Project Report: Application Prediction

1 Optimization

1.1 Conventional Matrix Completion

The latent factor model has an underlying assumption: the real exact rating matrix are low-rank matrix. By modelling target rating variable as

$$R_{ij} = U_i^T V_j \tag{1}$$

Then we are able to use least square cost to guide the optimization procedure:

$$\min_{U,V} \sum_{(i,j)\in\Omega} (A_{ij} - U_i^T V_j)^2 + \frac{\lambda}{2}(||U||_F^2 + ||V||_F^2) \tag{2}$$

where $A_{i,j}$ is the observed ground-truth entry and Ω is the index set of entries that are observed.

1.2 One-Class Matrix Completion

$$\min_{U,V} \sum_{(i,j)\in\Omega} (1 - U_i^T V_j)^2 + \frac{\lambda}{2}(||U||_F^2 + ||V||_F^2) \tag{3}$$

1.3 Conventional Inductive Matrix Completion

Formulate the problem as that of recovering a low-rank matrix W_* using observed entries $R_{ij} = x_i^T W_* y_j$ and the user/job feature vectors x_i, y_j. By factoring $W = UV^T$, we see that this scheme constitutes a bi-linear prediction $(x^T U_*)(V_* y)$ for a new user/job pair (x, y).

$$\min_{U,V} \sum_{(i,j)\in\Omega} (A_{ij} - x_i^T U V^T y_j)^2 + \frac{\lambda}{2}(||U||_F^2 + ||V||_F^2) \tag{4}$$

Remark 1 *According to [1], under standard set of assumptions, alternating minimization provably converges at a linear rate to the global optimum of two low-rank estimation problems: a) RIP measurements based general low-rank matrix sensing, and b) low-rank matrix completion. A more recent paper [2] in bioinformatics demonstrated successful application of such inductive matrix completion framework on gene-disease analytics.*

1.4 One-Class Inductive Matrix Completion

$$\min_{U,V} \sum_{(i,j)\in\Omega} (1 - x_i^T U V^T y_j)^2 + \frac{\lambda}{2}(||U||_F^2 + ||V||_F^2) \tag{5}$$

1.5 One-Class Inductive Matrix Completion with Biases

$$\min_{U,V} \quad (1-\alpha) \sum_{A_{ij}=1} (1 - x_i^T U V^T y_j)^2 + \alpha \sum_{A_{ij}=0} (0 - x_i^T U V^T y_j)^2 + \frac{\lambda}{2}(||U||_F^2 + ||V||_F^2) \tag{6}$$

1.6 One-Class Inductive Matrix Completion with kernel method

One possible enhancement for above inductive matrix completion is to extend our consideration from the linear association between features and latent factors to an version that accepts non-linear association. By this intuition, we can name this approach as *Kernel-based Inductive Matrix Completion*. By taking into account non-linear relations between designed features and hidden topics (shared latent factors), the space of latent factors can be largely expanded and then it would be more likely to automatically detect latent factors with higher quality.

1.7 One-Class Inductive Matrix Completion with pre-clustering

2 Algorithm

The algorithmic procedures for *One-class Inductive Matrix Completion* are shown in Algorithm 1.

Algorithm 1 Alternating Minimization for One-Class Inductive Matrix Completion with Biases

INPUT:
a) sparse matrices X and Y denote features of users and jobs.
b) matrix A denotes partial observation of application association with observed index set Ω

Initialize $U_{(0)}$ and V_0 by uniform randomization
Do
$V_{(k+1)}$ = argmin $(1-\alpha)\sum_{(i,j)\in\Omega}(1-\boldsymbol{x}_i^T U_{(k)} V_{(k)}^T \boldsymbol{y}_j)^2 + \alpha\sum_{(i,j)\notin\Omega}(0-\boldsymbol{x}_i^T U_{(k)} V_{(k)}^T \boldsymbol{y}_j)^2$
$U_{(k+1)}$ = argmin $(1-\alpha)\sum_{(i,j)\in\Omega}(1-\boldsymbol{x}_i^T U_{(k)} V_{(k+1)}^T \boldsymbol{y}_j)^2 + \alpha\sum_{(i,j)\notin\Omega}(0-\boldsymbol{x}_i^T U_{(k)} V_{(k+1)}^T \boldsymbol{y}_j)^2$
Until Convergence.
Predict values for missing entries: for some $(i,j)\notin\Omega$,
$R_{ij} = 1,$ if $\boldsymbol{x}_i^T U_* V_*^T \boldsymbol{y}_j >$ cutoff ϕ
$R_{ij} = 0,$ otherwise

OUTPUT:
a) Model Parameter U_* and V_*

Algorithm 2 Alternating Minimization for One-Class Inductive Matrix Completion with Kernel Methods

INPUT:
a) sparse matrices X and Y denote features of users and jobs.
b) matrix A denotes partial observation of application association with observed index set Ω

Initialize $U_{(0)}$ and V_0 by uniform randomization

OUTPUT:
a) Model Parameter U_* and V_*

Algorithm 3 Alternating Minimization for One-Class Inductive Matrix Completion with Pre-clustering

INPUT:
a) sparse matrices X and Y denote features of users and jobs.
b) matrix A denotes partial observation of application association with observed index set Ω

Initialize $U_{(0)}$ and V_0 by uniform randomization

OUTPUT:
a) Model Parameter U_* and V_*

3 Practical Issues

Due to limitation of acquired data, our first experiment is oriented to problem of application prediction. Specifically, given a set of featured users and featured jobs, the designed system should predict whether one user will apply for one particular job.

3.1 One-class Matrix Completion with Bias

Bias. Bias parameter $\alpha \in (0, 1)$ is the weight to balance optimization objective between observed and missing labels. Note that with unbiased version of inductive one-class matrix completion, this parameter should be set to 1, that is $\alpha = 1$.

Initialization. Randomizing all components of $U_{(0)}$ and $V_{(0)}$ to uniform distribution over $(0, 1)$ works well in our experimentation.

Optimization. During each stage of alternating minimization, standarded conjugate gradient method is employed to solve least square problem.

4 Experiments

4.1 Dataset and Preprocessed Data

4.2 Models for Comparison

(I) One-Class Conventional Matrix Completion

(II) One-Class Inductive Matrix Completion

(III) One-Class Inductive Matrix Completion with Biases

(IV) (TBD) One-Class Inductive Matrix Completion with Kernel Methods

(V) (TBD) One-Class Inductive Matrix Completion with Pre-Clustering

4.3 Results

Figure 1: Precision v.s. Recall Comparison between (I) (II) (III) with various number of topics (a) $K = K_1$, (b) $K = K_2$, (c) $K = K_3$, (d) $K = K_4$.

Figure 2: Precision v.s. Recall Comparison between models with $\lambda = \lambda_0, \lambda_1, \lambda_2, \lambda_3$

Figure 3: Precision v.s. Recall Comparison between various models with different preprocessed data

5 Conclusions

References

[1] Prateek Jain and Inderjit S Dhillon. Provable inductive matrix completion. *arXiv preprint arXiv:1306.0626*, 2013. 1

[2] Nagarajan Natarajan and Inderjit S Dhillon. Inductive matrix completion for predicting gene–disease associations. *Bioinformatics*, 30(12):i60–i68, 2014. 1