Energy-Based Factor Graphs for Prediction in Relational Data

Sumit Chopra and Yann LeCun
Courant Institute of Mathematical Sciences, New York University.

A majority of supervised learning algorithms process an input point independently of other points, based on the assumptions that the input data is sampled independently and identically from a fixed underlying distribution. However in a number of real-world problems the data inherently possesses a *relational structure*. The value of variables associated with each sample not only depends on features specific to that sample, but also on the features and variables of other related samples.

Prices of real estate properties possess such a relational structure. Price of a house is a function of features that are specific to that house, such as the number of bedrooms, etc. In addition it is also influenced by features of the neighborhood in which it lies, some of which are measurable, such as the quality of the local schools. However most of the features that make a particular neighborhood desirable are very difficult to measure directly, and are merely reflected in the price of houses in that neighborhood. Hence the "desirability" of a location/neighborhood can be modeled as a latent variable, that must be estimated as part of the learning process, and efficiently inferred for unseen samples.

A number of authors have recently proposed architectures and learning algorithms that make use of relational structure. The earlier techniques were based on the idea of influence propagation [1, 3, 6, 5]. Probabilistic Relational Models (PRMs) were introduced in [4, 2] as an extension of Bayesian networks to relational data. Their discriminative extensions, called Relational Markov Networks (RMN) were later proposed in [7].

This paper introduces a general framework for prediction in relational data called Energy-Based Relational Factor Graphs (EBRFG). An architecture is presented that allows efficient inference algorithms for continuous variables with relational dependencies. The class of models introduced is novel in several ways: 1. it pertains to *relational regression problems* in which the answer variable is continuous; 2. it allows inter-sample dependencies through hidden variables (which may also be continuous), as well as through the answer variable; 3. it allows log-likelihood functions that are non-linear in the parameters, which leads to non-convex loss functions but are considerably more flexible; 4. it allows the use the non-parametric models for the relational factors, while providing an efficient inference algorithm for new samples; 5. it eliminates the intractable partition function problem through appropriate design of the relational and non-relational factors.

The idea behind a relational factor graph is to have a single graph that models the entire collection of data samples. The relationships between samples is captured by the factors that connect the variables associated with multiple samples. We are given a set of N training samples, each of which is described by a sample-specific feature vector X^i and an answer to be predicted Y^i. Let the collection of input variables be denoted by $\mathbf{X} = \{X^i,\ i = 1 \ldots N\}$, the output variables by $\mathbf{Y} = \{Y^i,\ i = 1 \ldots N\}$, and the latent variables by $\mathbf{Z}$. The EBRFG is defined by an *energy function* of the form $E(W, \mathbf{Z}, \mathbf{Y}, \mathbf{X}) = E(W, \mathbf{Z}, Y^1, \ldots, Y^N, X^1, \ldots, X^N)$, in which W is the set of parameters to be estimated by learning. Given a test sample feature vector X^0, the model is used to predict the value of the corresponding answer variable Y^0. One way to do this is by minimizing the following energy function augmented with the test sample (X^0, Y^0)

$$Y^{0*} = \text{argmin}_{Y^0}\{\min_{\mathbf{Z}} E(W, \mathbf{Z}, Y^0, \ldots, Y^N, X^0, \ldots, X^N)\}. \quad (1)$$

For it to be usable on new test samples without requiring excessive work, the energy function must be carefully constructed is such a way that the addition of a new sample in the arguments will not require re-training the entire system, or re-estimating some high-dimensional hidden variables. Moreover, the parameterization must be designed in such a way that its estimation on the training sample will actually result in good prediction on test samples. Training an EBRFG can be performed by minimizing the negative log conditional probability of the answer variables with respect to the parameter W. We propose an efficient training and inference algorithm for the general model.

The architecture of the version of the EBRFG that was used for predicting the prices of real estate properties is shown in Figure 1 (top). The price of a house is modeled as a product of two quantities: 1. its "intrinsic" price which is dependent only on its individual features, and 2. the desirability of its location. A pair of factors E^i_{xyz} and E^i_{zz} are associated with every house. E^i_{xyz} is non-relational and captures the sample specific dependencies. It is modeled as a parametric function with learnable parameters W_{xyz}. The parameters W_{xyz} are shared across all the instances of E^i_{xyz}. The factor E^i_{zz} is relational and captures the dependencies between the samples via the "hidden" variables Z^i. These dependencies influence the answer for a sample through the intermediary hidden variable d^i. The variables Z^i can be interpreted as the desirability of the location of the i-th house, and d^i can be viewed as the estimated desirability of the house using the desirabilities of the houses related to it (those that lie in its vicinity). This factor is modeled as a non-parametric function. In particular we use a locally weighted linear regression, with weights given by a gaussian kernel.

The model is trained by maximizing the likelihood of the training data, which is realized by minimizing the negative log likelihood function with respect to W and $\mathbf{Z}$. However we show that this minimization reduces to

$$\mathcal{L}(W, \mathbf{Z}) = \sum_{i=1}^{n} \frac{1}{2}(Y^i - (G(W_{xyz}, X^i) + H(Z_{\mathcal{N}^i}, X^i)))^2 + R(\mathbf{Z}), \tag{2}$$

where $R(\mathbf{Z})$ is a regularizer on $\mathbf{Z}$ (an L_2 regularizer in the experiments). This is achieved by applying a type of deterministic generalized EM algorithm. It consists of iterating through two phases. Phase 1 involves keeping the parameters W fixed and minimizing the loss with respect to $\mathbf{Z}$. The loss is quadratic in $\mathbf{Z}$ and we show that its minimization reduced to solving a large scale sparse quadratic system. We used conjugate gradient method using an adaptive threshold to minimize it. Phase 2 involves fixing the hidden variables $\mathbf{Z}$ and minimizing with respect to W. Since the parameters W are shared among the factors, this can be done using gradient descent. Inference on a new sample X^o involves computing its neighboring training samples, and using the learnt values of their hidden variables $Z_{\mathcal{N}^o}$ to get an estimate of its "desirability" d^o; passing the house specific features X^o_h through the learnt parametric model to get its "intrinsic" price; and combining the two to get its predicted price.

The model was trained and tested on a very challenging and diverse real world dataset that included 42,025 sale transactions of houses in Los Angeles county in the year 2004. Each house was described using a set of 18 house specific attributes like gps coordinates, living area, year build, number of bedrooms, etc. In addition, for each house, a number of neighborhood specific attributes obtained from census tract data and the school district data were also used. It included attributes like average house hold income of that area, percentage of owner occupied homes etc. The performance of the proposed model was compared with a number of standard non-relational techniques that that have been used in literature for this problem, namely nearest neighbor, locally weighted regression, linear regression, and fully connected neural network. EBRFG gives the best prediction accuracy by far, compared to other models. In addition we also plot the "desirabilities" learnt by our model (Figure 1 (bottom)). The plot shows that the model is actually able to learn the "desirabilities" of various areas in a way that is reflective of the real world situation.

References

[1] S. Chakrabarti, B. Dom, and P. Indyk. Enhanced hypertext categorization using hyperlinks. *In. Proc. of ACM SIGMOD98*, pages 307 – 318, 1998.

[2] N. Friedman, L. Getoor, D. Koller, and A. Pfeffer. Learning probabilistic relational models. *In Proc. IJCAI99*, pages 1300 – 1309, 1999.

[3] J. M. Klienberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604 – 632, 1999.

[4] D. Koller and A. Pfeffer. Probabilistic frame-based systems. *In Proc. AAAI98*, pages 580 – 587, 1998.

[5] J. Neville and D. Jensen. Iterative classification in relational data. *In Proc. AAAI00 Workshop on Learning Statistical Models From Relational Data*, pages 13 – 20, 2000.

[6] S. Slattery and T. Mitchell. Discovering test set regularities in relational domain. *In. Proc. ICML00*, pages 895 – 902, 2000.

[7] B. Taskar, P. Abbeel, and D. Koller. Discriminative probabilistic models for relational data. *Eighteenth Conference on Uncertainty on Machine Intelligence (UAI02)*, 2002.

Table 1: *Prediction accuracies of various algorithms on the test set. Absolute Relative Forecasting Error (fe) was measured. The error (fe_i) on the i^{th} sample is defined as $fe_i = |A_i - Pr_i|/A_i$, where A_i is the actual selling price and Pr_i is the predicted price. Three performance quantities on the test set are reported; percentage of houses with a forecasting error of less than 5%, with less than 10% and with less 15%.*

MODEL CLASS	MODEL	$< 5\%$	$< 10\%$	$< 15\%$
NON-PARAMETRIC	NEAREST NEIGHBOR	25.41	47.44	64.72
NON-PARAMETRIC	LOCALLY WEIGHTED REGRESSION	32.98	58.46	75.21
PARAMETRIC	LINEAR REGRESSION	26.58	48.11	65.12
PARAMETRIC	FULLY CONNECTED NEURAL NETWORK	33.79	60.55	76.47
HYBRID	**RELATIONAL FACTOR GRAPH**	**39.47**	**65.76**	**81.04**

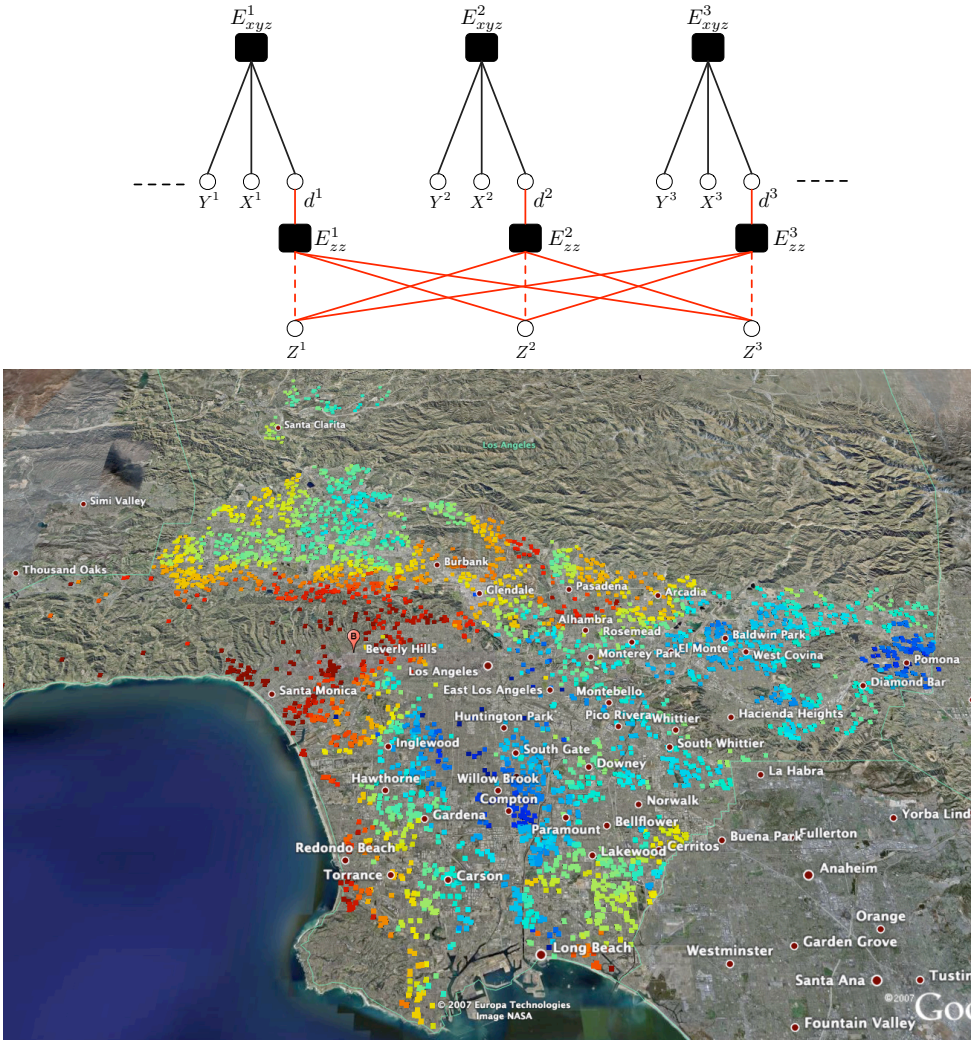

Figure 1: **(Top)** *A typical Energy Based Relational Factor Graph showing the connections between three samples. The factors E^i_{xyz} capture the dependencies between the features of individual samples and their answer variable Y^i, as well as the dependence on local latent variables d^i. The factors E^i_{zz} captures the dependencies between the hidden variables of multiple samples. The connection to these two factors may exist only from a subset of samples that are related to sample i. When the energy of factor E_{zz} is quadratic in d.* **(Bottom)** *The color coded values of the desirability surface at the location of the test samples. For every test sample, the estimate of its desirability is computed and is color coded according to its value. Blue color implies low desirability and red color implies high desirability.*