A Gradient-based Framework for Personalization

October 23, 2017

Liangjie Hong
Head of Data Science, Etsy Inc.

Liangjie Hong

- **Head of Data Science**
  - **Etsy Inc.** in NYC, NY (2016. – Present)
  - Search & Discovery; Personalization and Recommendation; Computational Advertising

- **Senior Manager of Research**
  - **Yahoo Research** in Sunnyvale, CA (2013 – 2016)

  Leading science efforts for personalization and search sciences

- Published papers in **SIGIR**, **WWW**, **KDD**, **CIKM**, **AAAI**, **WSDM**, **RecSys** and **ICML**

- **WWW 2011 Best Poster Paper Award**
  WSDM 2013 Best Paper Nominated
  RecSys 2014 Best Paper Award

- Program committee members in **KDD**, **WWW**, **SIGIR**, **WSDM**, **AAAI**, **EMNLP**, **ICWSM**, **ACL**, **CIKM**, **IJCAI** and various journal reviewers

- PhD in Computer Science from Lehigh University (2013)

About This Paper

- Authors
  Yue Ning, PhD Student from **Virginia Tech**
  Yue Shi, Research Scientist at **Facebook**
  Liangjie Hong, Head of Data Science at **Etsy Inc.**
  Huzefa Rangwala, Associate Professor at **George Mason University**
  Naren Ramakrishnan, Professor at **Virginia Tech**

- Paper Venue
  Full Research Paper in The 11th ACM Conference on Recommender Systems (**RecSys'17**)

Challenges in Personalized Recommender Systems

Challenges in Personalized Recommender Systems

- **“Average” Experiences for Users**

Challenges in Personalized Recommender Systems

- **“Average” Experiences for Users**

  1) Global objective functions

  2) Biased towards heavy features

Challenges in Personalized Recommender Systems

- **“Average” Experiences for Users**

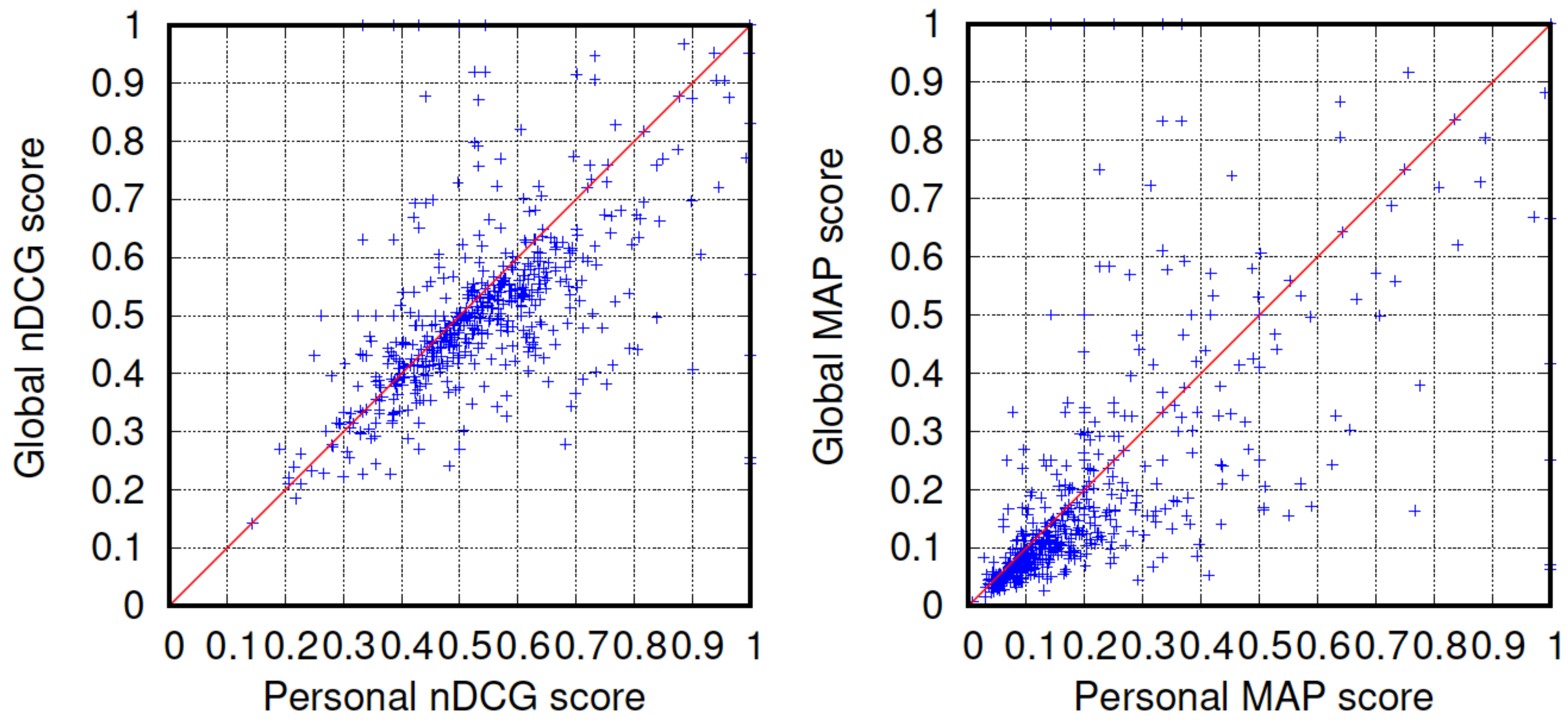


Figure 1: An example of global and personal models. Left figure showcases the nDCG score of users from global (y-axis) and personal (x-axis) models. (Right: MAP score).

Challenges in Personalized Recommender Systems

- **Lack of A Generic Framework for Personalization**

Challenges in Personalized Recommender Systems

- **Lack of A Generic Framework for Personalization**
  1) Beutel et al. **Beyond Globally Optimal: Focused Learning for Improved Recommendations**. WWW 2017.
  2) Zhang et al. **Generalized Linear Mixed Models For Large-Scale Response Prediction**. KDD 2016.
  3) Miao et al. **Distributed Personalization**. KDD 2015.

Challenges in Personalized Recommender Systems

- **Distributed Model Learning Requires Accessing Global Data**

Challenges in Personalized Recommender Systems

- **Distributed Model Learning Requires Accessing Global Data**
  1) Needs to access global data
  2) Sophisticated learning framework

Challenges in Personalized Recommender Systems

- **“Average” Experiences for Users**
- **Lack of A Generic Framework for Personalization**
- **Distributed Model Learning Requires Accessing Global Data**

Proposed Framework

A Gradient-based Adaptive Learning Framework

System Framework

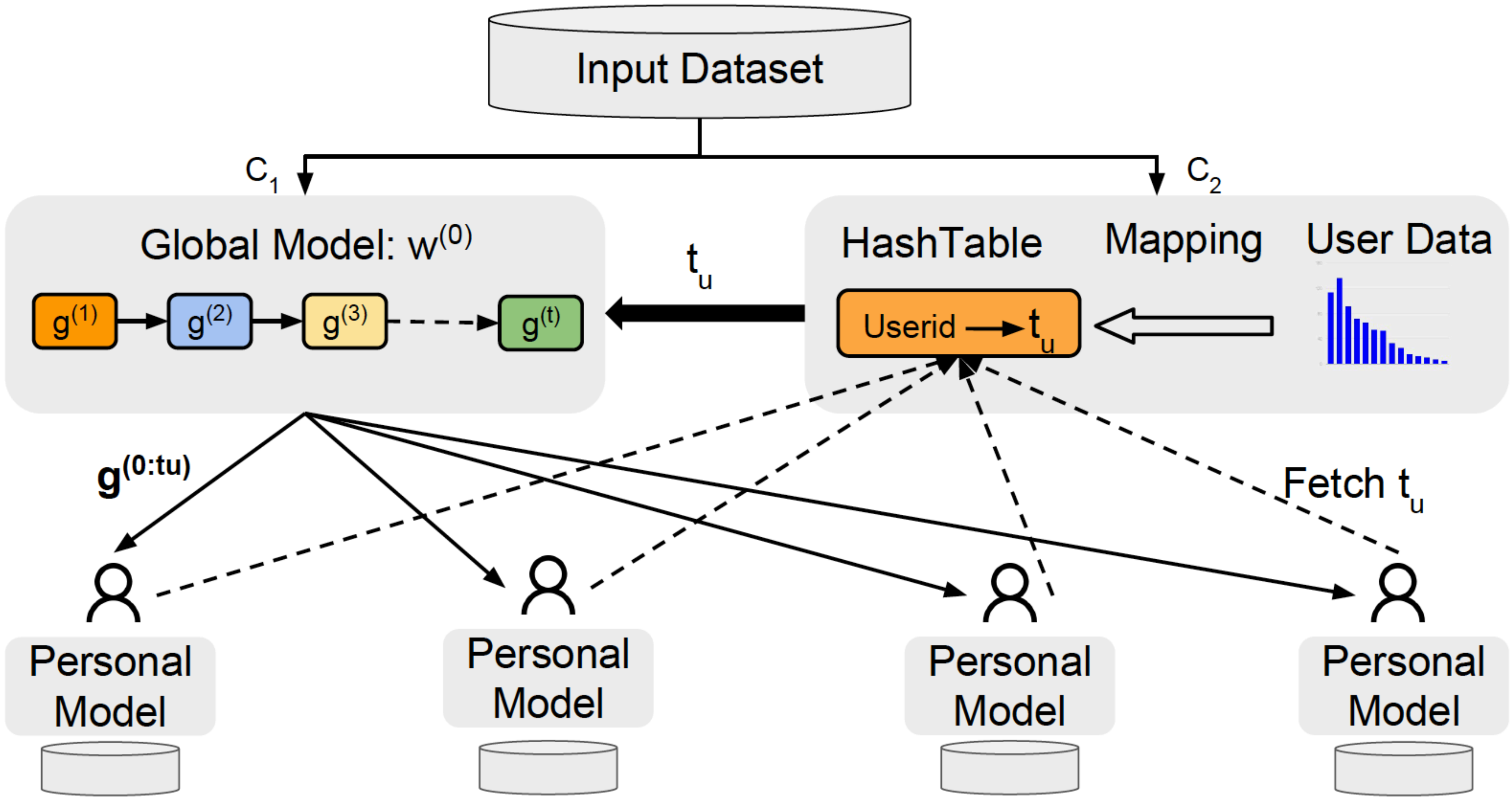


Figure 2: System Framework. Component C_1 trains a global model. Component C_2 generates a hashtable based on users' data distribution. Users request t_u from C_2 and C_1 returns a subsequence of gradients $g^{(0:t_u)}$ to users.

A Gradient-based Adaptive Learning Framework

Adaptation Mechanism

Global update $\rightarrow$

$$\boldsymbol{\theta}^{(T)} = \boldsymbol{\theta}^{(0)} - \eta \sum_{t=1}^{T} g^{(t)}(\boldsymbol{\theta})$$

Local update $\rightarrow$

$$\widetilde{\boldsymbol{\theta}}_u = \boldsymbol{\theta}^{(0)} - \eta_1 \sum_{t=1}^{t_u - 1} g^{(t)}(\boldsymbol{\theta}) - \eta_2 \sum_{t=t_u}^{T} g^{(t)}(\boldsymbol{\theta}_u)$$

- $\boldsymbol{\theta}$: the global model parameter.
- $\boldsymbol{\theta_u}$: the personal model parameter.
- u: the index for one user.
- t_u: the index of global gradients for user u.
- $g^{(t)}(\boldsymbol{\theta})$: global gradients
- $g^{(t)}(\boldsymbol{\theta}_u)$: personal gradients

A Gradient-based Adaptive Learning Framework

How do we choose the index?

- Group users into C groups based on their data sizes in descending order.
- Decide the position $p_u = \frac{i}{C}$,
  - C is # groups.
  - i is the group assignment for user u.
  - the first group (i=1) of users has the most data.
- Set $t_u = \lfloor T * p_u \rfloor$
  - T: total iterations in the global SGD algorithm
  - Users with the most data have the earliest stop for global gradients.

A Gradient-based Adaptive Learning Framework

Adaptive Logistic Regression

Objective:

$$\min_{\mathbf{w}} L(\mathbf{w}) = f(\mathbf{w}) + \lambda r(\mathbf{w}) \tag{1}$$

- $f(\mathbf{w})$ is the negative log-likelihood.
- $r(\mathbf{w})$ is a regularization function.

Adaptation Procedure:

- Global update $\rightarrow$

$$\widetilde{\mathbf{w}}_u^{(0)} = \mathbf{w}^{(0)} - \eta_1 \sum_{t=1}^{t_u-1} g^{(t)}(\mathbf{w}) \tag{2}$$

- Local update $\rightarrow$

$$\widetilde{\mathbf{w}}_u^{(T)} = \widetilde{\mathbf{w}}_u^{(0)} - \eta_2 \sum_{t=1}^{T-t_u} g^{(t)}(\mathbf{w}_u) \tag{3}$$

A Gradient-based Adaptive Learning Framework

Adaptive Gradient Boosting Decision Tree

Objective:

$$\begin{aligned} L^{(t)} &= \sum_{d}^{N} l(y_d, F_d^{(t-1)} + \rho h^{(t)}) + \Omega(h^{(t)}) \\ &= \sum_{d}^{N} l(y_d, F_d^{(0)} + \rho h^{(0:t)}) + \Omega(h^{(t)}) \end{aligned} \tag{4}$$

Adaptation Procedure:

$$\widetilde{F}_u^{(0)} = F^{(0)} + \rho h^{(0:t_u)} \tag{5}$$

$$\widetilde{F}_u^{(T)} = \widetilde{F}_u^{(0)} + \rho h_u^{(t_u:T)} \tag{6}$$

A Gradient-based Adaptive Learning Framework

Adaptive Matrix Factorization

Objective:

$$\min_{q_*,p_*,b_*} \sum_{u,i} (r_{ui} - \mu - b_u - b_i - \mathbf{q}_u^T \mathbf{p}_i) + \lambda(||\mathbf{q}_u||^2 + ||\mathbf{p}_i||^2 + b_u^2 + b_i^2) \quad (7)$$

Adaptation Procedure:

$$\widetilde{\mathbf{q}}_u^{(0)} = \mathbf{q}_u^{(0)} - \eta_1 \sum_{t=0}^{t_u} g^{(t)}(\mathbf{q}_u), \widetilde{\mathbf{q}}_u^{(T)} = \widetilde{\mathbf{q}}_u^{(0)} - \eta_2 \sum_{t=0}^{T-t_u} g^{(t)}(\widetilde{\mathbf{q}}_u) \quad (8)$$

$$\widetilde{b}_u^{(0)} = b_u^{(0)} - \eta_1 \sum_{k=0}^{t_u} g^{(t)}(b_u), \widetilde{b}_u^{(T)} = \widetilde{b}_u^{(0)} - \eta_2 \sum_{t=0}^{T-t_u} g^{(t)}(\widetilde{b}_u) \quad (9)$$

A Gradient-based Adaptive Learning Framework

Properties

- **Generality**: The framework is generic to a variety of machine learning models that can be optimized by gradient-based approaches.
- **Extensibility**: The framework is extensible to be used for more sophisticated use cases.
- **Scalability**: In this framework, the training process of a personal model for one user is independent of all the other users.

Experiments

Experiments

Datasets

Table: Dataset Statistics

News Portal	
# users	54845
# features	351
# click events	2,378,918
# view events	26,916,620
avg # click events per user	43
avg # events per user	534

Movie Ratings		
	Netflix	Movielens
# users	478920	1721
# items	17766	3331
sparsity	0.00942	0.039

- For LogReg and GBDT: News Portal dataset
- For Matrix Factorization: Movie rating datasets (Netflix, Movielens)

Experiments

Metrics

- MAP: Mean Average Precision.
- MRR: Mean Reciprocal Rank.
- AUC: Area Under (ROC) Curve.
- nDCG: Normalized Discounted Cumulative Gain.
- RMSE: Root Mean Square Error
- MAE: Mean Absolute Error

Experiments

Comparison Methods

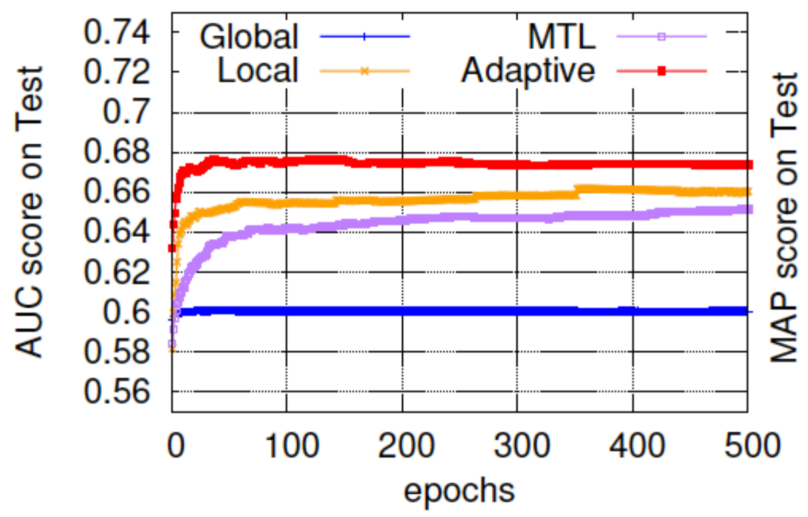
Table: Objective functions for different methods.

Model	LogReg
Global	$\sum_{d=1}^{N} f(\mathbf{w}) + \lambda\|\|\mathbf{w}\|\|_2^2$
Local	$\sum_{j=1}^{N_u} f(\mathbf{w}_u) + \lambda\|\|\mathbf{w}_u\|\|_2^2$
MTL	$\sum_{j}^{N_u} f(\mathbf{w}_u) + \frac{\lambda_1}{2}\|\|\mathbf{w}_u - \mathbf{w}\|\|^2 + \frac{\lambda_2}{2}\|\|\mathbf{w}_u\|\|^2$
Model	**GBDT**
Global	$\sum_{d}^{N} l(y_d, F_d^{(0)} + \rho h^{(0:t)}) + \Omega(h^{(t)})$
Local	$\sum_{j}^{N_u} l(y_j, F_j^{(0)} + \rho h^{(0:t)}) + \Omega(h^{(t)})$
MTL	-
Model	**MF**
Global	$\sum_{u,i}(r_{ui} - \mu - b_u - b_i - \mathbf{q}_u^T \mathbf{p}_i) + \lambda(\|\|\mathbf{q}_u\|\|^2 + \|\|\mathbf{p}_i\|\|^2 + b_u^2 + b_i^2)$
Local	$\sum_{i \in N_u}(r_{ui} - \mu - \widetilde{b}_u - \widetilde{b}_i - \widetilde{\mathbf{q}}_u^T \widetilde{\mathbf{p}}_i) + \lambda(\|\|\widetilde{\mathbf{q}}_u\|\|^2 + \|\|\widetilde{\mathbf{p}}_i\|\|^2 + \widetilde{b}_u^2 + \widetilde{b}_i^2)$
MTL	global$+\lambda_2[(\mathbf{q}_u - \mathbf{q})^2 + (\mathbf{p}_i - \mathbf{p})^2 + (b_u - A_u)^2 + (b_i - A_i)^2]$

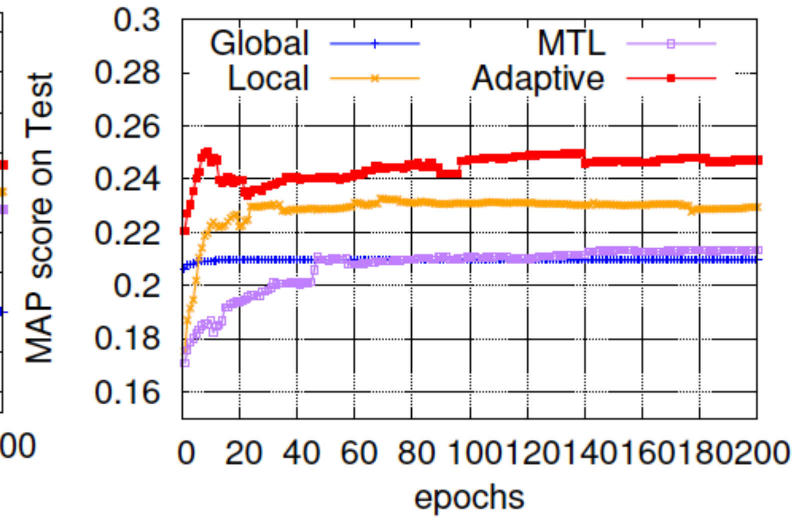
- Global: models are trained on all users' data
- Local: models are learned locally on per user's data
- MTL: users models are averaged by a global parameter.

Experiments

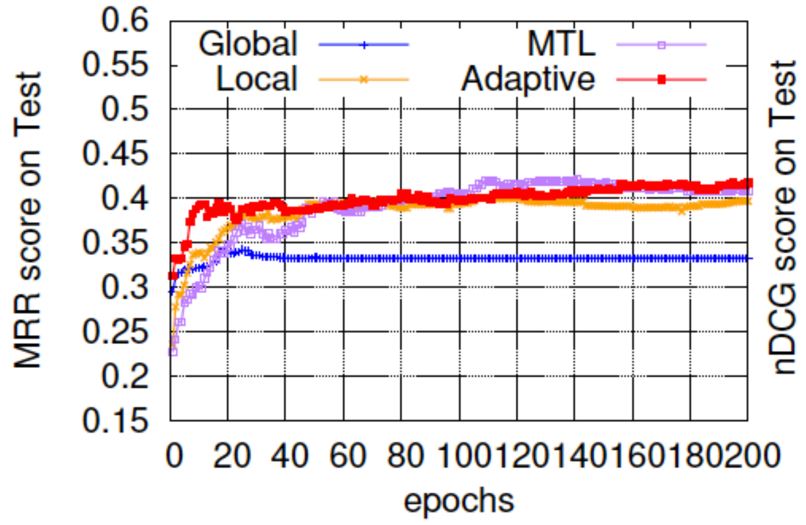
Ranking Performance – Logistic Regression



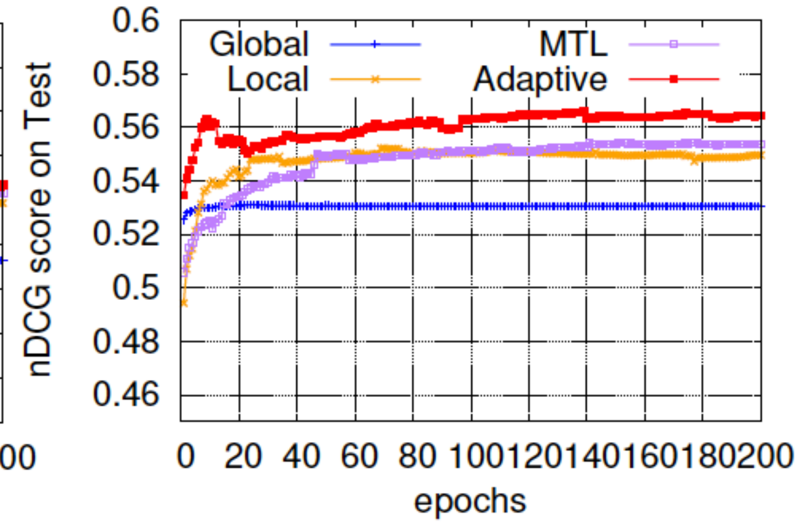
(a) AUC



(b) MAP



(c) MRR



(d) nDCG

- AUC, MAP, MRR and nDCG scores on the test dataset with varying training epochs.
- The proposed adaptive LogReg models achieve higher scores with fewer epochs.
- Global models perform the worst.

Experiments

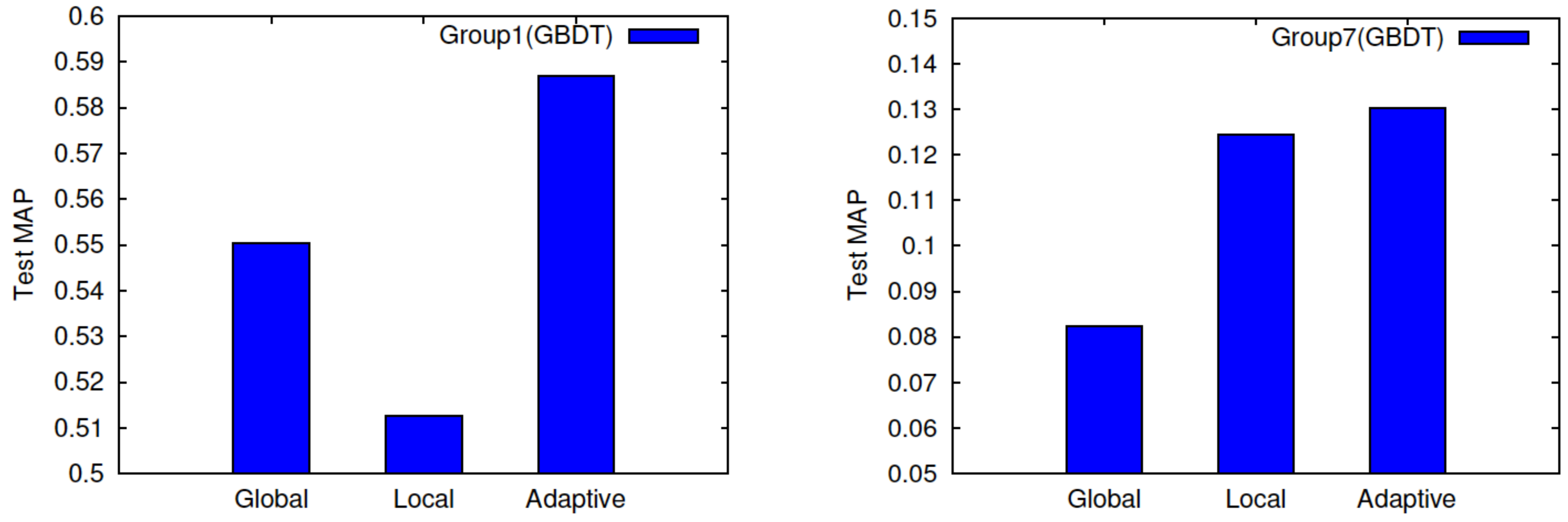
Ranking Performance – GBDT

Table: Performance comparison based on MAP, MRR, AUC and nDCG for GBDT. Each value is calculated from the average of 10 runs with standard deviation.

	Global-GBDT			
#Trees	MAP	MRR	AUC	nDCG
20	0.2094(1e-3)	0.3617(2e-3)	0.6290(1e-3)	0.5329(6e-4)
50	0.2137(1e-3)	0.3726(1e-3)	0.6341(1e-3)	0.5372(6e-4)
100	0.2150(8e-3)	0.3769(1e-3)	0.6356(8e-4)	0.5392(6e-4)
200	0.2161(5e-4)	0.3848(1e-3)	0.6412(6e-4)	0.5415(5e-4)
	Local-GBDT			
#Trees	MAP	MRR	AUC	nDCG
20	0.2262(2e-3)	0.4510(5e-3)	0.6344(3e-3)	0.5604(2e-3)
50	0.2319(2e-3)	0.4446(4e-3)	0.6505(2e-3)	0.5651(2e-3)
100	0.2328(1e-3)	0.4465(5e-3)	0.6558(2e-3)	0.5651(2e-3)
200	0.2322(2e-3)	0.4431(2e-3)	0.6566(1e-3)	0.5649(1e-3)
	Adaptive-GBDT			
#Trees	MAP	MRR	AUC	nDCG
20+50	**0.2343**(2e-3)	0.4474(4e-3)	0.6555(2e-3)	0.5661(2e-3)
50+50	0.2325(2e-3)	0.4472(1e-4)	0.6561(8e-4)	**0.5666**(6e-4)
10+100	0.2329(2e-3)	0.4423(3e-3)	**0.6587**(1e-3)	0.5650(3e-3)

Experiments

Ranking Performance – GBDT



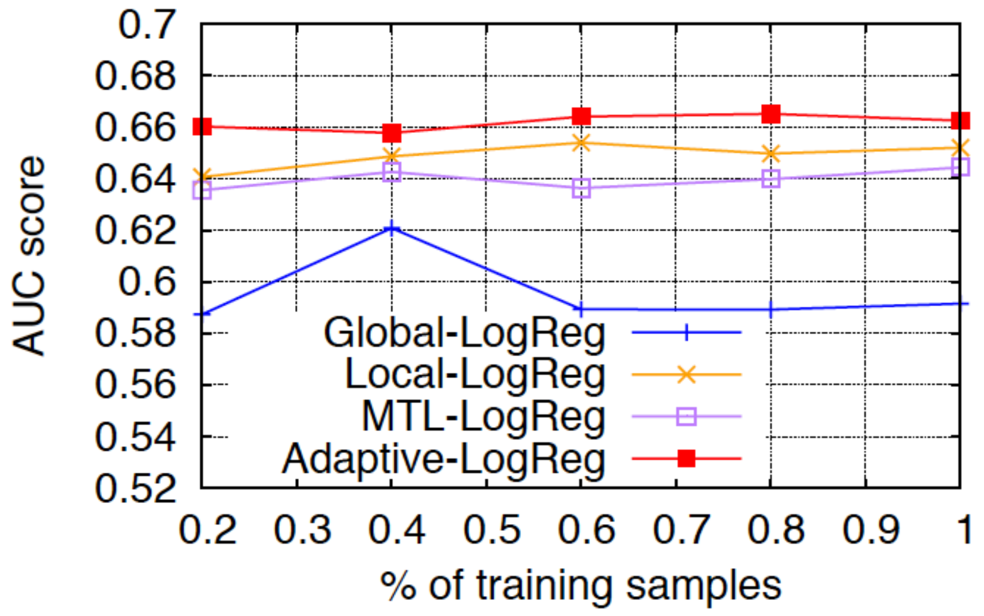
(a) Group 1 (b) Group 7

Figure: MAP Comparison of Group 1 (least) and Group 7 (most) for GBDT methods.

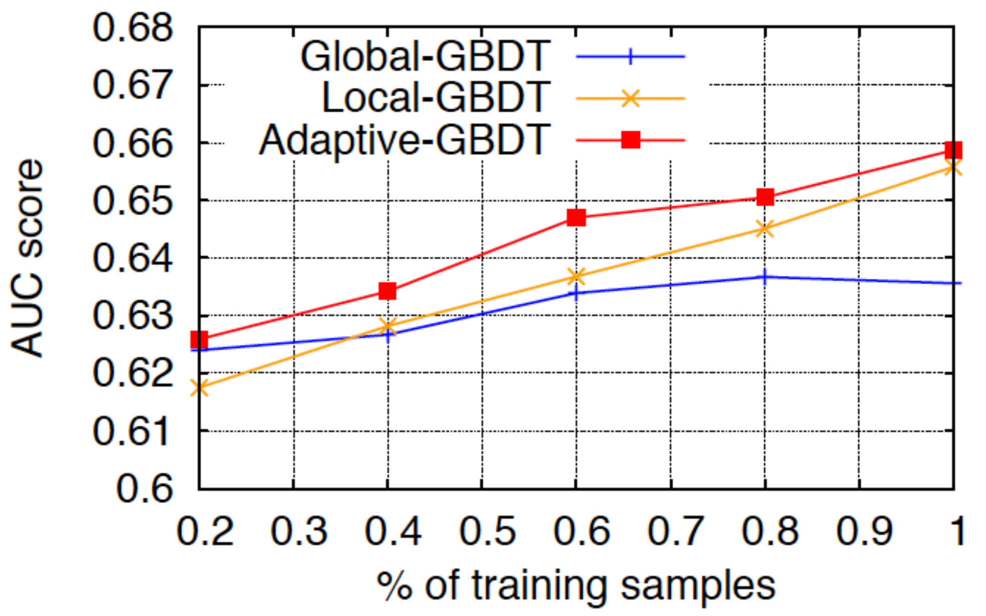
- MAP score for the groups of users with least data (Group 1) and most data (Group 7) for GBDT models.
- Adaptive-GBDT *outperform* both global and local GBDT models in terms of MAP for all groups of users.

Experiments

Ranking Performance – Logistic Regression v.s. GBDT



(a) LogReg

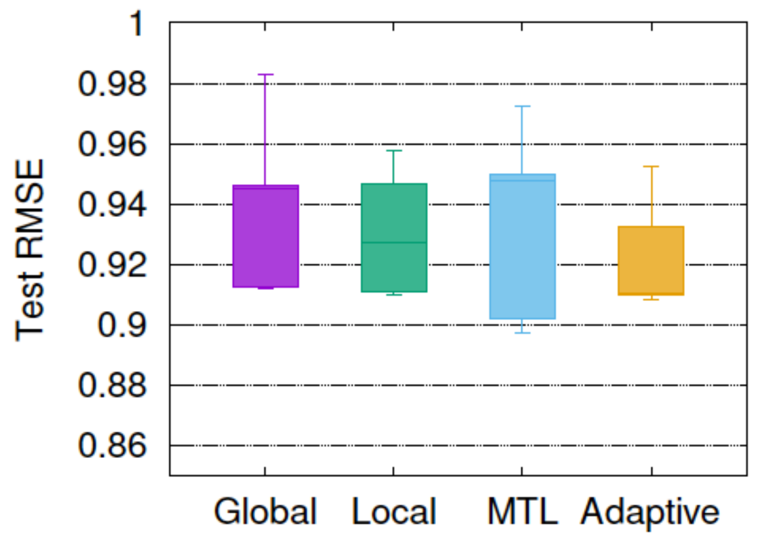


(b) GBDT

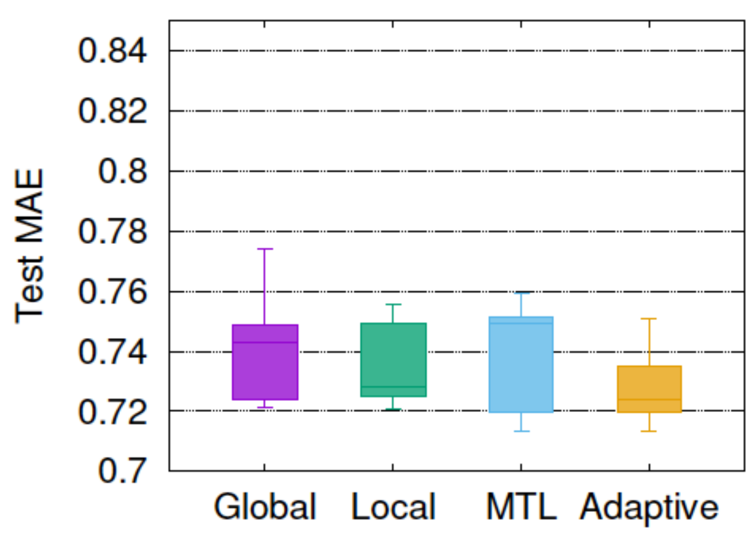
- AUC score for Global-GBDT, Local-GBDT, and Adaptive-GBDT with # of training samples from 20% to 100%.
- On average of AUC, Adaptive-GBDT performs better than other methods.
- With the increase of training samples, GBDT based methods tend to perform better while LogReg methods achieve relatively stable scores.

Experiments

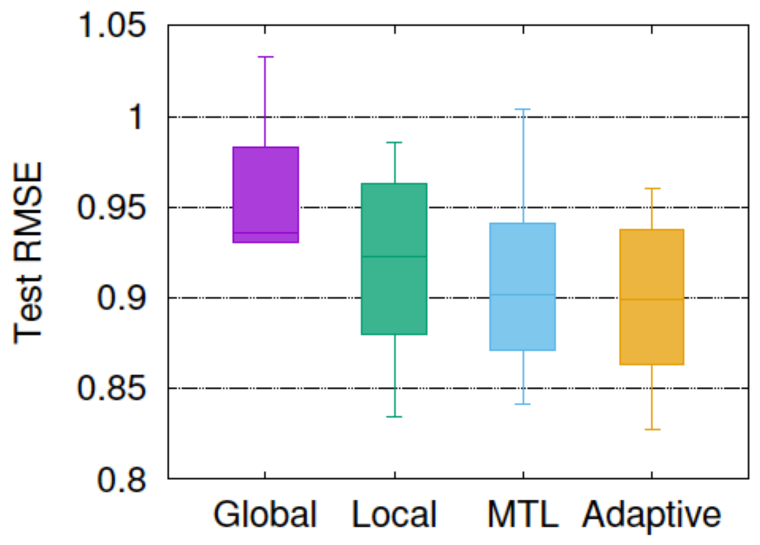
Results – Matrix Factorization



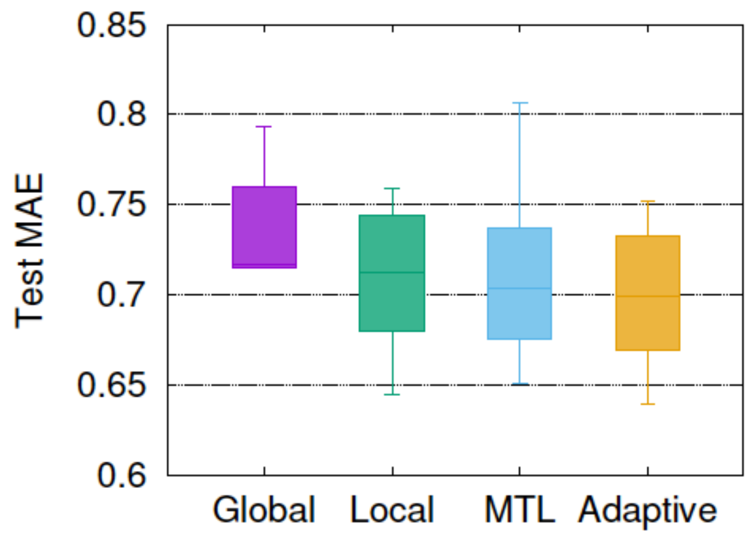
(a) ML-RMSE



(b) ML-MAE



(c) Netflix-RMSE



(d) Netflix-MAE

- RMSE and MAE on MovieLens(ML) and Netflix datasets.
- The quartile analysis of the group level RMSE and MAE for different MF models.
- Gold: Adaptive-MF

Summary

- *Effectively and efficiently* build personal models that lead to improved recommendation performance over either the global model or the local model.
- Adaptively learn personal models by **exploiting the global gradients** according to **individuals characteristic**.
- Our experiments demonstrate the usefulness of our framework across a wide scope, in terms of both model classes and application domains.

Questions