Functional lifted Bayesian networks

Statistical relational learning and reasoning with relative frequencies

Felix Quirin Weitkämper

Institut für Informatik der Ludwig-Maximilians-Universität München

September 27, 2022

Outline

What is Probability?

- What do probabilistic statements actually refer to?
- Rudolf Carnap in 1950 and later in the computer science community Halpern in 1990 distinguish two fundamentally different uses of probabilistic language.

Type I probability

Example

1% of the population are suffering from the disease.

- This is a statement about the *relative frequency* of an illness in the population, also known as a *statistical* probability.
- Such a statement is modelled by a *single* world of individuals, and a probability measure *on its domain*.
- If the domain is finite, one can assume the uniform probability measure on the domain.

Type II probability

Example

Considering his symptoms, the likelihood that this patient is suffering from the disease is 20%

- This is a statement about the *degree of confirmation* of the assertion that a particular patient has this illness
- Such a statement is modelled by a *set* of possible worlds, and a probability measure *on that set of possible worlds*.

Type III probability

Example

With a likelihood of 10%, more than 60% of the population will have been ill by the end of the year.

- This is a statement about the *degree of confirmation* of a statement that itself refers to relative frequencies
- Such a statement is modelled by a *set* of possible worlds, and a probability measure *on that set of possible worlds.*
- Additionally, each possible world is equipped with a measure on its domain.
- If the domains are finite, one can again assume the uniform probability measure on the individual worlds.

Probabilities in statistical relational artificial intelligence

- Despite the name, the semantics underlying most statistical relational approaches are based on a possible world semantics, encoding Type I probabilities.
- This includes *Markov logic networks*, *probabilistic logic programming under the distribution semantics* and *relational Bayesian networks*.
- Interesting exceptions include *stochastic logic programs*, which can be considered a Type I formalism, and the class-based semantics for *parametrised Bayesian networks*.

Type III Representations

- Recently, some frameworks have emerged that integrate relative frequencies into possible-worlds formalisms, representing Type III probabilities.
- *Lifted Bayesian networks based on conditional probability logic* encode *discrete* dependencies on conditional relative frequencies.
- *PASTA*, a probabilistic logic programming approach, incorporates relative frequencies as constraints under a credal semantics and provides upper and lower bounds rather than a unique probability measure.

Functional lifted Bayesian networks

- We introduce *functional lifted Bayesian networks* as a Type III probabilistic framework incorporating continuous dependencies on relative frequencies into the combination functions of a lifted Bayesian network.
- As in other approaches lifting Bayesian networks, a functional lifted Bayesian network has an underlying DAG whose node set are the predicates of a signature.
- The conditional probabilities of $R(\vec{a})$ for a predicate symbol R is represented as $f_R((\|\chi_{R,i}(\vec{a},\vec{y})\|_{\vec{y}})_{i \leq n_R})$ where f_R is a continuous function and $\|\chi_{R,i}(\vec{a},\vec{y})\|_{\vec{y}}$ refers to the relative frequency of tuples satisfying $\chi_{R,i}(\vec{a},\vec{y})$.

Syntax

Definition

A functional lifted Bayesian network (FLBN) over a relational signature σ consists of the following:

- A DAG G with node set σ.
- For each $R \in \sigma$ a finite tuple $(\chi_{R,i}(\vec{x}, \vec{y}))_{i \leq n_R}$ of first-order par(R)-formulas, where $\vec{x}$ is a sort-appropriate tuple whose length is the arity of R.
- For each $R \in \sigma$ a continuous function $f_R : [0,1]^{n_R} \to [0,1]$.

Semantics

- Consider an FLBN $\mathfrak{G}$ over σ and a finite domain D.
- Then the probability distribution induced by $\mathfrak{G}$ on the set of σ-structures with domain D is given by the following Bayesian network:
- The nodes are given by $R(\vec{a})$, where R is a relation symbol in σ and $\vec{a}$ is a tuple of elements of D of the right length and the right sorts for R.
- There is an edge between two nodes $R_1(\vec{b})$ and $R_2(\vec{a})$ if there is an edge between R_1 and R_2 in the DAG G underlying $\mathfrak{G}$.
- It remains to define a conditional probability table for every node $R(\vec{a})$: Given a choice of values for $P(\vec{b})$ for all $P \in \mathrm{par}(R)$ and appropriate tuples $\vec{b}$ from D, the probability of $R(\vec{a})$ is set as $f_R((\|\chi_{R,i}(\vec{a},\vec{y})\|_{\vec{y}})_{i \leq n_R})$.

Examples

Example

The signature σ has two unary relation symbols Q and R, and the underlying DAG G is $Q \longrightarrow R$. We model a relationship between $R(x)$ and those y that satisfy $Q(y)$. Consider $\chi_R := Q(y)$ and the following choices for f_R:

- The choice $f(x) = wx + c$ corresponds to *linear regression* on the proportion of y that satisfy Q.
- The choice $f(x) = \mathrm{sigmoid}(wx + c)$ corrresponds to *logistic regression*.
- The choice $f(x) = \alpha \mathrm{e}^{-\beta(x-p)^2}$ models a dependency on how far the proportion is from an optimal value p.

Learning from samples

- Parameter learning for statistical relational representations is encumbered by high complexity on large datasets.
- One approach to mitigate this would be to estimate the optimal parameters on sampled subsets of the whole domain
- In general, this does not lead to statistically consistent estimates.

Projective families

- For a certain class of families of distributions called *projective*, this approach has been shown to be statistically consistent.
- However, projectivity is a very limiting condition.
- The known projective fragments of common statistical relational frameworks are essentially propositional and cannot model any interaction between an individual and the population-at-large.

Asymptotic representation

Theorem

Let $\mathfrak{G}$ be an FLBN such that for all n-ary aggregation functions f_R, $f_R^{-1}\{0,1\} \subseteq \{0,1\}^n$.

Then $\mathfrak{G}$ is asymptotically equivalent to a quantifier-free lifted Bayesian network (i. e. an FLBN all of whose formulas $\chi_{R,i}$ are quantifier-free with $\vec{y} = \emptyset$).

Furthermore, quantifier-free lifted Bayesian networks are projective and support statistically consistent parameter estimation from sampling.

Estimating parameters from samples

- Let G_θ be the parametric family of corresponding FLBN models and G'_θ the parametric family of asymptotically equivalent quantifier-free lifted Bayesian networks.
- Sample substructures of a small fixed domain size
- Maximise the sum of the log-likelihoods of the samples on G'_θ.
- This is a statistically consistent estimate of the optimal parameters on the entire dataset.
- By the convergence result, if the original dataset is sufficiently large, this is a good estimate of the optimal parameters on G_θ.