210430 Special Lectures on Database (RecSys)

Factorization Machines

Heeseung Yun

heeseung.yun@vision.snu.ac.kr

Contents

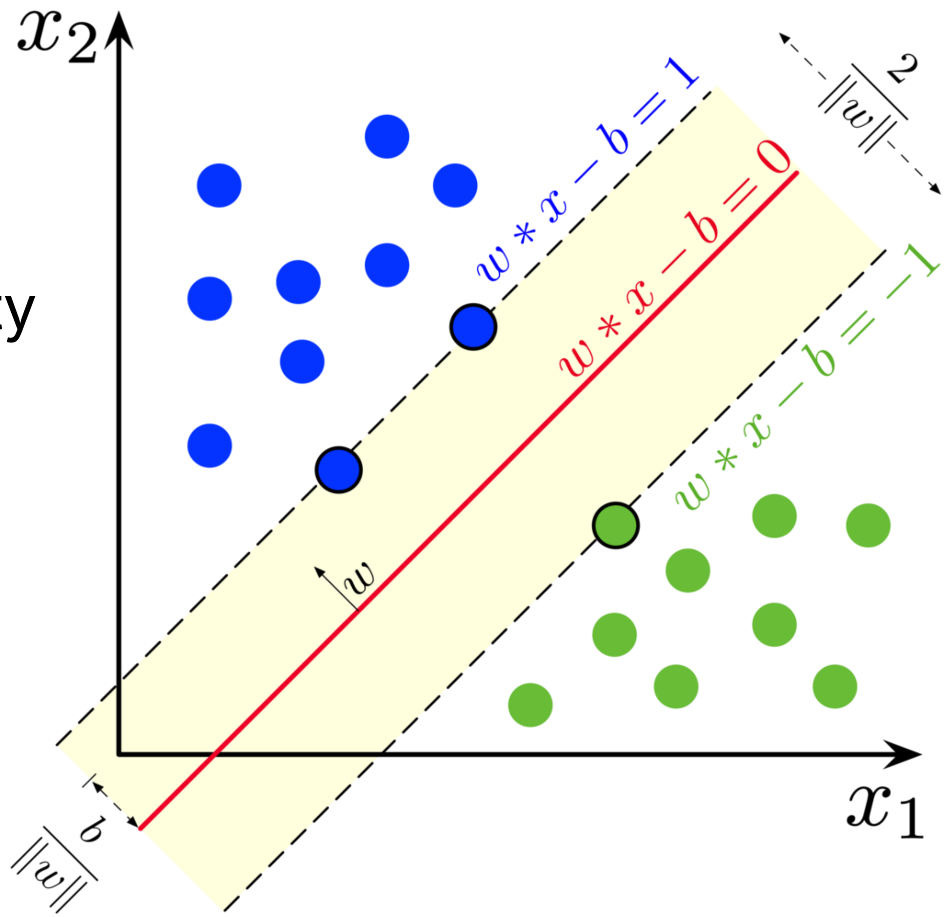
Background

Previous Recommender Systems

- Support Vector Machines
- Tensor Factorization

Previous Recommender Systems

- Support Vector Machines
  - (+) General predictor for any $\mathbb{R}^n$
  - (–) Susceptible to sparsity
  - (–) Susceptible to complex non-linearity



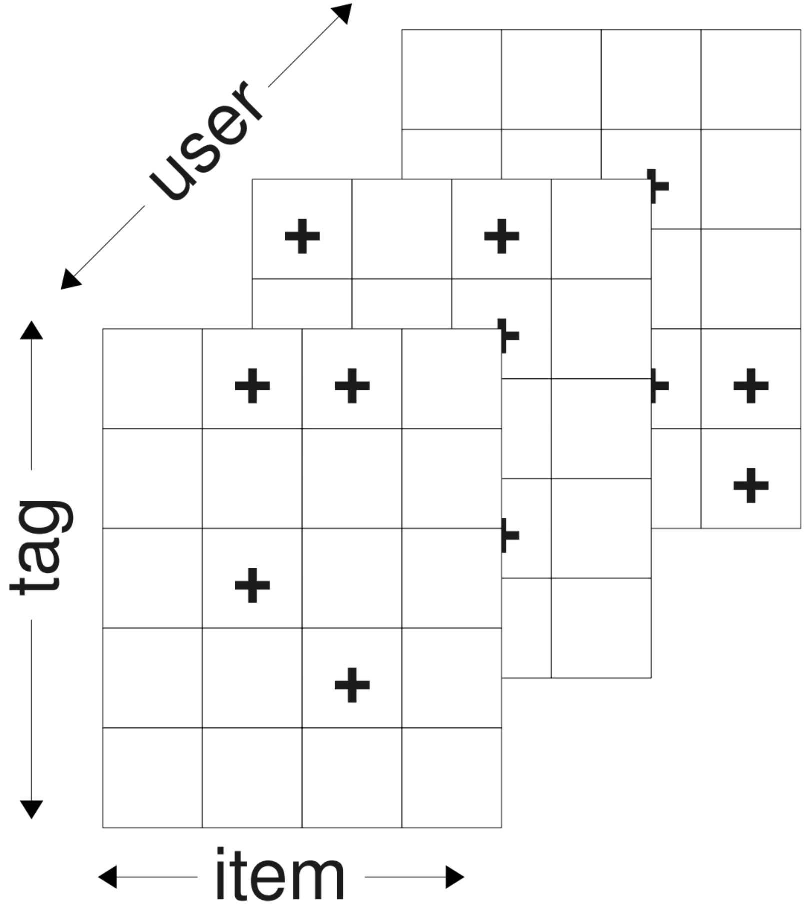
Credit: Wikipedia

Previous Recommender Systems

- Tensor Factorization
  - $(+)$ Sparsity-aware
  - $(-)$ General features ($\mathbb{R}^n$) not applicable
    - Mostly categorical
  - $(-)$ Task-specific modeling

Previous Recommender Systems

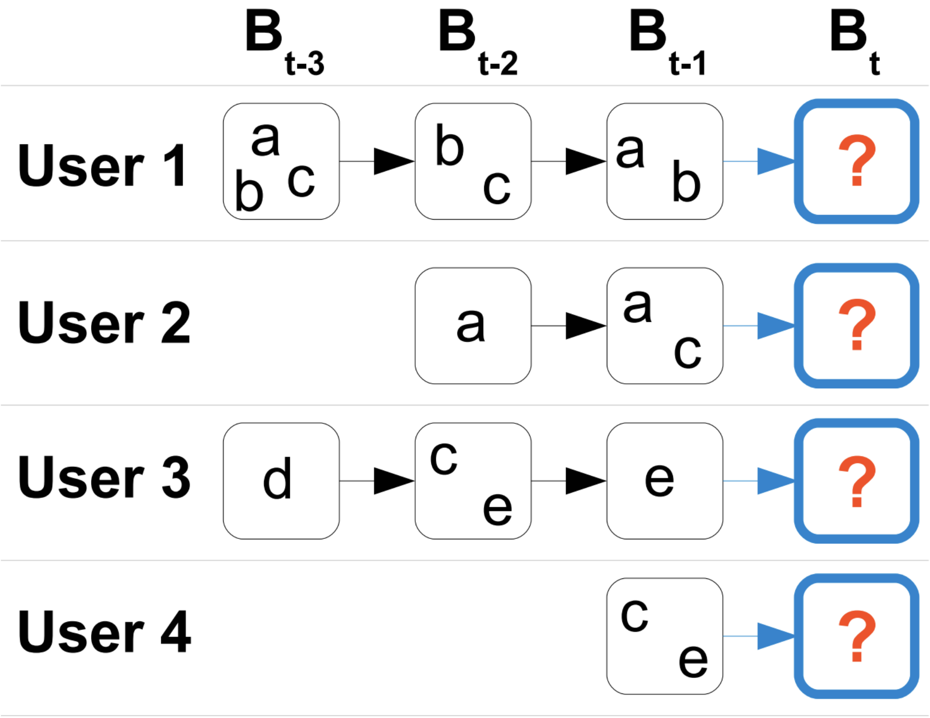
- Tensor Factorization
  - (ex1) SVD++ for movie rating prediction
    - Factorization with implicit feedback
  - (ex2) PITF for personalized tagging
    - Linearized Tucker decomposition



Koren. Factorization Meets the Neighborhood. In SIGKDD 2008.
Rendle et al. Pairwise Interaction Tensor Factorization for Personalized Tag Recommendation. In WSDM 2010.

Previous Recommender Systems

- Tensor Factorization
  - (ex3) FPMC for Next-Basket Recommendation
    - Factorization + Markov Chain
  - (ex4) PARAFAC
    - Parallel Factor Analysis



Rendle et al. Factorizing Personalized Markov Chains for Next-Basket Recommendation. In WWW 2010.
Harshman. Foundations of the PARAFAC Procedure. UCLA Working Papers in Phonetics. 1970.

Previous Recommender System

- Need for a unified framework
  - Take nested variable interaction into account
  - Efficient in terms of time & parameters
  - Work with various features and task-agnostic

Previous Recommender System

- Need for a unified framework
  - Take nested variable interaction into account → <u>Sparsity-Robustness</u>
  - Efficient in terms of time & parameters → <u>Scalability</u>
  - Work with various features and task-agnostic → <u>Generalizability</u>

- Solution: **Factorization Machines**

Factorization Machines

Definition

- Objective – predict $y : x \to T$ $(x \in \mathbb{R}^n)$
  - T depends on context
    - Real-valued, binary, ranking, etc.
  - x is generally categorical
    - BoW, transaction, etc.
    - Thereby highly sparse, i.e., $\bar{m}_D \ll n$
  - But not always categorical
    - Month, normalized indicator, $\mathbb{R}^n$

Definition

- Common tasks
  - Rating – regression with MSE
  - Binary – classification with hinge/logit loss (BCE)
  - Ranking – pairwise classification loss
    - Kendall's τ (Joachim, 2002)
    - Margin ranking loss $\max\left(0, f(x_{low}) + \alpha - f\left(x_{high}\right)\right)$

Definition

- Objective – predict $y : x \rightarrow T \; (x \in \mathbb{R}^n)$
  - T depends on context
    - Real-valued, binary, ranking, etc.
  - x is generally categorical
    - BoW, transaction, etc.
    - Thereby highly sparse, i.e., $\underbrace{\bar{m}_D}_{\text{average \# of nonzero entries}} \ll n$
  - But not always categorical
    - Month, normalized indicator, $\mathbb{R}^n$

Definition

- Example – movie rating

	Feature vector **x**																				Target y	
	User				Movie					Other Movies rated					Time	Last Movie rated						
	A	B	C	...	TI	NH	SW	ST	...	TI	NH	SW	ST	...	Time	TI	NH	SW	ST	...		
$\mathbf{x}^{(1)}$	1	0	0	...	1	0	0	0	...	0.3	0.3	0.3	0	...	13	0	0	0	0	...	5	$y^{(1)}$
$\mathbf{x}^{(2)}$	1	0	0	...	0	1	0	0	...	0.3	0.3	0.3	0	...	14	1	0	0	0	...	3	$y^{(2)}$
$\mathbf{x}^{(3)}$	1	0	0	...	0	0	1	0	...	0.3	0.3	0.3	0	...	16	0	1	0	0	...	1	$y^{(2)}$
$\mathbf{x}^{(4)}$	0	1	0	...	0	0	1	0	...	0	0	0.5	0.5	...	5	0	0	0	0	...	4	$y^{(3)}$
$\mathbf{x}^{(5)}$	0	1	0	...	0	0	0	1	...	0	0	0.5	0.5	...	8	0	0	1	0	...	5	$y^{(4)}$
$\mathbf{x}^{(6)}$	0	0	1	...	1	0	0	0	...	0.5	0	0.5	0	...	9	0	0	0	0	...	1	$y^{(5)}$
$\mathbf{x}^{(7)}$	0	0	1	...	0	0	1	0	...	0.5	0	0.5	0	...	12	1	0	0	0	...	5	$y^{(6)}$

Definition

- Example – movie rating

Feature vector **x**

	User				Movie					Other Movies rated					Time	Last Movie rated					Target y	
	A	B	C	...	TI	NH	SW	ST	...	TI	NH	SW	ST	...	Time	TI	NH	SW	ST	...		
$\mathbf{x}^{(1)}$	1	0	0	...	1	0	0	0	...	0.3	0.3	0.3	0	...	13	0	0	0	0	...	5	$y^{(1)}$
$\mathbf{x}^{(2)}$	1	0	0	...	0	1	0	0	...	0.3	0.3	0.3	0	...	14	1	0	0	0	...	3	$y^{(2)}$
$\mathbf{x}^{(3)}$	1	0	0	...	0	0	1	0	...	0.3	0.3	0.3	0	...	16	0	1	0	0	...	1	$y^{(2)}$
$\mathbf{x}^{(4)}$	0	1	0	...	0	0	1	0	...	0	0	0.5	0.5	...	5	0	0	0	0	...	4	$y^{(3)}$
$\mathbf{x}^{(5)}$	0	1	0	...	0	0	0	1	...	0	0	0.5	0.5	...	8	0	0	1	0	...	5	$y^{(4)}$
$\mathbf{x}^{(6)}$	0	0	1	...	1	0	0	0	...	0.5	0	0.5	0	...	9	0	0	0	0	...	1	$y^{(5)}$
$\mathbf{x}^{(7)}$	0	0	1	...	0	0	1	0	...	0.5	0	0.5	0	...	12	1	0	0	0	...	5	$y^{(6)}$

Categorical (User, Movie) | Real-valued (Other Movies rated) | Int (Time)

Definition

- Prediction

$$\hat{y}(\mathbf{x}) := \underbrace{w_0}_{\text{Bias}} + \sum_{i=1}^{n} \underbrace{w_i\, x_i}_{\text{Signal Strength}} + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \underbrace{\langle \mathbf{v}_i, \mathbf{v}_j \rangle\, x_i\, x_j}_{\text{Interaction}}$$

- Parameters: $O(kn)$

$$w_0 \in \mathbb{R}, \quad \mathbf{w} \in \mathbb{R}^n, \quad \mathbf{V} \in \mathbb{R}^{n \times k}$$

- Hyperparameter: factor dimension $k \in \mathbb{N}_0^+$
  - Larger k theoretically reconstructs better
  - Smaller k leads to better generalization

Definition

- Factorization counts
  - Latent factors reflect context

(Rating)	Titanics	Star Wars	Star Trek
Alice	5	1	??
Bob		4	5
Charlie	1		5

→ Ratings are not independent

Complexity

- Definition

$$\hat{y}(\mathbf{x}) := w_0 + \sum_{i=1}^{n} w_i \, x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle \mathbf{v}_i, \mathbf{v}_j \rangle \, x_i \, x_j$$

- Naïve approach: $O(kn^2)$

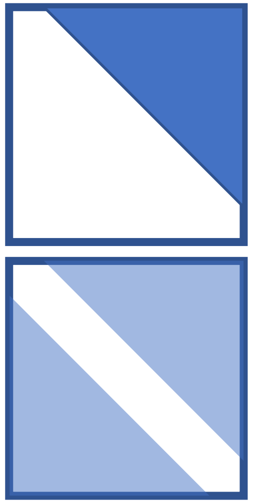
Complexity

- Optimization

$$\sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle \mathbf{v}_i, \mathbf{v}_j \rangle \, x_i \, x_j$$

$$= \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \langle \mathbf{v}_i, \mathbf{v}_j \rangle \, x_i \, x_j - \frac{1}{2} \sum_{i=1}^{n} \langle \mathbf{v}_i, \mathbf{v}_i \rangle \, x_i \, x_i$$

$$= \frac{1}{2} \left(\sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{f=1}^{k} v_{i,f} \, v_{j,f} \, x_i \, x_j - \sum_{i=1}^{n} \sum_{f=1}^{k} v_{i,f} \, v_{i,f} \, x_i \, x_i \right)$$



Complexity

- Optimization

$$=\frac{1}{2}\boxed{\sum_{f=1}^{k}}\left(\left(\sum_{i=1}^{n} v_{i,f}\, x_i\right)\left(\sum_{j=1}^{n} v_{j,f}\, x_j\right)-\sum_{i=1}^{n} v_{i,f}^2\, x_i^2\right)$$

$$=\frac{1}{2}\sum_{f=1}^{k}\left(\left(\boxed{\sum_{i=1}^{n}} v_{i,f}\, x_i\right)^2-\boxed{\sum_{i=1}^{n}} v_{i,f}^2\, x_i^2\right)$$

- Time complexity for inference: $O(kn)$
  - $O(k\overline{m}_D)$ for practical usage

Complexity

- Training
  - Stochastic Gradient Descent

$$\frac{\partial}{\partial\theta}\hat{y}(\mathbf{x}) = \begin{cases} 1, & \text{if } \theta \text{ is } w_0 \\ x_i, & \text{if } \theta \text{ is } w_i \\ x_i \underbrace{\sum_{j=1}^{n} v_{j,f} x_j}_{\text{Constant}} - v_{i,f} x_i^2, & \text{if } \theta \text{ is } v_{i,f} \end{cases}$$

  - Time complexity for learning step: $O(kn)$, i.e., $O(1)$ per parameter

Params & inference time & training time are asymptotically <u>linear</u>!

Extension

- d-way Factorization Machine

$$\hat{y}(x) := \overset{\text{Bias}}{w_0} + \sum_{i=1}^{n} \overset{\text{Signal Strength}}{w_i \, x_i} + \sum_{l=2}^{d} \sum_{i_1=1}^{n} \cdots \sum_{i_l=i_{l-1}+1}^{n} \left(\prod_{j=1}^{l} x_{i_j} \right) \left(\sum_{f=1}^{k_l} \overset{\text{(PARAFAC) Interaction}}{\prod_{j=1}^{l} v_{i_j,f}^{(l)}} \right)$$

  - ex. Interaction of five SNS hashtags

- Can be optimized as in 2-way FM
  - i.e., linear complexity

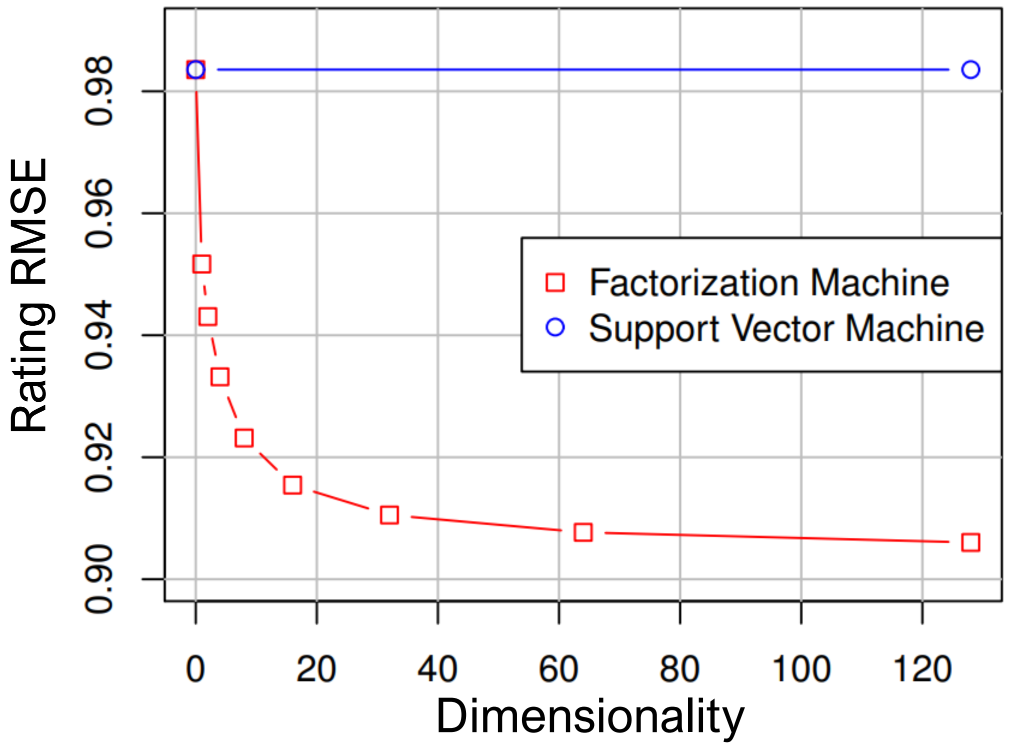
Comparison

SVM

- SVM with Linear Kernel $\phi(\mathbf{x}) := (1, x_1, \ldots, x_n)$

$$\hat{y}(\mathbf{x}) = w_0 + \sum_{i=1}^{n} w_i\, x_i = w_0 + w_u + w_i$$

  - Identical to degree=1 FM
  - Netflix rating prediction
    - Interaction does count!



SVM

- SVM with Polynomial Kernel $(1, x_1, \ldots, x_n)^d$
  - For $d = 2$

Independence

$$\hat{y}(\mathbf{x}) = w_0 + \sqrt{2}\sum_{i=1}^{n} w_i\, x_i + \sum_{i=1}^{n} w_{i,i}^{(2)}\, x_i^2 + \sqrt{2}\sum_{i=1}^{n}\sum_{j=i+1}^{n} \boxed{w_{i,j}^{(2)}}\, x_i\, x_j$$

    - FM without factorization

  - cf. Original FM

Interaction

$$\hat{y}(\mathbf{x}) := w_0 + \sum_{i=1}^{n} w_i\, x_i + \sum_{i=1}^{n}\sum_{j=i+1}^{n} \boxed{\langle \mathbf{v}_i, \mathbf{v}_j \rangle}\, x_i\, x_j$$

SVM

- SVM with Polynomial Kernel $(1, x_1, \dots, x_n)^d$
  - For degree=2 FM with $x_u = x_i = 1, x_{else} = 0$

$$\hat{y}(\mathbf{x}) = w_0 + \sqrt{2}\sum_{i=1}^{n} w_i\, x_i + \sum_{i=1}^{n} w_{i,i}^{(2)}\, x_i^2 + \sqrt{2}\sum_{i=1}^{n}\sum_{j=i+1}^{n} w_{i,j}^{(2)}\, x_i\, x_j$$

$$= w_0 + \sqrt{2}(w_u + w_i) + w_{u,u}^{(2)} + w_{i,i}^{(2)} + \sqrt{2} w_{u,i}^{(2)}$$

SVM

- SVM with Polynomial Kernel $(1, x_1, \ldots, x_n)^d$
  - For degree=2 FM with $x_u = x_i = 1, x_{else} = 0$

$$\hat{y}(\mathbf{x}) = w_0 + \sqrt{2}\sum_{i=1}^{n} w_i\, x_i + \sum_{i=1}^{n} w_{i,i}^{(2)}\, x_i^2 + \sqrt{2}\sum_{i=1}^{n}\sum_{j=i+1}^{n} w_{i,j}^{(2)}\, x_i\, x_j$$

$$= w_0 + \sqrt{2}(w_u + w_i) + w_{u,u}^{(2)} + w_{i,i}^{(2)} + \sqrt{2} w_{u,i}^{(2)}$$

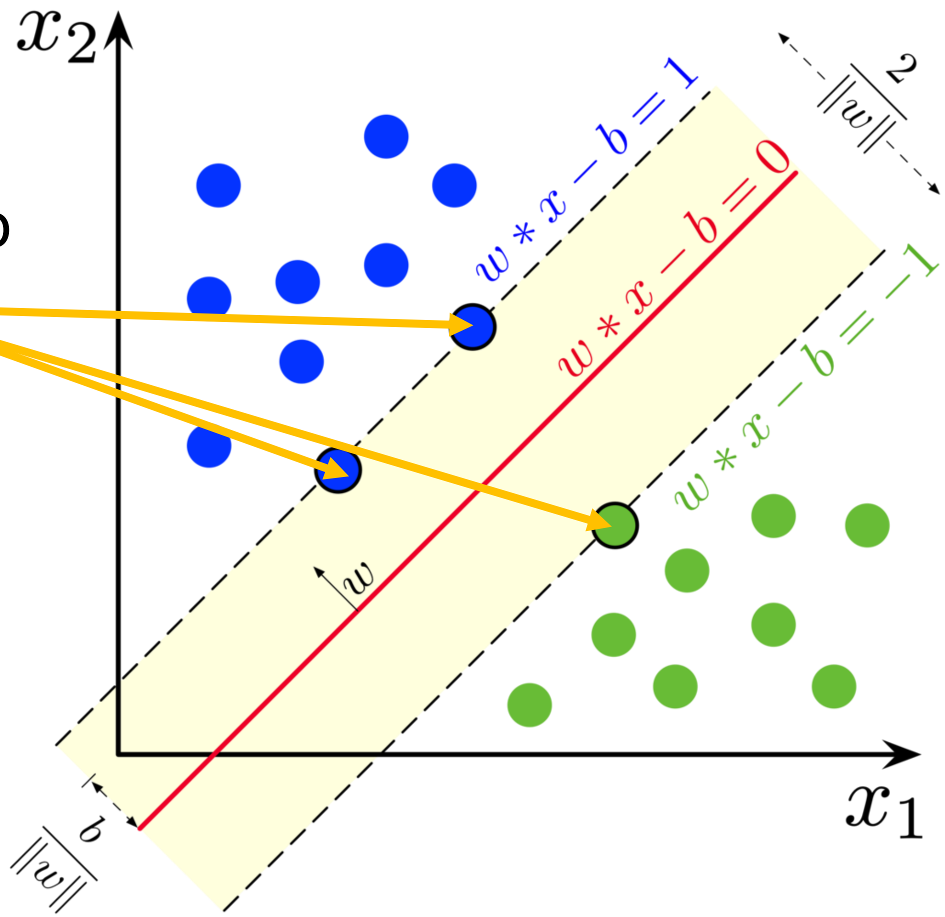
Identical to $w_u + w_i$

Zero for most combination $(\bar{m}_D \ll n)$

  - i.e., discards latent factors

SVM

- (Slide 5 revisited)
  - (–) Susceptible to sparsity
  - (–) Susceptible to complex relationship
  - (–) Dependent on support vectors
  - (·) Solved in dual, not primal



Credit: Wikipedia

Tensor Factorization

- Matrix Factorization (MF, SVD)

$$\hat{y}(\mathbf{x}) = w_0 + w_u + w_i + \langle \mathbf{v}_u, \mathbf{v}_i \rangle$$

  - Identical to FM with $x_u = x_i = 1, x_{else} = 0$

$$\hat{y}(\mathbf{x}) := w_0 + \sum_{i=1}^{n} w_i\, x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle \mathbf{v}_i, \mathbf{v}_j \rangle\, x_i\, x_j$$

- PARAFAC
  - Features with multiple categorical variables, i.e., d-degree FM
  - Both are limited to categorical variables

Tensor Factorization

- Movie rating (SVD++)

$$\hat{y}(\mathbf{x}) = w_0 + w_u + w_i + \langle \mathbf{v}_u, \mathbf{v}_i \rangle + \frac{1}{\sqrt{|N_u|}} \sum_{l \in N_u} \langle \mathbf{v}_i, \mathbf{v}_l \rangle$$

user-movie
+
movie-rated

  - Partial model of FM with $x_u = x_i = 1, x_l = \frac{1}{\sqrt{|N_u|}}, x_{else} = 0$

$$\hat{y}(\mathbf{x}) := w_0 + \sum_{i=1}^{n} w_i \, x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle \mathbf{v}_i, \mathbf{v}_j \rangle \, x_i \, x_j$$

Koren. Factorization Meets the Neighborhood. In SIGKDD 2008.

Tensor Factorization

- Movie rating (FM)

$$\hat{y}(\mathbf{x}) = w_0 + w_u + w_i + \langle \mathbf{v}_u, \mathbf{v}_i \rangle + \frac{1}{\sqrt{|N_u|}} \sum_{l \in N_u} \langle \mathbf{v}_i, \mathbf{v}_l \rangle$$

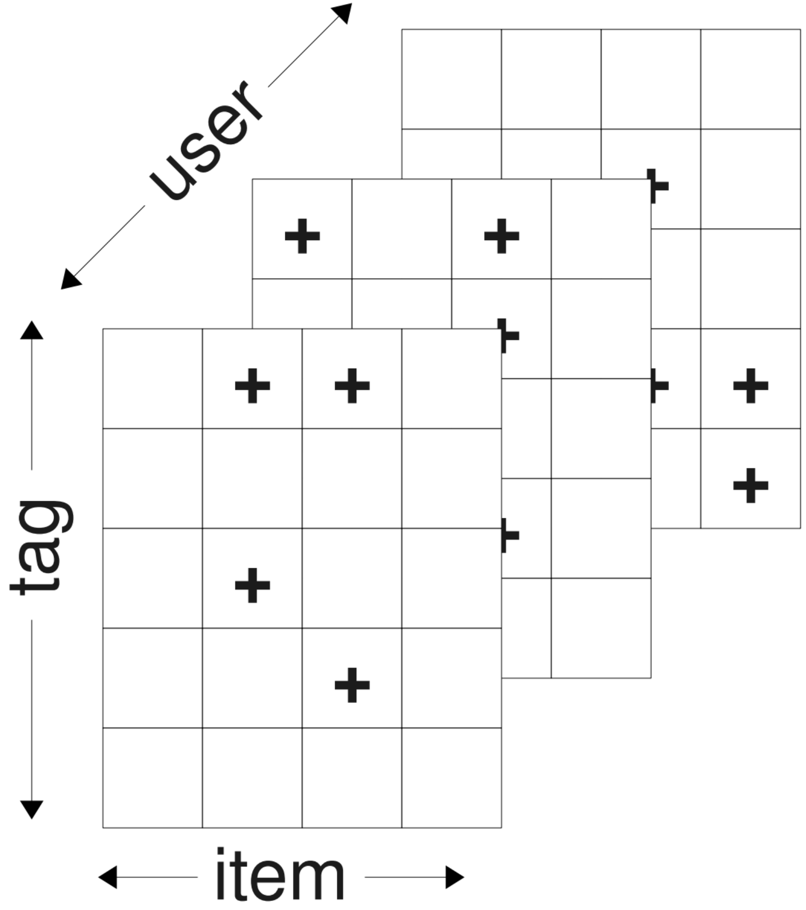
$$+ \frac{1}{\sqrt{|N_u|}} \sum_{l \in N_u} \left(w_l + \langle \mathbf{v}_u, \mathbf{v}_l \rangle + \frac{1}{\sqrt{|N_u|}} \sum_{l' \in N_u, l' > l} \langle \mathbf{v}_l, \mathbf{v}_l' \rangle \right)$$

user-movie
+
movie-rated
+
user-rated
+
rated-rated

Tensor Factorization

- Personal Tag Recommendation (PITF)

$$\hat{y}_{u,i,t} = \sum_f \underbrace{\hat{u}_{u,f} \cdot \hat{t}^U_{t,f}}_{\text{user-tag MF}} + \sum_f \underbrace{\hat{i}_{i,f} \cdot \hat{t}^I_{t,f}}_{\text{item-tag MF}}$$



Rendle et al. Pairwise Interaction Tensor Factorization for Personalized Tag Recommendation. In WSDM 2010.

Tensor Factorization

- Personal Tag Recommendation (PITF)

$$\hat{y}_{u,i,t} = \sum_f \hat{u}_{u,f} \cdot \hat{t}^U_{t,f} + \sum_f \hat{i}_{i,f} \cdot \hat{t}^I_{t,f}$$

  - Near-identical to FM optimized for ranking task with $x_u = x_i = x_t = 1$

$$\begin{aligned}\hat{y}(\mathbf{x}) &:= w_0 + \sum_{i=1}^{n} w_i\, x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle \mathbf{v}_i, \mathbf{v}_j \rangle\, x_i\, x_j \\ &= w_0 + w_u + w_i + w_t + \langle \mathbf{v}_u, \mathbf{v}_i \rangle + \langle \mathbf{v}_u, \mathbf{v}_t \rangle + \langle \mathbf{v}_i, \mathbf{v}_t \rangle\end{aligned}$$

Rendle et al. Pairwise Interaction Tensor Factorization for Personalized Tag Recommendation. In WSDM 2010.

Tensor Factorization

- Personal Tag Recommendation (PITF)

$$\hat{y}_{u,i,t} = \sum_f \hat{u}_{u,f} \cdot \hat{t}^U_{t,f} + \sum_f \hat{i}_{i,f} \cdot \hat{t}^I_{t,f}$$

  - Near-identical to FM optimized for ranking task with $x_u = x_i = x_t = 1$

$$\hat{y}(\mathbf{x}) := w_0 + \sum_{i=1}^{n} w_i \, x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle \mathbf{v}_i, \mathbf{v}_j \rangle \, x_i \, x_j$$

$$= w_0 + \cancel{w_u} + w_i + w_t + \cancel{\langle \mathbf{v}_u, \mathbf{v}_i \rangle} + \langle \mathbf{v}_u, \mathbf{v}_t \rangle + \langle \mathbf{v}_i, \mathbf{v}_t \rangle$$

Independent of tag t

Rendle et al. Pairwise Interaction Tensor Factorization for Personalized Tag Recommendation. In WSDM 2010.

Tensor Factorization

- Personal Tag Recommendation (PITF)

$$\hat{y}_{u,i,t} = \sum_f \hat{u}_{u,f} \cdot \hat{t}^U_{t,f} + \sum_f \hat{i}_{i,f} \cdot \hat{t}^I_{t,f}$$ Distinct tag factorization

  - Near-identical to FM optimized for ranking task with $x_u = x_i = x_t = 1$

$$\hat{y}(\mathbf{x}) := w_0 + \sum_{i=1}^{n} w_i\, x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle \mathbf{v}_i, \mathbf{v}_j \rangle\, x_i\, x_j$$

$$= \cancel{w_0 + w_u} + w_i + w_t + \cancel{\langle \mathbf{v}_u, \mathbf{v}_i \rangle} + \langle \mathbf{v}_u, \mathbf{v}_t \rangle + \langle \mathbf{v}_i, \mathbf{v}_t \rangle$$

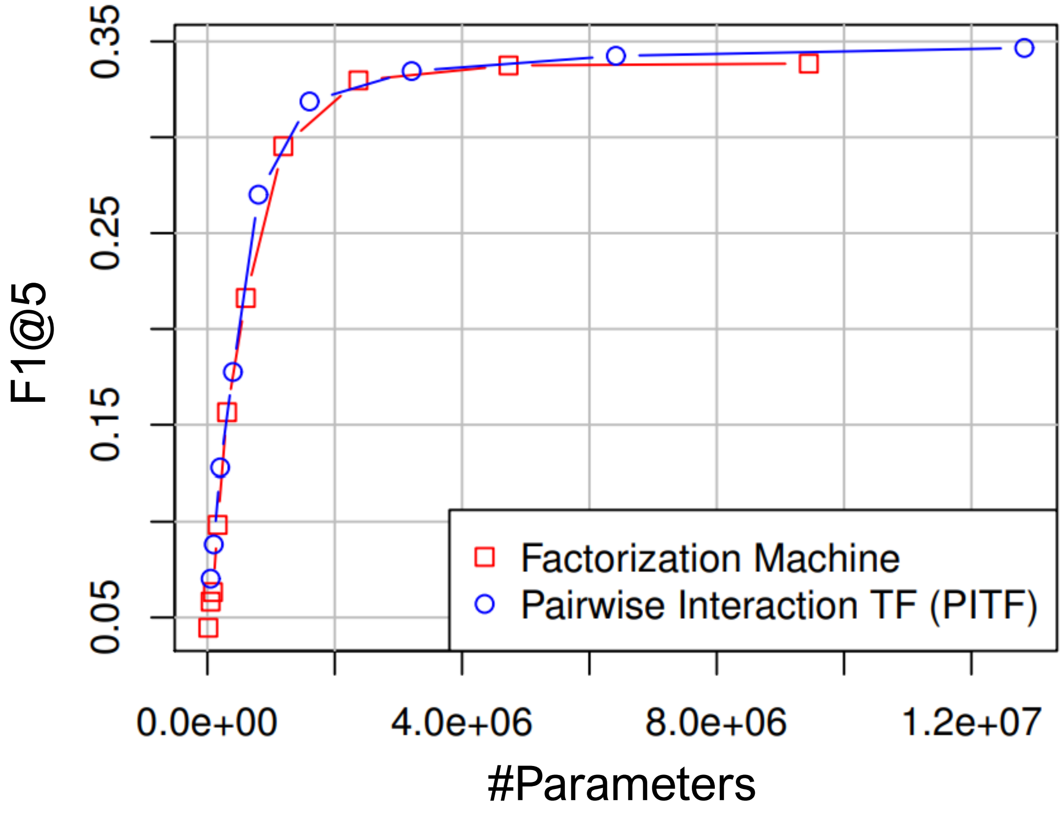
$$\hat{y}(\mathbf{x}) := w_t + \langle \mathbf{v}_u, \mathbf{v}_t \rangle + \langle \mathbf{v}_i, \mathbf{v}_t \rangle$$ Shared tag factorization

Bias

Rendle et al. Pairwise Interaction Tensor Factorization for Personalized Tag Recommendation. In WSDM 2010.

Tensor Factorization

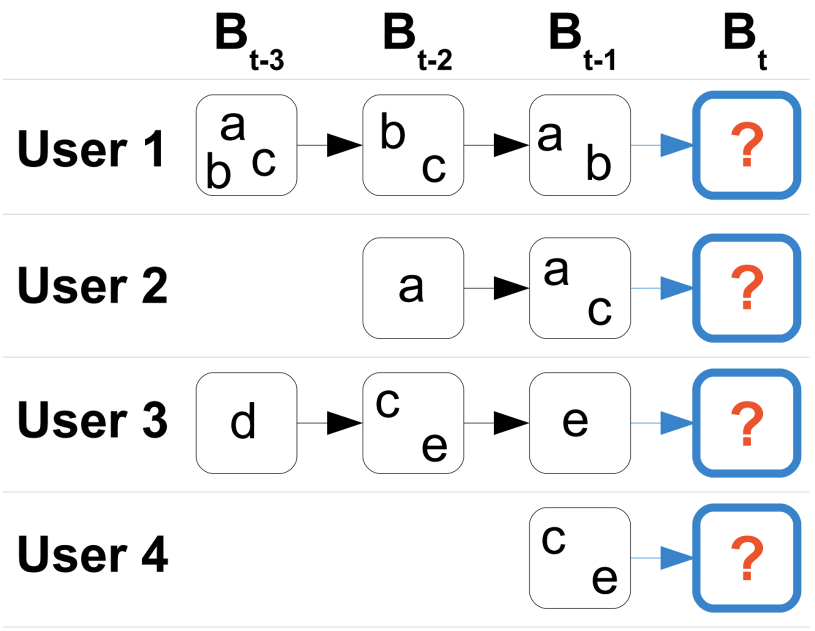
- Personal Tag Recommendation (PITF)
  - ECML/PKDD Discovery Challenge
    - Nearly identical behavior w.r.t params



Tensor Factorization

- Next-Basket Recommendation (FPMC)

$$\hat{x}_{u,t,i} := \underbrace{\langle v_u^{U,I}, v_i^{I,U} \rangle}_{\text{user-item MF}} + \underbrace{\frac{1}{|B_{t-1}^u|}}_{\text{Markov}} \sum_{l \in B_{t-1}^u} \underbrace{\langle v_i^{I,L}, v_l^{L,I} \rangle}_{\text{Item-bag MF}}$$



Rendle et al. Factorizing Personalized Markov Chains for Next-Basket Recommendation. In WWW 2010.

Tensor Factorization

- Next-Basket Recommendation (FPMC)

$$\hat{x}_{u,t,i} := \langle v_u^{U,I}, v_i^{I,U} \rangle + \frac{1}{|B_{t-1}^u|} \sum_{l \in B_{t-1}^u} \langle v_i^{I,L}, v_l^{L,I} \rangle$$

  - Near-identical to FM optimized for ranking ($x_u = x_i = 1, x_l = \frac{1}{\sqrt{|B_{t-1}^u|}}$)

$$\hat{y}(\mathbf{x}) = w_0 + w_u + w_i + \langle \mathbf{v}_u, \mathbf{v}_i \rangle + \frac{1}{|B_{t-1}^u|} \sum_{l \in B_{t-1}^u} \langle \mathbf{v}_i, \mathbf{v}_l \rangle$$
$$+ \frac{1}{|B_{t-1}^u|} \sum_{l \in B_{t-1}^u} \left(w_l + \langle \mathbf{v}_u, \mathbf{v}_l \rangle + \frac{1}{|B_{t-1}^u|} \sum_{l' \in B_{t-1}^u, l' > l} \langle \mathbf{v}_l, \mathbf{v}_l' \rangle \right)$$

Rendle et al. Factorizing Personalized Markov Chains for Next-Basket Recommendation. In WWW 2010.

Tensor Factorization

- Next-Basket Recommendation (FPMC)

$$\hat{x}_{u,t,i} := \langle v_u^{U,I}, v_i^{I,U} \rangle + \frac{1}{|B_{t-1}^u|} \sum_{l \in B_{t-1}^u} \langle v_i^{I,L}, v_l^{L,I} \rangle$$

  - Near-identical to FM optimized for ranking ($x_u = x_i = 1, x_l = \frac{1}{\sqrt{|B_{t-1}^u|}}$)

$$\hat{y}(\mathbf{x}) = \cancel{w_0 + w_u} + w_i + \langle \mathbf{v}_u, \mathbf{v}_i \rangle + \frac{1}{|B_{t-1}^u|} \sum_{l \in B_{t-1}^u} \langle \mathbf{v}_i, \mathbf{v}_l \rangle$$

$$\cancel{+ \frac{1}{|B_{t-1}^u|} \sum_{l \in B_{t-1}^u} \left(w_l + \langle \mathbf{v}_u, \mathbf{v}_l \rangle + \frac{1}{|B_{t-1}^u|} \sum_{l' \in B_{t-1}^u, l' > l} \langle \mathbf{v}_l, \mathbf{v}_l' \rangle \right)}$$

Independent of item i

Rendle et al. Factorizing Personalized Markov Chains for Next-Basket Recommendation. In WWW 2010.

Tensor Factorization

- Next-Basket Recommendation (FPMC)

$$\hat{x}_{u,t,i} := \langle v_u^{U,I}, v_i^{I,U} \rangle + \frac{1}{|B_{t-1}^u|} \sum_{l \in B_{t-1}^u} \langle v_i^{I,L}, v_l^{L,I} \rangle$$

Distinct item factorization

  - Near-identical to FM optimized for ranking ($x_u = x_i = 1, x_l = \frac{1}{\sqrt{|B_{t-1}^u|}}$)

$$\hat{y}(\mathbf{x}) = \cancel{w_0 + w_u} + w_i + \langle \mathbf{v}_u, \mathbf{v}_i \rangle + \frac{1}{|B_{t-1}^u|} \sum_{l \in B_{t-1}^u} \langle \mathbf{v}_i, \mathbf{v}_l \rangle$$

$$\cancel{+ \frac{1}{|B_{t-1}^u|} \sum_{l \in B_{t-1}^u} \left(w_l + \langle \mathbf{v}_u, \mathbf{v}_l \rangle + \frac{1}{|B_{t-1}^u|} \sum_{l' \in B_{t-1}^u, l' > l} \langle \mathbf{v}_l, \mathbf{v}_l' \rangle \right)}$$

$$= w_i + \langle \mathbf{v}_u, \mathbf{v}_i \rangle + \frac{1}{|B_{t-1}^u|} \sum_{l \in B_{t-1}^u} \langle \mathbf{v}_i, \mathbf{v}_l \rangle$$

Shared item factorization

Bias

Rendle et al. Factorizing Personalized Markov Chains for Next-Basket Recommendation. In WWW 2010.

Conclusion

Conclusion

- Factorization Machines
  - Factorization-oriented
  - Linear parameters & prediction time & learning time
  - Subsumes existing (task-specific) State-of-the-Art models

- Additional remarks
  - (+) Unified RS framework with clear theoretical explanation
  - (–) Lacks experimental evidence (SVD++, FPMC, time complexity)

Thank You

Steffen Rendle. Factorization Machines. In ICDM 2010.

Appendix

Lagrangian Duality

- Primal SVM

$$\zeta_i = \max\left(0, 1 - y_i(\mathbf{w}^T\mathbf{x}_i - b)\right)$$

$$\text{minimize } \frac{1}{n}\sum_{i=1}^{n}\zeta_i + \lambda\|\mathbf{w}\|^2$$

$$\text{subject to } y_i(\mathbf{w}^T\mathbf{x}_i - b) \geq 1 - \zeta_i \text{ and } \zeta_i \geq 0, \text{ for all } i$$

- Dual SVM

$$\text{maximize } f(c_1 \ldots c_n) = \sum_{i=1}^{n} c_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i c_i (\mathbf{x}_i^T\mathbf{x}_j) y_j c_j$$

$$\text{subject to } \sum_{i=1}^{n} c_i y_i = 0, \text{ and } 0 \leq c_i \leq \frac{1}{2n\lambda} \text{ for all } i$$