Relational Random Regression for Bayes Nets Supplementary Material

Oliver Schulte Hassan Khosravi Yuke Zhu Tianxiang Gao
School of Computing Science
Simon Fraser University
Vancouver-Burnaby, Canada

1 Empirical Observations of Weight Scaling Effects

We compare the extent to which different weight learning methods scale weights of count predictor variables. We first describe our comparison methods, then our benchmark databases, finally measurements of the different weight sizes produced by the comparison methods. Our code and datasets are available on the world-wide web [2].

1.0.1 Methods Compared. To obtain a Bayes net structure, which defines the predictors in the log-linear model, we applied the learn-and-join algorithm to each database [6]. We then convert the Parametrized Bayes net graph to a Markov Logic Network structure (set of clauses), using moralization, as described in Section 3.3. We declared attribute predicates as functional, as recommended by the Alchemy Group [1].

For a fixed MLN structure, we compared the two Bayes net parameter learning methods (log probabilities and log differences) and Markov net parameter learning. A Markov net model uses general weights w_{ijk}. To learn the w_{ijk} weights, we applied the default weight training procedure [5] of the Alchemy package [4]. (We added unit clauses for each node-value combination, see Section 6.2, as recommended by the Alchemy group.) We refer to this method as the **MBN** method, for "Moralized Bayes Net" [3]. We next introduce our benchmark databases.

1.0.2 Databases We used 5 benchmark real-world databases. For more details please see the references in [6] and on-line sources such as [2].

MovieLens Database. This is a standard dataset from the UC Irvine machine learning repository.

Mutagenesis Database. This dataset is widely used in ILP research. It contains information on Atoms, Molecules, and Bonds between them. We use the discretization of [6].

Hepatitis Database. This data is a modified version of the PKDD02 Discovery Challenge database. The database contains information on the laboratory examinations of hepatitis B and C infected patients.

Mondial Database. This dataset contains data from multiple geographical web data sources. We followed the modification of [7], and used a subset of the tables and features for fast inference.

UW-CSE database. This dataset lists facts about the Department of Computer Science and Engineering at the University of Washington (UW-CSE), such as entities (e.g., Student, Professor) and their relationships (i.e. AdvisedBy, Publication). The dataset was obtained by crawling pages in the department's Web site (www.cs.washington.edu).

1.1 Weight Size Observations The boxplots in Figure 1 examine weights in the typical scenario discussed in Section 6.1: weights for conjunctive formulas that correspond to conditional probabilities of the form $P(child - value|parent - values)$, where *child* represents a descriptive attribute of an entity (e.g., $gender(X)$). For each such formula, we distinguish *1-variable formulas* that contain only one population variable from *2-variable formulas* that contain more than one population variable. The 1-variable formulas have just one grounding for a given target node, whereas 2-variable formulas have many. If the weights include a a scaling component, we expect that the absolute size of weights will be smaller for 2-variable formulas. Figure 1 compares the weights assigned by different methods to 1-variable formulas vs. 2-variable formulas.

The optimized MBN-weights show clear scaling effects on every benchmark database. The strongest effect is in MovieLens and the weakest effect is in UW. This is evidence that optimal weights include a scaling component for balancing the different number of formula groundings. The log-difference weights also show scaling effects, which confirms our expectation

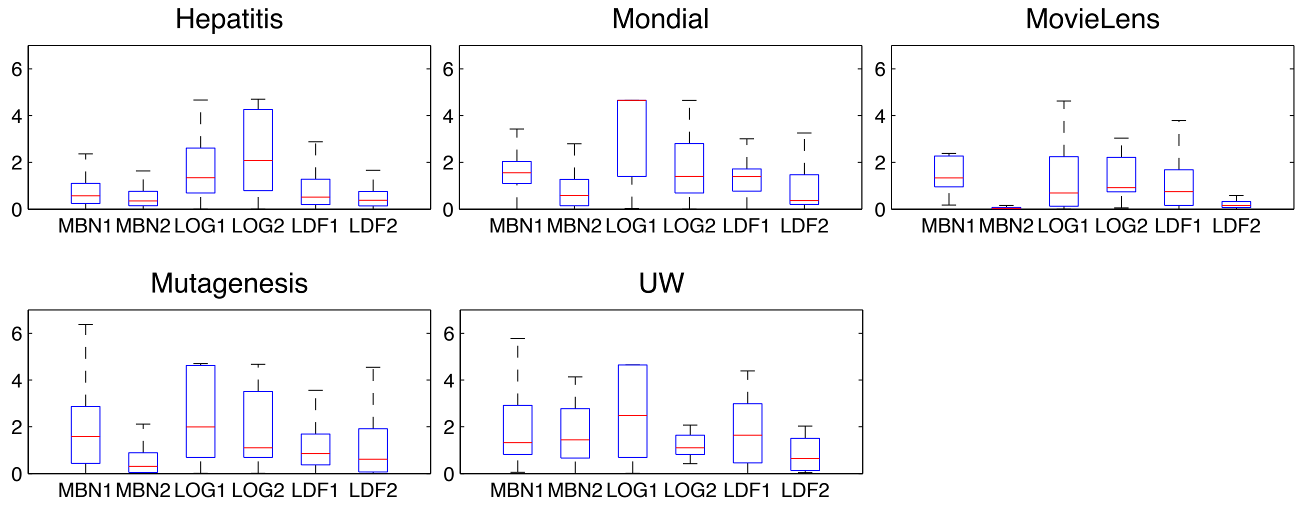

Figure 1: Boxplots of the absolute weight sizes from different weight learning methods. LOG=log(cp), LDF = log-diff. Method(i) is a box for (i)-variable formulas, for $i = 1, 2$. The box contains up to the 75th percentile of observed weights, the whisker up to the 95th percentile, and the line shows the median weight size.

that information from related entities is less important.

2 Proof of the Equivalence Theorem between frequency regression and random regression

We begin by giving a definition of random regression using formal notation. Let us write $\mathcal{P}_A$ for the set of entities in the domain of a population variable (e.g., the set of students for S). Let $A_1, \ldots, A_m$ be a list of *all* first-order variables that occur in the Markov blanket of target node Y in the regression graph for Y. We write γ to denote a simultaneous grounding of *each* variable A_i. We write Γ for the space of all possible groundings of the variables in the Markov blanket; so

$$\Gamma = \mathcal{P}_{A_1} \times \cdots \times \mathcal{P}_{A_m}.$$

In the regression graph of Figure 3, there are 2 possible groundings of the population variable Y, so $\Gamma = \mathcal{P}_Y$ and $|\Gamma| = 2$. Applying γ to a functor node $f(\boldsymbol{\tau})$ with population variables defines a ground functor node $\gamma(f(\boldsymbol{\tau}))$; to simplify notation, we write $\gamma(f)$ when the arguments to the functor are not relevant. The notation $[\gamma(f(\boldsymbol{\tau}))]_{\mathcal{D}}$ denotes the value determined by $\mathcal{D}$ when f is applied to ground term $\gamma(\boldsymbol{\tau})$. For instance, $[\mathit{gender}(\mathit{Anna})]_{\mathcal{D}} = W$ in the sample DB of Figure 1. When the notation is used with respect to a fixed database $\mathcal{D}$, we omit the $\mathcal{D}$ subscript. For a single γ, the unnormalized conditional Markov blanket probability is given by

(2.1)
$$\tilde{P}^{\gamma}_{B}(Y = y|\mathbf{X} = \mathbf{x})) = \prod_{i} P_B(f_i = [\gamma(f_i)]|\mathbf{pa}_i = [\gamma(\mathbf{pa}_i)])$$

where the index i runs over the target node and its children. In the regression graph of Figure 3, the index runs over the set $\{\mathit{gender}(\mathit{sam}), \mathit{coffee\_dr}(\mathit{sam})\}$.

Random regression is based on the expected value of the logarithm of this quantity, over all possible equiprobable groundings, which is given by

(2.2)
$$ln(\tilde{P}^{r}(Y = y|\mathbf{X} = \mathbf{x})) \equiv \frac{1}{|\Gamma|} \sum_{\gamma \in \Gamma} ln(P^{\gamma}_{B}(Y = y|\mathbf{X} = \mathbf{x})).$$

We are now ready to prove the equivalence of frequency regression and random regression.

Theorem 2.1. *The frequency regression value for a target node (Equation 5.2) equals the random regression value* $n(\tilde{P}^{r}(Y = y|\mathbf{X} = \mathbf{x}))$.

Proof. Notation. We use the following expressions. Fix a target node Y with Markov blanket population variables $A_1, \ldots, A_m$ and a database $\mathcal{D}$. Since Y is fixed, we omit the superscript Y from counting expressions like $n_{ijk}(\mathcal{D})$. For each family formula F_{ijk}, where i indexes the target node or a child of the target node, let $\gamma_{ijk}(\mathcal{D})$ be the number of simultaneous groundings of all Markov blanket population variables that satisfy F_{ijk}. Write m_i for the number of possible groundings of the Markov blanket population variables that occur in family F_{ijk}, and r_i for the number of possible groundings of the Markov blanket population variables that do *not* occur in F_{ijk} (if this set is empty, let $r_i := 1$.) For instance, if A_1, A_2 are the variables that occur in F_{ijk}, then $m_i = |\mathcal{P}_{X_1}| \times |\mathcal{P}_{X_2}|$, and $r_i = \prod_{l=3}^{m} |\mathcal{P}_{X_l}|$.

Then we have

$$\begin{aligned} |\Gamma| &= m_i \cdot r_i \\ \gamma_{ijk}(\mathcal{D}) &= n_{ijk}(\mathcal{D}) \cdot r_i \end{aligned}$$

Therefore $p_{ijk}(\mathcal{D}) = n_{ijk}(\mathcal{D})/m_i = \gamma_{ijk}(\mathcal{D})/|\Gamma|$ and the frequency regression equation 3.2 can be written as

$$(2.3)\quad \sum_{ijk} p_{ijk}(\mathcal{D}) \cdot ln(\theta_{ijk}) = \frac{1}{|\Gamma|} \sum_{ijk} \gamma_{ijk}(\mathcal{D}) \cdot ln(\theta_{ijk}).$$

Each factor $ln(\theta_{ijk})$ in Equation (2.2) appears in the sum $\sum_{\gamma\in\Gamma} ln(\tilde{P}_B^{\gamma}(Y = y|\mathbf{X} = \mathbf{x}))$ once for each simultaneous grounding $\gamma \in \Gamma$ that satisfies F_{ijk} in database $\mathcal{D}$. Therefore we have

$$(2.4)\quad \frac{1}{|\Gamma|} \sum_{\gamma\in\Gamma} ln(\tilde{P}_B^{\gamma}(\mathcal{D})) = \frac{1}{|\Gamma|} \sum_{ijk} \gamma_{ijk}(\mathcal{D}) \cdot ln(\theta_{ijk}).$$

Equations (2.4) and (2.3) together establish the identity of the random regression (2.2) and frequency regression 3.2.

3 Proof of the Equivalence proposition between log-conditional weights and log-difference weights

Proposition 3.1. *Frequency regression (Equation 5.2) returns the same result for log-conditional probability weights and for log-difference weights.*

Proof. In the following we fix an input database and therefore omit references to the database from the instance counts n_{ijk} and frequencies p_{ijk}. Similarly we fix a target node Y and omit the superscript Y. Then the frequency equation for the log-conditional parameters is

$$ln(\tilde{P}_{cp}(Y = y)) = \sum_{ijk} p_{ijk}\ ln(\theta_{ijk})$$

and for the log-difference probabilities it is

$$\begin{aligned} & ln(\tilde{P}_{\mathit{diff}}(Y = y)) \\ &= \sum_{ik} p_{i0k}\ ln(\theta_{i0k}) + \sum_{ijk} p_{ijk}\ (ln(\theta_{ijk}) - ln(\theta_{i0k})) \\ &= \sum_{ik} p_{i0k}\ ln(\theta_{i0k}) - \sum_{ijk} p_{ijk}\ ln(\theta_{i0k}) + ln(\tilde{P}_{cp}(Y = y)) \\ &= \sum_{ik} p_{i0k}\ ln(\theta_{i0k}) - \sum_{ik} ln(\theta_{i0k}) \sum_{j} p_{ijk} + ln(\tilde{P}_{cp}(Y = y)) \\ &= \sum_{ik} p_{i0k}\ ln(\theta_{i0k}) - \sum_{ik} ln(\theta_{i0k})\ p_{i0k} + ln(\tilde{P}_{cp}(Y = y)) \\ &= ln(\tilde{P}_{cp}(Y = y)) \end{aligned}$$

where the last-but-one equation follows because summing out the parent states from the joint probabilities of child state and parent state yields the marginal probability of the child state.

References

[1] Alchemy Group, *Frequently asked questions.* URL = http://alchemy.cs.washington.edu/.

[2] H. Khosravi, T. Man, J. Hu, E. Gao, and O. Schulte, *Learn and join algorithm code.* URL = http://www.cs.sfu.ca/~oschulte/jbn/.

[3] H. Khosravi, O. Schulte, T. Man, X. Xu, and B. Bina, *Structure learning for Markov logic networks with many descriptive attributes*, in AAAI, 2010, pp. 487–493.

[4] S. Kok, M. Summer, M. Richardson, P. Singla, H. Poon, D. Lowd, J. Wang, and P. Domingos, *The Alchemy system for statistical relational AI*, tech. report, University of Washington., 2009. Version 30.

[5] D. Lowd and P. Domingos, *Efficient weight learning for Markov logic networks*, in PKDD, 2007, pp. 200–211.

[6] O. Schulte and H. Khosravi, *Learning graphical models for relational data via lattice search*, Machine Learning, 88:3 (2012), pp. 331–368.

[7] R. She, K. Wang, and Y. Xu, *Pushing feature selection ahead of join*, in SIAM SDM, 2005.