FACTORIZATION MACHINES

Dinh Huu Nguyen, 09/22/2020

Abstract: what is learned from [1]

Contents

1. Symbols and terms

l	number of factors, with index i
m	number of features, with index j
n	number of samples, with index k
X	dataset of samples x
Y	dataset of labels y

2. Modeling

Consider dataset X of n samples $x_1, \ldots, x_n$ and m features $X_1, \ldots, X_m$

X	X_1	$\ldots$	X_m	Y
x_1	x_{11}	$\ldots$	x_{1m}	y_1
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_n	x_{n1}	$\ldots$	x_{nm}	y_n

2.1. **Degree-2 interaction.** Model equation to model degree-2 interaction between features is

$$\begin{aligned}\hat{y}(x) &= w^0 \\ &\quad + x_1 w_1^1 + \cdots + x_m w_m^1 \\ &\quad + x_1 x_2 \langle w_1^2, w_2^2 \rangle + \cdots + x_{m-1} x_m \langle w_{m-1}^2, w_m^2 \rangle \\ &= w^0 + \langle x, w^1 \rangle + \sum_{j_1 < j_2} x_{j_1} x_{j_2} \langle w_{j_1}^2, w_{j_2}^2 \rangle\end{aligned} \tag{1}$$

where

- $w^0 \in \mathbb{R}$ is global bias
- $w^1 = (w_1^1, \ldots, w_m^1) \in \mathbb{R}^m$ is bias vector and each w_j^1 is bias for feature X_j
- $w_j^2 \in \mathbb{R}^l$ are factor vectors and each w_j^2 is factor vector for feature X_j

Or in matrix form

$$\hat{y}(x) = w^0 + \langle x, w^1 \rangle + x^t W^2 x \tag{2}$$

where

- $W^2 = \begin{pmatrix} 0 & \langle w_1^2, w_2^2 \rangle & \langle w_1^2, w_3^2 \rangle & \ldots & \langle w_1^2, w_m^2 \rangle \\ 0 & 0 & \langle w_2^2, w_3^2 \rangle & \ldots & \langle w_2^2, w_m^2 \rangle \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & \ldots & \ldots & \langle w_{m-1}^2, w_m^2 \rangle \\ 0 & \ldots & \ldots & \ldots & 0 \end{pmatrix}$ is upper triangular interaction matrix and each $\langle w_{j_1}^2, w_{j_2}^2 \rangle$ is interaction between feature X_{j_1} and feature X_{j_2}

Note that this is different from using model equation

$$\hat{y}(x) = w^0 + \langle x, w^1 \rangle + x^t W^2 x \tag{3}$$

where

- $$W^2 = \begin{pmatrix} 0 & w_{12}^2 & w_{13}^2 & \dots & w_{1m}^2 \\ 0 & 0 & w_{23}^2 & \dots & w_{2m}^2 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & \dots & w_{m-1,m}^2 \\ 0 & \dots & \dots & \dots & 0 \end{pmatrix}$$

See subsection 3.3 for more details.

Computation of this model equation looks to have quadratic time $\mathcal{O}(lm^2)$, but [1, lemma 3.1] shows it has linear time $\mathcal{O}(lm)$, and indeed has something closer to $\mathcal{O}(l\mu_{nz})$ where μ_{nz} is the average of the number $nz(x)$ of nonzero feature values in x, for all $x \in X$.

The gradient ∇y consists of the following partial derivatives

$$\begin{aligned} \frac{\partial y}{\partial w^0} &= 1 \\ \frac{\partial y}{\partial w_j^1} &= x_j, 1 \leq j \leq m \\ \frac{\partial y}{\partial w_{i_\circ j_\circ}^2} &= \left(\sum_{j_1 < j_2} x_{j_1} x_{j_2} \langle w_{j_1}^2, w_{j_2}^2 \rangle \right)' \\ &= \left(\sum_{j_1 < j_2} x_{j_1} x_{j_2} \sum_{i=1}^{l} w_{ij_1}^2 w_{ij_2}^2 \right)' \\ &= \left(\sum_{j_1 < j_2} x_{j_1} x_{j_2} w_{i_\circ j_1}^2 w_{i_\circ j_2}^2 \right)' \\ &= x_{j_\circ} \sum_{j \neq j_\circ} x_j w_{i_\circ j}^2 \\ &= x_{j_\circ} \sum_{j=1}^{m} x_j w_{i_\circ j}^2 - x_{j_\circ}^2 w_{i_\circ j_\circ}^2 \end{aligned}$$

Computation of $\frac{\partial y}{\partial w_{i_\circ j_\circ}^2}$ looks to have linear time $\mathcal{O}(m)$. But the term $\sum_{j=1}^{m} x_j w_{i_\circ j}^2$ can be computed once for all $j_\circ$ so computation of of $\frac{\partial y}{\partial w_{i_\circ j_\circ}^2}$ has constant time $\mathcal{O}(1)$.

2.2. **Degree-d interaction.** Model equation to model degree-d interaction between features is

$$\begin{aligned}\hat{y}(x) = w^0 & \\ & + x_1 w_1^1 + \cdots + x_m w_m^1 \\ & + x_1 x_2 \langle w_1^2, w_2^2 \rangle + \cdots + x_{m-1} x_m \langle w_{m-1}^2, w_m^2 \rangle \\ & + \dots \\ & + x_1 \dots x_d \langle w_1^d, \dots, w_d^d \rangle + \cdots + x_{m-d+1} \dots x_m \langle w_{m-d+1}^d, \dots, w_m^d \rangle \\ = w^0 + \langle x, w^1 \rangle & + \sum_{a=2}^{d} \sum_{j_1 < \dots < j_a} x_{j_1} \dots x_{j_a} \langle w_{j_1}^a, \dots, w_{j_a}^a \rangle\end{aligned} \tag{4}$$

where $w_{j_1}^a, \dots, w_{j_a}^a \in \mathbb{R}^{l^a}$ and their "dot product" is defined as

$$\langle w_{j_1}^a, \dots, w_{j_a}^a \rangle = \sum_{i=1}^{l^a} w_{ij_1}^a \dots w_{ij_a}^a$$

Note: this "dot product" leaves something to be desired. Maybe a continuation of

- scalars for degree 0
- vectors for degree 1
- matrices for degree 2
- degree-d tensors for degree d in general? where dot product of tensors is just sum of entry-wise products. Especially if this generalizes (2)

Note that this model equation is just a sum of the first $d+1$ elementary symmetric polynomials in m variables $x_1, \dots, x_m$ with general coefficients

- $e_0(x_1, \dots, x_m) = 1$
- $e_1(x_1, \dots, x_m) = \sum_{j=1}^{m} x_i$
- $e_2(x_1, \dots, x_m) = \sum_{j_1 < j_2} x_{j_1} x_{j_2}$

- .
- .
- .
- $e_d(x_1, \ldots, x_m) = \sum_{j_1 < \cdots < j_d} x_{j_1} \ldots x_{j_e}$

Again computation of this model equation looks to have polynomial time $\mathcal{O}(l_d m^d)$, but a similar argument to [1, lemma 3.1] shows it has linear time.

The gradient ∇y consists of the following partial derivatives

$$\begin{aligned}
\frac{\partial y}{\partial w^0} &= 1 \\
\frac{\partial y}{\partial w_j^1} &= x_j, 1 \leq j \leq m \\
\frac{\partial y}{\partial w_{i_\circ j_\circ}^a} &= \left(\sum_{j_1 < \cdots < j_a} x_{j_1} \ldots x_{j_a} \langle w_{j_1}^a, \ldots, w_{j_a}^a \rangle \right)' \\
&= \left(\sum_{j_1 < \cdots < j_a} x_{j_1} \ldots x_{j_a} \sum_{i=1}^{l^a} w_{ij_1}^a \ldots w_{ij_a}^a \right)' \\
&= \left(\sum_{j_1 < \cdots < j_a} x_{j_1} \ldots x_{j_a} w_{i_\circ j_1}^a \ldots w_{i_\circ j_a}^a \right)' \\
&= x_{j_\circ} \sum_{j_1 < \cdots < j_{a-1}, j_\alpha \neq j_\circ} x_{j_1} \ldots x_{j_{a-1}} w_{i_\circ j_1}^a \ldots w_{i_\circ j_{a-1}}^a \\
&= x_{j_\circ} \sum_{j_1 < \cdots < j_a} x_{j_1} \ldots x_{j_a} w_{i_\circ j_1}^a \ldots w_{i_\circ j_a}^a - x_{j_\circ} \sum_{j_1 < \ldots j_\circ \cdots < j_a} x_{j_1} \ldots x_{j_\circ} \ldots x_{j_a} w_{i_\circ j_1}^a \ldots w_{i_\circ j_\circ}^a \ldots
\end{aligned}$$

Again computation of $\frac{\partial y}{\partial w_{i_\circ j_\circ}^a}$ looks to have time $\mathcal{O}(2(a-1)C(m, a-1))$. But the term $\sum_{j_1 < \cdots < j_a} x_{j_1} \ldots x_{j_a} w_{i_\circ j_1}^a \ldots w_{i_\circ j_a}^a$ can be computed once for all $j_\circ$ so computation of of $\frac{\partial y}{\partial w_{i_\circ j_\circ}^a}$ has constant time $\mathcal{O}(1)$.

3. Examples

Below are some examples.

3.1. **GMLs.** When d is 1 and g is a link function then factorization machines reduce to generalized linear models, including linear regression and logistic regression.

3.2. **SVMs.**

3.2.1. *Linear SVMs.* When a support vector machine has linear kernel

$$k_1(x, x') = 1 + \langle x, x' \rangle$$

and feature map

$$\phi(x) = (1, x_1, \ldots, x_m)$$

then its model equation is

$$\hat{y}(x) = w^0 + \langle x, w^1 \rangle$$

which is the same as (4) for a degree-1 factorization machine.

3.2.2. *Quadratic SVMs.* When a support vector machine as quadratic kernel

$$k(x, x') = (1 + \langle x, x' \rangle)^2$$

and feature map

$$\phi(x) = (1, \sqrt{2}x_1, \ldots, \sqrt{2}x_m, x_1^2, \ldots, x_m^2, \sqrt{2}x_1x_2, \ldots, \sqrt{2}x_{m-1}x_m)$$

then its model equation is

$$\hat{y}(x) = w^0 + \sqrt{2}\langle x, w^1 \rangle + \sqrt{2} \sum_{j_1 < j_2} x_{j_1} x_{j_2} w^2_{j_1 j_2} + \sum_{j=1}^{m} x_j^2 w^2_{jj}$$

which is the same as (1) except for the following

- the term $\sum_{j=1}^{m} x_j^2 w^2_{jj}$ (factorization machines do not model interaction between a feature with itself)
- the terms $w^2_{j_1 j_2}$ are independent (while the terms $\langle w_{j_1}, w_{j_2} \rangle$ and $\langle w_{j_1}, w_{j_3} \rangle$ in (1) are dependent).

3.3. **Recommender Systems.** When ratings of m_i items by m_u users are represented as

X	U_1				U_{m_u}	I_1				I_{m_i}	$rating$
x_1	1	0	...	...	0_{m_u}	0	...	1	...	$0_{m_u+m_i}$	y_1
x_2	1	0	...	...	0_{m_u}	0	...	1	...	$0_{m_u+m_i}$	y_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_{n-1}	0	...	...	0	1	0	...	1	...	$0_{m_u+m_i}$	y_{n-1}
x_n	0	...	...	0	1	0	...	1	...	$0_{m_u+m_i}$	y_n

then factorization machines reduce to recommendation systems, which can also be used to recommend items. Features are now users and items, and interaction between features is now interaction between user and item.

Model equation to model degree-2 interaction between features is now model equation to model interaction between user and item

$$\begin{aligned}\hat{y}(x) &= w^0 + \langle x, w^1\rangle + \sum_{j_1<j_2} x_{j_1}x_{j_2}\langle w^2_{j_1}, w^2_{j_2}\rangle \\ &= w^0 + \langle x, w^1\rangle + x^t W^2 x \\ &= w^0 + w^1_{j_u} + w^1_{j_i} + \langle w^2_{j_u}, w^2_{j_i}\rangle\end{aligned}$$

where

$$x = \begin{pmatrix} \dots & 1_{j_u} & \dots & \dots & 1_{j_i} & \dots \end{pmatrix}$$

$$w^1 = \begin{pmatrix} w^1_1 \\ \vdots \\ w^1_m \end{pmatrix}$$

$$W^2 = \begin{pmatrix} 0 & \langle w^2_1, w^2_2\rangle & \langle w^2_1, w^2_3\rangle & \dots & \langle w^2_1, w^2_m\rangle \\ 0 & 0 & \langle w^2_2, w^2_3\rangle & \dots & \langle w^2_2, w^2_m\rangle \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & \dots & \langle w^2_{m-1}, w^2_m\rangle \\ 0 & \dots & \dots & \dots & 0 \end{pmatrix}$$

We can rewrite this model equation as

$$\begin{aligned}\hat{y}(x) &= w^0 + \langle x^u, w^{1u} \rangle + \langle x^i, w^{1i} \rangle + x^{ut} W^{2ut} W^{2i} x^i \\ &= w^0 + w^{1u}_{j_u} + w^{1i}_{j_i} + \langle w^{2u}_{j_u}, w^{2i}_{j_i} \rangle\end{aligned}$$

where

- w^0 is global bias
- $w^{1u} = \begin{pmatrix} w^{1u}_1 \\ \vdots \\ w^{1u}_{m_u} \end{pmatrix}$ is bias vector and each entry $w^{1u}_{j_u}$ is bias for user U_{j_u}
- $w^{1i} = \begin{pmatrix} w^{1i}_1 \\ \vdots \\ w^{1i}_{m_i} \end{pmatrix}$ is bias vectors and each entry $w^{1i}_{j_i}$ is bias for item I_{j_i}
- $x_u = \begin{pmatrix} 0 \\ \vdots \\ 1_{j_u} \\ \vdots \\ 0_{m_u} \end{pmatrix}$ is user U_{j_u}
- $x_i = \begin{pmatrix} 0 \\ \vdots \\ 1_{j_i} \\ \vdots \\ 0_{m_i} \end{pmatrix}$ is item I_{j_i}
- $W^{2u} = \begin{pmatrix} w^{2u}_{11} & \dots & w^{2u}_{1m_u} \\ \vdots & \ddots & \vdots \\ w^{2u}_{l1} & \dots & w^{2u}_{lm_u} \end{pmatrix}$ is factor matrix and each column $w^{2u}_{j_u}$ is factor vector for user U_{j_u}

- $W^{2i} = \begin{pmatrix} w_{11}^{2i} & \dots & w_{1m_i}^{2i} \\ \vdots & \ddots & \vdots \\ w_{l1}^{2i} & \dots & w_{lm_i}^{2i} \end{pmatrix}$ is factor matrix and each column $w_{j_i}^{2i}$ is factor vector for item I_{j_i}

- each $\langle w_{j_u}^{2u}, w_{j_i}^{2i} \rangle$ models interaction between user U_{j_u} and item I_{j_i}

Note that this is different from using model equation

$$\hat{y}(x) = w^0 + \langle x^u, w^{1u} \rangle + \langle x^i, w^{1i} \rangle + x^{ut} W^2 x^i$$

where

- $W^2 = \begin{pmatrix} w_{11}^{2} & \dots & w_{1m_i}^{2} \\ \vdots & \ddots & \vdots \\ w_{m_u1}^{2} & \dots & w_{m_u,m_i}^{2} \end{pmatrix}$ is interaction matrix and each w_{j_u,j_i}^2 is interaction between user U_{j_u} and item I_{j_i}

In the first case, if user U_{j_u} and item I_{j_i} have not had interaction

- factor vector $w_{j_u}^{2u}$ for user U_{j_u} may be learned through interaction with other items

- factor vector $w_{j_i}^{2i}$ for item I_{j_i} may be learned through interaction with other users

- hence $\langle w_{j_u}^{2u}, w_{j_i}^{2i} \rangle$ may be learned

In the second case, if user U_{j_u} and item I_{j_i} have not had interaction then w_{j_u,j_i}^2 may not be learned.

Prediction of rating of item I_{j_i} by user U_{j_u} is now

$$\hat{y}(x) = w^0 + w_{j_u}^{1u} + w_{j_i}^{1i} + \langle w_{j_u}^{2u}, w_{j_i}^{2i} \rangle$$

Recommendation of items for user U_{j_u} now can be based on the ranking of such predictions.

3.4. **Inter-group interaction.** As a generalization to example 3.3 above, one can model degree-d interaction between d groups $G_1, \ldots, G_d$ of features with corresponding d groups $I_1, \ldots, I_d$ of indices in increasing order, where each feature in group G_a is to interact with features in groups $G_{a'}, a' \neq a$ with model equation

$$\hat{y}(x) = w^0 + \langle x, w^1 \rangle + \sum_{a=2}^{d} \sum_{j_1 < \ldots < j_a}^{j_1 \in I_1, \ldots, j_a \in I_a} x_{j_1} \ldots x_{j_a} \langle w^a_{j_1}, \ldots, w^a_{j_a} \rangle$$

3.5. **Intra-group interaction.** Or one can model degree-d interaction between features in some particular group G of features with corresponding group I of indices by restricting the indices in (4) to I.

4. Implementation

See Python code and application at github.com/dinhuun.

References

[1] Steffen Rendle, *Factorization Machines* (2010).