joint probability as the product of the conditional probability of the child node value given its parent values, for each child node in the network. This means that the log-joint probability can be *decomposed* as the node-wise sum

$$\ln P_B(\boldsymbol{X} = \boldsymbol{x}) = \sum_{i=1}^{n} \ln P_B(X_i = \boldsymbol{x}_i | \mathrm{Pa}(X_i) = \boldsymbol{x}_{\mathrm{Pa}(X_i)}) \tag{10}$$

where $\boldsymbol{x}_i$ resp. $\boldsymbol{x}_{\mathrm{Pa}(X_i)}$ is the assignment of values to node X_i resp. the parents of X_i determined by the assignment $\boldsymbol{x}$. To avoid difficulties with $\ln(0)$, here and below we assume that joint distributions are positive everywhere. The parameters of the Bayesian network are the conditional probabilities of child node values given assignments to parent nodes. A common compact notation for parameter values is the following abbreviation

$$\theta_{ijk} \equiv P_B(X_i = x_{ik} | \mathrm{Pa}(X_i) = \mathrm{pa}_j)$$

where x_{ik} denotes the k-th possible value of node X_i and pa_j denotes the j-th possible state (value assignment) for the parents of X_i. Since the parameter values for a Bayes net define a joint distribution over its nodes, they therefore entail a marginal, or unconditional, probability for a single node. We denote the **marginal probability** that node X_i has value x_{ik} as

$$\theta_{ik} \equiv P_B(X_i = x_{ik}).$$

5.2 First-Order Bayesian Networks

A **First-Order Bayesian Network** (FOB) is a Bayesian network whose nodes are first-order terms. Via equation 10, a FOB defines a distribution over first-order random variables. If a FOB is learned from a sample $\mathcal{D}$ drawn from world w, we can view the FOB as an estimator of the relational class-level probability distribution $P_w(\cdot)$.

A FOB does not determine a unique probability distribution $P_B^*(\cdot)$ over *all* ground random variables, because a set of ground random variables may instantiate the same FOB multiple times. The Halpern instantiation principle (9) for a FOB B constrains the ground and first-order distributions to agree on single groundings:

$$P_B^*(\boldsymbol{X}^* = \boldsymbol{x}) = P_B(\boldsymbol{X} = \boldsymbol{x}) \tag{11}$$

for any grounding $\boldsymbol{X}^*$ of the nodes in the FOB.

6 The Random Selection Pseudo-Likelihood

For i.i.d. data, the likelihood function quantifies how well a Bayesian network fits the data. The random selection semantics suggests a pseudo-likelihood function that plays

the same role for relational data. The term "pseudo" indicates that the likelihood function is not normalized, meaning that the sum of the pseudo-likelihoods of all possible worlds is not 1.

The expected log-likelihood of a randomly selected grounding is given by the sum

$$L_B(\mathcal{D}) \equiv \frac{1}{|\mathcal{I}_{\mathbb{A}_1}| \times \cdots \times |\mathcal{I}_{\mathbb{A}_k}|} \sum_{\boldsymbol{a}} \ln P_B(\boldsymbol{X} = \boldsymbol{x}_{\mathcal{D}}^{\mathbb{A}\backslash\boldsymbol{a}}) \tag{12}$$

where the summation ranges over complete groundings of B.

By the instantiation principle (11), the joint probability $P_B^*(\boldsymbol{X}^* = \boldsymbol{x}_{\mathcal{D}}^{\mathbb{A}\backslash\boldsymbol{a}})$ can be computed from class-level probabilities as shown in Equation (12). The sum ranges over an exponentially large domain of possible groundings. However, the next proposition provides a closed-form that sums over the nodes in the FOB instead. The closed-form can be expressed using the following notation.

1. Let $(X_i, \mathrm{Pa}(X_i)) = x_{ijk}$ be the joint assignment that assigns to node X_i its k-th possible value, and to its parents their j-possible state. This formula is associated with a particular BN structure, which will be fixed by context.

2. Let $\mathrm{p}_{ijk}(\mathcal{D}) \equiv P_{\mathcal{D}}((X_i, \mathrm{Pa}(X_i)) = x_{ijk})$ be the class-level frequency of the family assignment $(X_i, \mathrm{Pa}(X_i)) = x_{ijk}$ in sample $\mathcal{D}$.

Proposition 6.1 *Assume the Bayes net instantiation principle 11. Then*

$$L_B(\mathcal{D}) = \sum_{ijk} \ln \theta_{ijk} \cdot \mathrm{p}_{ijk}(\mathcal{D})$$

where the summation on the right ranges over all nodes, node values, and parent states.

We refer to $L_B(\cdot)$ as the **random selection pseudo log-likelihood function**, and to its exponent $\exp L_B(\cdot)$ as the **random selection pseudo likelihood function**.

The closed form of the random selection pseud log-likelihood is almost identical to the log-likelihood function for i.i.d. data, except that the latter replaces frequencies by counts. This implies that as with i.i.d. data, the random selection pseudo log-likelihood is maximized by using the observed empirical frequencies as parameter estimates.

Proposition 6.2 *The parameter values that maximize the random selection pseudo log-likelihood are the empirical conditional frequencies observed in a database:*

$$\widehat{\theta}_{ijk}(\mathcal{D}) = \frac{\mathrm{p}_{ijk}(\mathcal{D})}{\sum_{k'} \mathrm{p}_{ijk'}(\mathcal{D})}.$$

7 From Class-Level Probabilities to Instance-Level Probabilities

This section considers applying class-level frequencies to inference about ground random variables where the same class-level random variables may be instantiated several

times. We examine the classification version of this problem: Given a single ground term, called the query random variable, and values assigned to *all* other ground terms, what is the probability distribution over values of the query RV? We extend our notation as follows to define this question formally.

For a ground random variable X^*, and a database defined by a complete set of literals w^*, we write $w^*_{-X^*}$ for the set of ground literals that specify the values of all ground literals except for X^*. The problem is then to define a conditional distribution

$$P^*(Y^* = y | w^*_{-Y^*}).$$

Any conditional probability is always proportional to the joint probability:

$$P^*(Y = y | \boldsymbol{X} = \boldsymbol{x}) = P^*(y, \boldsymbol{x})/Z(\boldsymbol{x})$$

where the normalization constant depends on $\boldsymbol{x}$ but not y. A natural approach is therefore to define the query probability as proportional to the pseudo log-likelihood:

$$P^*(Y^* = y | w^*_{-Y^*}) \propto \exp L_B(Y^* = y, w^*_{-Y^*}) = \exp \sum_{ijk} \ln \theta_{ijk} \cdot \mathrm{p}_{ijk}(Y^* = y, w^*_{-Y^*}). \tag{13}$$

The problem with this approach is that often many of the θ_{ijk} terms involve $(X_i, \mathrm{Pa}(X_i)) = x_{ijk}$ conditions that are irrelevant for the target variable. For example, if we want to predict the box office receipts of a movie based on ratings and the age of the rater, the ages of users who did *not* rate the movie may well be irrelevant. Equation 13 can be restricted to features that are relevant to the target variable as follows.

Definition 7.1 *A family configuration is **relevant** if the prior probability of the child node value differs from its conditional probability given the parent node values. The event that a relevant family configuration obtains is defined by*

$$R_i = \{x_{ijk} : \theta_{ijk} \neq \theta_{ik}\}.$$

Now we can define the log-linear equation.

$$P^*(Y^*_{\mathbb{A}\setminus\boldsymbol{a}} = y | w^*_{-Y^*_{\mathbb{A}\setminus\boldsymbol{a}}}) \propto \exp \sum_{ijk} \ln \theta_{ijk} \cdot P_w((X_i, \mathrm{Pa}(X_i)) = x_{ijk} | R_i, \mathbb{A} = \boldsymbol{a}) \tag{14}$$

The relevant conditional family frequency $P_w((X_i, \mathrm{Pa}(X_i)) = x_{ijk} | R_i, \mathbb{A} = \boldsymbol{a})$ may be computed as follows.

$$P_w((X_i, \mathrm{Pa}(X_i)) = x_{ijk} | R_i, \mathbb{A} = \boldsymbol{a}) = \begin{cases} 0 & \text{if } x_{ijk} \notin R_i \\ \frac{\mathrm{n}[(X_i, \mathrm{Pa}(X_i)) = x_{ijk} | \mathbb{A} = \boldsymbol{a}; w]}{\sum_{x_{ij'k'} \in R_i} \mathrm{n}\left[(X_i, \mathrm{Pa}(X_i)) = x_{ij'k'} | \mathbb{A} = \boldsymbol{a}; w\right]} & \text{if } x_{ijk} \in R_i \end{cases} \tag{15}$$

8 Parameter and Structure Learning Algorithms

Definition 8.1 *The Inverse Fast Möbius Transform*

Input *A set of first-order random variables; a database.*

Output *A contingency table that lists, for each join assignment of values to the first-order random variables, the number of groundings that match the assignment in the database.*

We first compute the Möbius parameters of the joint distribution. These involve positive relationships only. The IFMT transforms these into joint probabilities *without* further data access.

Definition 8.2 *The Learn-and-Join Structure Learning Algorithm*

Input *A set of first-order random variables; a database.*

Output *A first-order Bayesian network whose nodes are the given set of first-order random variables.*

The algorithm performs a hierarchical search in the lattice of relationship chains. It learns a Bayesian multi-net, one Bayesian network for each point in the lattice. Edges learned on smaller relationship chains are propagated to larger relationship chains.

9 Questions

in the notation $\boldsymbol{x}_w^{\mathbb{A}\backslash\boldsymbol{a}}$, should we make the grounding the superscript or the subscript? Since this is a vector of values, not a ground object. DaCampos uses superscript for row indices. Need to check the machine learning literature.