Relational Learning via Collective Matrix Factorization

Prepared by Thomas Perry

School of Computational Science & Engineering
College of Computing
Georgia Institute of Technology

Introduction

Goal: To present a generalized approach to collective matrix factorization.

Rating based paradigm

- Users $u, v \in U$; Movies: $i, j \in V$; Genres $s, t \in Z$
- Ratings X_{ui} indicating preference of user u for movie i. Highers scores imply higher preference.
- Genres Y_{it} indicating movie i belongs to genre t. These are binary values.
- Problem: Ratings are not defined over all $U \times V$, need to predict missing ratings while utilizing Y.

Introduction: Unified View of Factorization

The building block for collective factorization is single-matrix factorization.

1. Prediction link $f : \Re^{m\times n} \rightarrow \Re^{m\times n}$
2. Loss function $D\left(X, f\left(UV^T\right)\right) \geq 0$
3. Optional Data Weights $W \in \Re_{+}^{m\times n}$
4. Hard Constraints on factors U, V.
5. Regularization penalty, $R\left(U, V\right) \geq 0$

Prediction Links

For the model $X \approx f(UV^T)$ we solve
For weighted SVD

$$argmin_{U,V} \left\| W \odot \left(X - UV^T \right) \right\|_{Fro}^2 \tag{1}$$

Generalized Form

$$argmin_{U,V} \left[D\left(X, f\left(UV^T \right) \right) + R\left(U, V \right) \right] \tag{2}$$

Choices for D and f are related to distributional assumptions on X.

Collective Factorization - SVD

One model of collective matrix factorization using regularized SVD.

$$L_1(U,V) = \sum_{u,i \in O} (X_{u,i} - U_u V_i)^2 + \lambda_1 \|U_u\|^2 + \lambda_2 \|V_i\|^2 \tag{3}$$

$$L_2(V,Z) = \sum_{i,s \in P} (Y_{i,s} - V_i Z_s)^2 + \lambda_2 \|V_i\|^2 + \lambda_3 \|Z_s\|^2 \tag{4}$$

$$L(U,V,Z) = \alpha L_1(U,V) + (1-\alpha) L_2(V,Z) \tag{5}$$

Possibly add additional weighting terms

Bregman Divergences

The approach is generalized for Bregman divergences.

$$D_F(Z \| Y) = F(Z) + F^*(Y) - Y \circ Z \tag{6}$$

F is a closed convex function, F:$\Re^{m \times n} \rightarrow \Re$
F^* the convex dual
$A \circ B$ is the matrix dot product $tr\left(AB^T\right)$

If F decomposes into a sum over components of Z, we can define the weighted divergence.

$$D_F(Z \| Y, W) = \sum_{ij} W_{ij}\left(F(Z_{ij}) + F^*(Y_{ij}) - Y_{ij} Z_{iz}\right) \tag{7}$$

Examples

- $F(x) = x^2$
- $F(x) = xlog(x) - x$

Exponential Family oF Distributions

$p_F(x|\theta)$ is a regular exponential family if each density has the form

$$log_{p_F(x|\theta)} = log_{p_0}(x) + \theta^T x - F(\theta) \tag{8}$$

where $F(\theta)$ is the log partition function

$$F(\theta) = log \int p_0(x) \cdot exp\left(\theta^T x\right) dx \tag{9}$$

For regular exponential families

$$log_{p_F(x|\theta)} = log_{p_0}(x) + F^*(x) - D_F(x \| f(\theta)) \tag{10}$$

where $f(x) = \nabla F(\theta)$
Minimizing the Bregman divergence under a matching link is equivalent to maximum likelihood for the cooresponding exponential family distribution.

Collective Factorization

One model of Bregman matrix factorization proposes the following decomposable loss function.

$$L_1(U, V|W) = D_{F1}\left(UV^T \| X, W\right) + R_1(U) + R_2(V) \quad (11)$$

$$L_2\left(V, Z|\tilde{W}\right) = D_{F2}\left(VZ^T \| Y, \tilde{W}\right) + R_2(V) + R_3(Z) \quad (12)$$

$$L\left(U, V, Z|W, \tilde{W}\right) = \alpha L_1(U, V|W) + (1-\alpha) L_2\left(V, Z|\tilde{W}\right) \quad (13)$$

Parameter Estimation

- Equation 13 is convex in any one of its arguments.
- Parameters can be estimated by fixing all but one argument of $L = L(U, V, Z|W, \tilde{W})$ and updating the free factor using a Newton-Raphson Step.
- The Newton update for the factors reduce to row-wise optimzation of U,V, and Z when L_1 and L_2 are decomposable functions.

$$U_i^{new} = U_i - \eta \cdot q\left(U_i\right)\left[q'\left(U_i\right)\right]^{-1} \tag{14}$$

Parameter Esimation cont.

Definitions of $q\left(U_i\right), q\left(V_i\right), q\left(Z_i\right)$

$$q\left(U_i\right) = \alpha\left(W \odot \left(f_1\left(UV^T\right) - X\right)\right) V + \nabla R_1\left(U\right) \tag{15}$$

$$q\left(V_i\right) = \alpha\left(W \odot \left(f_1\left(UV^T\right) - X\right)\right)^T U + \tag{16}$$

$$\left(1-\alpha\right)\left(\tilde{W} \odot \left(f_2\left(VZ^T\right) - Y\right)\right) Z - \nabla R_2\left(V\right) \tag{17}$$

$$q\left(Z_i\right) = \left(1-\alpha\right)\left(\tilde{W} \odot \left(f_2\left(VZ^T\right) - Y\right)\right)^T V - \nabla R_3\left(Z\right) \tag{18}$$

Generalizing to Arbitrary Schemas

- The three-factor model generalizes to any pair-wise relational schema.
- The binary relations are represented as a set of edges: $E = (i, j) : \epsilon_i \tilde{} \epsilon_j \wedge i < j$.
- $[U]$ denotes the set of latent factors and $[W]$ the weight matrices.

The loss of the model is.

$$L([U][W]) = \sum_{(i,j) \in E} \alpha^{ij} \left(D_{F^{(ij)}} U^{(i)} \left(U^{(j)} \right)^T \| X^{(ij)}, W^{(ij)} \right) + \tag{19}$$

$$\sum_{i=1}^{t} \left(\sum_{j:(i,j) \in E} \alpha^{(ij)} \right) D_{G^{(i)}} \left(0 \| U^{(i)} \right) \tag{20}$$

Experiments

Using the Netflix data and movie genre data from IMDB.
Goals

- Predict whether a user rated a particular movie.
- Predict the value of a rating for a particular movie.

Loss is calculated by MAE.
l_2 regularization is used throughout

Outcomes

- Presents collective matrix factorization as a model of pairwise relational data.
- Under the assumption of decomposable, twice differentiable loss, we derive a full Newton step.
- Practical on relational domains with large amounts of data.