Bayesian Personalized Feature Interaction Selection for Factorization Machines

Yifan Chen[1,2] Pengjie Ren[1] Yang Wang[3] Maarten de Rijke[1]

[1]University of Amsterdam, Amsterdam, the Netherlands

[2]Key Laboratory of Science and Technology on Information System Engineering, National University of Defense Technology, Changsha, China

[3]Hefei University of Technology, Hefei, China

SIGIR 2019

Factorization Machines

What is Factorization Machine?

- A generic supervised learning method
- Account for feature interactions with factored parameters
  - the combination of features

#Hashtag	Feature combinations
"comics"	("comics" , "marvel")
"marvel"	("comics" , "avengers")
"avengers"	

Factorization Machines

- Linear regression: $O(d)$

$$\hat{r}(\boldsymbol{x}) = b_0 + \sum_{i=1}^{d} w_i x_i$$

Factorization Machines

- Linear regression: $O(d)$

$$\hat{r}(\boldsymbol{x}) = b_0 + \sum_{i=1}^{d} w_i x_i$$

- Degree-2 polynomial regression: $O(d^2)$

$$\hat{r}(\boldsymbol{x}) = b_0 + \sum_{i=1}^{d} w_i x_i + \sum_{i=1}^{d} \sum_{j=i+1}^{d} w_{ij} \cdot x_i x_j$$

SIGIR 2019

Factorization Machines

- Linear regression: $O(d)$

$$\hat{r}(\boldsymbol{x}) = b_0 + \sum_{i=1}^{d} w_i x_i$$

- Degree-2 polynomial regression: $O(d^2)$

$$\hat{r}(\boldsymbol{x}) = b_0 + \sum_{i=1}^{d} w_i x_i + \sum_{i=1}^{d} \sum_{j=i+1}^{d} w_{ij} \cdot x_i x_j$$

- Factorization machine: $O(dk)$

$$\hat{r}(\boldsymbol{x}) = b_0 + \sum_{i=1}^{d} w_i x_i + \sum_{i=1}^{d} \sum_{j=i+1}^{d} \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle \cdot x_i x_j$$

Factorization Machines

Example

$$\hat{r}(\text{spider-man}) = b_0 + w_{\text{comics}} + w_{\text{marvel}} + w_{\text{avengers}} + \langle \boldsymbol{v}_{\text{comics}}, \boldsymbol{v}_{\text{marvel}} \rangle + \langle \boldsymbol{v}_{\text{comics}}, \boldsymbol{v}_{\text{avengers}} \rangle$$

#Hashtag

"comics"

"marvel"

"avengers"

Feature combinations

("comics" , "marvel")

("comics" , "avengers")

Introduction

Factorization Machines for Recommendation

Feature Interaction Selection

Factorization Machines for Recommendation

- Effective use of **historical interactions** between users and items
- Incorporate **additional information** associated with users or items
- High-dimensional feature space
  - #feature = #user + #item + #additional
  - not all features or feature interactions are helpful

SIGIR 2019

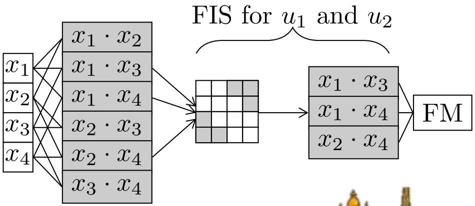
Factorization Machines for Recommendation

- Effective use of **historical interactions** between users and items
- Incorporate **additional information** associated with users or items
- High-dimensional feature space
  - #feature = #user + #item + #additional
  - not all features or feature interactions are helpful

Feature Interaction Selection (FIS)

Filter out useless feature interactions

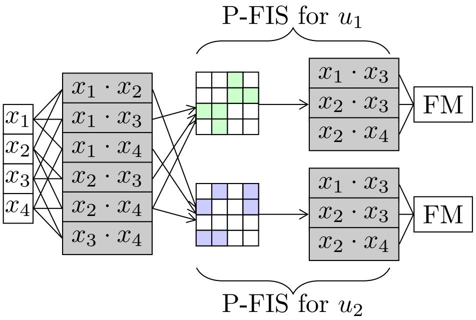
- P-FIS: Select feature interactions for users personally
- FIS: select a common set of interactions



Feature Interaction Selection (FIS)

Filter out useless feature interactions

- P-FIS: Select feature interactions for users personally
- FIS: select a common set of interactions



Personalized Factorization Machines (PFM)

FM

$$\hat{r}(\boldsymbol{x}) = b_0 + \sum_{i=1}^{d} w_i x_i + \sum_{i=1}^{d} \sum_{j=i+1}^{d} w_{ij} \cdot x_i x_j$$

PFM

$$\hat{r}(\boldsymbol{x}) = b_u + \sum_{i=1}^{d} w_{ui} x_i + \sum_{i=1}^{d} \sum_{j=i+1}^{d} w_{uij} \cdot x_i x_j$$

Select 1st-order and 2nd-order interactions by $\{w_{ui}\}$ and $\{w_{uij}\}$

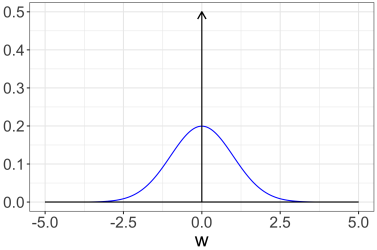
Bayesian Variable Selection (BVS)

- Apply BVS to select feature interactions
  - avoid expensive cross-validation
- Priors for BVS
  - sparsity priors
  - spike-and-slab

SIGIR 2019

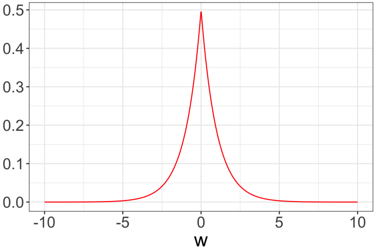
Bayesian Variable Selection

Spike-and-slab



- Spike (black arrow): $p(w = 0) = 0.5$
- Slab (blue line)

Sparsity priors



- $f(w) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$
- $p(w = 0) = 0$

SIGIR 2019

Hereditary Spike-and-Slab Priors

- Spike-and-slab

$$s \sim Bernoulli(\pi), \quad \tilde{w} \sim \mathcal{N}(0,1), \quad w = \tilde{w} \cdot s.$$

- Hereditary spike-and-slab
  - capture the relations between 1st-order and 2nd-order feature interactions

$$\begin{aligned} s_{ui}, s_{uj} &\sim Bernoulli(\pi_1) \\ p(s_{uij} = 1 \mid s_{ui}s_{uj} = 1) &= 1 & \quad (Strong\ heredity) \\ p(s_{uij} = 1 \mid s_{ui} + s_{uj} = 1) &= \pi_2 & \quad (Weak\ heredity) \\ p(s_{uij} = 1 \mid s_{ui} + s_{uj} = 0) &= 0 \end{aligned}$$

Hereditary Spike-and-Slab Priors

- Spike-and-slab

$$s \sim Bernoulli(\pi), \quad \tilde{w} \sim \mathcal{N}(0,1), \quad w = \tilde{w} \cdot s.$$

- Hereditary spike-and-slab
  - capture the relations between 1st-order and 2nd-order feature interactions

$$s_{ui}, s_{uj} \sim Bernoulli(\pi_1)$$

$$p(s_{uij} = 1 \mid s_{ui}s_{uj} = 1) = 1 \qquad (Strong\ heredity)$$

$$p(s_{uij} = 1 \mid s_{ui} + s_{uj} = 1) = \pi_2 \qquad (Weak\ heredity)$$

$$p(s_{uij} = 1 \mid s_{ui} + s_{uj} = 0) = 0$$

Hereditary Spike-and-Slab Priors

- Spike-and-slab

$$s \sim Bernoulli(\pi), \quad \tilde{w} \sim \mathcal{N}(0,1), \quad w = \tilde{w} \cdot s.$$

- Hereditary spike-and-slab
  - capture the relations between 1st-order and 2nd-order feature interactions

$$s_{ui}, s_{uj} \sim Bernoulli(\pi_1)$$

$$p(s_{uij} = 1 \mid s_{ui} s_{uj} = 1) = 1 \qquad (Strong\ heredity)$$

$$p(s_{uij} = 1 \mid s_{ui} + s_{uj} = 1) = \pi_2 \qquad (Weak\ heredity)$$

$$p(s_{uij} = 1 \mid s_{ui} + s_{uj} = 0) = 0$$

Hereditary Spike-and-Slab Priors

- Spike-and-slab

$$s \sim Bernoulli(\pi), \quad \tilde{w} \sim \mathcal{N}(0,1), \quad w = \tilde{w} \cdot s.$$

- Hereditary spike-and-slab
  - capture the relations between 1$^{\text{st}}$-order and 2$^{\text{nd}}$-order feature interactions

$$\begin{aligned} s_{ui}, s_{uj} &\sim Bernoulli(\pi_1) & \\ p(s_{uij} = 1 \mid s_{ui} s_{uj} = 1) &= 1 & (Strong\ heredity) \\ p(s_{uij} = 1 \mid s_{ui} + s_{uj} = 1) &= \pi_2 & (Weak\ heredity) \\ p(s_{uij} = 1 \mid s_{ui} + s_{uj} = 0) &= 0 & \end{aligned}$$

Hereditary Spike-and-Slab Priors

- Spike-and-slab

$$s \sim Bernoulli(\pi), \quad \tilde{w} \sim \mathcal{N}(0,1), \quad w = \tilde{w} \cdot s.$$

- Hereditary spike-and-slab
  - capture the relations between 1st-order and 2nd-order feature interactions

$$\begin{aligned} s_{ui}, s_{uj} &\sim Bernoulli(\pi_1) \\ p(s_{uij} = 1 \mid s_{ui} s_{uj} = 1) &= 1 && (Strong\ heredity) \\ p(s_{uij} = 1 \mid s_{ui} + s_{uj} = 1) &= \pi_2 && (Weak\ heredity) \\ p(s_{uij} = 1 \mid s_{ui} + s_{uj} = 0) &= 0 \end{aligned}$$

Generative Procedure of BP-FIS

Algorithm Generation procedure

Generative Procedure of BP-FIS

Algorithm Generation procedure

1: **for** each user $u \in \mathcal{U}$ **do**
2: **for** each feature $i \in \mathcal{F}$ **do**
3: draw first-order interaction selection variable $s_{ui} \sim Bernoulli(\pi_1)$;

SIGIR 2019

Generative Procedure of BP-FIS

Algorithm Generation procedure

1: **for** each user $u \in \mathcal{U}$ **do**
2: **for** each feature $i \in \mathcal{F}$ **do**
3: draw first-order interaction selection variable $s_{ui} \sim Bernoulli(\pi_1)$;
4: draw first-order interaction weight $\tilde{w}_i \sim \mathcal{N}(0, 1)$;

SIGIR 2019

Generative Procedure of BP-FIS

Algorithm Generation procedure

for each user $u \in \mathcal{U}$ **do**
  for each feature $i \in \mathcal{F}$ **do**
    draw first-order interaction selection variable $s_{ui} \sim Bernoulli(\pi_1)$;
    draw first-order interaction weight $\tilde{w}_i \sim \mathcal{N}(0, 1)$;
    $w_{ui} = s_{ui} \cdot \tilde{w}_i$.

Generative Procedure of BP-FIS

Algorithm Generation procedure

for each user $u \in \mathcal{U}$ **do**
  for each feature $i \in \mathcal{F}$ **do**
    draw first-order interaction selection variable $s_{ui} \sim Bernoulli(\pi_1)$;
    draw first-order interaction weight $\tilde{w}_i \sim \mathcal{N}(0, 1)$;
    $w_{ui} = s_{ui} \cdot \tilde{w}_i$.
  for each feature pair $i, j \in \mathcal{F}$ **do**
    draw second-order interaction selection variable $s_{uij} \sim p(s_{uij} \mid s_{ui}, s_{uj})$;

Generative Procedure of BP-FIS

Algorithm Generation procedure

for each user $u \in \mathcal{U}$ **do**
  for each feature $i \in \mathcal{F}$ **do**
    draw first-order interaction selection variable $s_{ui} \sim Bernoulli(\pi_1)$;
    draw first-order interaction weight $\tilde{w}_i \sim \mathcal{N}(0, 1)$;
    $w_{ui} = s_{ui} \cdot \tilde{w}_i$.
  for each feature pair $i, j \in \mathcal{F}$ **do**
    draw second-order interaction selection variable $s_{uij} \sim p(s_{uij} \mid s_{ui}, s_{uj})$;
    draw second-order interaction weight $\tilde{w}_{ij} \sim \mathcal{N}(0, 1)$;

SIGIR 2019

Generative Procedure of BP-FIS

Algorithm Generation procedure

```
for each user $u \in \mathcal{U}$ do
    for each feature $i \in \mathcal{F}$ do
        draw first-order interaction selection variable $s_{ui} \sim Bernoulli(\pi_1)$;
        draw first-order interaction weight $\tilde{w}_i \sim \mathcal{N}(0, 1)$;
        $w_{ui} = s_{ui} \cdot \tilde{w}_i$.
    for each feature pair $i, j \in \mathcal{F}$ do
        draw second-order interaction selection variable $s_{uij} \sim p(s_{uij} \mid s_{ui}, s_{uj})$;
        draw second-order interaction weight $\tilde{w}_{ij} \sim \mathcal{N}(0, 1)$;
        $w_{uij} = s_{uij} \cdot \tilde{w}_{ij}$.
```

SIGIR 2019

Generative Procedure of BP-FIS

Algorithm Generation procedure

```
for each user $u \in \mathcal{U}$ do
    for each feature $i \in \mathcal{F}$ do
        draw first-order interaction selection variable $s_{ui} \sim Bernoulli(\pi_1)$;
        draw first-order interaction weight $\tilde{w}_i \sim \mathcal{N}(0, 1)$;
        $w_{ui} = s_{ui} \cdot \tilde{w}_i$.
    for each feature pair $i, j \in \mathcal{F}$ do
        draw second-order interaction selection variable $s_{uij} \sim p(s_{uij} \mid s_{ui}, s_{uj})$;
        draw second-order interaction weight $\tilde{w}_{ij} \sim \mathcal{N}(0, 1)$;
        $w_{uij} = s_{uij} \cdot \tilde{w}_{ij}$.
for each feature vector $\boldsymbol{x} \in \mathcal{X}$ do
    calculate the rating prediction $\hat{r}(\boldsymbol{x})$ by PFM;
```

SIGIR 2019

Generative Procedure of BP-FIS

Algorithm Generation procedure

```
for each user $u \in \mathcal{U}$ do
    for each feature $i \in \mathcal{F}$ do
        draw first-order interaction selection variable $s_{ui} \sim Bernoulli(\pi_1)$;
        draw first-order interaction weight $\tilde{w}_i \sim \mathcal{N}(0, 1)$;
        $w_{ui} = s_{ui} \cdot \tilde{w}_i$.
    for each feature pair $i, j \in \mathcal{F}$ do
        draw second-order interaction selection variable $s_{uij} \sim p(s_{uij} \mid s_{ui}, s_{uj})$;
        draw second-order interaction weight $\tilde{w}_{ij} \sim \mathcal{N}(0, 1)$;
        $w_{uij} = s_{uij} \cdot \tilde{w}_{ij}$.
for each feature vector $\boldsymbol{x} \in \mathcal{X}$ do
    calculate the rating prediction $\hat{r}(\boldsymbol{x})$ by PFM;
    draw $r(\boldsymbol{x}) \sim p(r \mid \hat{r}(\boldsymbol{x}))$.
```

SIGIR 2019

Optimization

Maximum A Posteriori: $\arg\max_{\tilde{W},S} p(\tilde{W}, S \mid \mathcal{R}, \mathcal{X})$

Optimization

Maximum A Posteriori: $\arg\max_{\tilde{W},S} p(\tilde{W}, S \mid \mathcal{R}, \mathcal{X})$

Infeasible exact inference

- space complexity: $O(md^2)$
- time complexity: $O(2^{md^2})$

Optimization

Maximum A Posteriori: $\arg\max_{\tilde{W},S} p(\tilde{W}, S \mid \mathcal{R}, \mathcal{X})$

Infeasible exact inference

- space complexity: $O(md^2)$
- time complexity: $O(2^{md^2})$

Variational inference

- approximate $p(\tilde{W}, S \mid \mathcal{R}, \mathcal{X})$ by $q(\tilde{W}, S)$
  - space complexity: $O(md)$
- Stochastic Gradient Variational Bayes
  - time complexity: $O(dk)$, same as FMs

Experimental Setup

Datasets

HetRec: Information Heterogeneity and Fusion in Recommender Systems

- **MovieLens**: rating and tagging
- **LastFM**: rating, tagging, social networking
- **Delicious**: rating, tagging, social networking

Baselines

- Factorization Machine (FM)
- Sparse Factorization Machine (SFM)
- Attentional Factorization Machine (AFM)
- Neural Factorization Machine (NFM)

SIGIR 2019

Experimental Setup

Our methods

apply BP-FIS to a linear and a non-linear FMs

- BP-FM
- BP-NFM

Evaluation

Top-N recommendation

- Leave-One-Out-Cross-Validation (LOOCV)
- ranking among 100 items
- metrics: HR@N and ARHR@N

Overall Performances

Table: Delicious

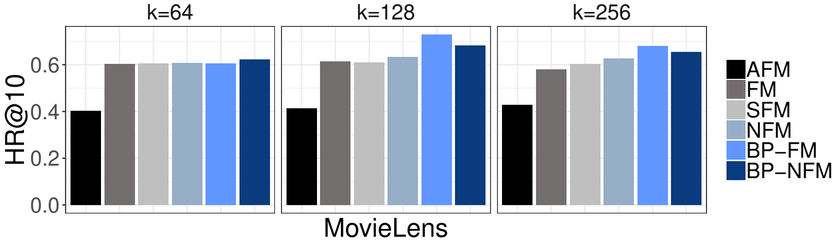
Method	HR@1	HR@10	ARHR@10
FM	0.0202	0.1147	0.0440
SFM	0.0229	0.1212	0.0465
AFM	0.0274	0.1169	0.0494
BP-FM	**0.0278**	**0.1240****	**0.0509***
NFM	0.0229	0.1065	0.0426
BP-NFM	**0.0268**	**0.1289****	**0.0504****

* and ** indicate that the best score is significantly better than the second best score with $p < 0.1$ and $p < 0.05$, respectively.

- SFM outperforms FM and AFM on HR@10: need for FIS
- BP-FM and BP-NFM significantly outperforms FMs and NFM, respectively: effect of P-FIS

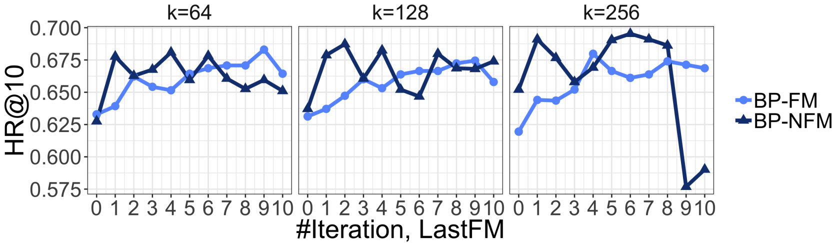
SIGIR 2019

Impact of Embedding Size



- $k = 64$: P-FIS has insignificant effect of FMs
- $k = 128, 256$:
  - BP-FM and BP-NFM significantly outperform FMs and NFM
  - BP-NFM does not outperform BP-FM

Impact of Training



- $k = 64$: BP-FM plays competitively with BP-NFM
- $k = 128$: BP-FM grows constantly, while BP-NFM fluctuate
- $k = 256$: BP-NFM performs better with less iterations, but unstable

SIGIR 2019

Conclusion

1. We study personalized feature interaction selection (P-FIS) for Factorization Machines.

Conclusion

2. We propose a Bayesian personalized feature interaction selection (BP-FIS) method based on the Bayesian variable selection.
   - We propose hereditary spike-and-slab as priors to achieve P-FIS.
   - BP-FIS is a plug-and-play framework for FMs

Conclusion

3. We design an efficient optimization algorithm based on Stochastic Gradient Variational Bayes (SGVB).

Future Work

1. Extend BP-FIS to select higher-order feature interactions
2. Consider group-level personalization via clustering to speed up training