

Click-Through-Rate Prediction

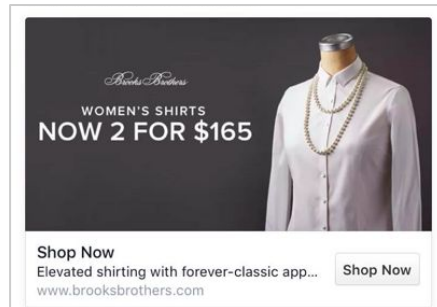
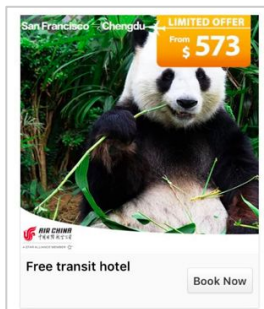
Scalable Factorization Machines

Binjie Lai, **M**iao Lu, **W**anyan Xie
MSAN, **USF**



Click-Through-Rate Prediction

$$\text{CTR} = \frac{\text{Clicks} \quad (\# \text{ times of ads clicked})}{\text{Impressions} \quad (\# \text{ times of ads shown})}$$



Kaggle Dataset :Avazu CTR Prediction

Train: 5.88 GB Test: 629 MB

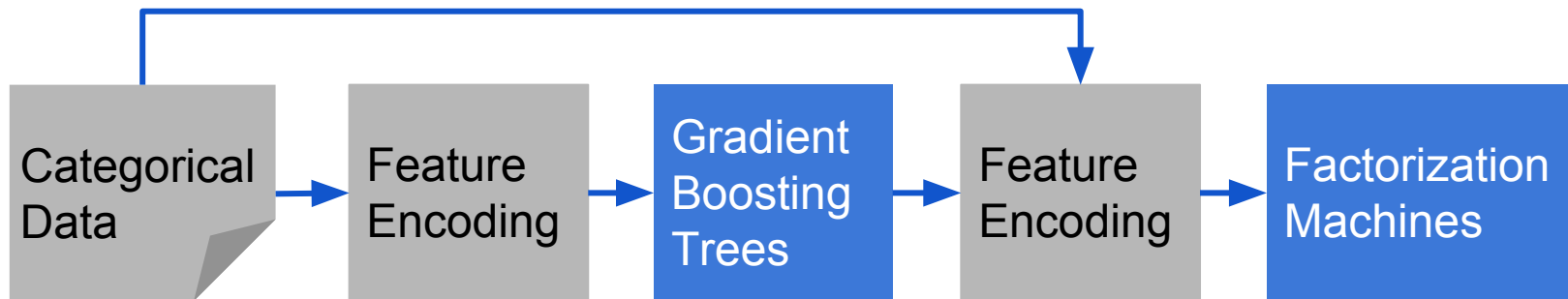
Click	Banner _position	App _id	App _category	Device _id	Device _model	...	Impressions
0	0	ecad2386	07d7df22	a99f214a	44956a24	...	~ 36 million
1	0	ecad2386	07d7df22	a99f214a	44956a24	...	
0	2	7e091613	f028772b	6afc734f	e352da7	...	
?	1	85f751fd	c4e18dd6	6afc734f	e352da7	...	~ 4 million

label

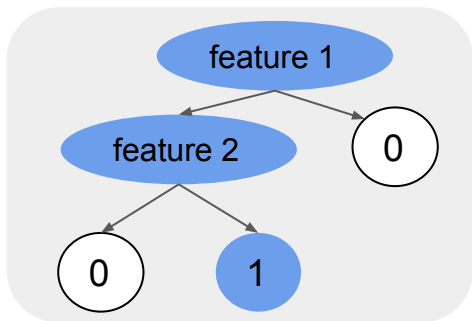
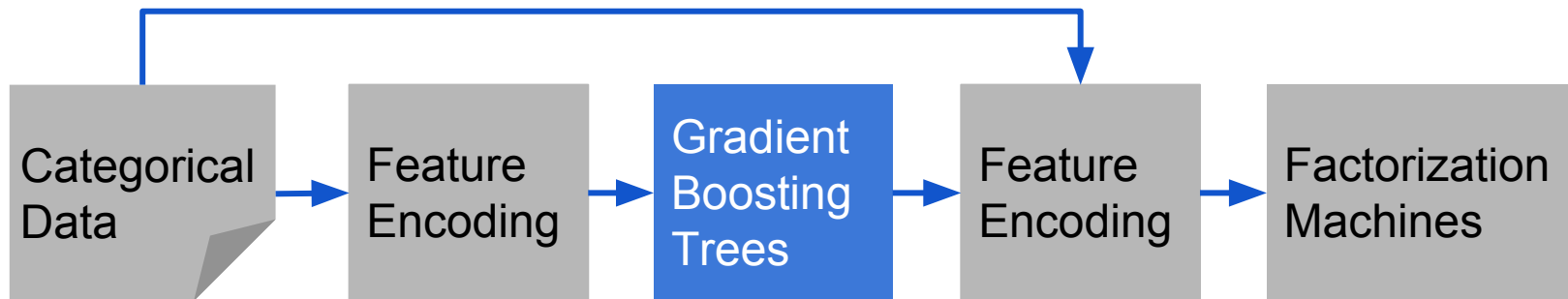
23 categorical features

Data Source: <https://www.kaggle.com/c/avazu-ctr-prediction>

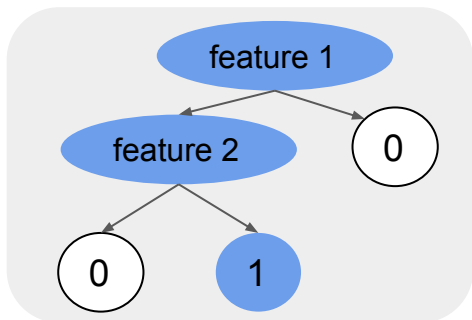
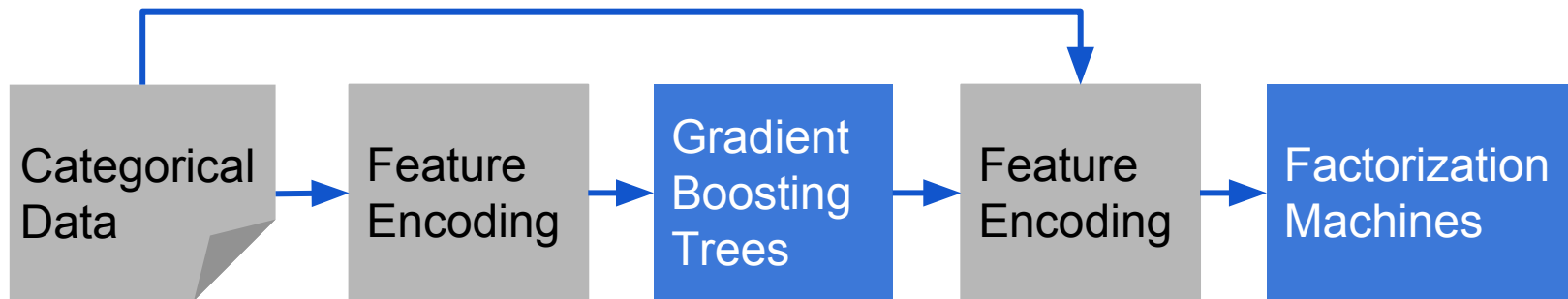
Gradient Boosting Trees + Factorization Machines



Gradient Boosting Trees + Factorization Machines



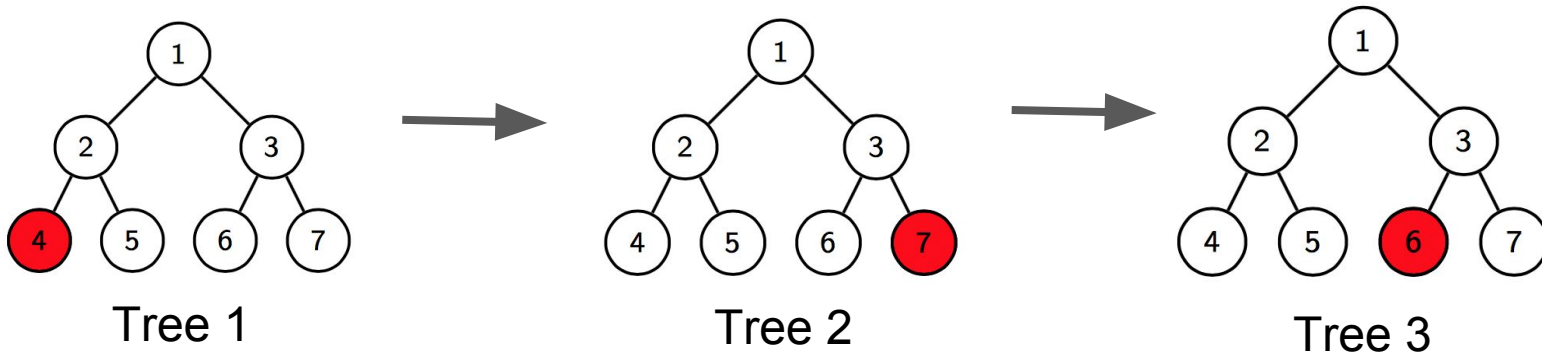
Gradient Boosting Trees + Factorization Machines



Feature vector \mathbf{x}																				Target y	
$\mathbf{x}^{(1)}$	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 $y^{(6)}$
A B C ... User				TI NH SW ST ... Movie				TI NH SW ST ... Other Movies rated				Time				TI NH SW ST ... Last Movie rated					

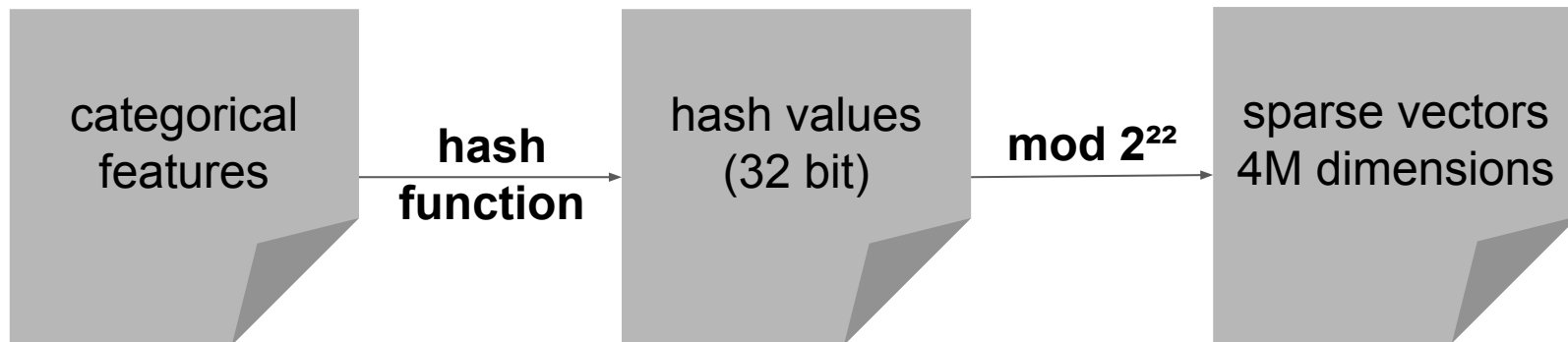
Feature Engineering: Gradient Boosting Trees Encoding

x



new features: tree1: 4, tree 2: 7, tree 3: 6

Feature Encoding



Factorization Machines

logistic regression:

$$y(x) = \sigma\left(\omega_0 + \sum_{i=1}^n \omega_i x_i + \sum_{i=1}^n \sum_{j=i+1}^n w_{i,j} x_i x_j\right)$$

Factorization Machines

logistic regression:

$$y(x) = \sigma(\omega_0 + \sum_{i=1}^n \omega_i x_i + \sum_{i=1}^n \sum_{j=i+1}^n \boxed{w_{i,j}} x_i x_j)$$

$\mathbf{W} : 4\text{M} \times 4\text{M}$

Factorization Machines

logistic regression:

$$y(x) = \sigma\left(\omega_0 + \sum_{i=1}^n \omega_i x_i + \sum_{i=1}^n \sum_{j=i+1}^n \omega_{i,j} x_i x_j\right)$$

$W : 4M \times 4M$

$$\omega_{i,j} = \langle v_i, v_j \rangle$$

$V : 4M \times k \ (k \ll 4M)$

**break down
independence of W**

factorization machines:

$$y(x) = \sigma\left(\omega_0 + \sum_{i=1}^n \omega_i x_i + \sum_{i=1}^n \sum_{j=i+1}^n \langle v_i, v_j \rangle x_i x_j\right)$$

Factorization Machines Optimization

Markov Chain Monte Carlo

-- Sampling model parameters



Automatic regularization



Prediction in training time

Built on a single machine

Stochastic Gradient Descent (SGD)

-- Minimizing loss function



Fast prediction, scalable



Lots of hyperparameters

Built on Spark

Distributed Mini-Batch SGD

Comparison Among Different Methods

Method	AUC	Log loss
FM + GBT	0.7456	0.3933
FM	0.7431	0.3952
LR + GBT	0.7432	0.3945
LR	0.7408	0.4006

FM: Factorization Machines with Markov Chain Monte Carlo

GBT: Gradient Boosting Trees

LR: Logistic Regression

Comparison Among Different Methods

Method	AUC	Log loss
FM + GBT	0.7456	0.3933
FM	0.7431	0.3952
LR + GBT	0.7432	0.3945
LR	0.7408	0.4006

FM: Factorization Machines with Markov Chain Monte Carlo

GBT: Gradient Boosting Trees

LR: Logistic Regression

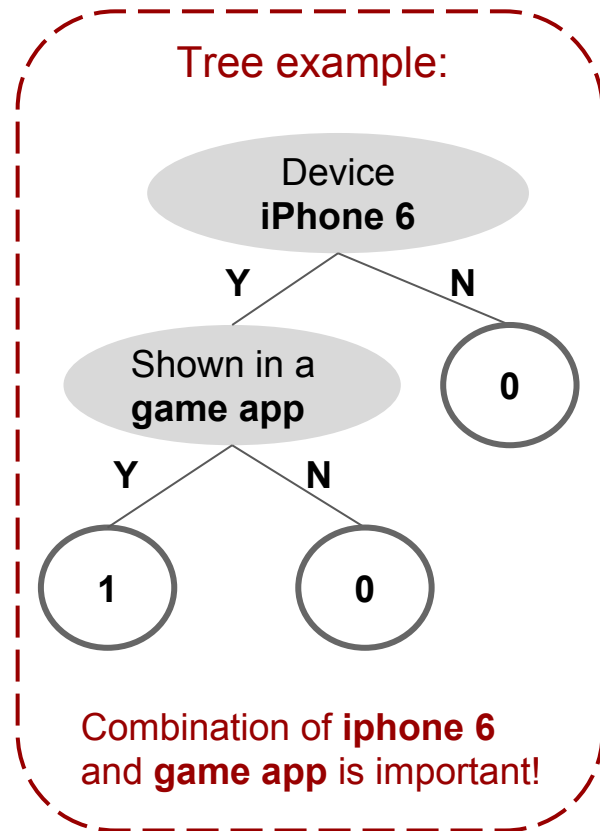
Comparison Among Different Methods

Method	AUC	Log loss
FM + GBT	0.7456	0.3933
FM	0.7431	0.3952
LR + GBT	0.7432	0.3945
LR	0.7408	0.4006

FM: Factorization Machines with Markov Chain Monte Carlo

GBT: Gradient Boosting Trees

LR: Logistic Regression



Comparison Among Different Methods

Method	AUC	Log loss
FM + GBT	0.7456	0.3933
FM	0.7431	0.3952
LR + GBT	0.7432	0.3945
LR	0.7408	0.4006

FM: Factorization Machines with Markov Chain Monte Carlo

GBT: Gradient Boosting Trees

LR: Logistic Regression

Running Time

Step	Single Machine
Encoding	~ 5 hrs
GBT	~ 55 hrs
FM / LR	~ 2 hrs



Single machine: EC2 with 160 GB memory

Distributed System: Spark EC2 with 10 nodes

Running Time

Step	Single Machine	Distributed System
Encoding	~ 5 hrs	~ 40 min
GBT	~ 55 hrs	~ 6 hrs
FM / LR	~ 2 hrs	~ 30 min




Single machine: EC2 with 160 GB memory
Distributed System: Spark EC2 with 10 nodes

Summary


Gradient Boosting Trees + Factorization Machines

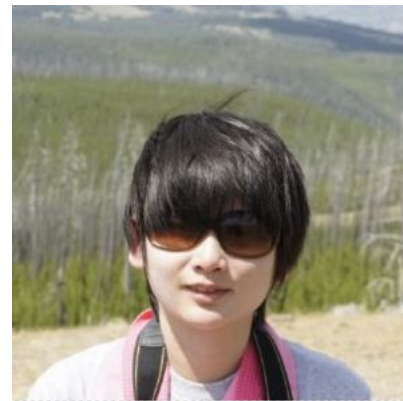
- 😊 **Automatic** feature engineering
- 😊 Capturing high order interactions under **sparsity**
- 😊 **Scalable** algorithm
- 😊 Potential for **production**




Binjie Lai
Binjie.lai@gmail.com
 /in/binjielai



Miao Lu
lumiao1105@gmail.com
 /in/lumiao1105



Wanyanxie
wanyanxie@gmail.com
 /in/wanyanxie