Data Intelligence

Field-aware Factorization Machine

U Kang
Seoul National University

In This Lecture

- Field-aware Factorization Machine
  - Click-Through Rate (CTR) prediction problem
  - Comparison with other methods (LM, Poly2, FM)
  - Optimization
  - Experiments

Outline

- **CTR Prediction**
- Models
  - Linear Model
  - Degree-2 Polynomial Mapping
  - Factorization Machine
  - Field-aware Factorization Machine
- Optimization
- Experiments
- Conclusion

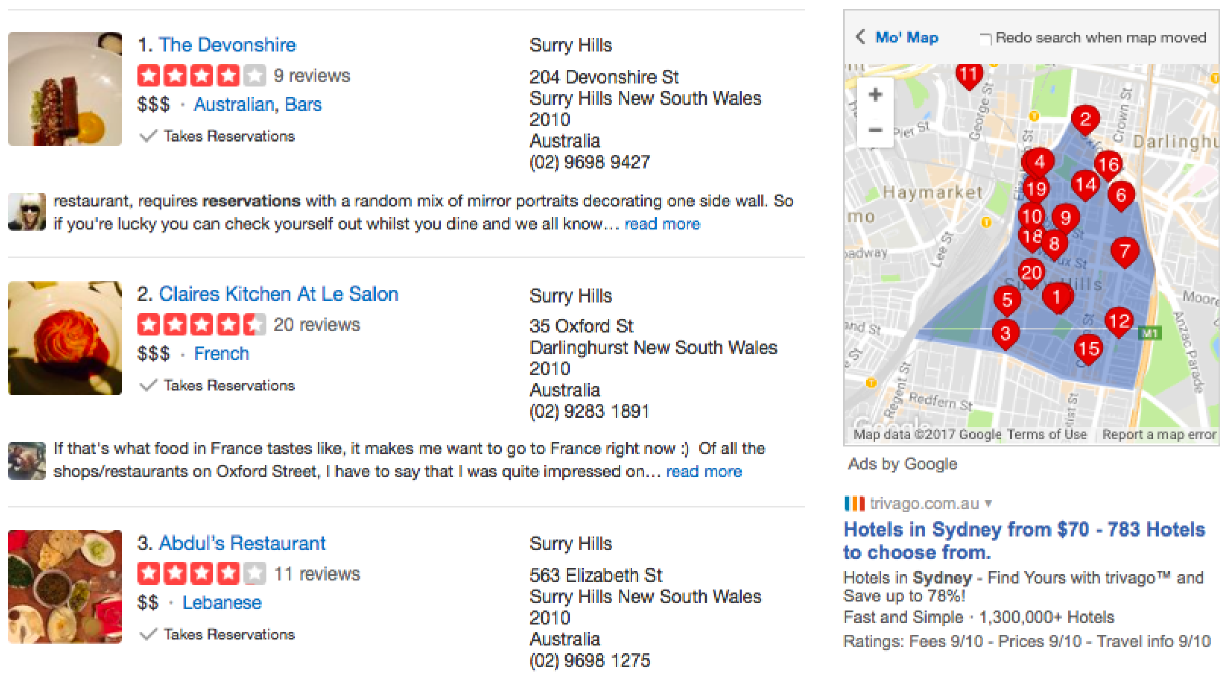
Click-Through Rate Prediction (1)

- Learn: $P(y|x)$
  - x – features
  - y – click or not

9.2%

3.1%

0.7%



Click-Through Rate Prediction (2)

- An example

Clicked (Y)	Publisher (X1)	Advertiser (X2)	Gender (X3)
Yes	ESPN	Nike	Male
No	NBC	Adidas	Female
…	…	…	…

Learn

$$P(y|x)$$

Predict

Clicked (Y)	Publisher (X1)	Advertiser (X2)	Gender (X3)
?	ESPN	Adidas	Male
?	NBC	Nike	Female
…	…	…	…

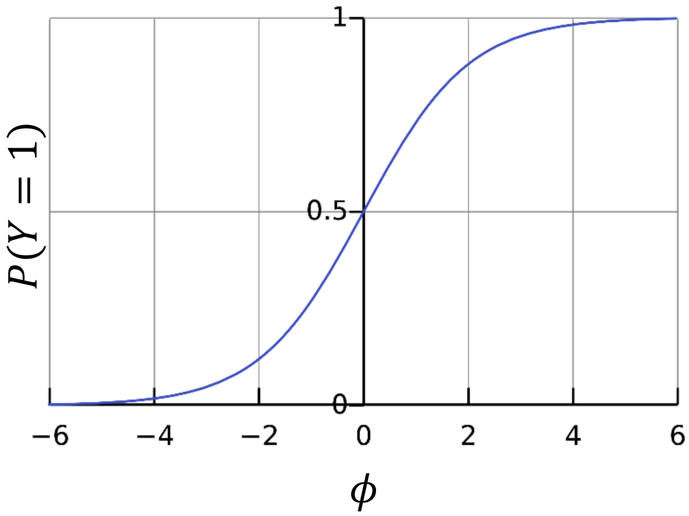
Outline

- ☑ CTR Prediction
- ☐ **Models**
  - ☐ Linear Model
  - ☐ Degree-2 Polynomial Mapping
  - ☐ Factorization Machine
  - ☐ Field-aware Factorization Machine
- ☐ Optimization
- ☐ Experiments
- ☐ Conclusion

Logistic Regression

- How can we define the CTR prediction?
  - The logistic regression

$$P(y=1|x) = \frac{1}{1+e^{-\phi(w,x)}}$$



where

x is the feature
y is the label
w is a parameter of the model
ϕ is a logit

Optimization

- Given a dataset with m instances $\{(y_i, x_i)\}_{i=1}^{m}$
  - where x_i is the feature
  - $y_i \in \{1, -1\}$ is the label
  - m is the number of instances
- The model is trained by the optimization problem:

$$\min_{w} \frac{\lambda}{2} \|w\|_2^2 + \sum_{i=1}^{m} \log(1 + \exp(-y_i \phi(w, x_i)))$$

Models (1)

- How can we define the model?

$$P(y=1|x) = \frac{1}{1+e^{-\phi(w,x)}}$$

- Linear Model (slide 11~14 pages)
  - $\phi(w,x) = w^T x = \sum_{j \in C_1} w_j x_j$
    - where C_1 is the non-zero elements in x
- Poly2 (slide 15~17 pages)
  - $\phi(w,x) = \sum_{j_1, j_2 \in C_2} w_{j_1, j_2} x_{j_1} x_{j_2}$
    - where C_2 is the 2-combination of non-zero elements in x

Models (2)

- How can we define the model?

$$P(y = 1|x) = \frac{1}{1 + e^{-\phi(w,x)}}$$

- Factorization Machine (slide 18~21 pages)
  - $\phi(w,x) = \sum_{j_1,j_2 \in C_2} \langle w_{j_1}, w_{j_2} \rangle x_{j_1} x_{j_2}$
    - where w_{j_1} and w_{j_2} are two vectors with length k
- Field-aware Factorization Machine (slide 22~38 pages)
  - $\phi(w,x) = \sum_{j_1,j_2 \in C_2} \langle w_{j_1,f_2}, w_{j_2,f_1} \rangle x_{j_1} x_{j_2}$
    - where f_1 and f_2 are respectively fields of j_1 and j_2
    - w_{j_1,f_2} and w_{j_2,f_1} are two vectors with length k

Outline

- ☑ CTR Prediction
- ☑ Models
  - ☐ **Linear Model**
  - ☐ Degree-2 Polynomial Mapping
  - ☐ Factorization Machine
  - ☐ Field-aware Factorization Machine
- ☐ Optimization
- ☐ Experiments
- ☐ Conclusion

Linear Model (1)

- Linear Model (LM) captures the patterns in a linear way
- For the LM:
  - $\phi(w, x) = w^T x = \sum_{j \in C_1} w_j x_j$
    - where C_1 is the non-zero elements in x

Toy Example

- Consider a data instance:

Publisher	Advertiser
ESPN	NIKE

  - where we have two kinds of features ***Publisher*** and ***Advertiser***
  - An advertisement from ***NIKE*** displayed on ***ESPN***

Linear Model (2)

Publisher	Advertiser
ESPN	NIKE

- $\phi(w, x) = w_{ESPN} + w_{NIKE}$
  - where $w_{ESPN}, w_{NIKE} \in \mathbb{R}$
- Two weights learn the average behavior of ESPN and NIKE, respectively
- LM cannot learn the effect of a **feature conjunction**
  - If the CTR of the ads from NIKE on ESPN is particularly high or low, it cannot learn this effect well

Outline

- ☑ CTR Prediction
- ☑ Models
  - ☑ Linear Model
  - ☐ **Polynomial-2**
  - ☐ Factorization Machine
  - ☐ Field-aware Factorization Machine
- ☐ Optimization
- ☐ Experiments
- ☐ Conclusion

Degree-2 Polynomial Mapping (1)

- Degree-2 Polynomial Mapping (Poly2) learns the effect of a feature conjunction
- For the Poly2:
  - $\phi(w, x) = \sum_{j_1, j_2 \in C_2} w_{j_1, j_2} x_{j_1} x_{j_2}$
    - where C_2 is the 2-combination of non-zero elements in x

Degree-2 Polynomial Mapping (2)

Publisher	Advertiser
ESPN	NIKE

- $\phi(w, x) = w_{ESPN,NIKE}$
  - where $w_{ESPN,NIKE} \in \mathbb{R}$
  - The weight learns the conjunction pattern between ESPN and NIKE
  - Poly2 cannot learn any patterns in **unseen pairs**
    - It cannot infer any information about a pair (ESPN, Gucci) even if it has learned pairs (ESPN, NIKE) and (Vogue, Gucci)

Outline

- ☑ CTR Prediction
- ☑ Models
  - ☑ Linear Model
  - ☑ Polynomial-2
  - ☐ **Factorization Machine**
  - ☐ Field-aware Factorization Machine
- ☐ Optimization
- ☐ Experiments
- ☐ Conclusion

Factorization Machine (1)

- Factorization Machine (FM) learns the feature conjunction in a latent space
- For the FM:
  - $\phi(w, x) = \sum_{j_1, j_2 \in C_2} \langle w_{j_1}, w_{j_2} \rangle x_{j_1} x_{j_2}$
    - where w_{j_1} and w_{j_2} are two vectors with length user-defined k

Factorization Machine (2)

Publisher	Advertiser
ESPN	NIKE

- $\phi(w, x) = w_{ESPN}^T w_{NIKE}$
  - where $w_{ESPN}, w_{NIKE} \in \mathbb{R}^k$
  - k is a pre-defined dimensionality
- Factorization Machine (FM) has a benefit in the case of predicting on unseen data (see the next slide for the concrete example)
- It shares a **single latent vector** for a feature in all fields

Poly2 vs. Factorization Machine

	Publisher	Advertiser	Poly2	FM
Train	ESPN	NIKE	$w_{\text{ESPN,NIKE}}$	$\mathbf{w}_{\text{ESPN}} \cdot \mathbf{w}_{\text{NIKE}}$
	Vogue	Gucci	$w_{\text{Vogue,Gucci}}$	$\mathbf{w}_{\text{Vogue}} \cdot \mathbf{w}_{\text{Gucci}}$
Test	ESPN	Gucci	$w_{\text{ESPN,Gucci}}$	$\mathbf{w}_{\text{ESPN}} \cdot \mathbf{w}_{\text{Gucci}}$

- There is no (ESPN, Gucci) pair in the training data
- For Poly2, there is no way to learn the weight $w_{ESPN,Gucci}$
- However, FM is able to do reasonable prediction on $w_{EPSN}^{T} w_{Gucci}$ since it learns w_{ESPN} from the (ESPN, NIKE) pair and w_{Gucci} from the (Vogue, Gucci) pair

Outline

- ☑ CTR Prediction
- ☑ Models
  - ☑ Linear Model
  - ☑ Polynomial-2
  - ☑ Factorization Machine
  - ☐ **Field-aware Factorization Machine**
- ☐ Optimization
- ☐ Experiments
- ☐ Conclusion

Field-aware Factorization Machine

- Field-aware Factorization Machine (FFM) learns the feature conjunction in a field-wise latent space
  - FFM splits the latent space into many smaller latent spaces
  - Latent effect of a feature (e.g., male) could be different for different fields (e.g., publisher and advertiser)
- For the FFM:
  - $\phi(w, x) = \sum_{j_1, j_2 \in C_2} \langle w_{j_1, f_2}, w_{j_2, f_1} \rangle x_{j_1} x_{j_2}$
    - where f_1 and f_2 are respectively fields of j_1 and j_2
    - w_{j_1, f_2} and w_{j_2, f_1} are two vectors with length user-defined k

Toy Example

- Consider a data instance:

Publisher (P)	Advertiser (A)	Gender (G)
ESPN	NIKE	Male

  - where we have three kinds of features ***Publisher, Advertiser*** and ***Gender***
  - An advertisement from ***NIKE*** displayed on ***ESPN*** was clicked by a **male**

FM vs. FFM

Publisher (P)	Advertiser (A)	Gender (G)
ESPN	NIKE	Male

- For FM:
  - $\phi(w,x) = w_{ESPN}^T w_{NIKE} + w_{ESPN}^T w_{Male} + w_{NIKE}^T w_{Male}$
- For FFM:
  - $\phi(w,x) = w_{ESPN,A}^T w_{NIKE,P} + w_{ESPN,G}^T w_{Male,P} + w_{NIKE,G}^T w_{Male,A}$
  - A feature (e.g., male) is associated with two different latent vectors for different fields (e.g., publisher and advertiser)

Poly2 vs. FM vs. FFM (1)

- Illustration of the toy example (Poly2):
  - A dedicated weight is learned for each feature pair

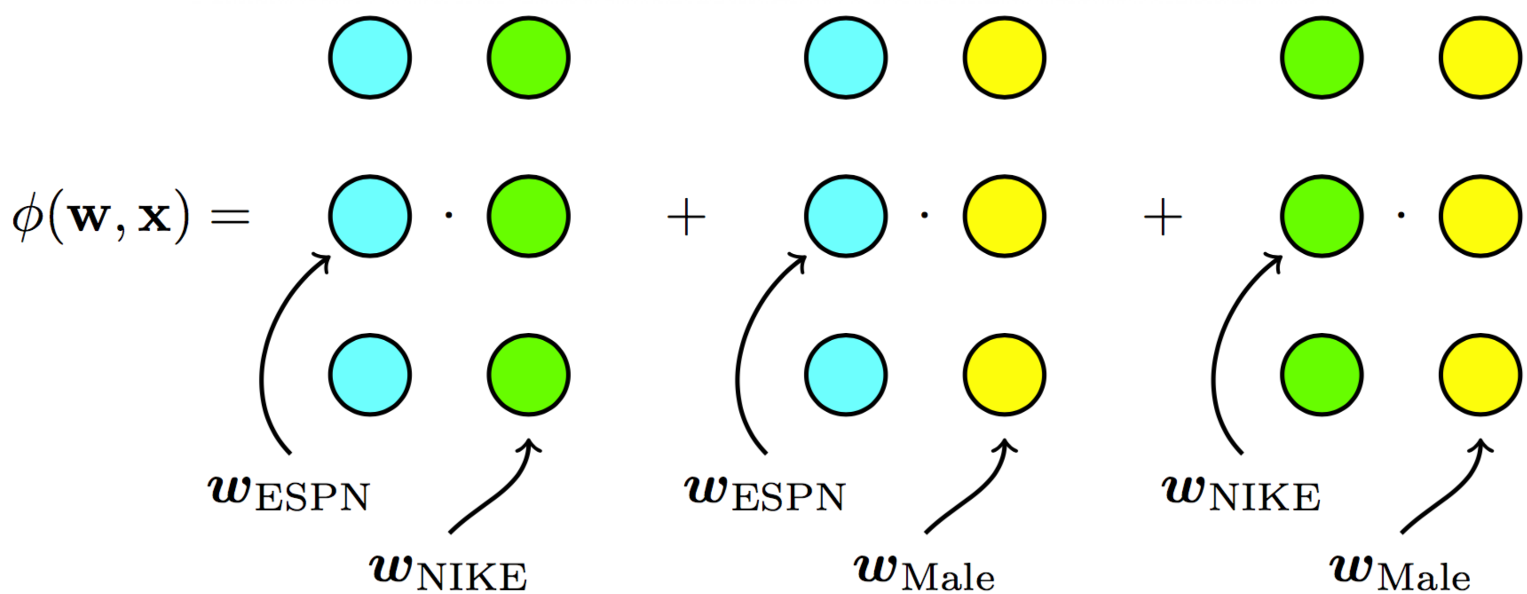
Publisher (P)	Advertiser (A)	Gender (G)
ESPN	NIKE	Male

$$\phi(\mathbf{w}, \mathbf{x}) = w_{\text{ESPN},\text{NIKE}} + w_{\text{ESPN},\text{Male}} + w_{\text{NIKE},\text{Male}}$$

Poly2 vs. FM vs. FFM (2)

- Illustration of the toy example (FM):
  - Each feature has one latent vector, which is used to interact with other latent vectors

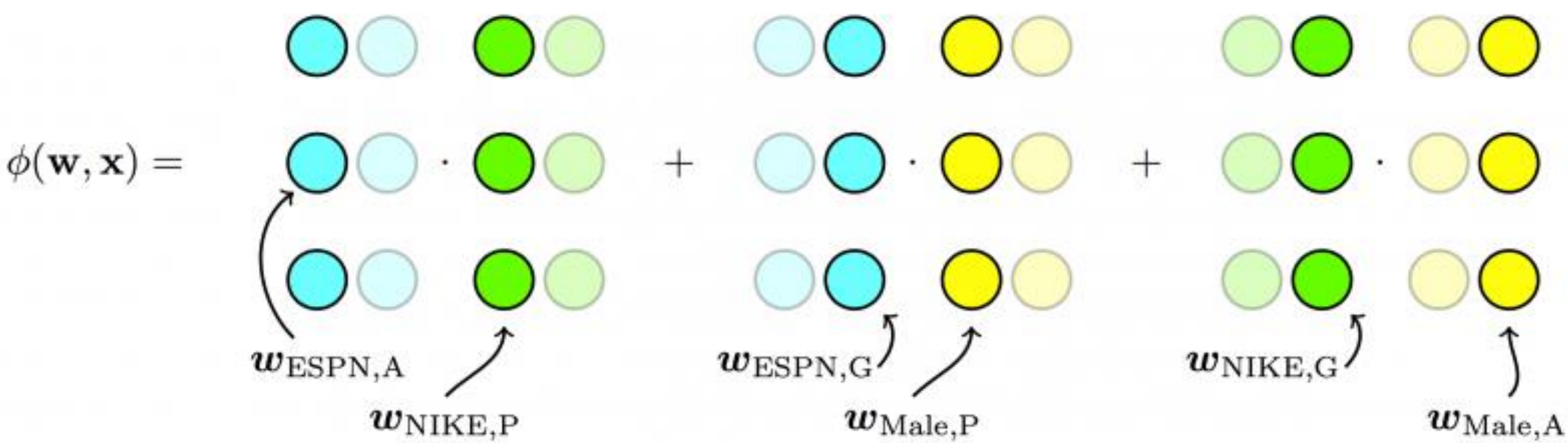
Publisher (P)	Advertiser (A)	Gender (G)
ESPN	NIKE	Male



Poly2 vs. FM vs. FFM (3)

- Illustration of the toy example (FFM):
  - Each feature has several latent vectors, one of them is used depending on the field of the other feature

Publisher (P)	Advertiser (A)	Gender (G)
ESPN	NIKE	Male



Further Example

- Consider the example:

User (Us)	Movie (Mo)	Genre (Ge)	Pr (Pr)
YuChin (YC)	3Idiots (3I)	Comedy,Drama (Co,Dr)	$9.99

  - Note that ***User***, ***Movie***, and ***Genre*** are categorical variables, and ***Price*** is a numerical variable

Further Example

- For LM, $\phi(w, x)$ is:

$$w_{\text{Us-Yu}} \cdot x_{\text{Us-Yu}} + w_{\text{Mo-3I}} \cdot x_{\text{Mo-3I}} + w_{\text{Ge-Co}} \cdot x_{\text{Ge-Co}} + w_{\text{Ge-Dr}} \cdot x_{\text{Ge-Dr}} + w_{\text{Pr}} \cdot x_{\text{Pr}}$$

  - where $x_{Us-Yu} = x_{Mo-3I} = x_{Ge-Co} = x_{Ge-Dr} = 1$ and $x_{Pr} = 9.99$
  - Note that because ***User***, ***Movie***, and ***Genre*** are categorical variables, the values are all ones

Further Example

- For Poly2, $\phi(w, x)$ is:

$$
\begin{aligned}
& w_{\text{Us-Yu-Mo-3I}} \cdot x_{\text{Us-Yu}} \cdot x_{\text{Mo-3I}} + w_{\text{Us-Yu-Ge-Co}} \cdot x_{\text{Us-Yu}} \cdot x_{\text{Ge-Co}} + w_{\text{Us-Yu-Ge-Dr}} \cdot x_{\text{Us-Yu}} \cdot x_{\text{Ge-Dr}} + w_{\text{Us-Yu-Pr}} \cdot x_{\text{Us-Yu}} \cdot x_{\text{Pr}} \\
& + w_{\text{Mo-3I-Ge-Co}} \cdot x_{\text{Mo-3I}} \cdot x_{\text{Ge-Co}} + w_{\text{Mo-3I-Ge-Dr}} \cdot x_{\text{Mo-3I}} \cdot x_{\text{Ge-Dr}} + w_{\text{Mo-3I-Pr}} \cdot x_{\text{Mo-3I}} \cdot x_{\text{Pr}} \\
& + w_{\text{Ge-Co-Ge-Dr}} \cdot x_{\text{Ge-Co}} \cdot x_{\text{Ge-Dr}} + w_{\text{Ge-Co-Pr}} \cdot x_{\text{Ge-Co}} \cdot x_{\text{Pr}} \\
& + w_{\text{Ge-Dr-Pr}} \cdot x_{\text{Ge-Dr}} \cdot x_{\text{Pr}}
\end{aligned}
$$

Further Example

- For FM, $\phi(w, x)$ is:

$$
\begin{aligned}
&\langle \mathbf{w}_{\text{Us-Yu}}, \mathbf{w}_{\text{Mo-3I}} \rangle \cdot x_{\text{Us-Yu}} \cdot x_{\text{Mo-3I}} + \langle \mathbf{w}_{\text{Us-Yu}}, \mathbf{w}_{\text{Ge-Co}} \rangle \cdot x_{\text{Us-Yu}} \cdot x_{\text{Ge-Co}} + \langle \mathbf{w}_{\text{Us-Yu}}, \mathbf{w}_{\text{Ge-Dr}} \rangle \cdot x_{\text{Us-Yu}} \cdot x_{\text{Ge-Dr}} + \langle \mathbf{w}_{\text{Us-Yu}}, \mathbf{w}_{\text{Pr}} \rangle \cdot x_{\text{Us-Yu}} \cdot x_{\text{Pr}} \\
&+ \langle \mathbf{w}_{\text{Mo-3I}}, \mathbf{w}_{\text{Ge-Co}} \rangle \cdot x_{\text{Mo-3I}} \cdot x_{\text{Ge-Co}} + \langle \mathbf{w}_{\text{Mo-3I}}, \mathbf{w}_{\text{Ge-Dr}} \rangle \cdot x_{\text{Mo-3I}} \cdot x_{\text{Ge-Dr}} + \langle \mathbf{w}_{\text{Mo-3I}}, \mathbf{w}_{\text{Pr}} \rangle \cdot x_{\text{Mo-3I}} \cdot x_{\text{Pr}} \\
&+ \langle \mathbf{w}_{\text{Ge-Co}}, \mathbf{w}_{\text{Ge-Dr}} \rangle \cdot x_{\text{Ge-Co}} \cdot x_{\text{Ge-Dr}} + \langle \mathbf{w}_{\text{Ge-Co}}, \mathbf{w}_{\text{Pr}} \rangle \cdot x_{\text{Ge-Co}} \cdot x_{\text{Pr}} \\
&+ \langle \mathbf{w}_{\text{Ge-Dr}}, \mathbf{w}_{\text{Pr}} \rangle \cdot x_{\text{Ge-Dr}} \cdot x_{\text{Pr}}
\end{aligned}
$$

Further Example

- For FFM, $\phi(w, x)$ is:

$$
\begin{aligned}
&\langle \mathbf{w}_{\text{Us-Yu},\text{Mo}}, \mathbf{w}_{\text{Mo-3I},\text{Us}} \rangle \cdot x_{\text{Us-Yu}} \cdot x_{\text{Mo-3I}} + \langle \mathbf{w}_{\text{Us-Yu},\text{Ge}}, \mathbf{w}_{\text{Ge-Co},\text{Us}} \rangle \cdot x_{\text{Us-Yu}} \cdot x_{\text{Ge-Co}} + \langle \mathbf{w}_{\text{Us-Yu},\text{Ge}}, \mathbf{w}_{\text{Ge-Dr},\text{Us}} \rangle \cdot x_{\text{Us-Yu}} \cdot x_{\text{Ge-Dr}} + \langle \mathbf{w}_{\text{Us-Yu},\text{Pr}}, \mathbf{w}_{\text{Pr},\text{Us}} \rangle \cdot x_{\text{Us-Yu}} \cdot x_{\text{Pr}} \\
&+ \langle \mathbf{w}_{\text{Mo-3I},\text{Ge}}, \mathbf{w}_{\text{Ge-Co},\text{Mo}} \rangle \cdot x_{\text{Mo-3I}} \cdot x_{\text{Ge-Co}} + \langle \mathbf{w}_{\text{Mo-3I},\text{Ge}}, \mathbf{w}_{\text{Ge-Dr},\text{Mo}} \rangle \cdot x_{\text{Mo-3I}} \cdot x_{\text{Ge-Dr}} + \langle \mathbf{w}_{\text{Mo-3I},\text{Pr}}, \mathbf{w}_{\text{Pr},\text{Mo}} \rangle \cdot x_{\text{Mo-3I}} \cdot x_{\text{Pr}} \\
&+ \langle \mathbf{w}_{\text{Ge-Co},\text{Ge}}, \mathbf{w}_{\text{Ge-Dr},\text{Ge}} \rangle \cdot x_{\text{Ge-Co}} \cdot x_{\text{Ge-Dr}} + \langle \mathbf{w}_{\text{Ge-Co},\text{Pr}}, \mathbf{w}_{\text{Pr},\text{Ge}} \rangle \cdot x_{\text{Ge-Co}} \cdot x_{\text{Pr}} \\
&+ \langle \mathbf{w}_{\text{Ge-Dr},\text{Pr}}, \mathbf{w}_{\text{Pr},\text{Ge}} \rangle \cdot x_{\text{Ge-Dr}} \cdot x_{\text{Pr}}
\end{aligned}
$$

Further Example

- In practice we need to map these features into numbers. Consider the following mapping:

Field name		Field index	Feature name		Feature index
User	$\rightarrow$	field 1	User-YuChin	$\rightarrow$	feature 1
Movie	$\rightarrow$	field 2	Movie-3Idiots	$\rightarrow$	feature 2
Genre	$\rightarrow$	field 3	Genre-Comedy	$\rightarrow$	feature 3
Price	$\rightarrow$	field 4	Genre-Drama	$\rightarrow$	feature 4
			Price	$\rightarrow$	feature 5

- After transforming to the LIBFFM format, the data becomes: 1:1:1 2:2:1 3:3:1 3:4:1 4:5:9.99
  - A red number is an index of field, a blue number is an index of feature, and a green number is the value of the feature

Further Example

- For LM, $\phi(w, x)$ is:

$$w_1 \cdot 1 + w_2 \cdot 1 + w_3 \cdot 1 + w_4 \cdot 1 + w_5 \cdot 9.99$$

Further Example

- For Poly2, $\phi(w, x)$ is:

$$
\begin{aligned}
w_{1,2} \cdot 1 \cdot 1 + w_{1,3} \cdot 1 \cdot 1 + w_{1,4} \cdot 1 \cdot 1 + w_{1,5} \cdot 1 \cdot 9.99 \\
+ w_{2,3} \cdot 1 \cdot 1 + w_{2,4} \cdot 1 \cdot 1 + w_{2,5} \cdot 1 \cdot 9.99 \\
+ w_{3,4} \cdot 1 \cdot 1 + w_{3,5} \cdot 1 \cdot 9.99 \\
+ w_{4,5} \cdot 1 \cdot 9.99
\end{aligned}
$$

Further Example

- For FM, $\phi(w, x)$ is:

$$
\begin{aligned}
&\langle \mathbf{w}_1, \mathbf{w}_2 \rangle \cdot 1 \cdot 1 + \langle \mathbf{w}_1, \mathbf{w}_3 \rangle \cdot 1 \cdot 1 + \langle \mathbf{w}_1, \mathbf{w}_4 \rangle \cdot 1 \cdot 1 + \langle \mathbf{w}_1, \mathbf{w}_5 \rangle \cdot 1 \cdot 9.99 \\
&+ \langle \mathbf{w}_2, \mathbf{w}_3 \rangle \cdot 1 \cdot 1 + \langle \mathbf{w}_2, \mathbf{w}_4 \rangle \cdot 1 \cdot 1 + \langle \mathbf{w}_2, \mathbf{w}_5 \rangle \cdot 1 \cdot 9.99 \\
&+ \langle \mathbf{w}_3, \mathbf{w}_4 \rangle \cdot 1 \cdot 1 + \langle \mathbf{w}_3, \mathbf{w}_5 \rangle \cdot 1 \cdot 9.99 \\
&+ \langle \mathbf{w}_4, \mathbf{w}_5 \rangle \cdot 1 \cdot 9.99
\end{aligned}
$$

Further Example

- For FFM, $\phi(w, x)$ is:

$$
\begin{aligned}
&\langle \mathbf{w}_{1,2}, \mathbf{w}_{2,1} \rangle \cdot 1 \cdot 1 + \langle \mathbf{w}_{1,3}, \mathbf{w}_{3,1} \rangle \cdot 1 \cdot 1 + \langle \mathbf{w}_{1,3}, \mathbf{w}_{4,1} \rangle \cdot 1 \cdot 1 + \langle \mathbf{w}_{1,4}, \mathbf{w}_{5,1} \rangle \cdot 1 \cdot 9.99 \\
&\quad + \langle \mathbf{w}_{2,3}, \mathbf{w}_{3,2} \rangle \cdot 1 \cdot 1 + \langle \mathbf{w}_{2,3}, \mathbf{w}_{4,2} \rangle \cdot 1 \cdot 1 + \langle \mathbf{w}_{2,4}, \mathbf{w}_{5,2} \rangle \cdot 1 \cdot 9.99 \\
&\quad + \langle \mathbf{w}_{3,3}, \mathbf{w}_{4,3} \rangle \cdot 1 \cdot 1 + \langle \mathbf{w}_{3,4}, \mathbf{w}_{5,3} \rangle \cdot 1 \cdot 9.99 \\
&\quad + \langle \mathbf{w}_{4,4}, \mathbf{w}_{5,3} \rangle \cdot 1 \cdot 9.99
\end{aligned}
$$

Outline

- ☑ CTR Prediction
- ☑ Models
  - ☑ Linear Model
  - ☑ Polynomial-2
  - ☑ Factorization Machine
  - ☑ Field-aware Factorization Machine
- ☐ **Optimization**
- ☐ Experiments
- ☐ Conclusion

Optimization (1)

- Given a dataset with m instances $\{(y_i, x_i)\}_{i=1}^{m}$
  - where x_i is the feature
  - $y_i \in \{1, -1\}$ is the label
  - m is the number of instances
- FFM is trained by the optimization problem:

$$\min_{w} f(w) = \min_{w} \frac{\lambda}{2} \|w\|_2^2 + \sum_{i=1}^{m} \log(1 + \exp(-y_i \phi_{FFM}(w, x_i)))$$

Optimization (2)

- How can we train a Field-aware Factorization Machine (FFM) to optimize the following problem?

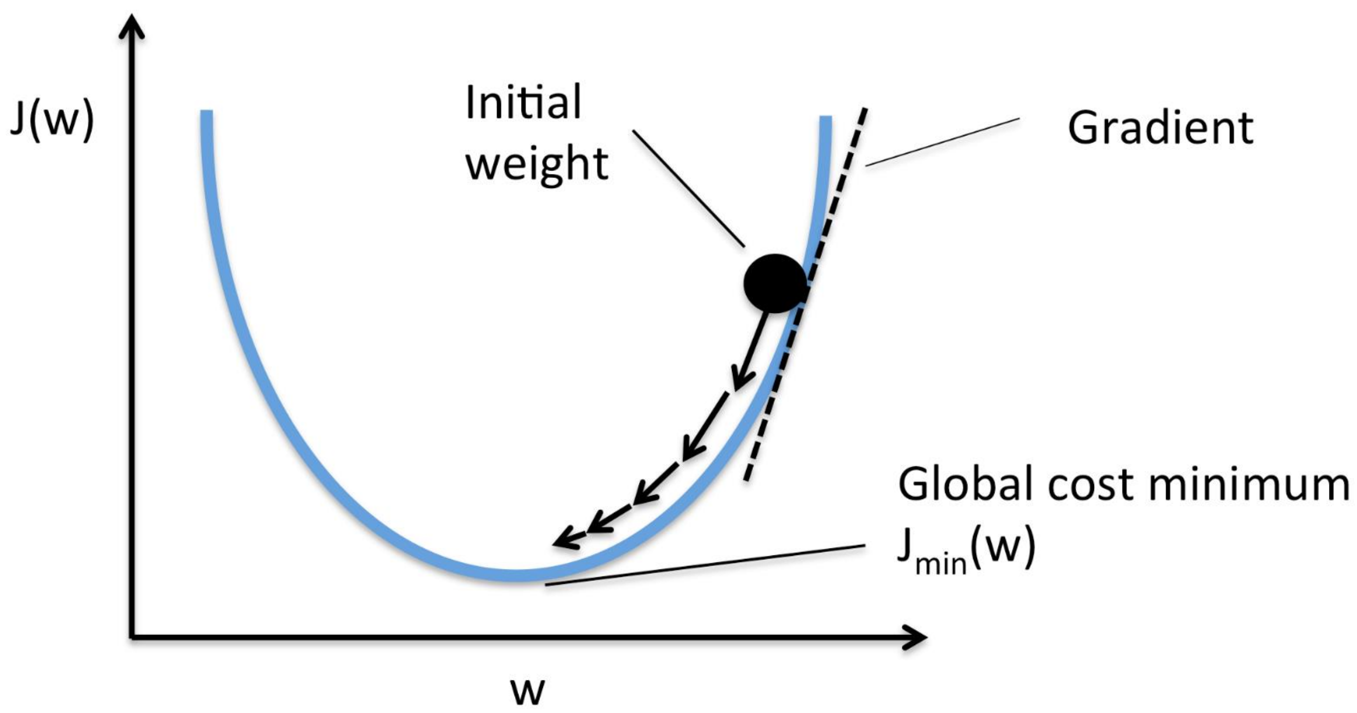
$$\min_w f(w) = \min_w \frac{\lambda}{2} \|w\|_2^2 + \sum_{i=1}^{m} \log(1 + \exp(-y_i \phi_{FFM}(w, x_i)))$$

where $\phi_{FFM}(w, x) = \sum_{j_1, j_2 \in C_2} \langle w_{j_1, f_2}, w_{j_2, f_1} \rangle x_{j_1} x_{j_2}$

  - ***Stochastic Gradient Descent***

Stochastic Gradient Descent

- Weights are updated incrementally based on their gradients about the loss



Solving the Optimization Problem (1)

$$\min_{w} f(w) = \min_{w} \frac{\lambda}{2}\|w\|_2^2 + \sum_{i=1}^{m} \log(1 + \exp(-y_i \phi_{FFM}(w, x_i)))$$

- Gradients:

$$\boldsymbol{g}_{j_1,f_2} \equiv \nabla_{\boldsymbol{w}_{j_1,f_2}} f(\boldsymbol{w}) = \lambda \cdot \boldsymbol{w}_{j_1,f_2} + \kappa \cdot \boldsymbol{w}_{j_2,f_1} x_{j_1} x_{j_2}$$

$$\boldsymbol{g}_{j_2,f_1} \equiv \nabla_{\boldsymbol{w}_{j_2,f_1}} f(\boldsymbol{w}) = \lambda \cdot \boldsymbol{w}_{j_2,f_1} + \kappa \cdot \boldsymbol{w}_{j_1,f_2} x_{j_1} x_{j_2}$$

  - where

$$\kappa = \frac{\partial \log(1 + \exp(-y\phi_{\mathrm{FFM}}(\boldsymbol{w}, \boldsymbol{x})))}{\partial \phi_{\mathrm{FFM}}(\boldsymbol{w}, \boldsymbol{x})} = \frac{-y}{1 + \exp(y\phi_{\mathrm{FFM}}(\boldsymbol{w}, \boldsymbol{x}))}$$

Solving the Optimization Problem (2)

- For each coordinate $d = 1, \dots, k$:

$$(G_{j_1,f_2})_d \leftarrow (G_{j_1,f_2})_d + (g_{j_1,f_2})_d^2$$

$$(G_{j_2,f_1})_d \leftarrow (G_{j_2,f_1})_d + (g_{j_2,f_1})_d^2$$

- Finally, the parameters are updated by:

$$(w_{j_1,f_2})_d \leftarrow (w_{j_1,f_2})_d - \frac{\eta}{\sqrt{(G_{j_1,f_2})_d}}(g_{j_1,f_2})_d$$

$$(w_{j_2,f_1})_d \leftarrow (w_{j_2,f_1})_d - \frac{\eta}{\sqrt{(G_{j_2,f_1})_d}}(g_{j_2,f_1})_d$$

  - where η is a user-specified learning rate

Outline

- CTR Prediction
- Models
  - Linear Model
  - Polynomial-2
  - Factorization Machine
  - Field-aware Factorization Machine
- Optimization
- **Experiments**
- Conclusion

Evaluation Criterion

- For the evaluation criterion:

$$\text{logloss} = \frac{1}{m}\sum_{i=1}^{m}\log(1+\exp(-y_i\phi(\boldsymbol{w},\boldsymbol{x}_i)))$$

  - where m is the number of test instances

Experimental Results

- Performance comparison between LM, Poly2, FM, FFM
  - Lower logloss indicates the better performance

	statistics			logloss			
Data set	# instances	# features	# fields	LM	Poly2	FM	FFM
KDD2010-bridge	20,012,499	651,166	9	0.27947	0.2622	0.26372	0.25639
KDD2012	149,639,105	54,686,452	11	0.15069	0.15099	0.15004	0.14906
phishing	11,055	100	30	0.14211	0.11512	0.09229	0.1065
adult	48,842	308	14	0.3097	0.30655	0.30763	0.30565
cod-rna (dummy fields)	331,152	8	8	0.13829	0.12874	0.12580	0.12914
cod-rna (discretization)	331,152	2,296	8	0.16455	0.17576	0.16570	0.14993
ijcnn (dummy fields)	141,691	22	22	0.20093	0.08981	0.07087	0.0692
ijcnn (discretization)	141,691	69,867	22	0.21588	0.24578	0.20223	0.18608

Table 4: Comparison between LM, Poly2, FM, and FFMs. The best logloss is underlined.

Outline

- ☑ CTR Prediction
- ☑ Models
  - ☑ Linear Model
  - ☑ Polynomial-2
  - ☑ Factorization Machine
  - ☑ Field-aware Factorization Machine
- ☑ Optimization
- ☑ Experiments
- ☐ **Conclusion**

What you should know

- CTR prediction problem
  - To learn $P(y|x)$ where x is a feature and y is a response
- Four methods for the problem
  - Linear Model
  - Degree-2 Polynomial Mapping
  - Factorization Machine
  - Field-aware Factorization Machine

Questions?