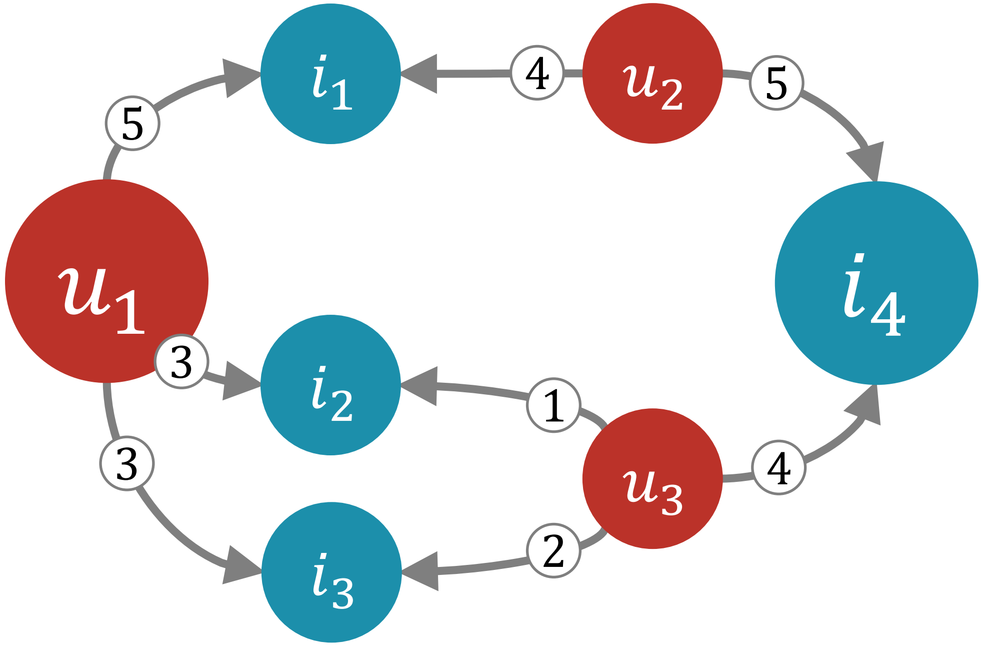
UFMG UNIVERSIDADE FEDERAL
DE MINAS GERAIS

Recommender Systems

Factorization Machines

Rodrygo L. T. Santos
rodrygo@dcc.ufmg.br

Interaction modeling



	i_1	i_2	i_3	i_4
u_1	5	3	3	
u_2	4			5
u_3		1	2	4

Interaction modeling

Distinct spaces in neighborhood models

- Users as n-dimensional vectors over items
- Items as m-dimensional vectors over users

Unified space in latent factor models

- Users and items as k-dimensional vectors

Highly effective in practice!

Interaction modeling

Hard to incorporate additional information

- User features (age, gender, income, ...)
- Item features (description, image, ...)
- Contextual features (location, time, ...)

Ad-hoc adaptations

- Handcrafted hypotheses and algorithms

Feature-based modeling

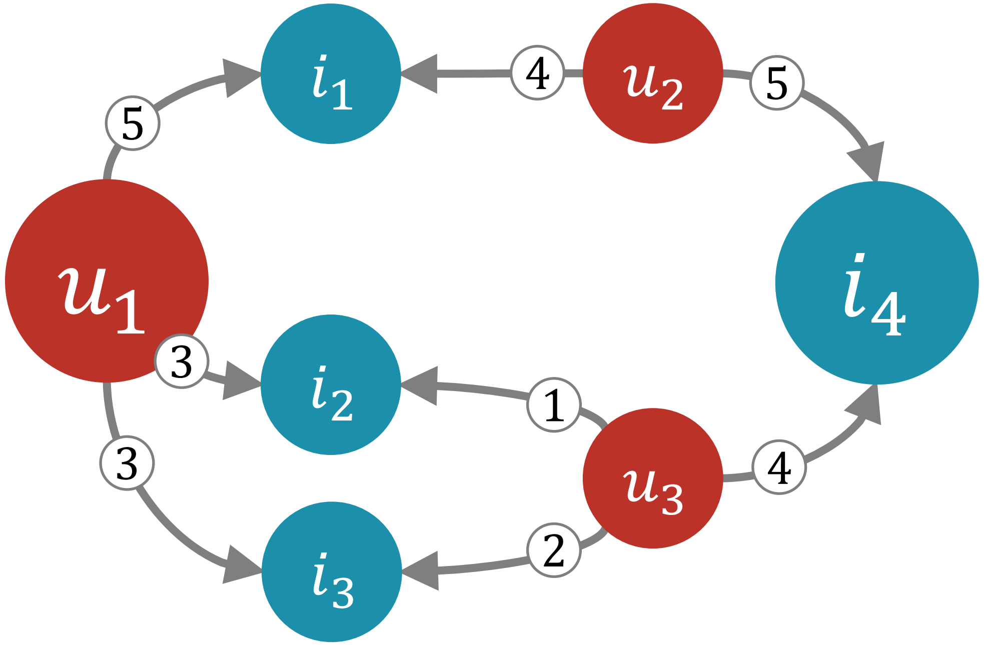
Standard representation via feature vectors

- Categorical features
- Numerical features

Advantages

- Allows modeling any number of variables
- Enables a variety of machine learning approaches

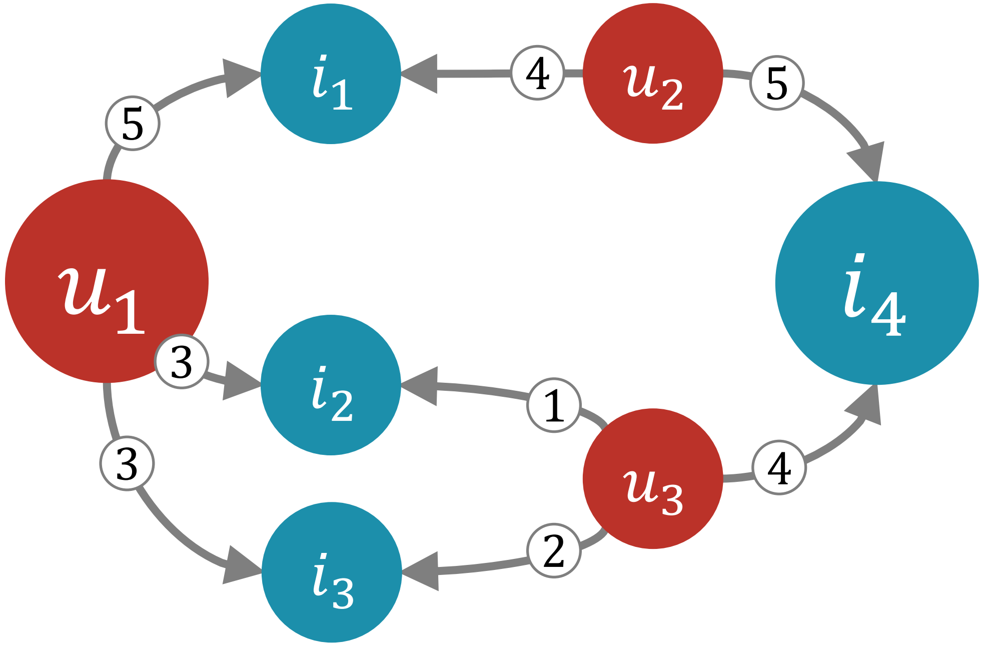
Feature-based modeling



$$\begin{array}{cccc} & u & i & y \\ \mathbf{x}_1^T & \boxed{1} & \boxed{1} & \boxed{5} \end{array}$$

user / item id are categorical features
→ one-hot encoding

Feature-based modeling



	u_1	u_2	u_3	i_1	i_2	i_3	i_4	y
$\mathbf{x}_1^T$	1	0	0	1	0	0	0	5
$\mathbf{x}_2^T$	1	0	0	0	1	0	0	3
$\mathbf{x}_3^T$	1	0	0	0	0	1	0	3
$\mathbf{x}_4^T$	0	1	0	1	0	0	0	4
$\mathbf{x}_5^T$	0	1	0	0	0	0	1	5
$\mathbf{x}_6^T$	0	0	1	0	1	0	0	1
$\mathbf{x}_7^T$	0	0	1	0	0	1	0	2
$\mathbf{x}_8^T$	0	0	1	0	0	0	1	4

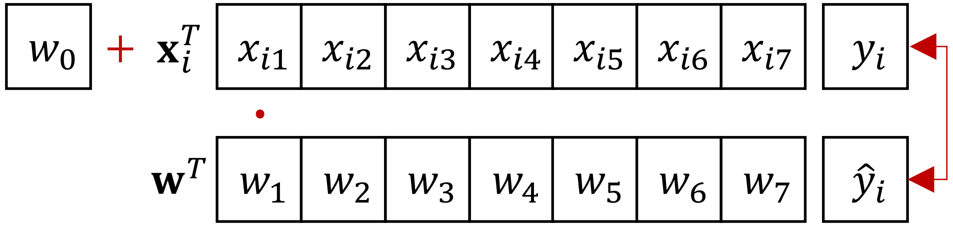
Linear regression

Model equation

- $\hat{y}_i = h(\mathbf{x}_i)$

  $= w_0 + \mathbf{w}^T \mathbf{x}_i$

  $= w_0 + \sum_{j=1}^{p} w_j x_{ij}$



w_0	$+$	$\mathbf{x}_i^T$	x_{i1}	x_{i2}	x_{i3}	x_{i4}	x_{i5}	x_{i6}	x_{i7}	y_i
		$\cdot$								
		$\mathbf{w}^T$	w_1	w_2	w_3	w_4	w_5	w_6	w_7	$\hat{y}_i$

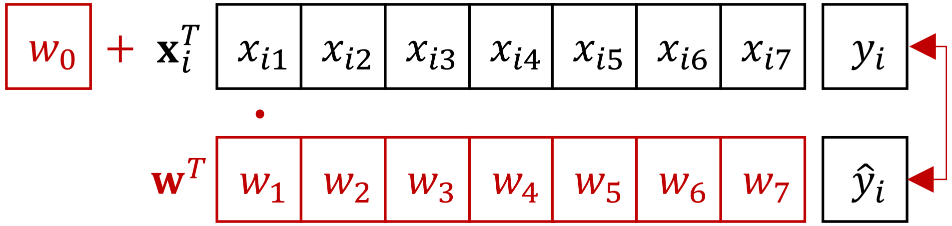
Linear regression

Model equation

- $\hat{y}_i = h(\mathbf{x}_i)$

$$= w_0 + \mathbf{w}^T \mathbf{x}_i$$

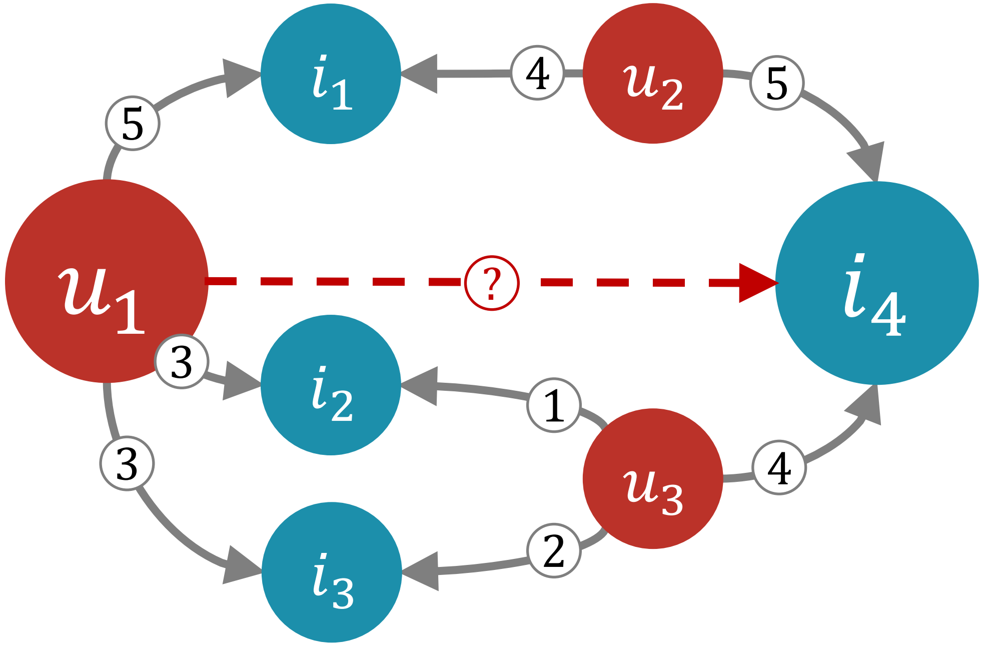
$$= w_0 + \sum_{j=1}^{p} w_j x_{ij}$$



Model parameters ($\mathcal{O}(p)$)

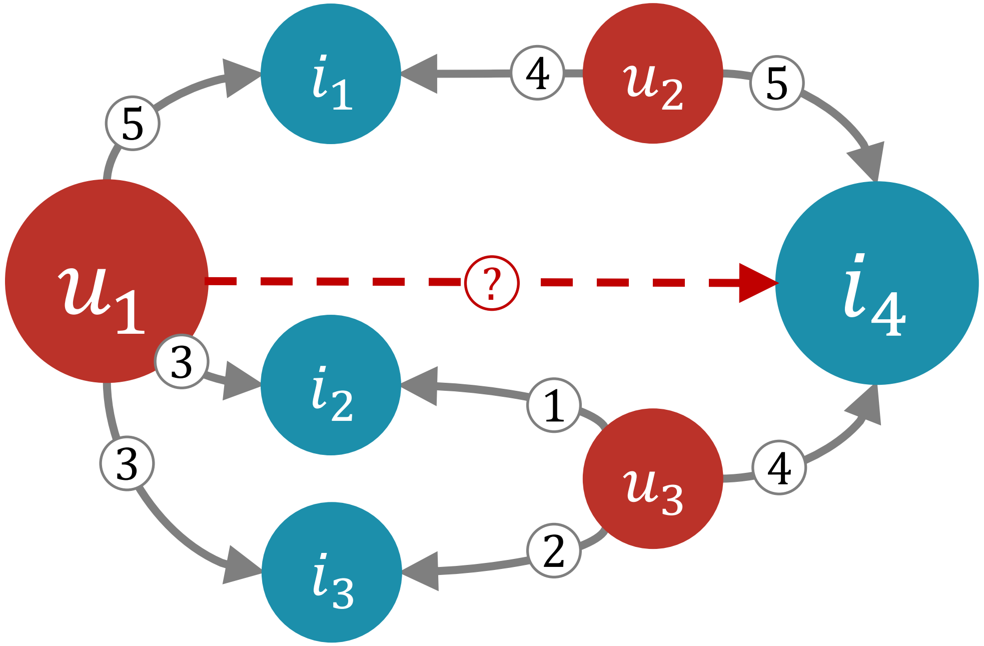
- $w_0 \in \mathbb{R}, \mathbf{w} \in \mathbb{R}^p$

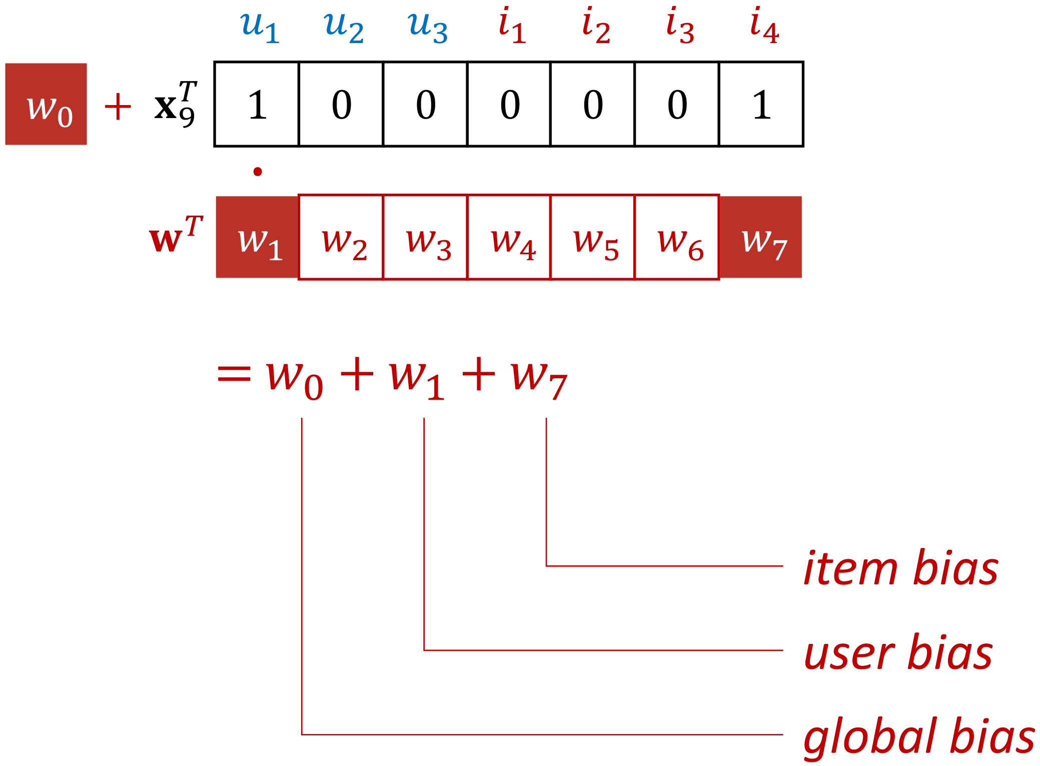
Linear regression



$$\begin{array}{rccccccc} & u_1 & u_2 & u_3 & i_1 & i_2 & i_3 & i_4 \\ \boxed{w_0} + \mathbf{x}_9^T & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ & \cdot & & & & & & \\ \mathbf{w}^T & w_1 & w_2 & w_3 & w_4 & w_5 & w_6 & w_7 \end{array}$$

Linear regression





$$w_0 + \mathbf{x}_9^T \begin{array}{|c|c|c|c|c|c|c|} u_1 & u_2 & u_3 & i_1 & i_2 & i_3 & i_4 \\ \hline 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ \hline \end{array}$$

$$\cdot$$

$$\mathbf{w}^T \begin{array}{|c|c|c|c|c|c|c|} \hline w_1 & w_2 & w_3 & w_4 & w_5 & w_6 & w_7 \\ \hline \end{array}$$

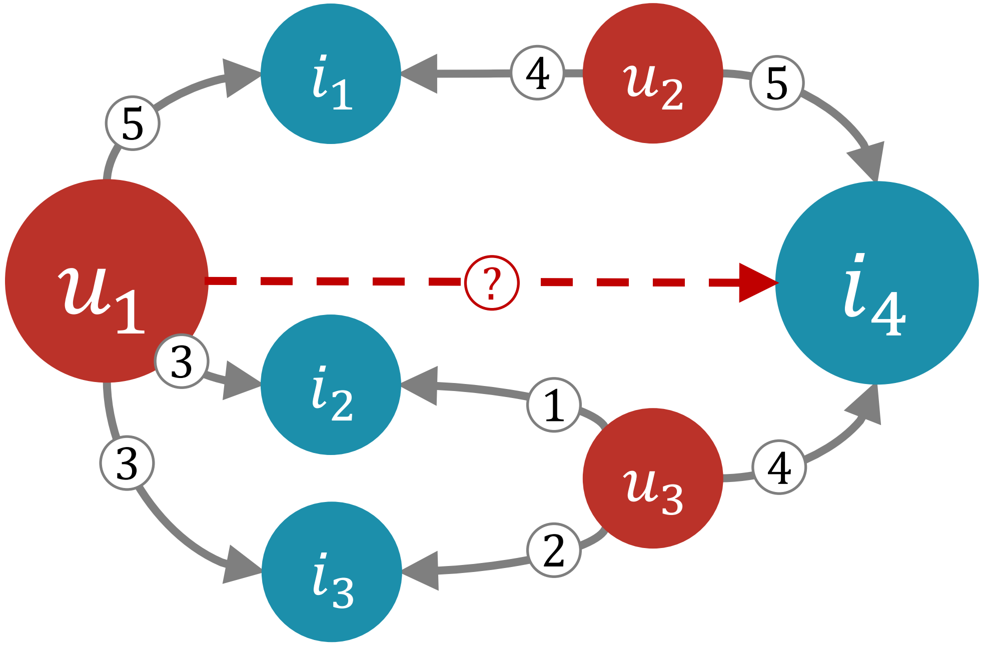
$$= w_0 + w_1 + w_7$$

item bias

user bias

global bias

Linear regression



$$\begin{array}{rccccccc} & u_1 & u_2 & u_3 & i_1 & i_2 & i_3 & i_4 \\ \boxed{w_0} + \mathbf{x}_9^T & \boxed{1} & \boxed{0} & \boxed{0} & \boxed{0} & \boxed{0} & \boxed{0} & \boxed{1} \\ & \cdot & & & & & & \\ \mathbf{w}^T & \boxed{w_1} & \boxed{w_2} & \boxed{w_3} & \boxed{w_4} & \boxed{w_5} & \boxed{w_6} & \boxed{w_7} \end{array}$$

$$= w_0 + w_1 + w_7$$

no interaction term

→ no personalization

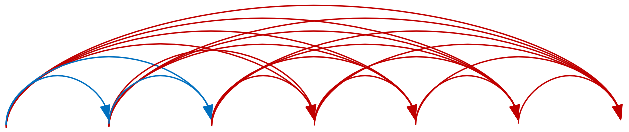
Feature interaction

Engineered interactions

- $\mathrm{CF}(u_1, i_1)$
- $\mathrm{CB}(u_1, i_1)$
- $\mathrm{KB}(u_1, i_1)$
- …

Problem: costly!

- Lots of potential interactions
- ***Mix of art and science***

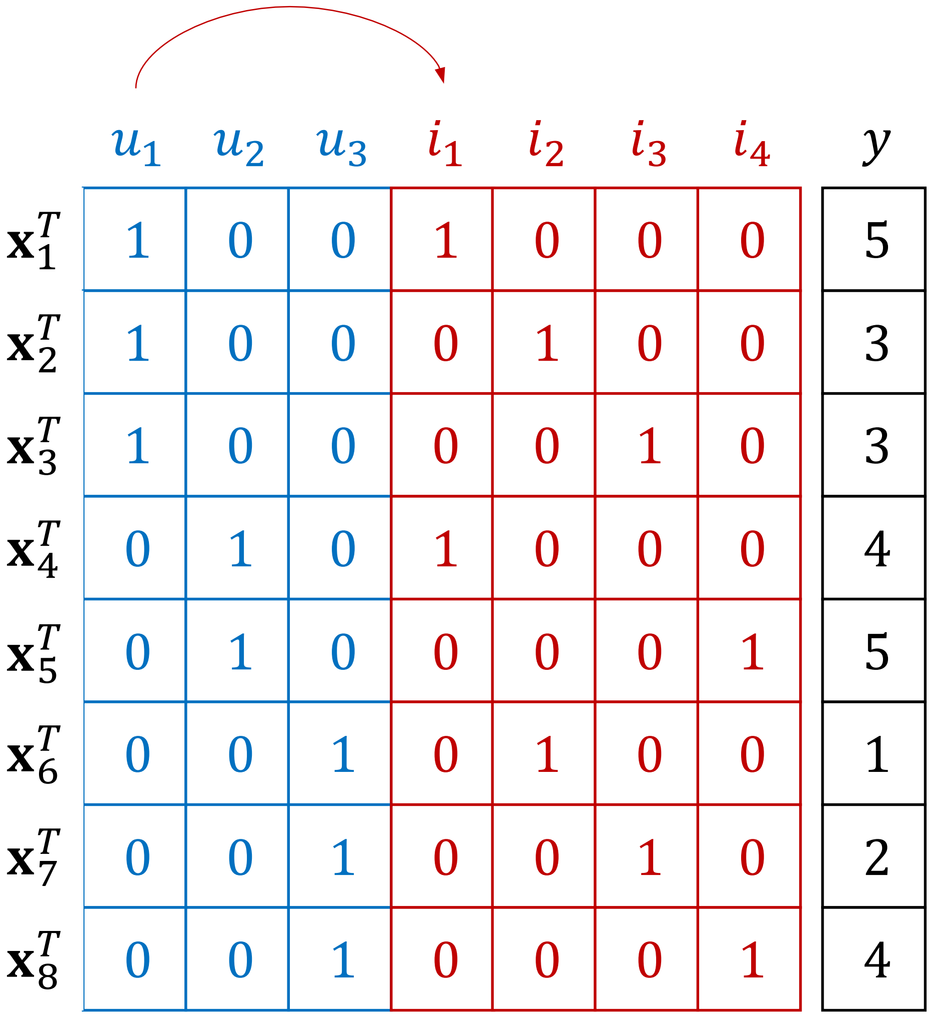


	u_1	u_2	u_3	i_1	i_2	i_3	i_4	y
$\mathbf{x}_1^T$	1	0	0	1	0	0	0	5
$\mathbf{x}_2^T$	1	0	0	0	1	0	0	3
$\mathbf{x}_3^T$	1	0	0	0	0	1	0	3
$\mathbf{x}_4^T$	0	1	0	1	0	0	0	4
$\mathbf{x}_5^T$	0	1	0	0	0	0	1	5
$\mathbf{x}_6^T$	0	0	1	0	1	0	0	1
$\mathbf{x}_7^T$	0	0	1	0	0	1	0	2
$\mathbf{x}_8^T$	0	0	1	0	0	0	1	4

Feature interaction

Learned interactions

- e.g. feature crosses

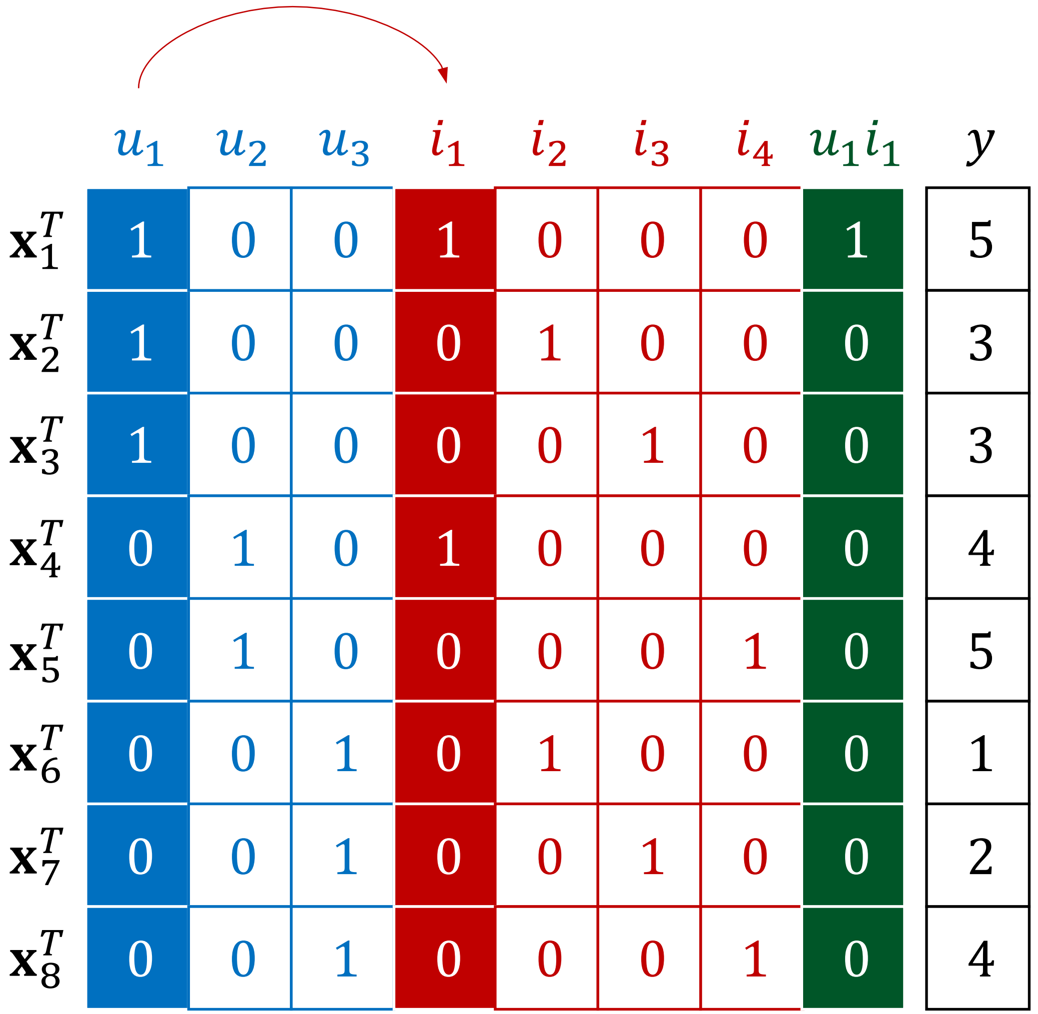


	u_1	u_2	u_3	i_1	i_2	i_3	i_4	y
$\mathbf{x}_1^T$	1	0	0	1	0	0	0	5
$\mathbf{x}_2^T$	1	0	0	0	1	0	0	3
$\mathbf{x}_3^T$	1	0	0	0	0	1	0	3
$\mathbf{x}_4^T$	0	1	0	1	0	0	0	4
$\mathbf{x}_5^T$	0	1	0	0	0	0	1	5
$\mathbf{x}_6^T$	0	0	1	0	1	0	0	1
$\mathbf{x}_7^T$	0	0	1	0	0	1	0	2
$\mathbf{x}_8^T$	0	0	1	0	0	0	1	4

Feature interaction

Learned interactions

- e.g. feature crosses



	u_1	u_2	u_3	i_1	i_2	i_3	i_4	$u_1 i_1$	y
$\mathbf{x}_1^T$	1	0	0	1	0	0	0	1	5
$\mathbf{x}_2^T$	1	0	0	0	1	0	0	0	3
$\mathbf{x}_3^T$	1	0	0	0	0	1	0	0	3
$\mathbf{x}_4^T$	0	1	0	1	0	0	0	0	4
$\mathbf{x}_5^T$	0	1	0	0	0	0	1	0	5
$\mathbf{x}_6^T$	0	0	1	0	1	0	0	0	1
$\mathbf{x}_7^T$	0	0	1	0	0	1	0	0	2
$\mathbf{x}_8^T$	0	0	1	0	0	0	1	0	4

Feature interaction

Learned interactions

- e.g. feature crosses

p features

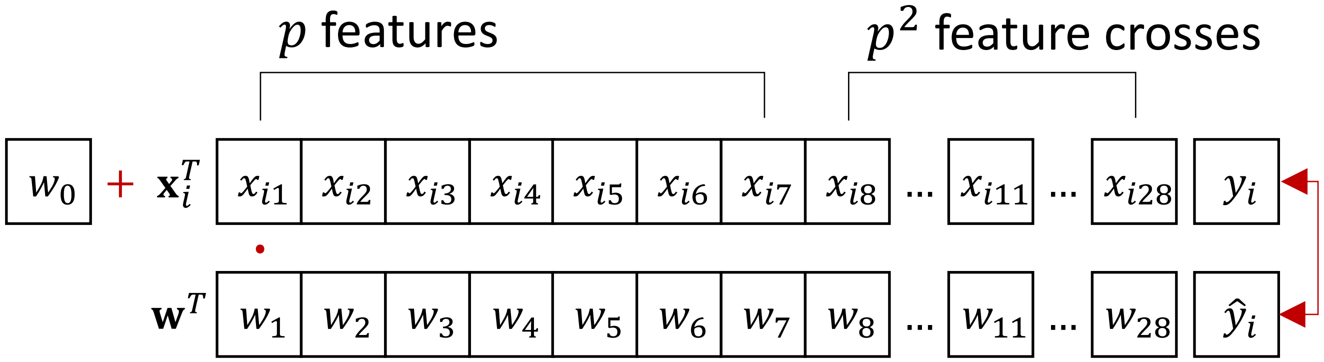
p^2 feature crosses

	u_1	u_2	u_3	i_1	i_2	i_3	i_4	$u_1 u_2$		$i_3 i_4$	y
$\mathbf{x}_1^T$	1	0	0	1	0	0	0	0		0	5
$\mathbf{x}_2^T$	1	0	0	0	1	0	0	0		0	3
$\mathbf{x}_3^T$	1	0	0	0	0	1	0	0		0	3
$\mathbf{x}_4^T$	0	1	0	1	0	0	0	0	…	0	4
$\mathbf{x}_5^T$	0	1	0	0	0	0	1	0		0	5
$\mathbf{x}_6^T$	0	0	1	0	1	0	0	0		0	1
$\mathbf{x}_7^T$	0	0	1	0	0	1	0	0		0	2
$\mathbf{x}_8^T$	0	0	1	0	0	0	1	0		0	4

Polynomial regression

Model equation (degree 2)

- $$\begin{aligned}\hat{y}_i &= h(\mathbf{x}_i)\\ &= w_0 + \mathbf{w}^T\mathbf{x}_i\\ &\quad +\mathbf{x}_i^T\mathbf{W}\mathbf{x}_i\\ &= w_0 + \sum_{j=1}^{p} w_j x_{ij}\\ &\quad + \sum_{j=1}^{p}\sum_{k=j+1}^{p} w_{jk}x_{ij}x_{ik}\end{aligned}$$



p features

p^2 feature crosses

w_0 + $\mathbf{x}_i^T$ x_{i1} x_{i2} x_{i3} x_{i4} x_{i5} x_{i6} x_{i7} x_{i8} … x_{i11} … x_{i28} y_i

$\mathbf{w}^T$ w_1 w_2 w_3 w_4 w_5 w_6 w_7 w_8 … w_{11} … w_{28} $\hat{y}_i$

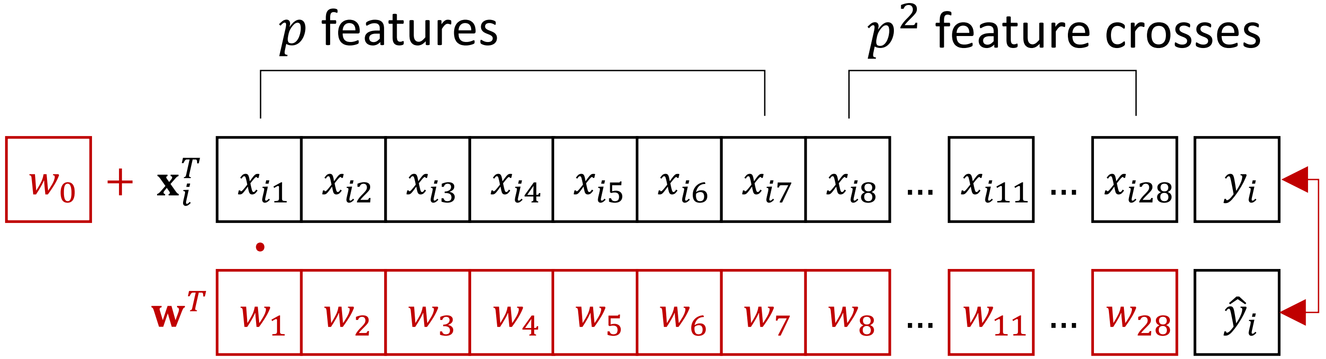
Polynomial regression

Model equation (degree 2)

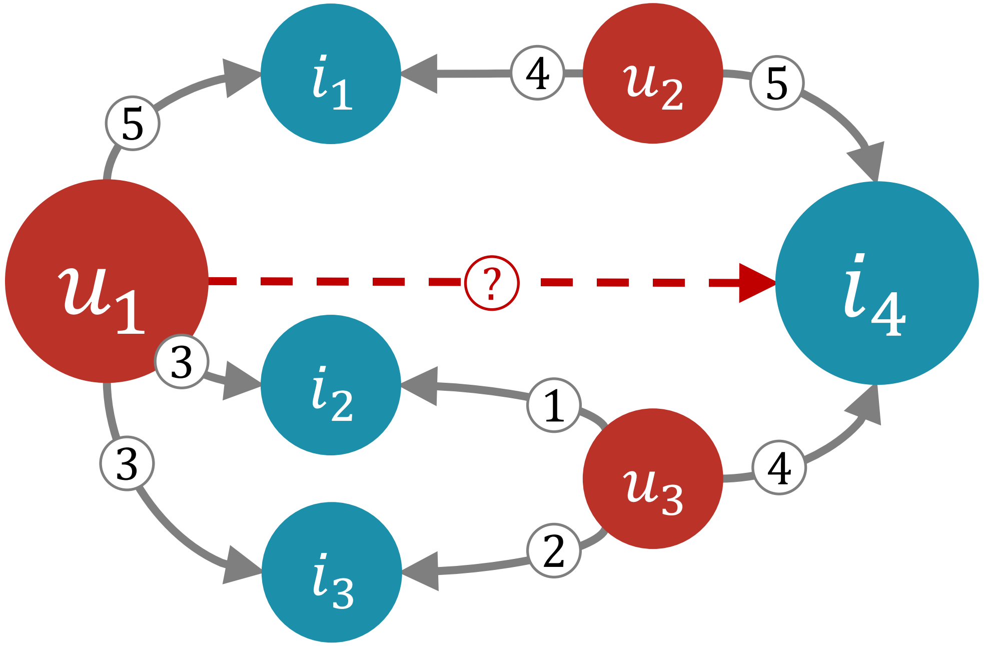
- $$\begin{aligned}\hat{y}_i &= h(\mathbf{x}_i)\\ &= w_0 + \mathbf{w}^T\mathbf{x}_i\\ &\quad +\mathbf{x}_i^T\mathbf{W}\mathbf{x}_i\\ &= w_0 + \sum_{j=1}^{p} w_j x_{ij}\\ &\quad + \sum_{j=1}^{p}\sum_{k=j+1}^{p} w_{jk} x_{ij} x_{ik}\end{aligned}$$

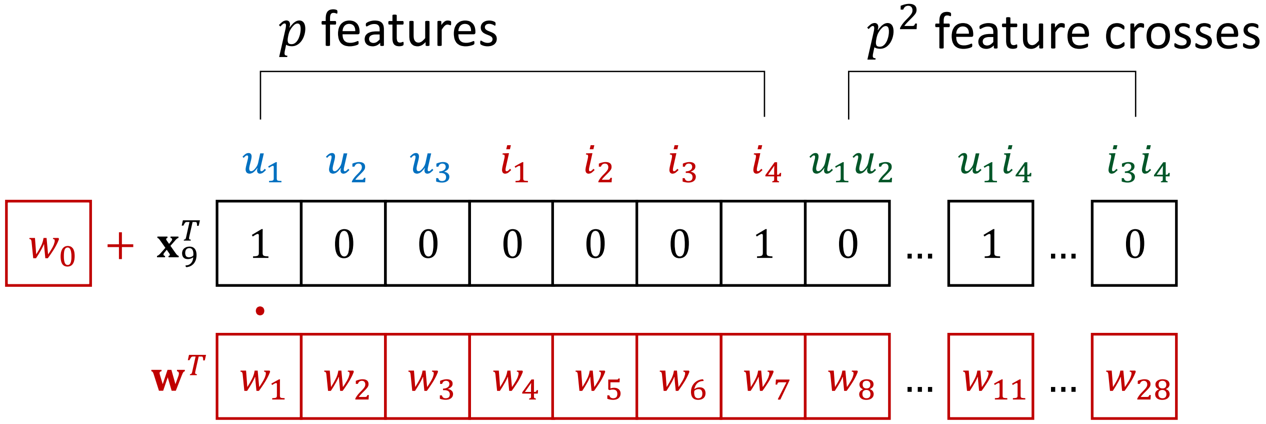
Model parameters ($\mathcal{O}(p^2)$)

- $w_0 \in \mathbb{R}, \mathbf{w} \in \mathbb{R}^p, \mathbf{W} \in \mathbb{R}^{p\times p}$



Polynomial regression



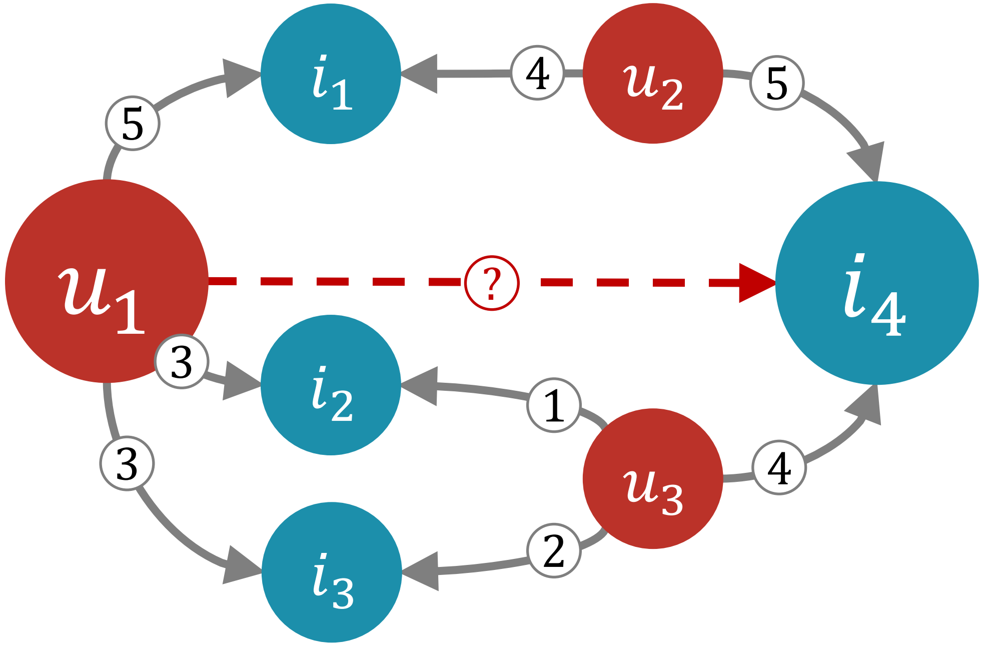


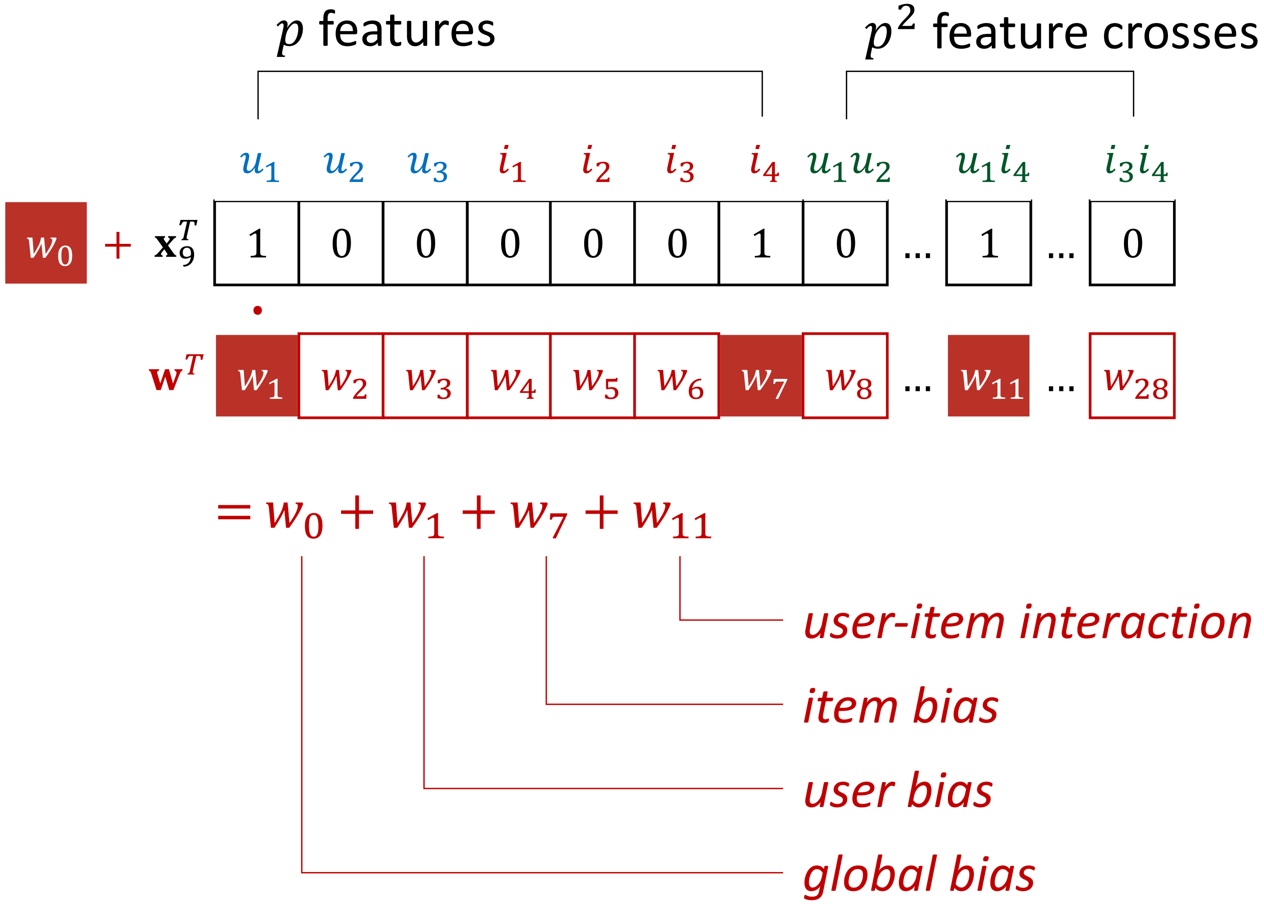
p features

p^2 feature crosses

$$\begin{array}{cccccccccccccc} & & u_1 & u_2 & u_3 & i_1 & i_2 & i_3 & i_4 & u_1u_2 & & u_1i_4 & & i_3i_4 \\ \boxed{w_0} & + \; \mathbf{x}_9^T & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & \dots & 1 & \dots & 0 \\ & & \cdot & & & & & & & & & & & \\ & \mathbf{w}^T & w_1 & w_2 & w_3 & w_4 & w_5 & w_6 & w_7 & w_8 & \dots & w_{11} & \dots & w_{28} \end{array}$$

Polynomial regression





p features

p^2 feature crosses

$$\begin{array}{c|cccccccccccc} & u_1 & u_2 & u_3 & i_1 & i_2 & i_3 & i_4 & u_1u_2 & \ldots & u_1i_4 & \ldots & i_3i_4 \\ w_0 + \mathbf{x}_9^T & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & \ldots & 1 & \ldots & 0 \\ \cdot & & & & & & & & & & & & \\ \mathbf{w}^T & w_1 & w_2 & w_3 & w_4 & w_5 & w_6 & w_7 & w_8 & \ldots & w_{11} & \ldots & w_{28} \end{array}$$

$$= w_0 + w_1 + w_7 + w_{11}$$

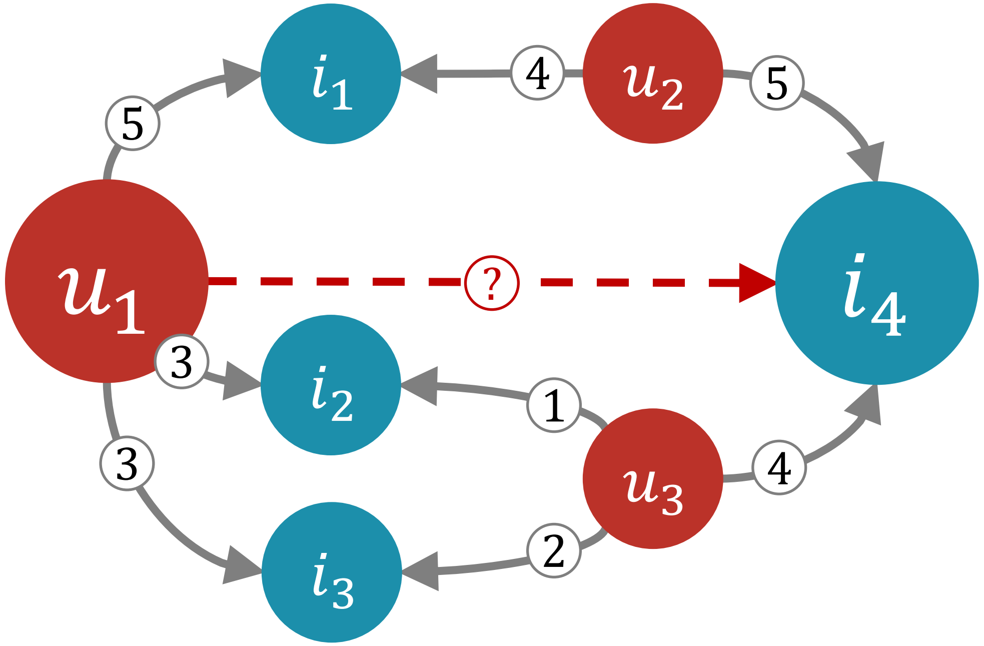
user-item interaction

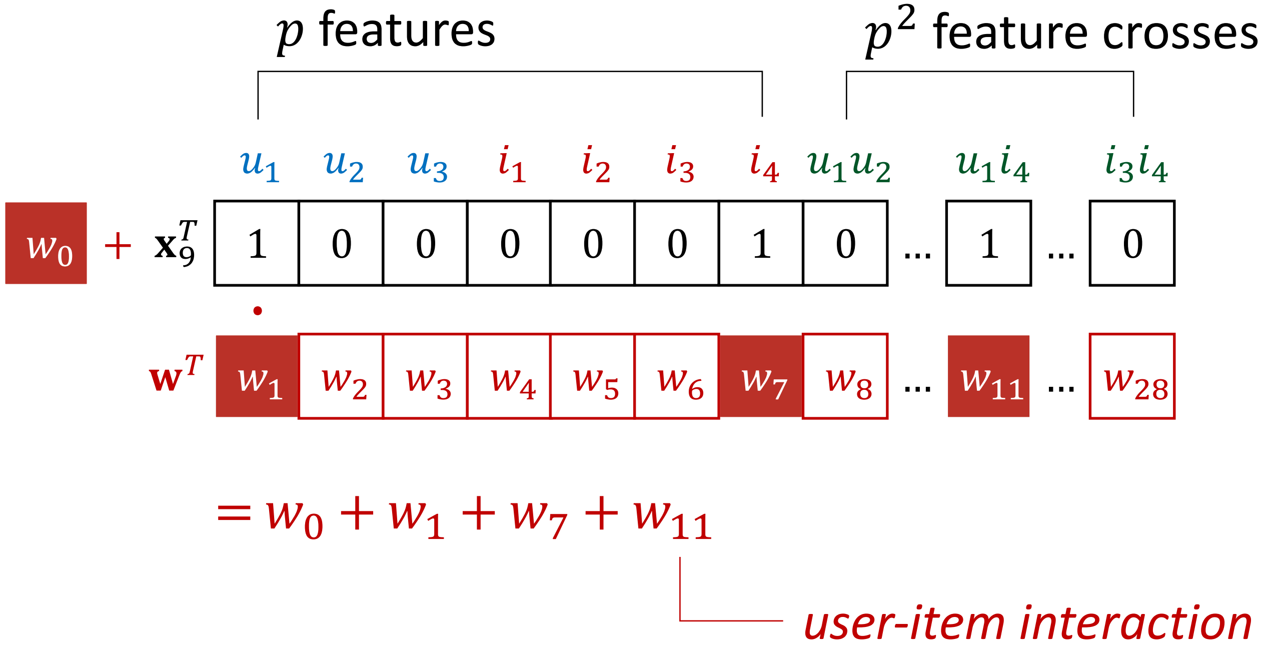
item bias

user bias

global bias

Polynomial regression





p features

p^2 feature crosses

	u_1	u_2	u_3	i_1	i_2	i_3	i_4	u_1u_2		u_1i_4		i_3i_4
$w_0 + \mathbf{x}_9^T$	1	0	0	0	0	0	1	0	...	1	...	0
·												
$\mathbf{w}^T$	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	...	w_{11}	...	w_{28}

$$= w_0 + w_1 + w_7 + w_{11}$$

user-item interaction

$u_1 i_4$ previously unseen

→ w_{11} undefined

Overfitting

Many more features ($\mathcal{O}(p^2)$) than instances ($\mathcal{O}(|R|)$)

- Model will overfit to available training

In practice

- $w_{ui} = \begin{cases} y - w_0 - w_u - w_i & \text{if } \langle u, i, y \rangle \in R \\ \text{undefined} & \text{otherwise} \end{cases}$

Bridging the gap

Latent factor models (e.g. matrix factorization)

- Effective for large categorical domains

Feature-based models (e.g. regression models)

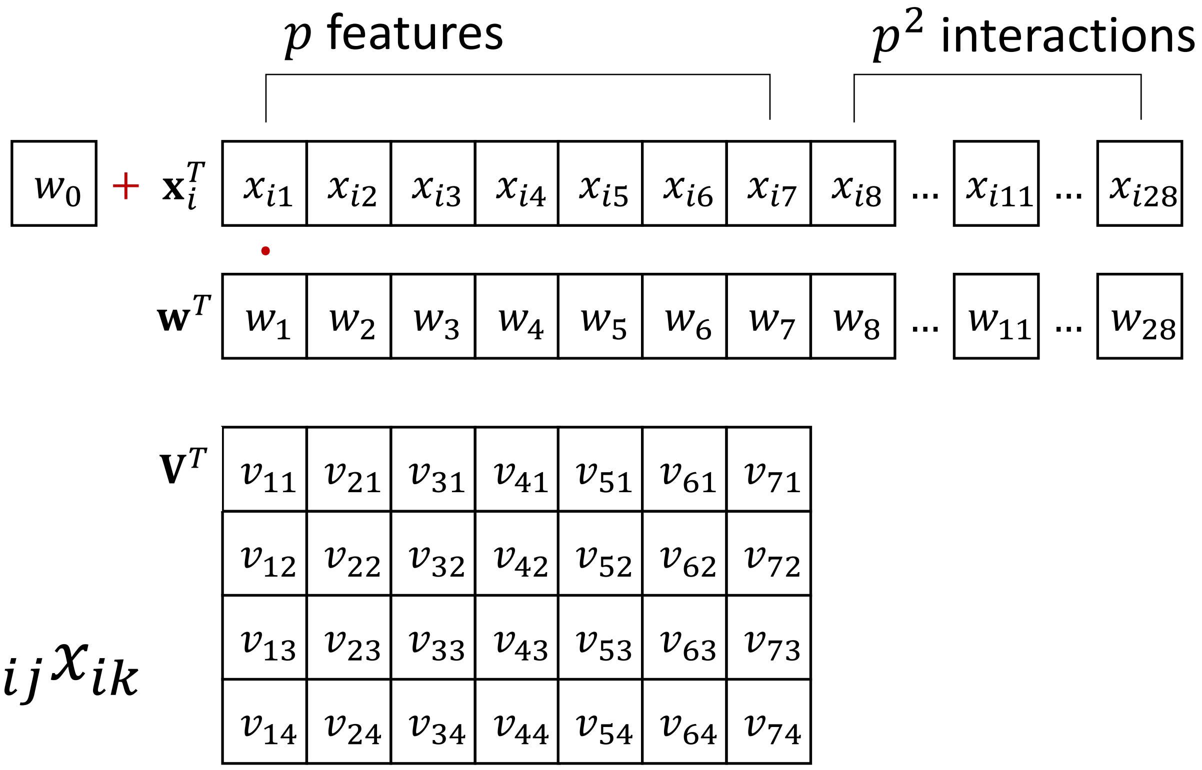
- Flexible to allow arbitrary features

How can these advantages be combined?

Factorization machines [Rendle, 2010]

Model equation (degree 2)

- $$\begin{aligned} \hat{y}_i &= h(\mathbf{x}_i) \\ &= w_0 + \mathbf{w}^T\mathbf{x}_i \\ &\quad +\mathbf{x}_i^T\mathbf{V}\,\mathbf{V}^T\mathbf{x}_i \\ &= w_0 + \textstyle\sum_{j=1}^{p} w_j x_{ij} \\ &\quad + \textstyle\sum_{j=1}^{p}\sum_{k=j+1}^{p}\langle v_j, v_k\rangle\, x_{ij}x_{ik} \end{aligned}$$



p features

p^2 interactions

$$\boxed{w_0} + \mathbf{x}_i^T \; \boxed{x_{i1}}\boxed{x_{i2}}\boxed{x_{i3}}\boxed{x_{i4}}\boxed{x_{i5}}\boxed{x_{i6}}\boxed{x_{i7}}\boxed{x_{i8}} \dots \boxed{x_{i11}} \dots \boxed{x_{i28}}$$

$$\mathbf{w}^T \; \boxed{w_1}\boxed{w_2}\boxed{w_3}\boxed{w_4}\boxed{w_5}\boxed{w_6}\boxed{w_7}\boxed{w_8} \dots \boxed{w_{11}} \dots \boxed{w_{28}}$$

$$\mathbf{V}^T \; \begin{array}{|c|c|c|c|c|c|c|} \hline v_{11} & v_{21} & v_{31} & v_{41} & v_{51} & v_{61} & v_{71} \\ \hline v_{12} & v_{22} & v_{32} & v_{42} & v_{52} & v_{62} & v_{72} \\ \hline v_{13} & v_{23} & v_{33} & v_{43} & v_{53} & v_{63} & v_{73} \\ \hline v_{14} & v_{24} & v_{34} & v_{44} & v_{54} & v_{64} & v_{74} \\ \hline \end{array}$$

Factorization machines [Rendle, 2010]

Model equation (degree 2)

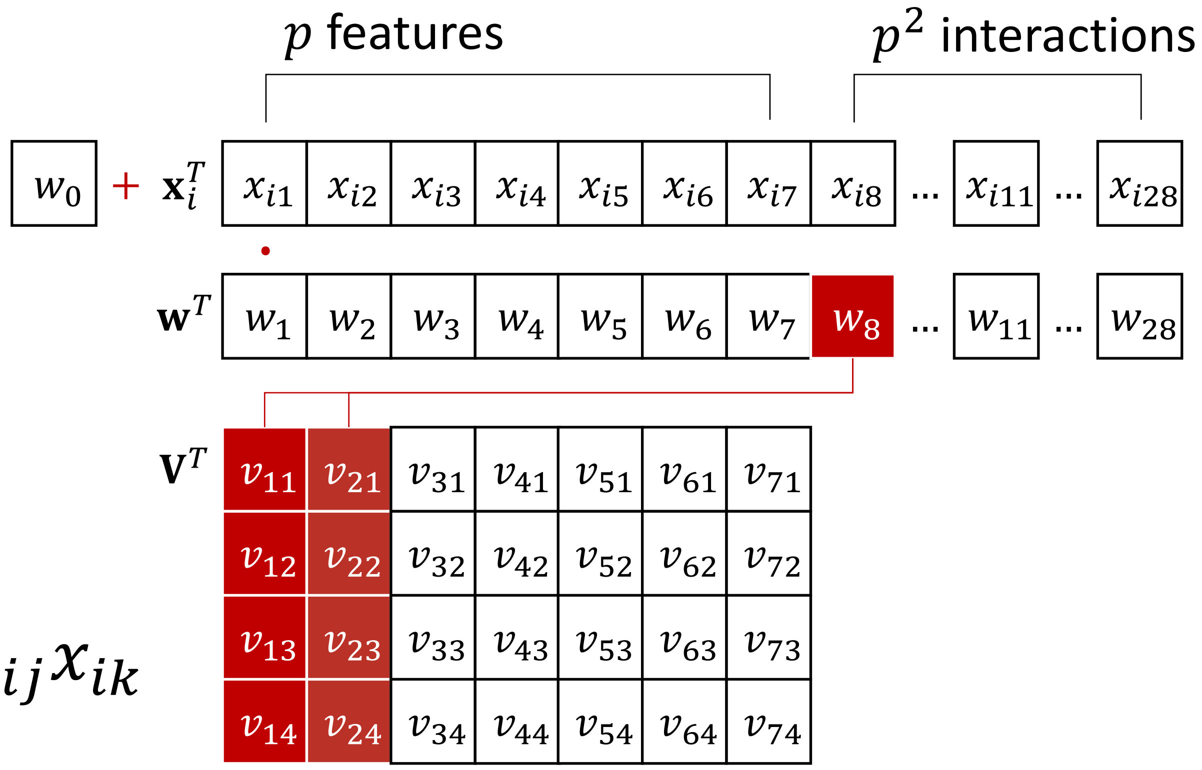
- $\hat{y}_i = h(\mathbf{x}_i)$

$$= w_0 + \mathbf{w}^T \mathbf{x}_i$$

$$+ \mathbf{x}_i^T \mathbf{V}\, \mathbf{V}^T \mathbf{x}_i$$

$$= w_0 + \sum_{j=1}^{p} w_j x_{ij}$$

$$+ \sum_{j=1}^{p} \sum_{k=j+1}^{p} \langle v_j, v_k \rangle\, x_{ij} x_{ik}$$



$$w_0 + \mathbf{x}_i^T \begin{array}{|c|c|c|c|c|c|c|c|c|c|c|c|} \hline x_{i1} & x_{i2} & x_{i3} & x_{i4} & x_{i5} & x_{i6} & x_{i7} & x_{i8} & \ldots & x_{i11} & \ldots & x_{i28} \\ \hline \end{array}$$

$$\cdot$$

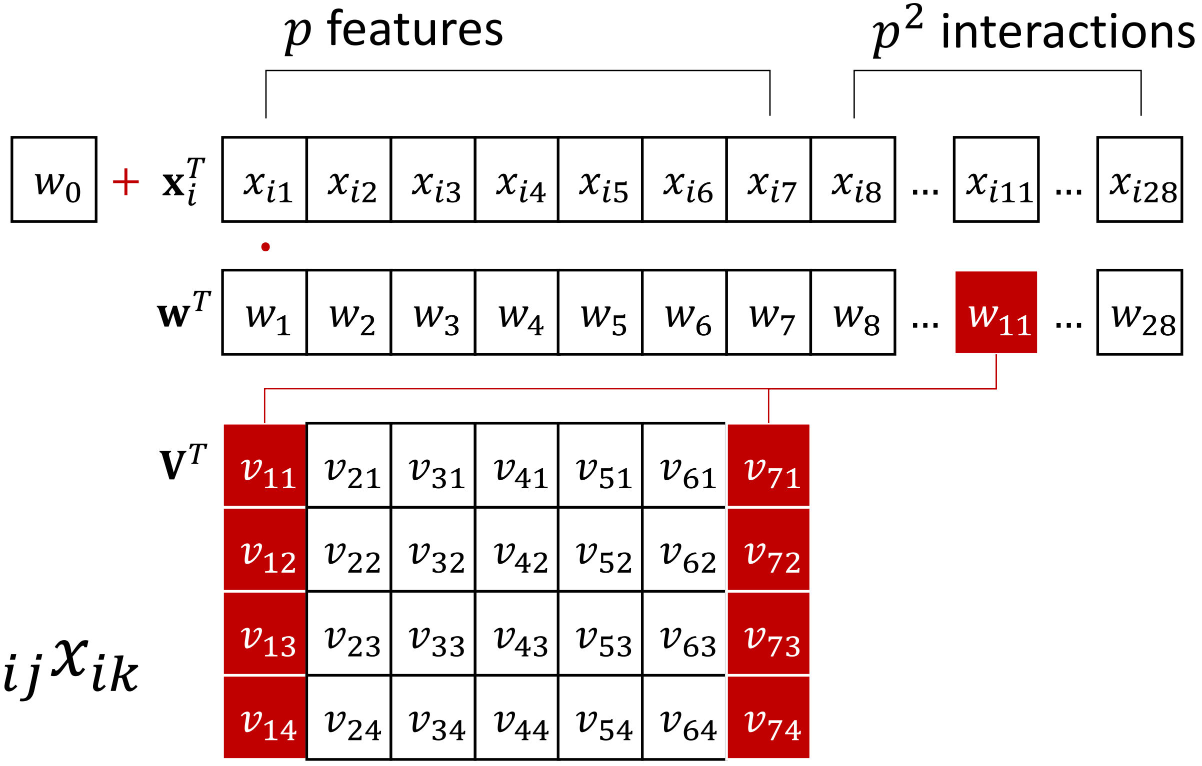
$$\mathbf{w}^T \begin{array}{|c|c|c|c|c|c|c|c|c|c|c|c|} \hline w_1 & w_2 & w_3 & w_4 & w_5 & w_6 & w_7 & w_8 & \ldots & w_{11} & \ldots & w_{28} \\ \hline \end{array}$$

$$\mathbf{V}^T \begin{array}{|c|c|c|c|c|c|c|} \hline v_{11} & v_{21} & v_{31} & v_{41} & v_{51} & v_{61} & v_{71} \\ \hline v_{12} & v_{22} & v_{32} & v_{42} & v_{52} & v_{62} & v_{72} \\ \hline v_{13} & v_{23} & v_{33} & v_{43} & v_{53} & v_{63} & v_{73} \\ \hline v_{14} & v_{24} & v_{34} & v_{44} & v_{54} & v_{64} & v_{74} \\ \hline \end{array}$$

Factorization machines [Rendle, 2010]

Model equation (degree 2)

- $$\begin{aligned}\hat{y}_i &= h(\mathbf{x}_i) \\ &= w_0 + \mathbf{w}^T\mathbf{x}_i \\ &\quad + \mathbf{x}_i^T \mathbf{V}\,\mathbf{V}^T \mathbf{x}_i \\ &= w_0 + \textstyle\sum_{j=1}^{p} w_j x_{ij} \\ &\quad + \textstyle\sum_{j=1}^{p} \sum_{k=j+1}^{p} \langle v_j, v_k \rangle\, x_{ij} x_{ik}\end{aligned}$$



p features

p^2 interactions

w_0	+ $\mathbf{x}_i^T$	x_{i1}	x_{i2}	x_{i3}	x_{i4}	x_{i5}	x_{i6}	x_{i7}	x_{i8}	…	x_{i11}	…	x_{i28}
	$\mathbf{w}^T$	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	…	w_{11}	…	w_{28}

$\mathbf{V}^T$	v_{11}	v_{21}	v_{31}	v_{41}	v_{51}	v_{61}	v_{71}
	v_{12}	v_{22}	v_{32}	v_{42}	v_{52}	v_{62}	v_{72}
	v_{13}	v_{23}	v_{33}	v_{43}	v_{53}	v_{63}	v_{73}
	v_{14}	v_{24}	v_{34}	v_{44}	v_{54}	v_{64}	v_{74}

Factorization machines [Rendle, 2010]

Model equation (degree 2)

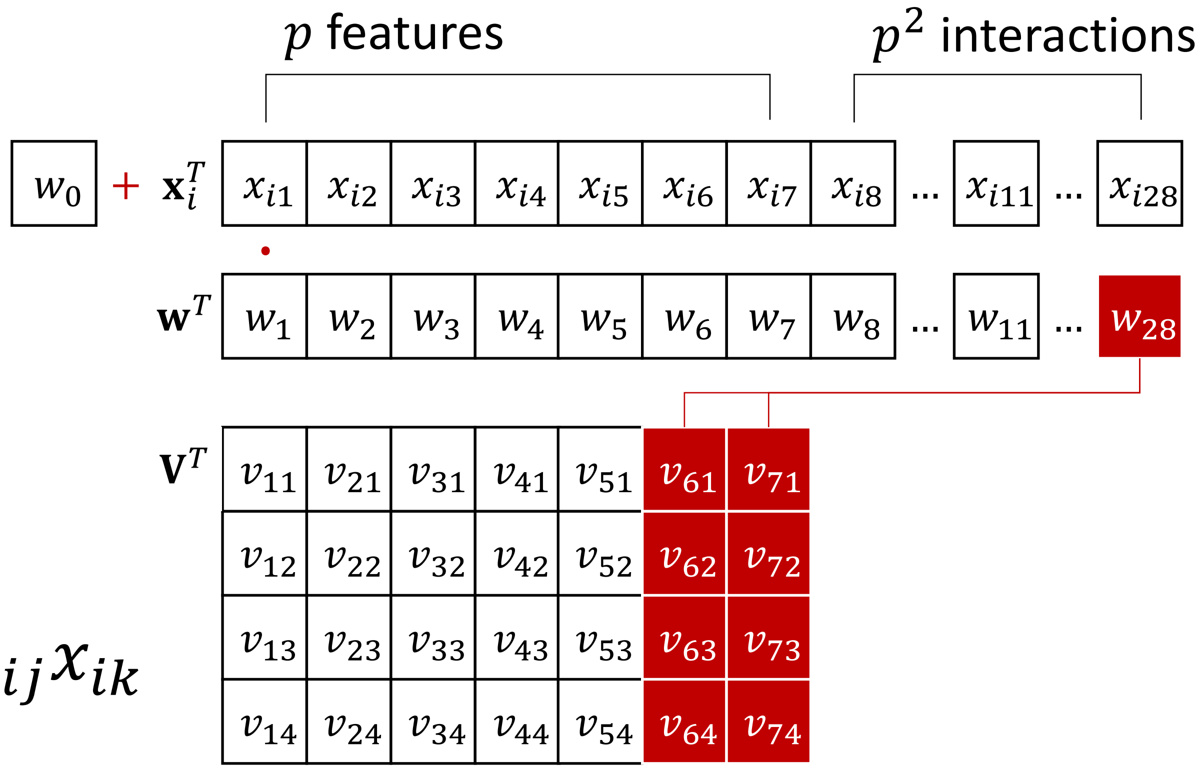
- $\hat{y}_i = h(\mathbf{x}_i)$

$$= w_0 + \mathbf{w}^T\mathbf{x}_i$$

$$+\mathbf{x}_i^T\mathbf{V}\,\mathbf{V}^T\mathbf{x}_i$$

$$= w_0 + \sum_{j=1}^{p} w_j x_{ij}$$

$$+ \sum_{j=1}^{p}\sum_{k=j+1}^{p}\langle v_j, v_k\rangle\, x_{ij}x_{ik}$$



p features

p^2 interactions

w_0	+ $\mathbf{x}_i^T$	x_{i1}	x_{i2}	x_{i3}	x_{i4}	x_{i5}	x_{i6}	x_{i7}	x_{i8}	…	x_{i11}	…	x_{i28}
	$\mathbf{w}^T$	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	…	w_{11}	…	w_{28}

$\mathbf{V}^T$	v_{11}	v_{21}	v_{31}	v_{41}	v_{51}	v_{61}	v_{71}
	v_{12}	v_{22}	v_{32}	v_{42}	v_{52}	v_{62}	v_{72}
	v_{13}	v_{23}	v_{33}	v_{43}	v_{53}	v_{63}	v_{73}
	v_{14}	v_{24}	v_{34}	v_{44}	v_{54}	v_{64}	v_{74}

Factorization machines [Rendle, 2010]

Model equation (degree 2)

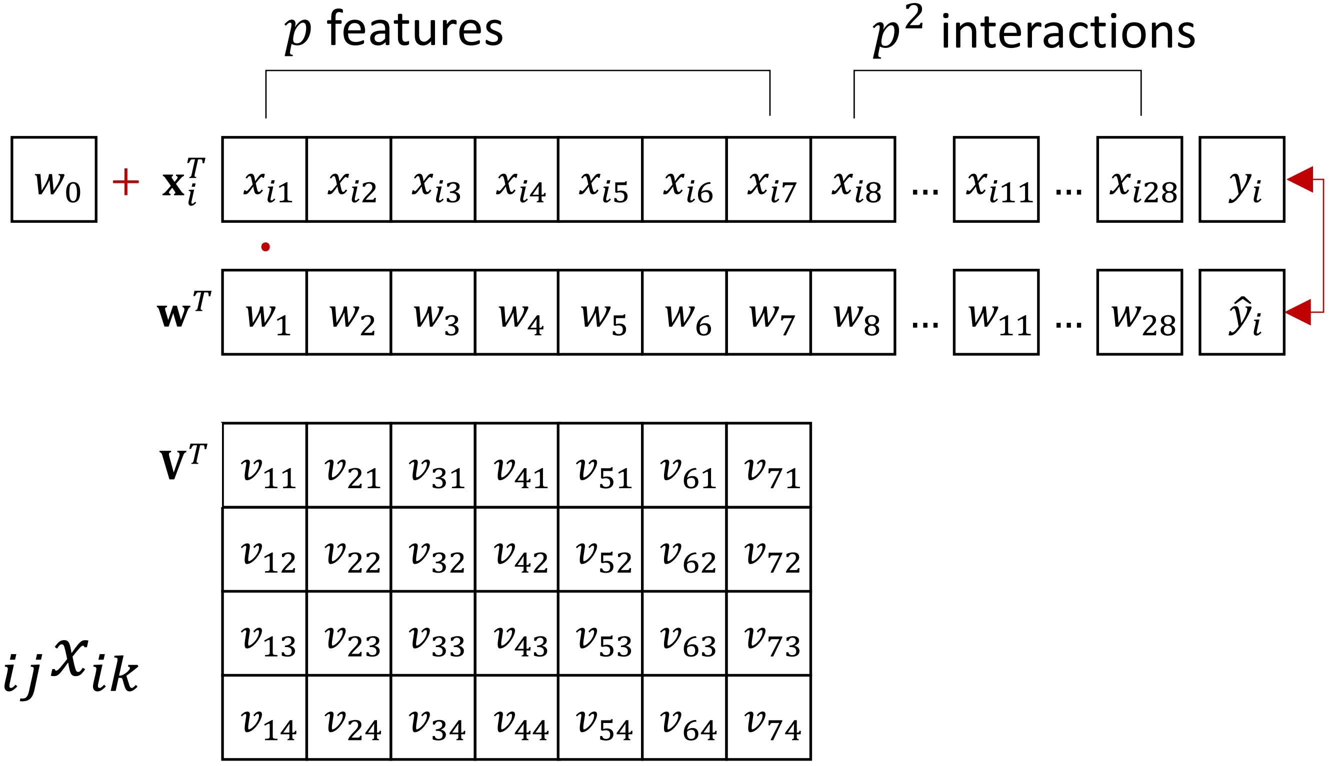
- $\hat{y}_i = h(\mathbf{x}_i)$

$$= w_0 + \mathbf{w}^T \mathbf{x}_i$$

$$+ \mathbf{x}_i^T \mathbf{V}\,\mathbf{V}^T \mathbf{x}_i$$

$$= w_0 + \sum_{j=1}^{p} w_j x_{ij}$$

$$+ \sum_{j=1}^{p} \sum_{k=j+1}^{p} \langle v_j, v_k \rangle \, x_{ij} x_{ik}$$



$$\boxed{w_0} + \mathbf{x}_i^T \begin{array}{|c|c|c|c|c|c|c|c|c|c|c|c|c|c|}\hline x_{i1} & x_{i2} & x_{i3} & x_{i4} & x_{i5} & x_{i6} & x_{i7} & x_{i8} & \ldots & x_{i11} & \ldots & x_{i28} & & y_i \\ \hline \end{array}$$

$$\mathbf{w}^T \begin{array}{|c|c|c|c|c|c|c|c|c|c|c|c|c|c|}\hline w_1 & w_2 & w_3 & w_4 & w_5 & w_6 & w_7 & w_8 & \ldots & w_{11} & \ldots & w_{28} & & \hat{y}_i \\ \hline \end{array}$$

$$\mathbf{V}^T \begin{array}{|c|c|c|c|c|c|c|}\hline v_{11} & v_{21} & v_{31} & v_{41} & v_{51} & v_{61} & v_{71} \\ \hline v_{12} & v_{22} & v_{32} & v_{42} & v_{52} & v_{62} & v_{72} \\ \hline v_{13} & v_{23} & v_{33} & v_{43} & v_{53} & v_{63} & v_{73} \\ \hline v_{14} & v_{24} & v_{34} & v_{44} & v_{54} & v_{64} & v_{74} \\ \hline \end{array}$$

Factorization machines [Rendle, 2010]

Model equation (degree 2)

- $\hat{y}_i = h(\mathbf{x}_i)$

$$= w_0 + \mathbf{w}^T\mathbf{x}_i$$

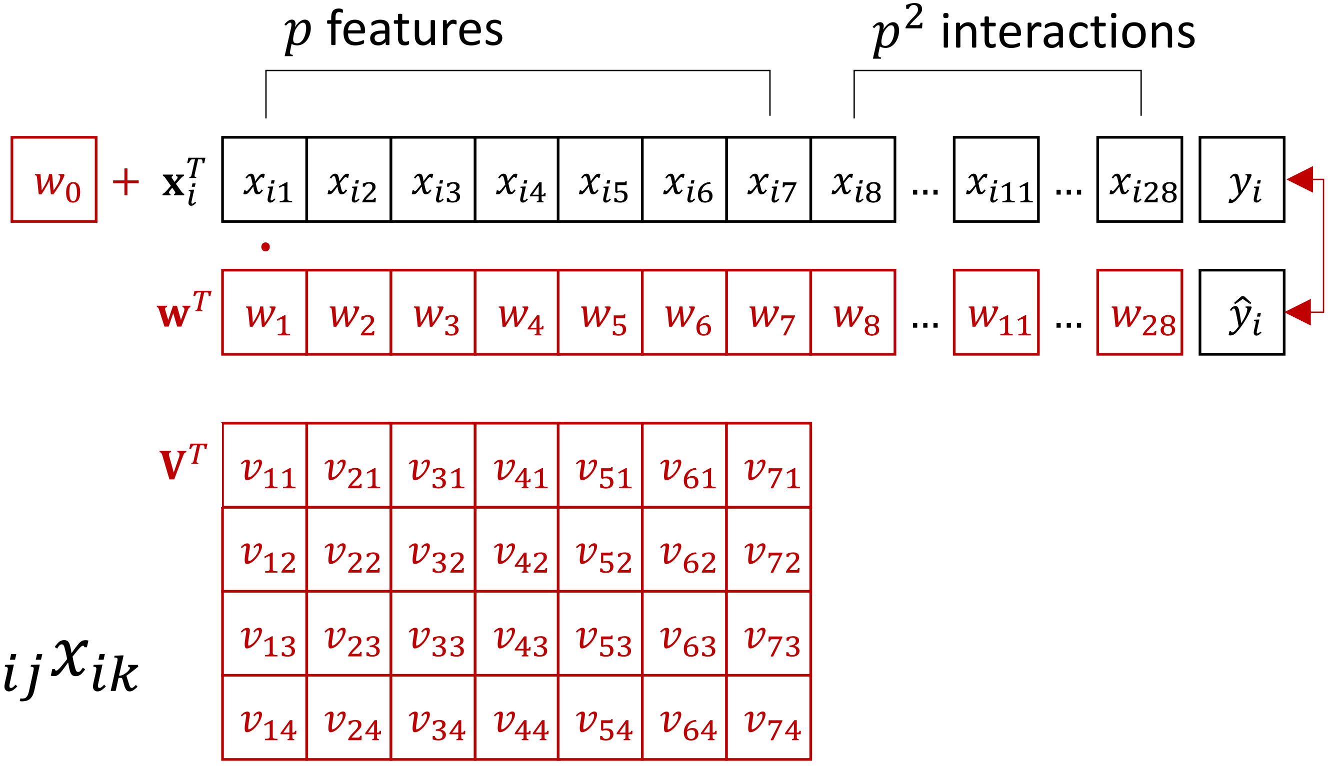
$$+\mathbf{x}_i^T \mathbf{V}\,\mathbf{V}^T \mathbf{x}_i$$

$$= w_0 + \sum_{j=1}^{p} w_j x_{ij}$$

$$+ \sum_{j=1}^{p} \sum_{k=j+1}^{p} \langle v_j, v_k \rangle \, x_{ij} x_{ik}$$

Model parameters ($\mathcal{O}(pd)$)

- $w_0 \in \mathbb{R}, \mathbf{w} \in \mathbb{R}^p, \mathbf{V} \in \mathbb{R}^{p \times d}$



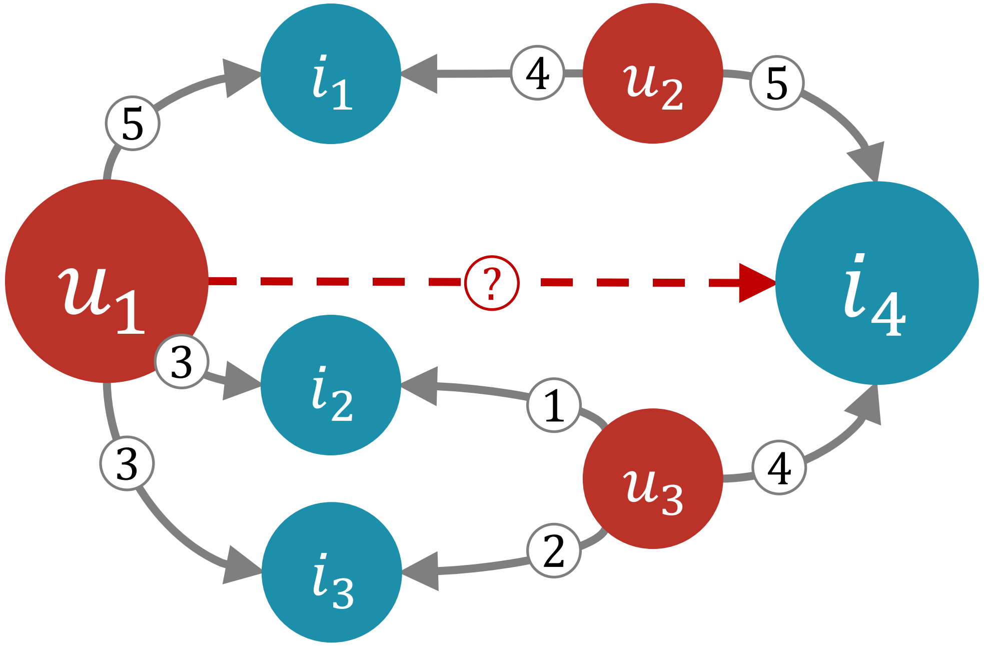
p features

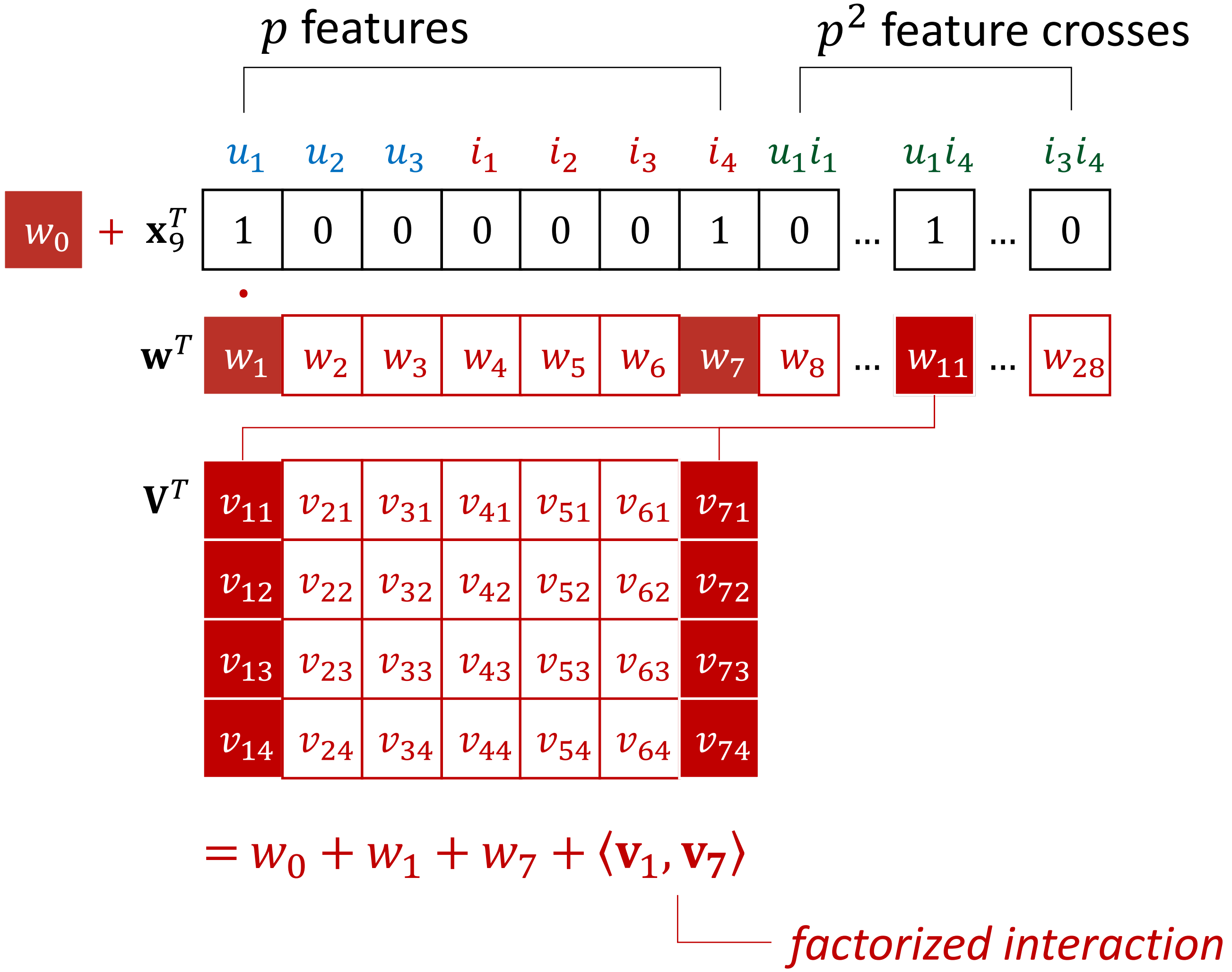
p^2 interactions

w_0	+ $\mathbf{x}_i^T$	x_{i1}	x_{i2}	x_{i3}	x_{i4}	x_{i5}	x_{i6}	x_{i7}	x_{i8}	…	x_{i11}	…	x_{i28}	y_i
	$\mathbf{w}^T$	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	…	w_{11}	…	w_{28}	$\hat{y}_i$

$\mathbf{V}^T$	v_{11}	v_{21}	v_{31}	v_{41}	v_{51}	v_{61}	v_{71}
	v_{12}	v_{22}	v_{32}	v_{42}	v_{52}	v_{62}	v_{72}
	v_{13}	v_{23}	v_{33}	v_{43}	v_{53}	v_{63}	v_{73}
	v_{14}	v_{24}	v_{34}	v_{44}	v_{54}	v_{64}	v_{74}

Factorization machines [Rendle, 2010]





p features

p^2 feature crosses

	u_1	u_2	u_3	i_1	i_2	i_3	i_4	$u_1 i_1$		$u_1 i_4$		$i_3 i_4$
$w_0 + \mathbf{x}_9^T$	1	0	0	0	0	0	1	0	...	1	...	0
$\mathbf{w}^T$	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	...	w_{11}	...	w_{28}
$\mathbf{V}^T$	v_{11}	v_{21}	v_{31}	v_{41}	v_{51}	v_{61}	v_{71}					
	v_{12}	v_{22}	v_{32}	v_{42}	v_{52}	v_{62}	v_{72}					
	v_{13}	v_{23}	v_{33}	v_{43}	v_{53}	v_{63}	v_{73}					
	v_{14}	v_{24}	v_{34}	v_{44}	v_{54}	v_{64}	v_{74}					

$$= w_0 + w_1 + w_7 + \underbrace{\langle \mathbf{v}_1, \mathbf{v}_7 \rangle}$$

factorized interaction

FM vs Poly2

Factorization machines

- $\hat{y}_i = h(\mathbf{x}_i)$

  $= w_0 + \mathbf{w}^T \mathbf{x}_i$

  $+\mathbf{x}_i^T \mathbf{V} \mathbf{V}^T \mathbf{x}_i$

  $= w_0 + \sum_{j=1}^{p} w_j x_{ij}$

  $+ \sum_{j=1}^{p} \sum_{k=j+1}^{p} \langle v_j, v_k \rangle x_{ij} x_{ik}$

Model parameters ($\mathcal{O}(pd)$)

- $w_0 \in \mathbb{R}, \mathbf{w} \in \mathbb{R}^p, \mathbf{V} \in \mathbb{R}^{p \times d}$

Polynomial regression

- $\hat{y}_i = h(\mathbf{x}_i)$

  $= w_0 + \mathbf{w}^T \mathbf{x}_i$

  $+\mathbf{x}_i^T \mathbf{W} \mathbf{x}_i$

  $= w_0 + \sum_{j=1}^{p} w_j x_{ij}$

  $+ \sum_{j=1}^{p} \sum_{k=j+1}^{p} w_{jk} x_{ij} x_{ik}$

Model parameters ($\mathcal{O}(p^2)$)

- $w_0 \in \mathbb{R}, \mathbf{w} \in \mathbb{R}^p, \mathbf{W} \in \mathbb{R}^{p \times p}$

Flexibility

FMs subsume previous factorization models

- e.g. matrix factorization, tensor factorization

FMs are much more flexible

- Handles also non-categorical variables (e.g. context)
- Handles higher-order dependencies (e.g. degree 3+)

No further requirement beyond raw representation

Learning

L2-regularized regression and classification

- Stochastic gradient descent [Rendle, ICDM 2010]
- Alternating least squares [Rendle, SIGIR 2011]
- Markov chain Monte Carlo [Rendle, TIST 2012]

L2-regularized ranking (pairwise loss)

- Stochastic gradient descent [Rendle, ICDM 2010]

Efficient prediction

FM trains $\mathcal{O}(pd)$ parameters

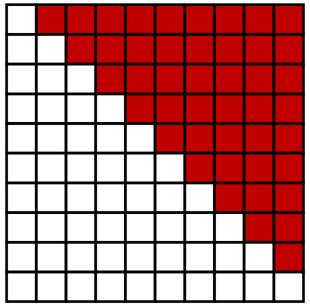
- Trivial prediction is still $\mathcal{O}(p^2 d)$

Simple optimization makes it $\mathcal{O}(pd)$

- $\hat{y} = h(x_i) = w_0 + \sum_{j=1}^{p} w_j x_{ij}$

$$+ \frac{1}{2} \sum_{f=1}^{d} \left(\left(\sum_{j=1}^{p} v_{jf} x_{ij} \right)^2 - \sum_{j=1}^{p} v_{jf}^2 x_{ij}^2 \right)$$

Efficient prediction



$$\sum_{j=1}^{p}\sum_{k=j+1}^{p}\langle \mathbf{v}_j, \mathbf{v}_k\rangle\, x_{ij}x_{ik}$$

$$= \underline{\frac{1}{2}\sum_{j=1}^{p}\sum_{k=1}^{p}\langle \mathbf{v}_j, \mathbf{v}_k\rangle\, x_{ij}x_{ik}} - \underline{\frac{1}{2}\sum_{j=1}^{p}\langle \mathbf{v}_j, \mathbf{v}_j\rangle\, x_{ij}x_{ij}}$$

$$= \underline{\frac{1}{2}}\left(\sum_{j=1}^{p}\sum_{k=1}^{p}\underline{\sum_{f=1}^{d} v_{jf}v_{kf}}\; x_{ij}x_{ik} - \sum_{j=1}^{p}\underline{\sum_{f=1}^{d} v_{jf}v_{jf}}\; x_{ij}x_{ij}\right)$$

$$= \frac{1}{2}\underline{\sum_{f=1}^{d}}\left(\left(\underline{\sum_{j=1}^{p} v_{jf}x_{ij}}\right)\left(\underline{\sum_{k=1}^{p} v_{kf}x_{ik}}\right) - \sum_{j=1}^{p} v_{jf}^{2}x_{ij}^{2}\right)$$

$$= \frac{1}{2}\sum_{f=1}^{d}\left(\left(\underline{\sum_{j=1}^{p} v_{jf}x_{ij}}\right)^{2} - \sum_{j=1}^{p} v_{jf}^{2}x_{ij}^{2}\right)$$

Summary

FMs are flexible

- Easy to leverage raw categorical features
- Easy to incorporate additional features

FMs are effective

- Automatic feature interactions via factorization

FMs are efficient

Extensions

Beyond a single latent representation per feature

- FFM [Juan, RecSys 2016]

Beyond linear, 2nd order interactions

- DeepFM [Guo, IJCAI 2017]
- xDeepFM [Lian, KDD 2018]

References

Factorization machines
Rendle, ICDM 2010

Factorization models for recommender systems and other applications
Schmidt-Thieme and Rendle, KDD 2012 tutorial

Recommender Systems: The Textbook (Sec. 8.5.2)