CSE 8803RS: Recommendation Systems

Maximum-Margin Matrix Factorization

Prashant Gaurav

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

Collaborative Prediction

Predict the unobserved entries of target matrix based on the subset of observed entries

Problem Definition

- Y is mxn matrix with m users ratings about n movies s.t. $y_{ij} = +1$ if user i likes movie j, and $y_{ij} = -1$ if he/she dislikes it. Y is partially observed and other entries are missing.
- The main goal is to find matrix X such that it predicts the value of its unknown entries based on the observed values and no other external information.

Matrix Factorization

- Fit matrix $X = UV^T$ to the observed entries such that the rank of each factor (U and V) is low.
  — Minimize the loss function versus a partially observed matrix
  — Use X to predict the unobserved entries
- Problems with minimizing loss over low-rank matrices
  — non-convex optimization problem
  — multiple local minima possible
- Instead use *Frobenius Norm* as the regularization term
  — $||X||_F = \sum x_{ij}^2 = tr(XX^T)$

Matrix Factorization as Linear Classification Problem

- As per problem definition, classify each entry of a matrix into either 1 or -1
- Suppose U is fixed, then fitting each column is a linear classification problem
  — each row of U is a feature vector
  — each column of V^T is a linear classifier
- In collaborative prediction, both U and V are unknown.
  — Learning features (rows in U) across all classifiers(columns of V^T) concurrently

Maximum-Margin Matrix Factorization

- Recall that in SVM maximizing the margin M is equivalent to minimizing the L_2 norm $||\beta||^2$ of the linear classifier.
- The problem addressed here (collaborative prediction) requires to predict U and V together.
  — When U is fixed, each column of V^T is SVM
  — So, predicting X with maximum margin is equivalent to minimizing the $||V||_F$ and $||U||_F$ together.

Optimization Problem and Trace norm

$minimize_{X=UV^T}(||U||_F^2 + ||V||_F^2) + C\sum_{ij \in S} h(Y_{ij}, X_{ij})$,
where C is a trade-off constant.

Lemma 1

$||X||_{\sum} = min_{X=UV^T}||U||_{Fro}||V||_{Fro} = min_{X=UV^T}\frac{1}{2}(||U||_{Fro} + ||V||_{Fro})$
where $||X||_{\sum}$ is trace norm of X and is defined as:
$||X||_{\sum} = \sum|\lambda_i| = Tr(\sqrt{XX^T})$
Based on the Lemma 1, we can rewrite the formulation as
$minimize_X||X||_{\sum} + C\sum_{ij \in S} h(Y_{ij}, X_{ij})$,

Optimization Problem and Trace norm

$minimize_{X=UV^T}(||U||_F^2 + ||V||_F^2) + C\sum_{ij \in S} h(Y_{ij}, X_{ij})$,
where C is a trade-off constant.

Lemma 1

$||X||_{\sum} = min_{X=UV^T}||U||_{Fro}||V||_{Fro} = min_{X=UV^T}\frac{1}{2}(||U||_{Fro} + ||V||_{Fro})$
where $||X||_{\sum}$ is trace norm of X and is defined as:
$||X||_{\sum} = \sum|\lambda_i| = Tr(\sqrt{XX^T})$
Based on the Lemma 1, we can rewrite the formulation as
$minimize_X||X||_{\sum} + C\sum_{ij \in S} h(Y_{ij}, X_{ij})$,

Proof of Lemma 1

- Let $X = UV^T$, and $X = PSQ^T$ is the SVD of X with $S = \mathrm{diag}(\sigma_1, \ldots, \sigma_N)$
- Let i-th row of U and V be u_i and v_i, respectively. Then

$$\sigma_i = u_i v_i^T \leq \|u_i\|_2 \|v_i\|_2$$

and

$$\sigma_1 + \cdots + \sigma_N \leq \|u_1\|_2 \|v_1\|_2 + \cdots + \|u_N\|_2 \|v_N\|_2$$

- The result follows by noticing,

$$\sum_i \|u_i\|_2 \|v_i\|_2 \leq \left(\sum_i \|u_i\|_2^2 \right)^{1/2} \left(\sum_i \|v_i\|_2^2 \right)^{1/2} = \|U\|_F \|V\|_F$$

Experiments

- Preliminary experiments was performed on a subset of the 100K MovieLens Dataset, consisting of the 100 users and 100 movies with the most ratings.
- CSDP was to solve the resulting SDPs.

Limitations

- The observed entries are assumed to be uniformly sampled which is unrealistic. For example, Users tend to rate items they like.
- The current SDP solvers can only handle MMMF problems on matrices of dimentionality of few hundreds.

Fast Maximum Margin Matrix Factorization

- A direct gradient-based optimization method for MMMF
- Suitable for large collaborative prediction problems

Fast Maximum Margin Matrix Factorization

- It is shown that trace-norm is a convex function.
  — minimizing the trace-norm combined with any covex loss funcion is a convex optimization problem.
- Using hinge-loss and a generalization of hinge-loss appropriate for discrete ordinal rating, the optimization problem results as follows :
  — $minimize \, ||X||_{\sum} + C \sum_{ij \in S} \sum_{r=1}^{R-1} h(T_{ij}^{r}(\theta_r - X_{ij}))$
  where $T_{ij}^{r} = \begin{cases} +1 & for\ r \geq Y_{ij} \\ -1 & for\ r < Y_{ij} \end{cases}$
  Here ordinal ratings is taken into account, $Y_{ij} \in 1, 2, \cdots, R$. To relate the real-valued X_{ij} to discrete Y_{ij}, $R-1$ thresholds $\theta_1, \cdots, \theta_{R-1}$ are used.

Fast MMMF Optimization Method

- Original objective
  — $minimize ||X||_{\sum} + C \sum_{ij \in S} h(Y_{ij}, X_{ij})$
  — complicated and non-differentiable
  — finding good descent direction is not easy
- Factorized objective
  — $minimize \frac{1}{2}(||U||_{Fro}^2 + ||V||_{Fro}^2) + C \sum_{ij \in S} h(Y_{ij}, U_i V_j^T)$
  — For smooth optimization function, Smooth Hinge is used instead of the Hinge loss.
  — gradient is easy to compute, we can use gradient descent method

Experiments

- Experiments were conducted on Movielens (1M ratings) and EachMovie(2.6M ratings) data sets.
- Tests were conducted were of both types - Weak Generalization and Strong Generalization.

Summary

- MMMF can be scaled to large problems by optimizing the Factorized Objective
- Empirical analysis shows that local minima are rare.

References

- Maximum Margin Matrix Factorization. Nathan Srebro, Jason Rennie and Tommi Jaakkola
  Advances in Neural Information Processing Systems (NIPS) 17, 2005 (December 2004 conference)
- Fast Maximum Margin Matrix Factorization for Collaborative Prediction. Jason Rennie and Nathan Srebro
  22nd International Conference on Machine Learning (ICML), August 2005.
- Learning with Matrix Factorizations. Nathan Srebro
  PhD Thesis, Massachusetts Institute of Technology, August 2004.