Collaborative Filtering via Rating Concentration

Bert Huang and Tony Jebara, Computer Science Department
Columbia University, {bert, jebara}@cs.columbia.edu

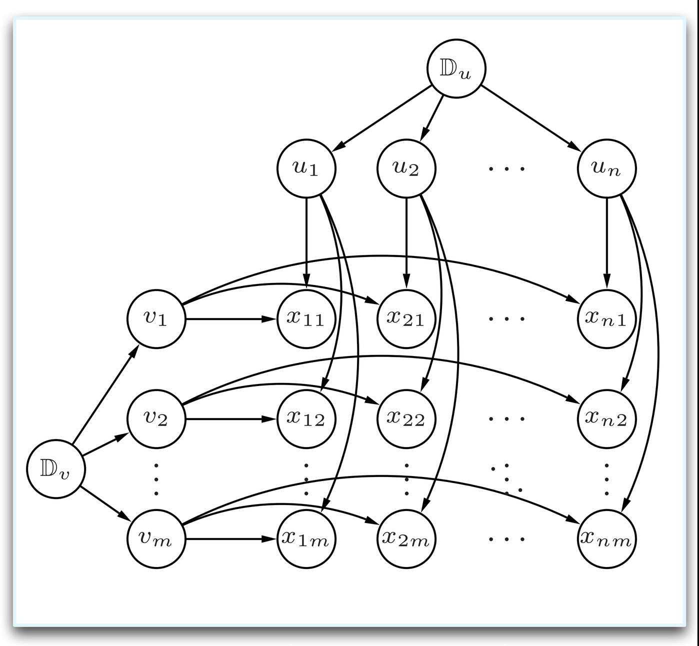
Overview

- Goal: predict users ratings for items from observed training ratings.
- Given simple assumptions about sampling process, rating statistics must concentrate.
- Expected averages of predicted rating statistics must be close to empirical averages.
- Enforcing only concentration, least informative (maximum entropy) distribution yields state-of-the-art performance
- No need for low-rank or spectral assumptions

Sampling Assumptions

- Assume users u and items v are sampled *iid* from stationary distributions.
- Ratings x are then sampled from distributions dependent on the rating user and rated item.

$$\prod_{ij} p(x_{ij}|u_i, v_j)p(u_i)p(v_j)$$



- Given training ratings, $\{x_{ij}|(i,j) \in T\}$ predict query rating probabilities $\{p(x_{ij}|u_i, v_j)|(i,j) \in Q\}$
- Typically, the priors are parameterized, as in fMMMF (RenSre05), PMF + BPMF (SalMni08)

Rating Features

- Consider bounded functions of rating values,

$$f_k(x) \mapsto [0,1] \quad \text{e.g.,} \quad \begin{aligned} f_i(x) &= I(x = i), \\ f_{i,j}(x) &= I(x = i \vee x = j). \end{aligned}$$

- How much can the empirical averages

$$\mu_{ik} = \frac{1}{m_i} \sum_{j|(i,j)\in T} f_k(x_{ij}), \quad \nu_{jk} = \frac{1}{n_j} \sum_{i|(i,j)\in T} f_k(x_{ij})$$

deviate from the expected averages over query ratings?

$$\frac{1}{\hat{m}_i} \sum_{j|(i,j)\in Q} \mathbb{E}_{p(x_{ij}|u_i,v_j)}[f_k(x_{ij})],$$

$$\frac{1}{\hat{n}_j} \sum_{i|(i,j)\in Q} \mathbb{E}_{p(x_{ij}|u_i,v_j)}[f_k(x_{ij})]$$

Selected References

- Dudík, M., Blei, D., & Schapire, R. (2007) Hierarchical maximum entropy density estimation. ICML
- Rennie, J., & Srebro, N. (2005). Fast maximum margin matrix factorization for collaborative prediction. ICML
- Marlin, B., Zemel, R., Roweis, S., & Slaney, M. (2007) Collaborative filtering and the missing at random assumption. UAI
- Salakhutdinov, R., & Mnih, A. (2008). Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. ICML

Rating Feature Concentration

Theorem: For the ratings of user i, the difference

$$\epsilon_{ik} = \mu_{ik} - \frac{1}{\hat{m}_i} \sum_{j|(i,j)\in Q} \mathbb{E}_{p(x_{ij}|u_i,v_j)}[f_k(x_{ij})]$$

between the average of $f_k(x) \mapsto [0,1]$ over the observed ratings and the average of the expected value of $f_k(x)$ over the query ratings is bounded above by

$$\epsilon_{ik} \leq \sqrt{\frac{\ln \frac{2}{\delta}}{2m_i}} + \sqrt{\frac{(m_i + \hat{m}_i)\ln \frac{2}{\delta}}{2m_i \hat{m}_i}}$$

with probability $1 - \delta$.

- Similar bound for item ratings. Both are also bounded above and below.

Maximum Entropy Method

- For fixed confidence value, concentration bounds form linear constraints on probability
- Predict least informative distribution by minimizing KL to simple prior s.t. constraints

$$\max_p \sum_{ij\in Q} H(p_{ij}(x_{ij})) + \sum_{ij\in Q, x_{ij}} p_{ij}(x_{ij}) \ln p_0(x_{ij})$$

$$\text{s.t.} \quad \left| \frac{1}{\hat{m}_i} \sum_{j|ij\in Q} \sum_{x_{ij}} p_{ij}(x_{ij}) f_k(x_{ij}) - \mu_{ik} \right| \leq \alpha_i, \forall i,k$$

$$\left| \frac{1}{\hat{n}_j} \sum_{i|ij\in Q} \sum_{x_{ij}} p_{ij}(x_{ij}) f_k(x_{ij}) - \nu_{jk} \right| \leq \beta_j, \forall j,k.$$

- Set α_i and β_j constraints using concentration bounds with fixed confidence δ
- Solve dual form using LBFGS

Synthetic Experiment

- Real data doesn't include true rating probabilities
- Generate user and item vectors $u_i, v_j \in [0,1]^5$ and draw ratings from multinomial

$$p(x_{ij} = r|u_i, v_j) \propto u_i(r)v_j(r)$$

- Draw 100k ratings btw. 500 users and 500 items, split in half training, half query
- With true distributions, can compare KL-divergence against logistic threshold likelihood from fast Max-Margin Matrix Factorization

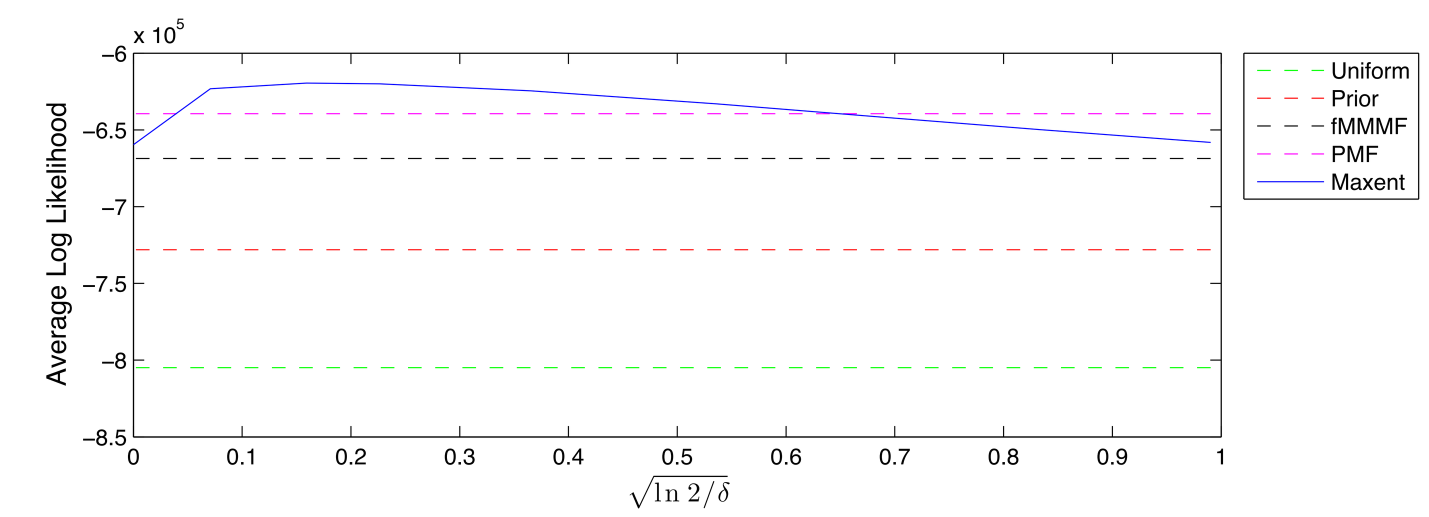
	fMMMF	Maxent
Log-Likelihood	-39690 ± 214	$\mathbf{-35732 \pm 216}$
KL-divergence	11254 ± 315	$\mathbf{4954 \pm 154}$

Movielens Experiment

- Movielens-million movie rating data set:
- 1m ratings by 6k+ users, for 3k+ movies, with three random splits of half training, half query ratings
- Compare likelihood with fMMMF (logistic threshold likelihood), Probabilistic Matrix Factorization (PMF) (discretized Gaussian likelihood)

	Uniform	Prior	fMMMF Distrib.	PMF	Maxent
Split 1	-8.0489e+05	-7.2800e+05	-6.6907e+05	-6.3904e+05	-6.1952e+05
Split 2	-8.0489e+05	-7.2796e+05	-6.6859e+05	-6.3936e+05	-6.1977e+05
Split 3	-8.0489e+05	-7.2809e+05	-6.6819e+05	-6.3987e+05	-6.1931e+05
Average	-8.0489e+05	-7.2802e+05	-6.6862e+05	-6.3942e+05	**-6.1953e+05**

Log-likelihood



Log-likelihood for various confidence parameters

- Compare RMS error with fMMMF, PMF, Bayesian PMF and some simple combinations of different algorithms

	fMMMF	PMF	Maxent	BPMF	Maxent+fMMMF	Maxent+PMF
Split 1	0.9585	0.9166	0.9168	0.8717	0.9079	0.8963
Split 2	0.9559	0.9175	0.9162	0.8710	0.9052	0.8965
Split 3	0.9583	0.9186	0.9166	0.8723	0.9065	0.8973
Average	0.9575	0.9176	0.9165	0.8717	0.9065	0.8967

Root mean squared error

Summary

- Concentration guarantees on functions over ratings predict with state-of-the-art performance
- Requires sampling assumptions but guarantees hold for any arbitrary probability functions; needs no parametric assumptions
- Future work: concentration constraints in parametric models may yield further improvement

Notation Glossary

Symbol	Meaning
x_{ij}	rating of user i for item j
u_i, v_j	user i and item j
m_i, n_j	counts of training ratings for user i and item j
$\hat{m}_i$, $\hat{n}_j$	counts of query ratings for user i and item j
T, Q	**t**raining and **q**uery rating sets
f_k	k'th bounded function of ratings
μ_{ik}, ν_{jk}	empirical averages of f_k for user i,item j
δ	failure probability of bound, or confidence param.
H	entropy
p_0	simple max-likelihood rating prior
α_i, β_j	allowed deviation of estimated expectation from empirical averages