On Approximate Reasoning Capabilities of Low-Rank Vector Spaces

Guillaume Bouchard
Xerox Research Center Europe
Grenoble, France
guillaume.bouchard@xrce.xerox.com

Sameer Singh
University of Washington
Seattle WA
sameer@cs.washington.edu

Théo Trouillon
Xerox Research Center Europe
LIG, Univ. Grenoble
theo.trouillon@xrce.xerox.com

Abstract

In relational databases, relations between objects, represented by binary matrices or tensors, may be arbitrarily complex. In practice however, there are recurring relational patterns such as transitive, permutation, and sequential relationships, that have a regular structure which is not captured by the classical notion of matrix rank or tensor rank. In this paper, we show that factorizing the relational tensor using a logistic or hinge loss instead of the more standard squared loss is more appropriate because it can accurately model many common relations with a fixed-size embedding (depends sub-linearly on the number of entities in the knowledge base). We illustrate this fact empirically by being able to efficiently predict missing links in several synthetic and real-world experiments. Further, we provide theoretical justification for logistic loss by studying its connection to a complexity measure from the field of information complexity called *sign rank*. Sign rank is a more appropriate complexity measure as it is low for transitive, permutation, or sequential relationships, while being suitably large, with a high probability, for uniformly sampled binary matrices/tensors.

Introduction

In statistical relational learning, embedding models, or distributed representations, are gaining attention due to their ability to efficiently learn and predict generic data models. They can be viewed as direct extensions of factorization techniques, such as Probabilistic Matrix Factorization (Salakhutdinov and Mnih 2008). Relational databases are often used to store relations between objects that are transitive, such as `isAncestorOf`. Permutation and inclusion relations are also commonly used, for example to represent one-to-one or one-to-many connections between tables. There are also many relations that are sequential, such as in temporal data. It is commonly believed that factorization models learned on such relations are unlikely to perform well, since the underlying matrix has nearly full, or close to full, rank. For example, n-object transitivity can be represented as an $n \times n$ upper-triangular matrix of rank n. Similarly, n-object permutation matrices have rank n, and sequences represented by the `nextTo` relation is upper-band-diagonal with rank $n-1$.

As an alternative to the rank as a complexity measure for relations, we first present the *sign-rank*, an existing measure of complexity for binary matrices. In short, the sign-rank of a binary matrix M is the minimal rank of a real matrix that, once thresholded, gives M. A nice property of the sign rank is that it does not depend on the dimension for many common matrices mentioned above, such as permutation matrices, triangular matrices, or band-diagonal matrices. This means that a constant size-embedding could be used in principle, independent of the number of objects in the database, and the sign rank is the minimal size of such embeddings to get a perfect recovery. In practice, the sign rank is hard to compute, but an approximate value can be obtained by factorizing the tensor under a binary loss, such as the logistic negative log-likelihood or the hinge loss. Such approximations do not lead exactly to a constant rank, but to a rank that seems to scale logarithmically (or at least sub-linearly) with the number of entities. As a consequence, we can compactly represent common logical relationships using embeddings, since such models are both easy to learn and scale using continuous optimization techniques. Thus using an embedding model as well as minimizing a binary loss are good ingredients for solving statistical relational learning problems.

Based on this theoretical insight, we show empirically that the reconstruction is better with logistic loss. We apply it on the permutation matrices (equivalent to identity matrices), on transitivity relations that include partial order relationships, and finally on sequences and spatial relationship. We do not consider the relation types that corresponds to clusters or set inclusion relationship in this work, because it has been already well-studied in the spectral clustering community, and therefore is known to have low-rank structure.

Statistical Relational Learning

Problem Setup

Assume we have K relations between entities. We make a closed world assumption, i.e. we know that there are E entities in the world, denoted $\mathcal{E} := \{e_1, \cdots, e_E\}$. Each relation is uniquely identified by a binary tensor of as many dimensions as the arity of the relation, for example binary matrices $E \times E$ for binary relations. Ones and zeros represent whether the relation holds, or does not hold, between the two entities.

We assume we observe T pairs of grounded atoms $(\mathcal{R}_{k_t}(\bar{x}_t), y_t)_{t=1}^{T}$ where k_t is the index of the relation, $\bar{x}_t = (x_{t1}, \cdots, x_{tA(k_t)})$ represents the t-th atom, where $A(k_t)$ is the arity of the k_t-th relation, and y_t is the truth value of this atom (true=1, false=0). Our goal is to answer a simple query that returns the probability of the truth value of an unobserved grounded atom $\mathcal{R}_{k_{T+1}}(\bar{x}_{T+1})$, denoted by the binary variable $y^{(T+1)}$.

Classical relational databases do not handle such queries and return usually a special symbol indicating a missing value (usually the NULL value). Instead, probabilistic databases define a probability distribution over tuples in the database, that has a well defined interpretation under the possible world semantics model. Statistical relational learning approaches are machine learning techniques that compute probabilities for the truth value of every unobserved tuple x_{T+1}. The most widely used approach is based on Markov Logic Networks (Richardson and Domingos 2004), but they face difficulty in scaling to massive datasets. Instead, we focus on factorization methods that model atoms as multi-linear dot products between entity embeddings.

Joint Factorization of Relational Data

Collective matrix and tensor factorizations are probabilistic models of embeddings that have been independently proposed by many authors over the last few years. The main difference between the approaches is the way the multilinear dot-product is defined, including Tucker decomposition (Acar, Kolda, and Dunlavy 2011), Rescal model (Nickel, Tresp, and Kriegel 2011), or the simpler but more scalable CP/Parafac model (Singh and Gordon 2008).

Embeddings: Entities are entirely determined by their R-dimensional embedding matrix $\Theta := (\theta_1, \cdots, \theta_E) \in \Re^{E \times R}$, where $\theta_i \in \Re^R$ is the embedding of the i-th entity.

Atom probabilities: For the k-th relation, the probability of an atom $\mathcal{R}(x_1, \cdots, x_A)$ to hold is given by $\sigma(s) := \frac{1}{1+e^{-\tau-s}}$, where τ is a threshold learned with the embeddings. The score s is the multi-linear dot-product across entities $\left\langle C_k; \theta_{x_1}, \theta_{x_2} \cdots, \theta_{x_{A(k)}} \right\rangle$ where C_k represents the core-tensor of the $k-$th relation. The core-tensors were constrained to be diagonal in our experiments because its complexity scales linearly with the size of the embeddings, and is equivalent to the Parafac/CP decomposition (Harshman and Lundy 1984). The R diagonal values of the C_k are sometimes called the embedding of the relations, as it has the same dimensionality as the entities embeddings θ_i, $i \in \{1, \cdots, E\}$.

Learning

The embedding matrix Θ and the core tensors $C := (C_1, \cdots, C_K)$ are learned by maximizing the likelihood of the observed atoms:

$$(\hat{\Theta}, \hat{C}, \hat{\tau}) \in \arg \max_{\Theta, C, \tau} \sum_{t=1}^{T} \log p(\mathcal{R}_{k_t}(\bar{x}_t) = y_t) \qquad (1)$$

which is often obtained using stochastic gradient descent (SGD) on the observations.

Complexity measures for Relations

Sign rank

The sign-rank is a measure of complexity for binary matrices that dates back to early work on communication complexity, a subfield of computational complexity (Alon, Frankl, and Rodl 1985). The sign-rank $\mathrm{rk}_{\pm}(Y)$ of a binary matrix $Y \in \{0,1\}^{n \times m}$ is defined as the smallest rank of a real matrix X such that if thresholded, we obtain the matrix Y. Formally,

$$\mathrm{rk}_{\pm}(Y) = \min_{\tau \in \Re, X \in \Re^{n \times m}} \{\mathrm{rk}(X); \mathbb{I}_{\{X + \tau \geq 0\}} = Y\} \qquad (2)$$

Link between the sign rank and loss minimization

A natural way to compute the sign-rank of a $E \times E$ matrix Y, is the following: for multiple embedding sizes $r = 1, 2, \cdots, E$, try to find the matrix of embeddings Θ and the threshold τ such that $\mathbb{I}_{\{\Theta\Theta^T + \tau \geq 0\}} = Y$. The sign rank is the smallest embedding size for which this problem has at least one solution. Finding such embeddings can be done by minimizing the empirical risk $\sum_{(i,j) \in \mathcal{E}^2} \mathbb{I}_{\{\delta_{ij}(\Theta;\tau) \neq Y_{ij}\}}$ where $\delta_{ij}(\Theta, \tau) = 1$ if $< \theta_i, \theta_j > +\tau > 0$ and 0 otherwise. This non-continuous function is hard to minimize in general, but can be approximated by the minimization of a continuous relation of the empirical loss. In this paper, we use logistic loss, but the hinge loss would work as well. When the matrix Y is of sign-rank r, there exist embeddings of size r that give an arbitrary small loss, but the magnitude of the embeddings can grow arbitrarily large, similarly to logistic regression estimated on separable data, so in practice we a use penalized estimator, and recover the rank only approximately.

Sign rank of identity matrices

In many applications, multiple knowledge bases have the same entities, but the correspondence between them is not known. One of the most common scenarios is to match relations extracted from texts on one side (e.g. using semantic role labeling and named entity recognition) to a static knowledge base (such as Freebase) on the other side. This amounts to estimating the `sameAs` relation, i.e. the equality relation, using supervised information and prior assumptions about the possible matches across the two data sources. In a simplified setting, each entity appears only once in each of the knowledge bases, which amounts to find an optimal assignment of entities. This can be modeled as finding a permutation matrix that maximizes the fit of the observed data. Rather than using a dedicated algorithm such as the Hungarian method to find this optimal matching, we show here that permutation matrices can be efficiently modeled by low-dimensional embeddings. Up to a permutation of rows or columns, permutation matrices are related to identity matrices, so we start by illustrating a surprising fact about identity matrices.

The low-rank representation of the identity matrix using the logistic model can be theoretically understood under the sign rank umbrella: it is known for some time that identity matrices can be uniquely represented by latent dimensions of length 2 (Paturi and Simon 1986), meaning that the sign-rank of identity matrices is two. To see this, one can represent entities in two dimensions equally spaced on the unit circle. The dot-product between two embeddings θ_i and θ_j will be 1 if the corresponding entities are the sames, i.e. if $i = j$, and will be strictly smaller that 1 if the entities are different. It is therefore enough to threshold the hyperplane at $1 - \epsilon$, where $\epsilon > 0$ is arbitrary small. In practice by minimizing the logistic

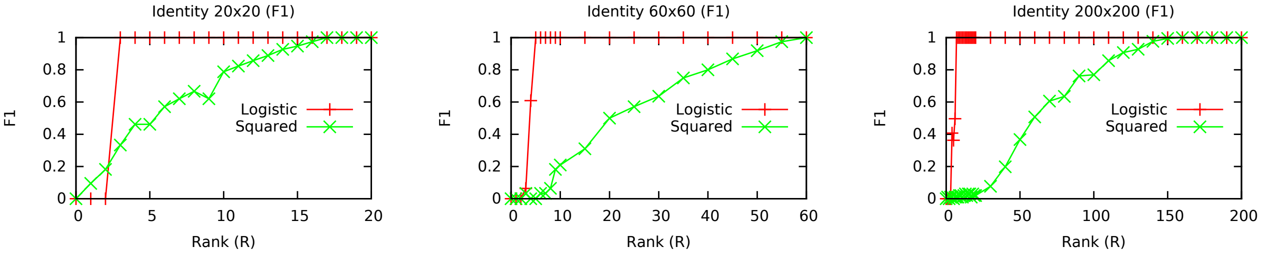

Figure 1: **Identity matrix:** F1 measure of the reconstruction of fully observed identity matrices of varying sizes, as function of the embedding size (rank, $\mathcal{R}$). The two curves correspond to the minimization of the squared loss and the logistic loss.

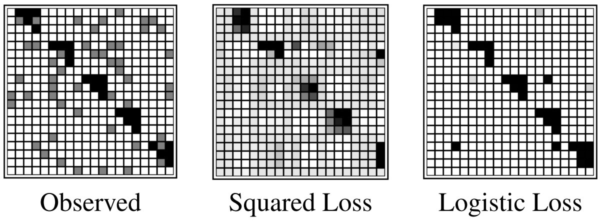

Figure 2: Transitive relation learning with rank-3 embeddings. Gray cells represent missing values in the training set (left). We see that they are much more accurately recovered by logistic loss (right) than linear loss (center).

loss, we recover the sign rank only approximately (Linial et al. 2007), and this is shown in the next section.

Embeddings of Standard Relations

Ontology matching and `sameAs` relations

Low-rank identity matrix: When approximating the identity matrix, standard low-rank approaches based on Singular Value Decomposition (SVD) give poor reconstruction error because the identity matrix is full-rank, and thus embeddings must have the same size as the number of entities in the database. However, by minimizing the logistic loss instead of the squared loss, we can represent the identity matrix using a constant number of dimension, i.e. the embedding size is nearly independent to the number of entities in the database. This fact is illustrated on Figure 1 where the identity matrix has been factorized using the squared loss (green curve) and the logistic loss (red curve). One can see that SGD finds an accurate reconstruction of the identity matrix for ranks that are smaller than 3. In theory, the sign rank is two, but since we converge to a local minima that is not the global one, we can see that the rank remains small, even when the the number of entities increases. Despite its simplicity, this experiments shows that using embedding models on identity can be meaningful, and since the ordering of the rows of columns is arbitrary, this result extends directly to permutation matrices.

Transitive and Hierarchical Relations

We now show that low-rank representations with logistic links can model transitive relationships accurately. A binary relation $\mathcal{R}$ is called transitive when the literals $\mathcal{R}(a, b)$ and $\mathcal{R}(b, c)$ are true implies that $\mathcal{R}(a, c)$ is also true. For example, the relations `isOlderThan`, `isInsideOf`, `isAncestorOf`, or `isLarger` are transitive. Transitivity often relates to the fact that there exists a partial ordering of the entities with respect to an attribute that is not always explicit: the relation `isOlderThan` implicitly assumes the entities have an age. It seems therefore natural that this relation can be represented by a low-dimensional embedding representing this partial ordering. We start by considering the simpler case of total ordering, which can be understood a transitive relation that is also anti-symmetric.

Total ordering: When representing the relation $\mathcal{R}$ representing a total ordering as a $E \times E$ matrix, we obtain, an upper triangular matrix, up to a permutation of rows and columns. The rank of such matrix is naturally E, i.e. the maximal rank, as no row or column can be expressed as a linear combination of the others, but the sign rank is only one, as it is enough to represent the entities on a uni-dimensional scale on which the order is preserved. Again, we see that the rank is not an appropriate measure of the complexity of the relation: it has the maximum possible value when the relation implicitly assumes there is a one-dimensional scale that could uniquely represent the individuals with respect to this relation, i.e. it has a sign rank of one.

Group ordering: We consider the more interesting example of a transitivity relation that only holds within groups of entities. In this setting, entities are members of n unrelated families. Each family has exactly k members, for a total of $E = nk$ entities. For a fixed set of families, we observe the full set of possible relations, which are transitive within this group. By a proper ordering of the entities, relational matrix is a block-diagonal matrix in which blocks are upper triangular, as shown in Figure 2. In this example, we removed some of the transitivity relations for 10% of the relations, and learned the model by treating these removed relations as missing (grey cells). Once learned, the embedding model, even if the transitivity was not constrained a priori, should use statistical redundancies to correctly find that the relation is true when the transitivity rule should apply. In such a case, we can safely say that the model is able to perform transitivity reasoning. The learning process is illustrated on Figure 2: on the left, the training data assumes that we observed all but few relation in the last two families (we used n=5 families and $k = 4$ members per family). When predicting the relations according to the logistic model (right figure), we see that most of the training data are recovered, but the model was

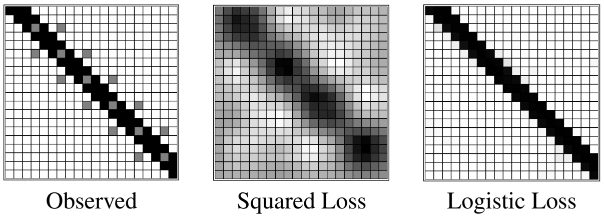

Figure 3: Sequential relation learning with rank-3 embeddings. The predictive accuracy of logistic-loss model is much higher.

	Squared			Logistic	
Rank, R	5	10	20	5	10
Inside	0.0	33.3	0.0	88.9	100.0
Neighbor	0.0	50.0	72.3	50.0	83.3
Overall	0.0	42.9	50.0	70.6	90.9

Table 1: F1 on the heldout atoms of the Countries dataset.

also able to predict the missing entries which illustrates the fact that transitivity has a low-complexity using logistic loss. The squared loss (middle matrix), as expected in such block-diagonal matrices, is not able to generalize transitivity to the last two families.

Hierarchies: Hierarchies are more complex in the sense that transitivity is true along every ancestor part. Example of such relations are `isSubpartOf` or `isInsideOf`. Such relations have similar structure as groups. This will be presented in the country relationship experiment below.

Sequential and Spatial Relationship

Sequences can be easily represented by an upper band-diagonal matrix that represents the existence of a forward relation, such as `isNext`. In this example as well, if a positive entry is missing, it is recovered by link prediction using a rank-3 embedding. This is illustrated in a missing data experiment, in which we generate a 1-hop or 2-hop sequential relationship by setting the relation $\mathcal{R}(x_i, x_j)$ to true when $j = i + 1$ or $j = i + 2$. This is illustrated in Figure 3, where we see that the squared loss is unable to identify the exact relation, and gives non-calibrated probabilities (the black or white region indicate predictions that were > 1 or < 0).

Experiments with Country Relationships

We also experiment on a dataset of countries generated from public geographical data[1]. We created two relations: `isInside(`e_1, e_2`)`, and `isNeighbor(`e_1, e_2`)`. Entities include the four continents (we merged Asia and Europe), 22 subcontinents and 245 countries. Each country has exactly one continent and one subcontinent for which the `isInside` relation is true, and similarly each subcontinent has exactly one continent. The `isNeighbor` relation holds for land borders only. We divided the data in four train and test sets, where train sets contain all true and false relations involving at least one entity from three continents, while relations involving two entities from the remaining continent are unknown. The test set contains queries for both relations on the remaining continent, that were chosen to accurately test the reasoning abilities of the model. We will make the dataset, and our train/test splits, publicly available.

Results of the prediction are shown in Table 1, where we compare the accuracy measures for different embedding sizes and losses. We see that the squared loss is not able to correctly predict the relations, even for larger embeddings size, while the use of logistic loss give nearly perfect prediction accuracy for embeddings of only size 10.

Conclusions

For a long time, practitioners have been reluctant to use embedding models because many common relationships, including the `sameAs` relation modeled as an identity matrix, were not trivially seen as low-rank. In this paper we showed that when sign-rank based binary loss is minimized, many common relations such as permutation matrices, sequential relationships, and transitivity can be represented by surprisingly small embeddings

References

Acar, E.; Kolda, T. G.; and Dunlavy, D. M. 2011. All-at-once optimization for coupled matrix and tensor factorizations. In *Proceedings of Mining and Learning with Graphs*.

Alon, N.; Frankl, P.; and Rodl, V. 1985. Geometrical realization of set systems and probabilistic communication complexity. In *Foundations of Computer Science*.

Harshman, R. A., and Lundy, M. E. 1984. The parafac model for three-way factor analysis and multidimensional scaling. *Research methods for multimode data analysis* 122–215.

Linial, N.; Mendelson, S.; Schechtman, G.; and Shraibman, A. 2007. Complexity measures of sign matrices. *Combinatorica* 27(4):439–463.

Nickel, M.; Tresp, V.; and Kriegel, H. 2011. A three-way model for collective learning on multi-relational data. In *International Conference on Machine Learning (ICML)*.

Paturi, R., and Simon, J. 1986. Probabilistic communication complexity. *Journal of Computer and System Sciences* 33(1):106–123.

Richardson, M., and Domingos, P. 2004. Markov logic networks. Technical report, U. Washington CS Dept.

Salakhutdinov, R., and Mnih, A. 2008. Probabilistic matrix factorization. In *Neural Information Processing Systems (NIPS)*.

Singh, A. P., and Gordon, G. J. 2008. Relational learning via collective matrix factorization. In *Knowledge discovery and data mining (KDD)*. ACM. 650658.

[1] `https://github.com/mledoze/countries`