Click-Through-Rate Prediction

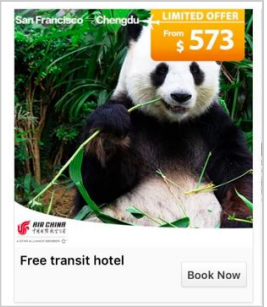
Scalable Factorization Machines

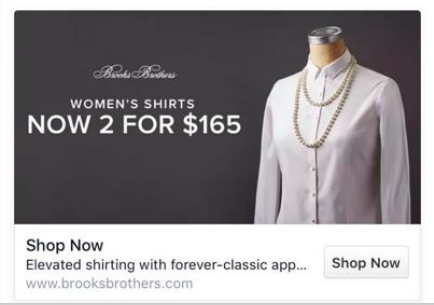
Binjie Lai, **M**iao Lu, **W**anyan Xie
MSAN, **USF**

USF

Click-Through-Rate Prediction

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





USF

Kaggle Dataset ：Avazu CTR Prediction

Train: **5.88 GB** Test: **629 MB**

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

Impressions

~ 36 million

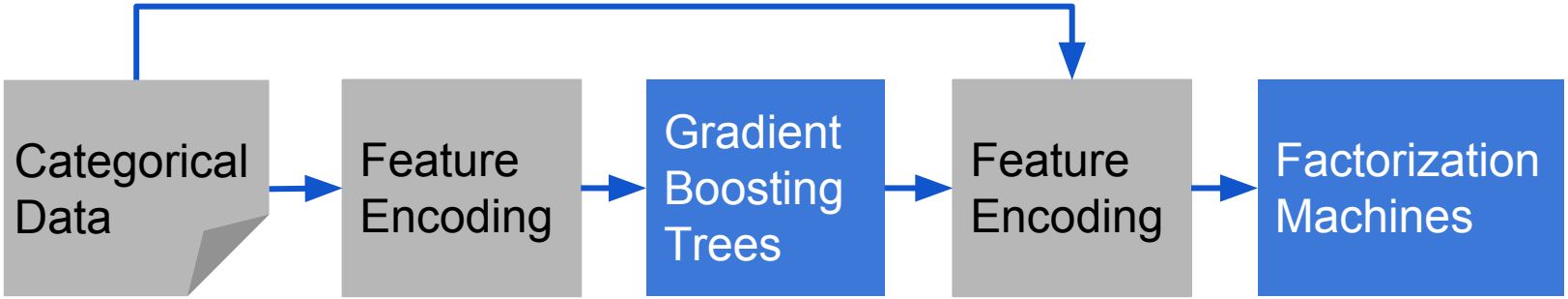
~ 4 million

label

23 categorical features

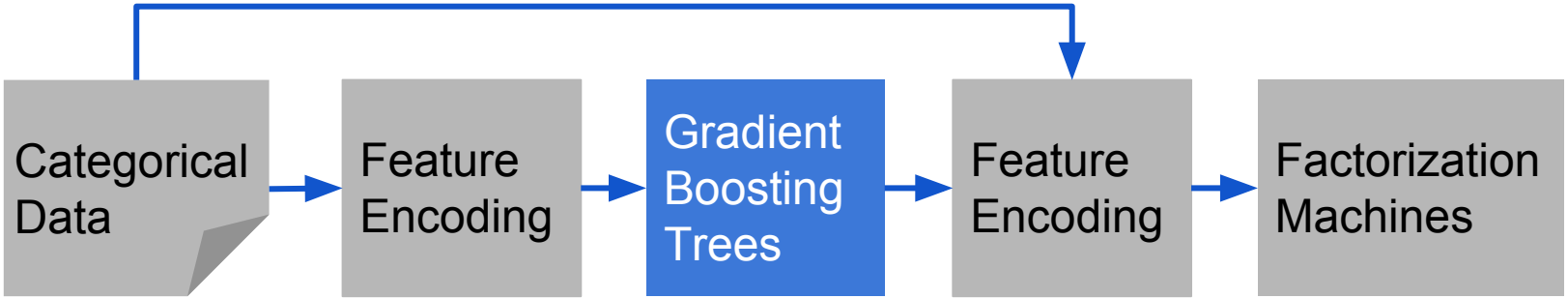
Data Source:https://www.kaggle.com/c/avazu-ctr-prediction

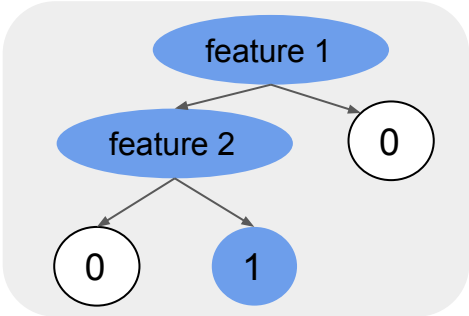
Gradient Boosting Trees + Factorization Machines



USF

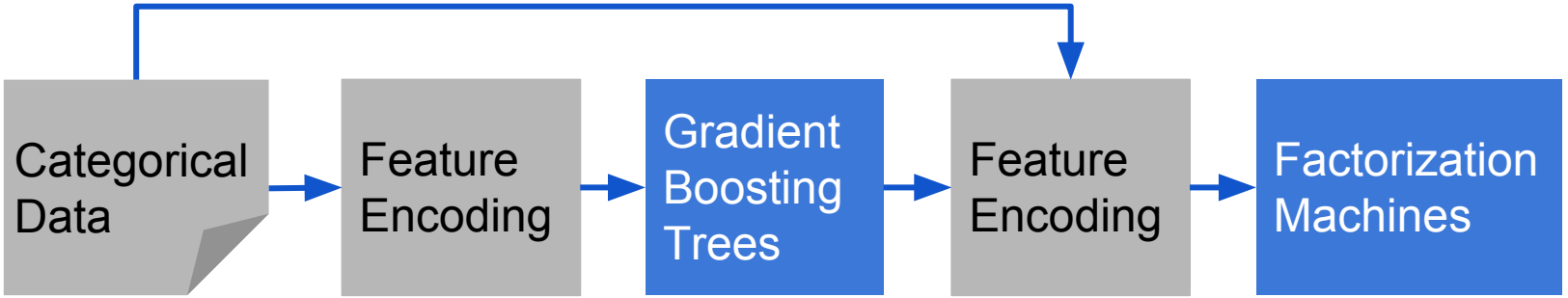
Gradient Boosting Trees + Factorization Machines

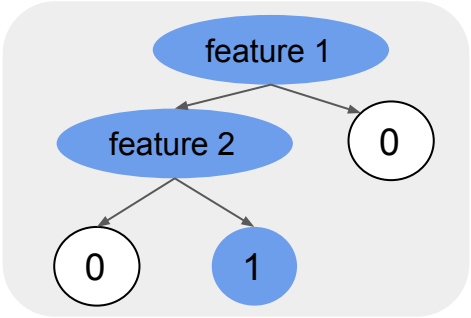


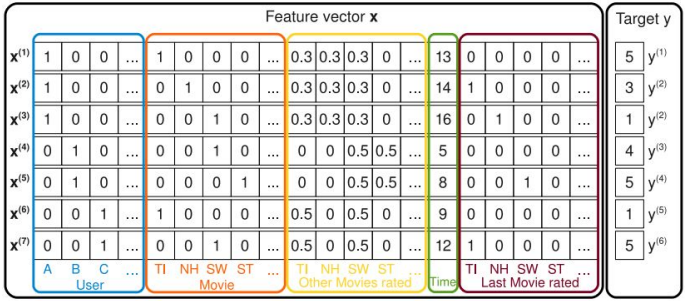


USF

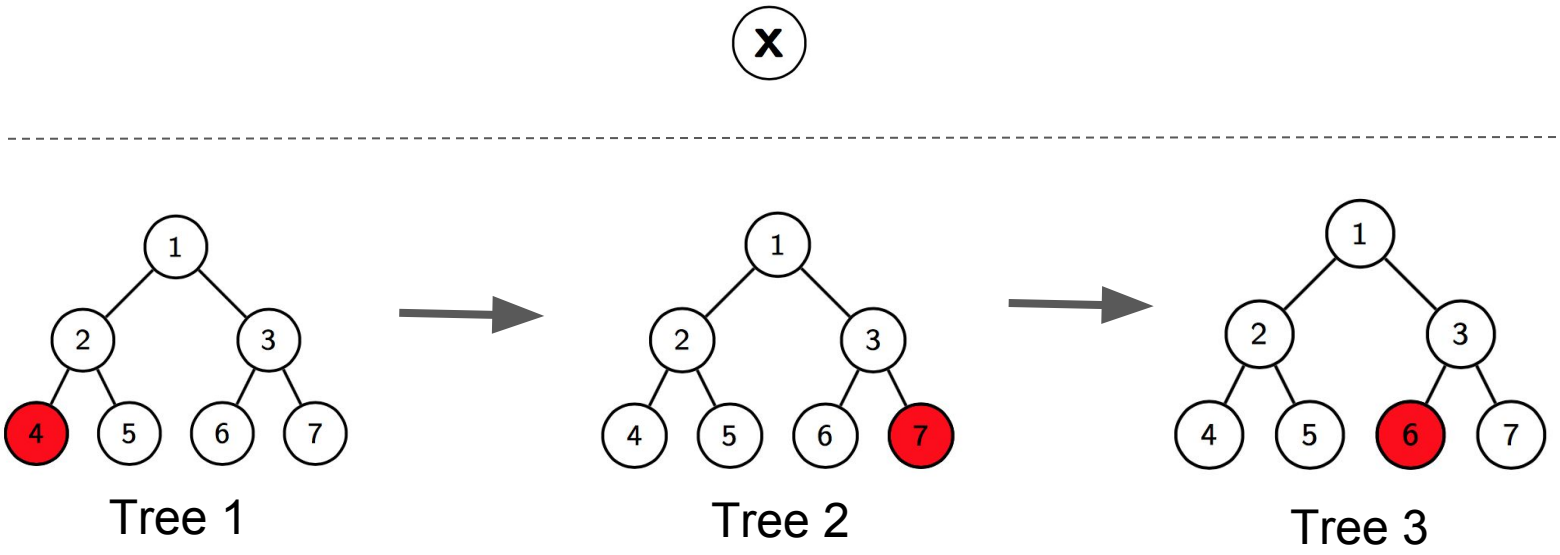
Gradient Boosting Trees + Factorization Machines







Feature Engineering: Gradient Boosting Trees Encoding

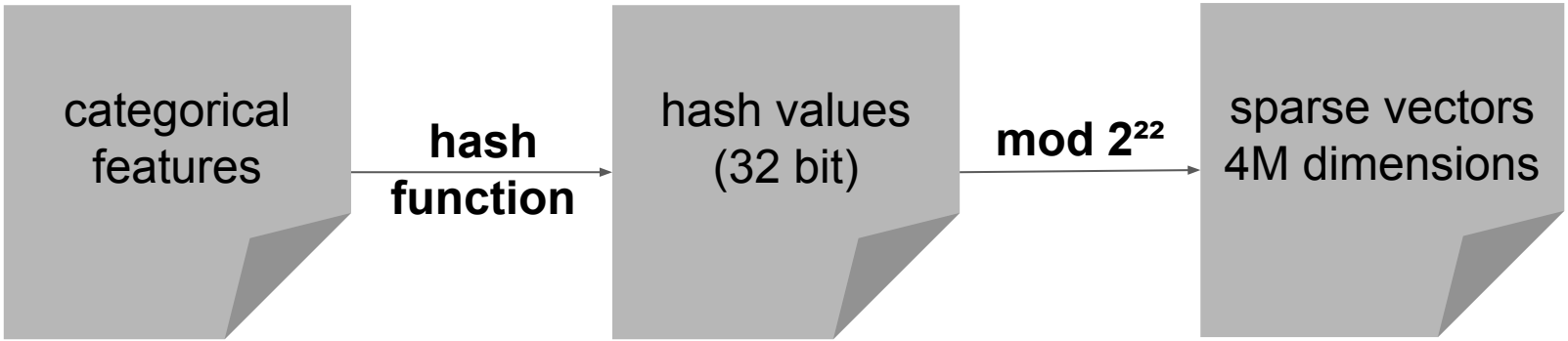


new features: tree1: 4, tree 2: 7, tree 3: 6

Xinran He, etc. *Practical Lessons from Predicting Clicks on Ads at Facebook*

USF

Feature Encoding



Factorization Machines

logistic regression:

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

Factorization Machines

logistic regression:

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

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

Factorization Machines Model

logistic regression:

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

$$\omega_{i,j} =< v_i, v_j >$$

$\mathrm{W} : 4\mathrm{M} \times 4\mathrm{M}$

$\mathrm{V} : 4\mathrm{M} \times \mathrm{k} \ (\mathrm{k} << 4\mathrm{M})$

break down independence of W

factorization machines:

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

S Rendle, Factorization Machines

USF

Factorization Machines Model Optimization

Markov Chain Monte Carlo

-- Sampling model parameters

 Automatic regulization

 Prediction in training time

Built on a single machine

Stocastic Gradient Desent (SGD)

-- Minimizing loss function

 Fast prediction, scalable

 Lots of hyperparameters

Built on Spark
Distributed Mini-Batch SGD

USF

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

USF

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

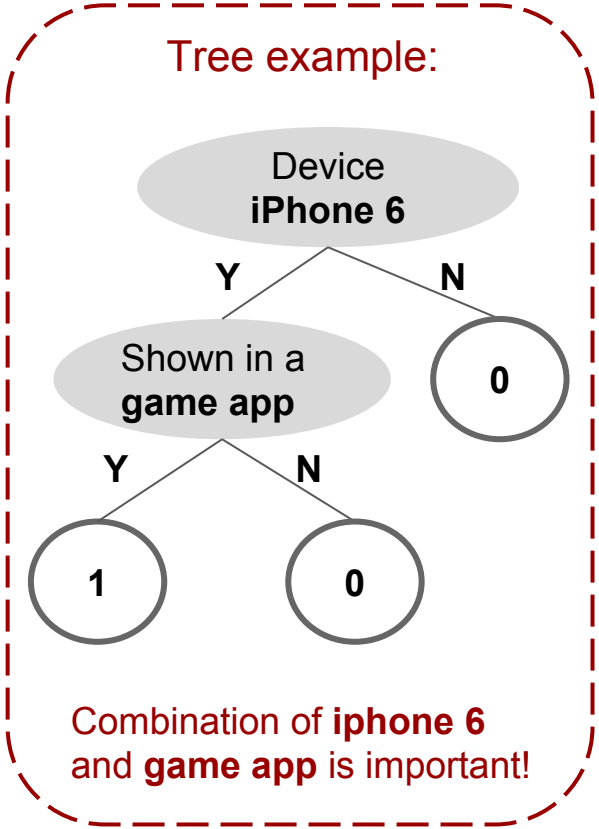
FM: Factorization Machines with Markov Chain Monte Carlo
GBT: Gradient Boosting Trees
LR: Logistic Regression

USF

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

USF

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

USF

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

USF

Summary

Gradient Boosting Trees + Factorization Machines


Automatic feature engineeing


Capturing high order interactions under **sparsity**


Scalable algorithm


Potential for **production**

USF

Thank you!

Binjie.lai@gmail.com
luMiao1105@gmail.com
Wanyanxie@gmail.com

USF