效果广告中的机器学习技术

赵学敏
xueminzhao@tencent.com

主要内容

1	什么是机器学习
2	监督学习：转化预估
3	非监督学习：文本主题建模
4	小结

主要内容

机器学习定义

- **Machine learning** is a scientific discipline that explores the construction and study of algorithms that can learn from data. Such algorithms operate by building a model from example inputs and using that to make predictions or decisions, rather than following strictly static program instructions.

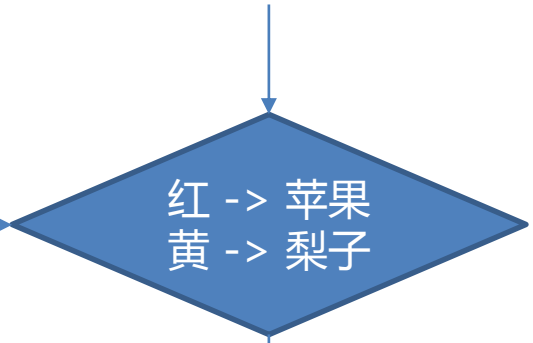
—— from wikipedia

(x, y)

一个例子：二分类

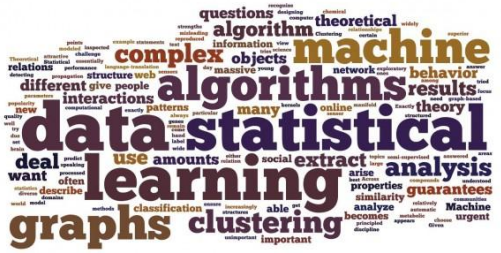
图片	颜色	形状	类别
	红	圆	苹果
	红	圆	苹果
	红	圆	苹果
	黄	椭圆	梨子
	黄	圆	梨子
	黄	圆	梨子
	x		y

黄 & 圆



梨子

一堆名词 …



计算广告学

人工智能

数据挖掘

机器学习

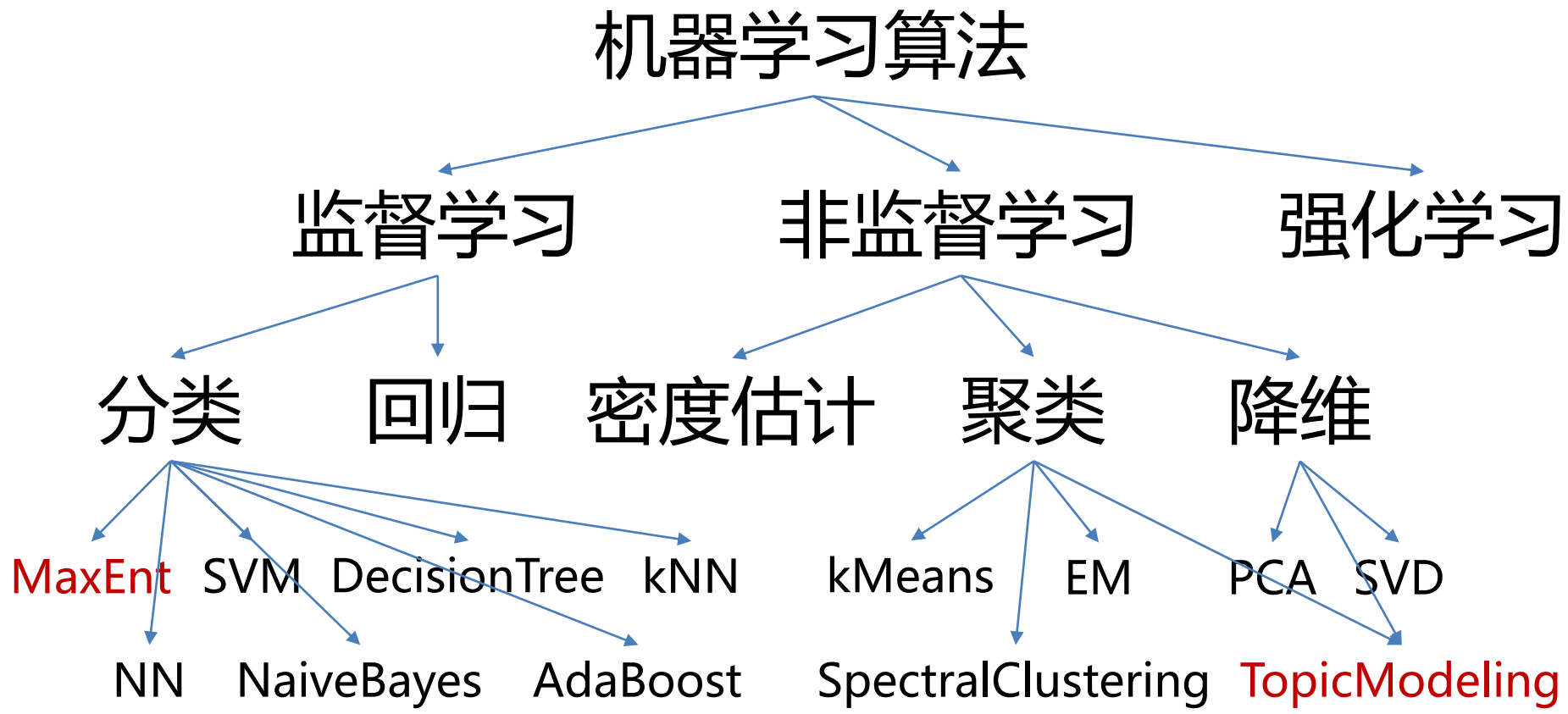
统计学

模式识别

信息检索

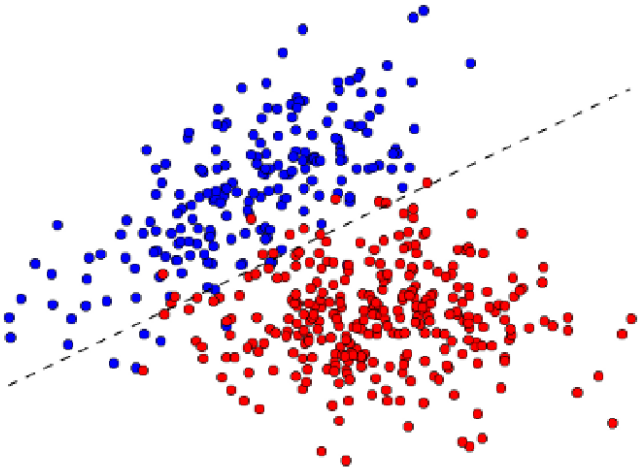
自然语言处理

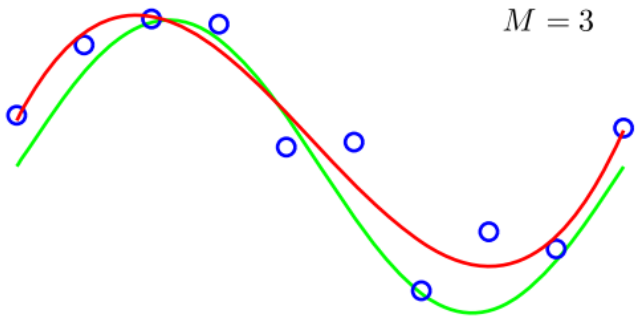
算法分类


机器学习算法
监督学习
非监督学习
强化学习
分类
回归
密度估计
聚类
降维
MaxEnt
SVM
DecisionTree
kNN
kMeans
EM
PCA
SVD
NN
NaiveBayes
AdaBoost
SpectralClustering
TopicModeling

区别

- 分类 vs. 回归
- 分类 vs. 聚类
- 生成模型 vs. 判别模型
  - ✓ $p(x, y)$
  - ✓ $p(y|x)$、$y = f(x)$

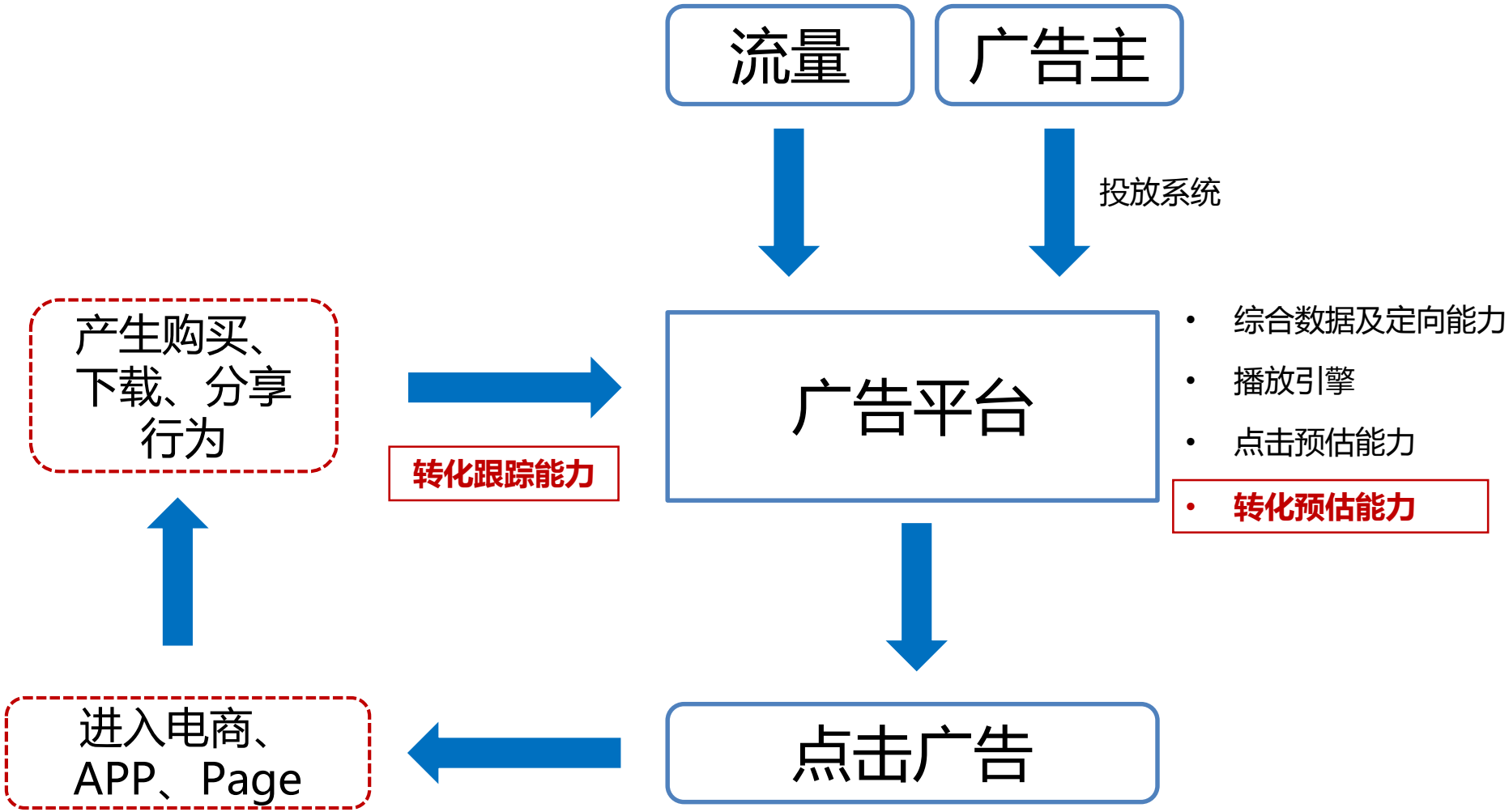




主要内容

1	什么是机器学习
2	监督学习：转化预估
3	非监督学习：文本主题建模
4	小结

广点通闭环生态系统


流量
广告主
投放系统
产生购买、下载、分享行为
转化跟踪能力
广告平台
综合数据及定向能力
播放引擎
点击预估能力
转化预估能力
进入电商、APP、Page
点击广告

为什么需要转化预估？

- CPA 广告
  - ✓ eCPM = CPA × CTR × CVR
- 打造闭环，兼顾广告主价值
  - ✓ 广告平台价值：点击率预估
  - ✓ 广告主价值：转化率、ROI
  - ✓ 点击 ≠ 转化

点击预估 vs. 转化预估

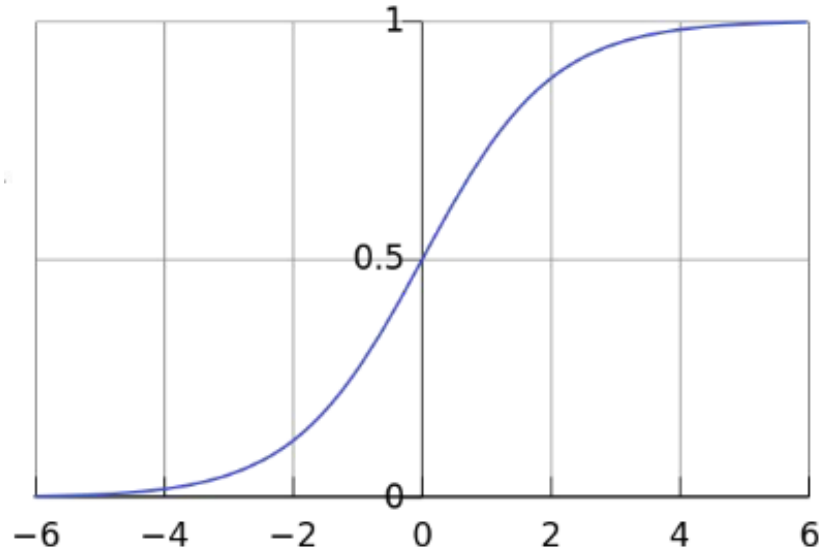
	点击预估	转化预估
物理含义	单一明确	多样，几十种
数据收集	容易	困难
数据量	巨大	大
回流时间	< 分钟	天级别
模型时效性	强	较强
数据噪声	误点击	数据“丢失”

转化预估建模

- Logistic Regression (LR) 模型
  - ✓ 判别、MaxEnt (C=2) 线性分类器

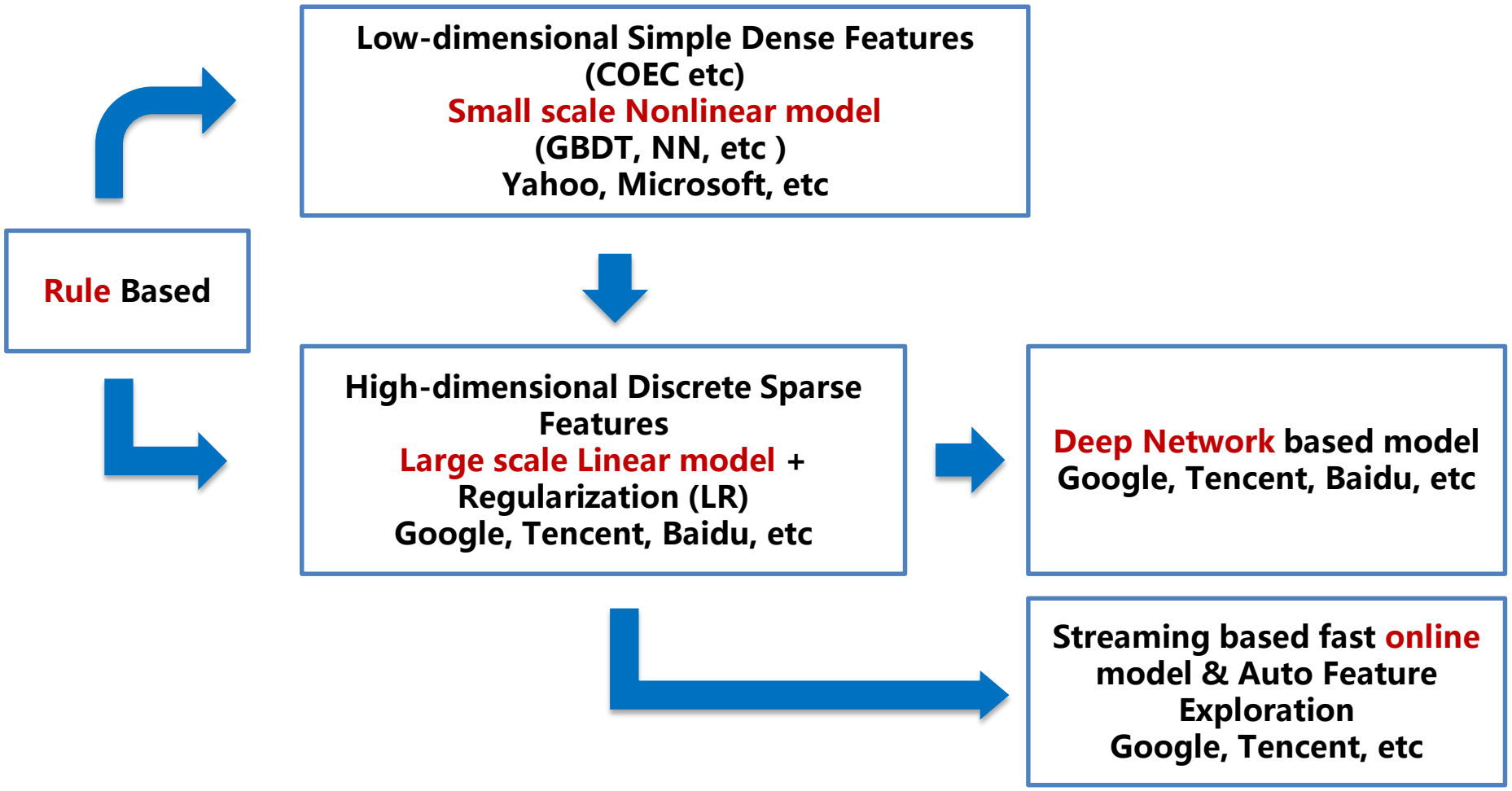
$$p(y=1|\boldsymbol{x}) = \frac{1}{1+e^{-\boldsymbol{wx}}}$$

$$\ln \frac{p(y=1|\boldsymbol{x})}{p(y=0|\boldsymbol{x})} = \boldsymbol{wx}$$



- 正则化：L_1 vs. L_2
- 参数估计
  - ✓ 最大似然估计
  - ✓ 优化方法 L-BFGS、SGD、FTRL …

技术演变

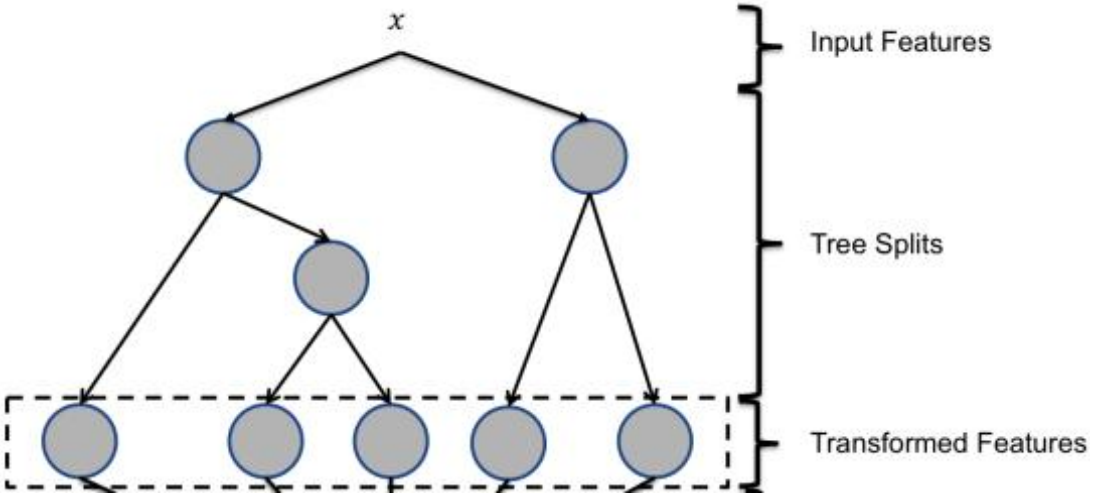

Rule Based
Low-dimensional Simple Dense Features
(COEC etc)
Small scale Nonlinear model
(GBDT, NN, etc)
Yahoo, Microsoft, etc
High-dimensional Discrete Sparse Features
Large scale Linear model + Regularization (LR)
Google, Tencent, Baidu, etc
Deep Network based model
Google, Tencent, Baidu, etc
Streaming based fast online model & Auto Feature Exploration
Google, Tencent, etc

为什么使用 LR 模型？

- 简单、容易并行化
- 可扩展性好：特征、样本
- 较好的处理不均衡数据
- 概率输出：适合 GSP 拍卖机制
  - ✓ $b_n = p_{n+1} b_{n+1} / p_n$
- Online 算法
- Exploration 策略

特征

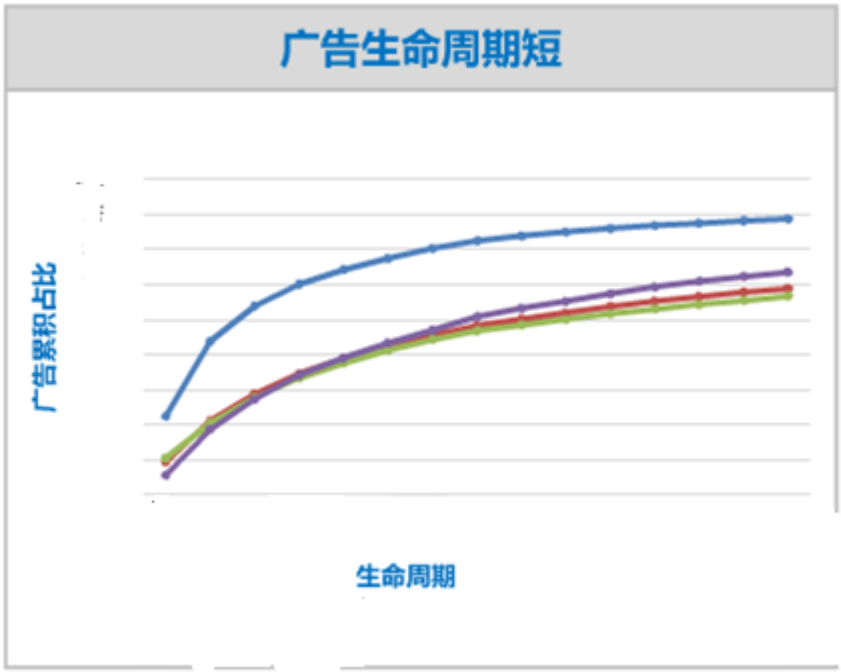
- 三类重要特征
  - ✓ 用户、广告、场景
  - ✓ 场景：流量、广告位、LBS、时间 …
- 特征类型
  - ✓ 类别型（性别）、连续值型（商品好评度）
- 引入非线性
  - ✓ Log
  - ✓ 离散化
  - ✓ 交叉特征

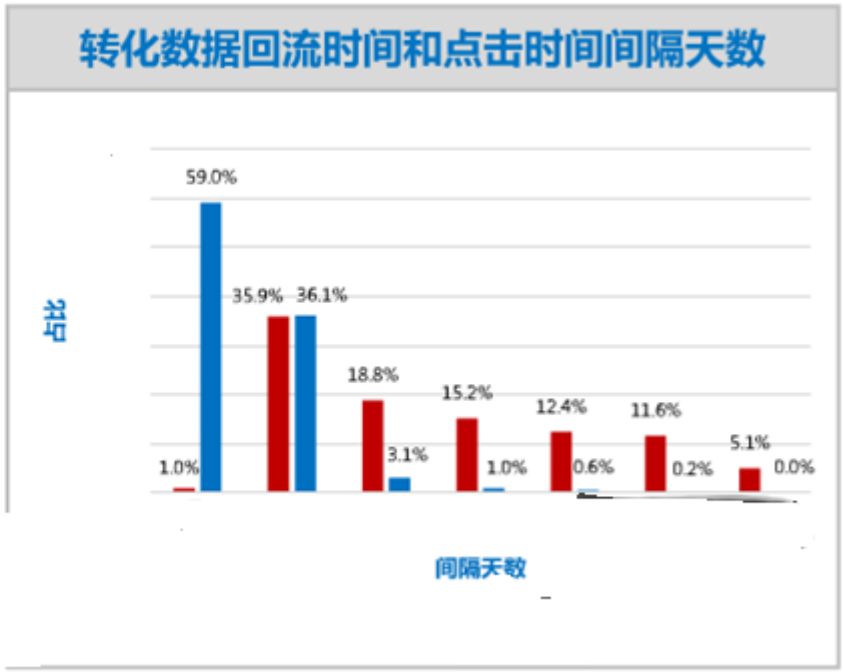


模型评估

- 离线评估：
  - ✓ 模型排序能力：AUC
  - ✓ 模型拟合能力：Log Likelihood、预估-真实CVR 散点图
- 在线实验：
  - ✓ Online A-B Test 实验，对比线上业务指标，包括CTR、CPC、CPM、CVR、ROI等。

广告生命周期 vs. 转化回流延迟





- 模型不使用最新日志数据：丢失新广告数据
- 模型使用最新日志数据：丢失延迟回流的转化数据
- PU-Learning：没有转化的点击数据，离现在越久，越有可能是负样本

其他

- 并行训练系统
- 特征选择
- 自动学习高阶组合特征
- Deep Learning
- 点击预估可能更关心的问题：
  - ✓ 大量图片广告，图像特征
  - ✓ 实时系统

主要内容

广告相关性的重要性

三星S6 PK 苹果6+：谁的拍照好？

苹果

- 苹果2015春季新品发布会专题报道
- 苹果发布12英寸MacBook 机身13.1mm

MWC

- 三星GALAXY S6/S6 edge上手体验
- HTC One M9动手试玩 系统调整丰富

CES

- LG发布第二代曲面屏手机 擦伤可自我修复
- 你没有看错！柯达推出旗下首款智能手机

专访

- 对话马麟：乐视EUI会开放 要做手机生态
- 华硕沈振来：智能手机销量目标2500万台

- 优质媒体有准入要求；
- 提升用户体验，提高广告点击率、转化率等。

短文本相关性

- 基于词 TF-IDF 特征，Cosine 相似度：

Q1 **apple** pie

Q2 iphone crack

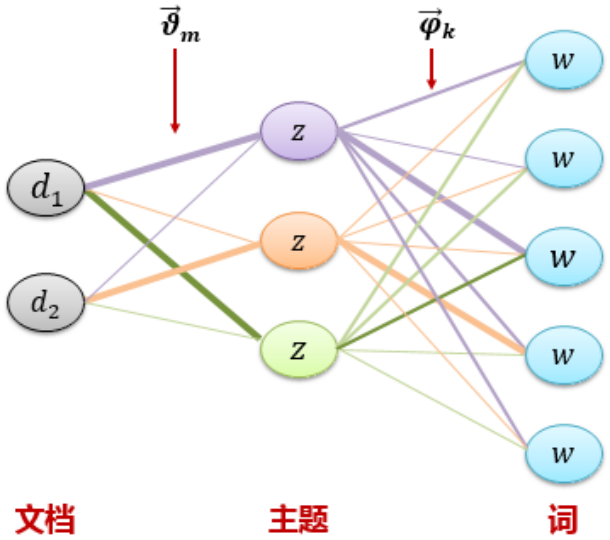
D1 **Apple** Computer Inc. is a well known company located in California, USA.

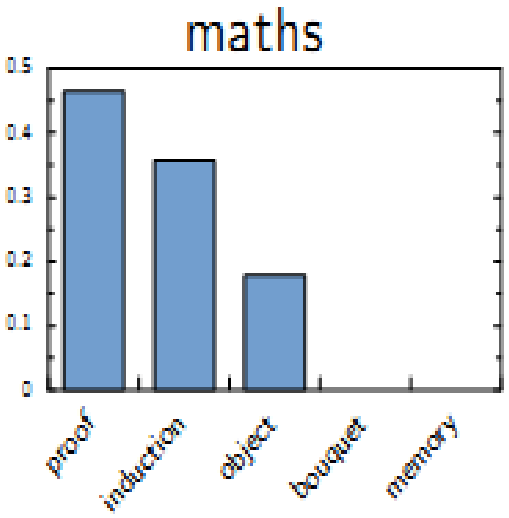
D2 The **apple** is the pomaceous fruit of the apple tree…

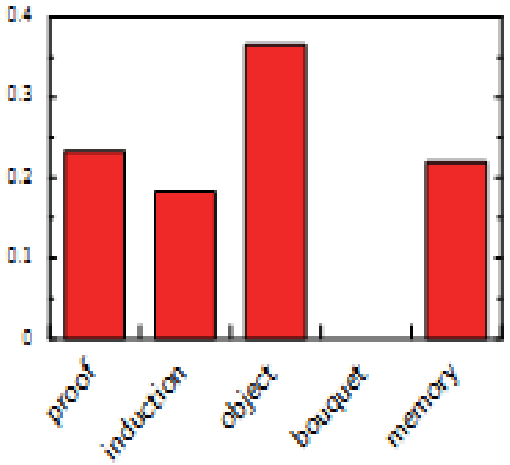
$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

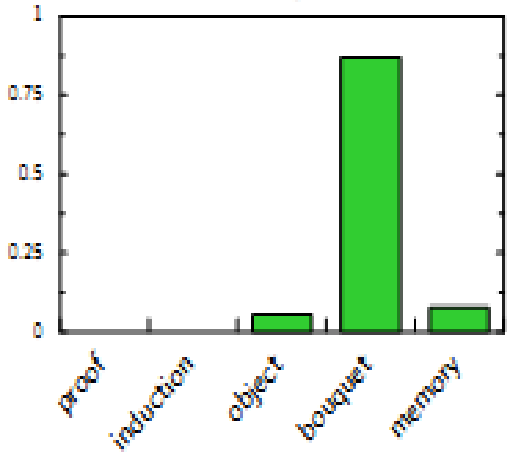
文本主题建模

- 文档是由主题组成的；
- 主题是词表上的概率分布。

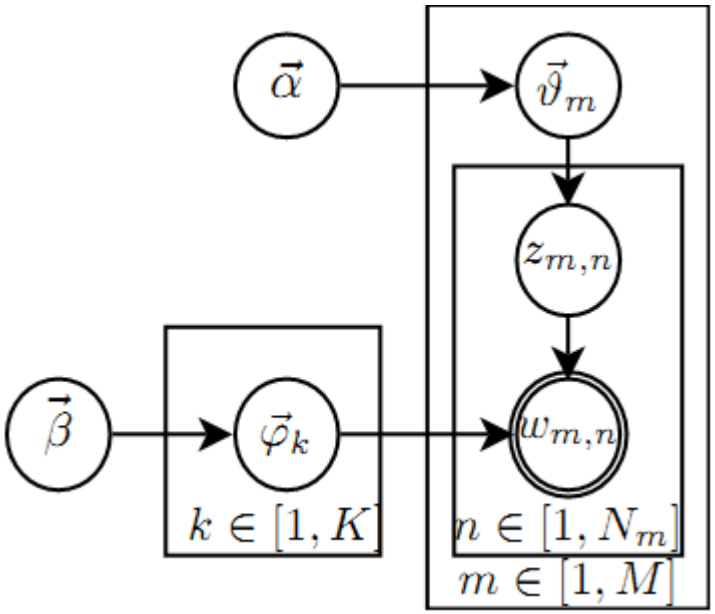


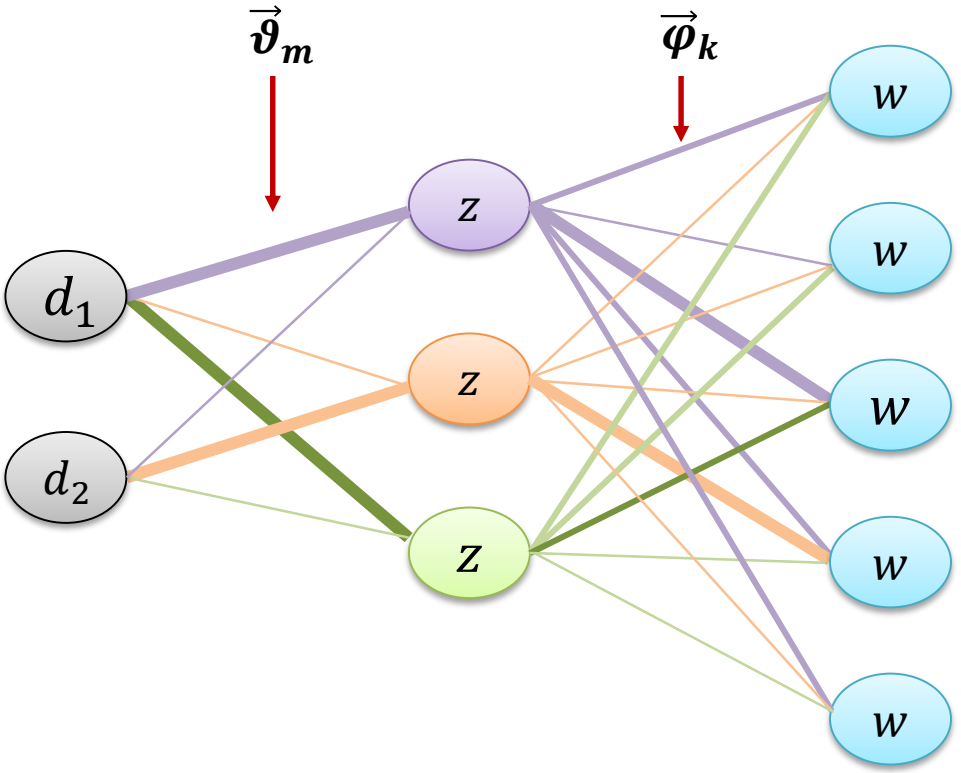






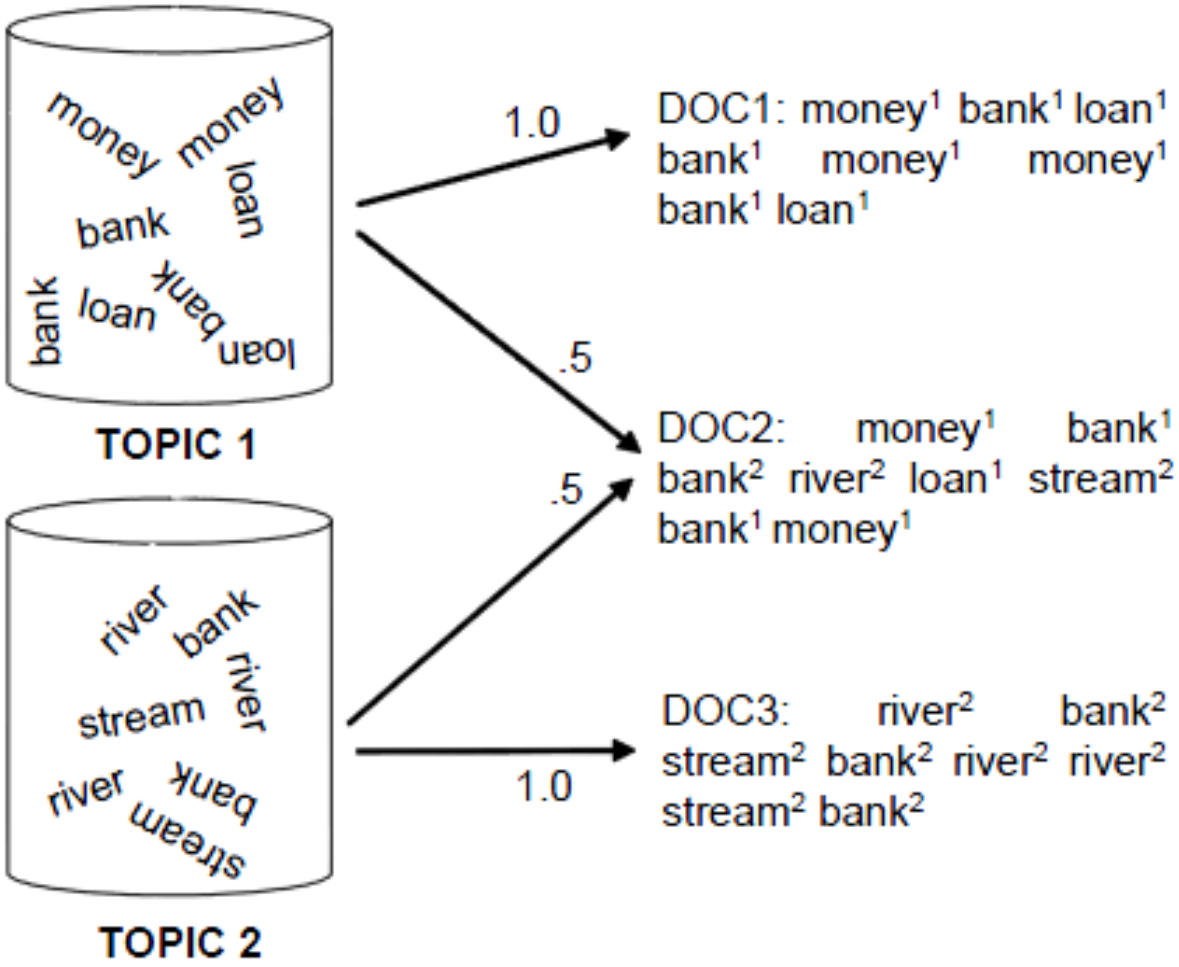
Latent Dirichlet Allocation



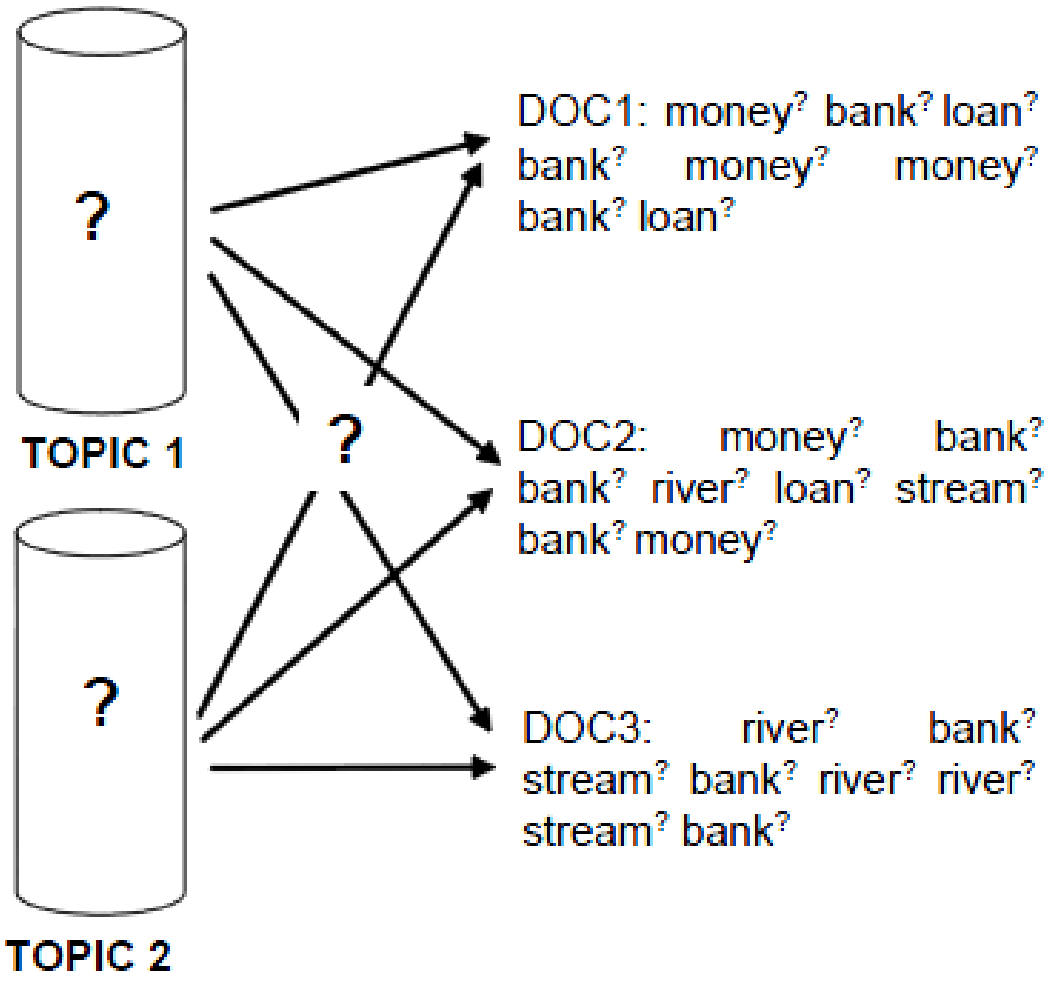


- 生成模型 $p(\vec{w}_m, \vec{z}_m, \vec{\vartheta}_m, \underline{\Phi}|\vec{\alpha}, \vec{\beta}) = \overbrace{\underbrace{\prod_{n=1}^{N_m} p(w_{m,n}|\vec{\varphi}_{z_{m,n}}) p(z_{m,n}|\vec{\vartheta}_m)}_{\text{word plate}} \cdot p(\vec{\vartheta}_m|\vec{\alpha}) \cdot \underbrace{p(\underline{\Phi}|\vec{\beta})}_{\text{topic plate}}}^{\text{document plate (1 document)}}$

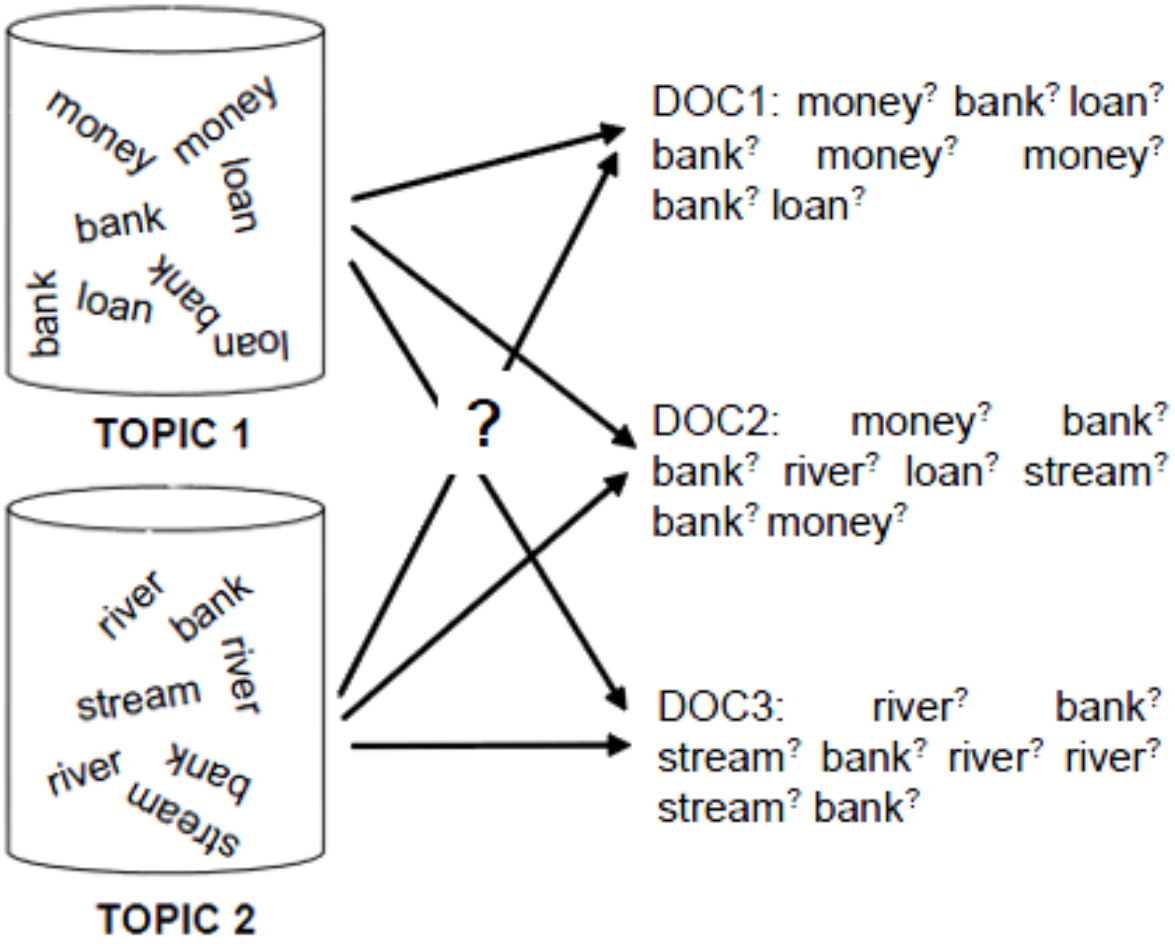
生成过程


money money loan bank bank loan bank loan
TOPIC 1
river bank river stream river bank stream
TOPIC 2
1.0
.5
.5
1.0
DOC1: money¹ bank¹ loan¹ bank¹ money¹ money¹ bank¹ loan¹
DOC2: money¹ bank¹ bank² river² loan¹ stream² bank¹ money¹
DOC3: river² bank² stream² bank² river² river² stream² bank²

训练过程


?
TOPIC 1
?
?
TOPIC 2
DOC1: money? bank? loan? bank? money? money? bank? loan?
DOC2: money? bank? bank? river? loan? stream? bank? money?
DOC3: river? bank? stream? bank? river? river? stream? bank?

推断过程


money money loan bank bank loan bank loan
TOPIC 1
river bank river stream river bank stream
TOPIC 2
?
DOC1: money? bank? loan? bank? money? money? bank? loan?
DOC2: money? bank? bank? river? loan? stream? bank? money?
DOC3: river? bank? stream? bank? river? river? stream? bank?

lda inference demo

text:

红酒木瓜汤

输入文档

submit

model information: num_topics = 10000, num_markov_chains = 10, num_total_iters

gibbs-sampling p(z|d): 文档表达的主要主题

一个主题一行

0.397815	6147	丰胸(0.164235) 产品(0.0776686) 减肥(0.0645986) 木瓜(0.0464668) 效果(0.0351566)
0.182650	3904	饭后(0.125137) 饭前(0.0757139) 服用(0.0263615) 减肥(0.0227619) 孕妇(0.0201505)
0.162571	3527	功效(0.0435682) 山药(0.0390381) 作用(0.0379043) 做法(0.0264466) 中药(0.0189896)
0.095631	6338	糖尿病(0.0811359) 血糖(0.0336291) 高血压(0.0285981) 孕妇(0.0218026) 血压(0.021
0.050685	4926	蜂蜜(0.0801284) 牛奶(0.0427497) 面膜(0.0303977) 好处(0.0256547) 鸡蛋(0.0238551)
0.044947	4515	做法(0.0598369) 萝卜(0.0569484) 排骨(0.0213306) 牛肉(0.017572) 腌制(0.0169023)
0.019126	8009	奇迹(0.238411) 世界(0.0786965) 木瓜(0.0362741) 加点(0.0362741) mu(0.0352766) 战
0.001914	4742	葡萄酒(0.128271) 干红(0.0887913) 价格(0.0739765) 红酒(0.0350377) 长城(0.032467
0.000956	5800	怀孕(0.142018) 肚子(0.130775) 孕妇(0.0953052) 初期(0.0334886) 征兆(0.0127122)

主题权重

使用词来描述主题的主要含义

gibbs-sampling p(w|d): 使用词来描述文档的主要含义

0.068293	丰胸
0.034068	减肥
0.032192	产品
0.023794	饭后
0.022052	木瓜
0.015636	效果
0.014349	饭前
0.009740	孕妇
0.008413	糖尿病
0.008080	作用

格式：词的权重 词

gibbs-sampling p(z|d):

```
0.170434   4998      苹果(0.230855) 手机(0.12452) iphone(0.0251039) 电脑(0.0171699)
0.086149   6261      范冰冰(0.114506) 苹果(0.0857748) 电影(0.0592054) 视频(0.0344978
0.058666   5642      iphone(0.166094) 手机(0.0703596) 3gs(0.0395753) 苹果(0.0330713)
0.025660   2134      千克(0.198335) 苹果(0.0784678) 重量(0.0269041) 大米(0.0199628)
0.014673   4966      手机(0.182923) 步步高(0.0826732) 电池(0.0434243) 下载(0.0414058
0.012849   4926      蜂蜜(0.0801284) 牛奶(0.0427497) 面膜(0.0303977) 好处(0.0256547)
0.011031   3898      圣诞节(0.0992735) 圣诞(0.0514906) 礼物(0.0351853) 祝福语(0.0344
0.011005   9480      下载(0.164783) mp4(0.103005) 电影(0.0636672) 视频(0.05098) 格式
0.009197   805       windows(0.0892736) xp(0.088124) 系统(0.0509883) 下载(0.0424573)
0.009190   8787      水果(0.0966501) 蔬菜(0.076337) 批发(0.0591047) 市场(0.0500198)
```

gibbs-sampling p(w|d):

```
0.054041   苹果
0.041369   手机
0.014812   下载
0.014351   iphone
0.010200   电脑
0.009983   范冰冰
0.008309   电影
0.007954   视频
0.007053   价格
0.005963   软件
```

图 4 Peacock 文档语义推断示例 1：“苹果”

gibbs-sampling p(z|d):

0.286885 2134 千克(0.198335) 苹果(0.0784678) 重量(0.0269041) 大米(0.0199628)
0.286885 532 桃子(0.0500153) 苹果(0.0364393) 李子(0.0283832) 南方(0.0261827)
0.104240 6338 糖尿病(0.0811359) 血糖(0.0336291) 高血压(0.0285981) 孕妇(0.021
0.095629 5691 咳嗽(0.0890044) 宝宝(0.0802794) 止咳(0.0497114) 小孩(0.0314726)
0.095629 3000 开花(0.029144) 果树(0.0262511) 修剪(0.0252923) 技术(0.0240111)
0.073634 1666 年级(0.0567557) 数学(0.0522898) 难题(0.0430788) 答案(0.0418692)
0.011476 177 奥比岛(0.261746) 奥比(0.0410186) 邮递员(0.0381663) 考试(0.0272
0.000959 6156 移动(0.0850844) 套餐(0.0559763) 联通(0.0540878) 动感地带(0.034
0.000957 4998 苹果(0.230855) 手机(0.12452) iphone(0.0251039) 电脑(0.0171699)

gibbs-sampling p(w|d):

0.058426 千克
0.036056 苹果
0.015995 桃子
0.008970 李子
0.008635 咳嗽
0.008599 糖尿病
0.008419 宝宝
0.008340 水果
0.008181 南方
0.007925 重量

图 5 Peacock 文档语义推断示例 2：“苹果 梨子”

gibbs-sampling p(z|d):

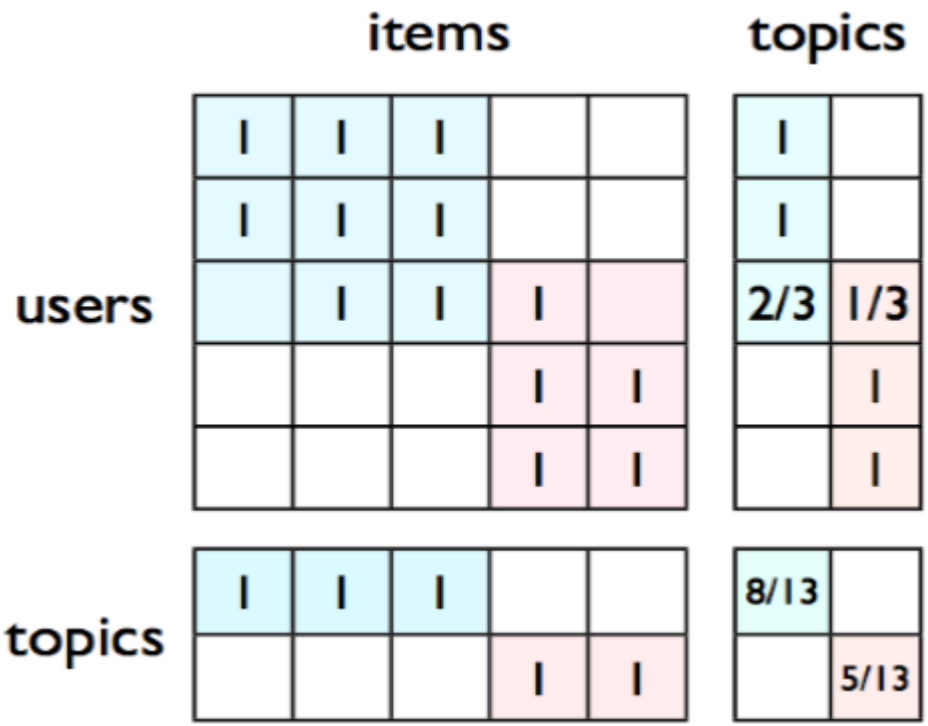
0.465717	6261	范冰冰(0.114506) 苹果(0.0857748) 电影(0.0592054) 视频(0.0344978)
0.206565	8602	家具(0.178903) 红木(0.0262665) 价格(0.0161301) 图片(0.0147936) 北
0.095628	4853	医生(0.0711662) 医院(0.0379315) 全身(0.0236984) 儿科(0.0210252) :
0.095628	56	减肥(0.112415) 噪声(0.0355498) 运动(0.0313898) 污染(0.0256226) c1
0.020083	2010	电影(0.16116) 情色(0.0753046) 在线(0.0676254) av(0.0481669) 伦理
0.015301	7563	边城(0.101929) 陆风(0.0856327) 沈从文(0.0575565) x8(0.0409549) x1
0.013388	2743	把握(0.157607) 机会(0.0667866) 作文(0.0167649) 教材(0.0152458) 分
0.010520	5275	价值(0.195102) 药用(0.111382) 收藏(0.0256952) 人生(0.0154372) 植
0.004781	6264	帝王(0.14446) 慵懒(0.0376641) 超女(0.0317217) 妖孽(0.0275368) su
0.002870	7876	饮料(0.127533) 食品(0.0703096) 公司(0.0695525) 有限(0.040378) 碳i

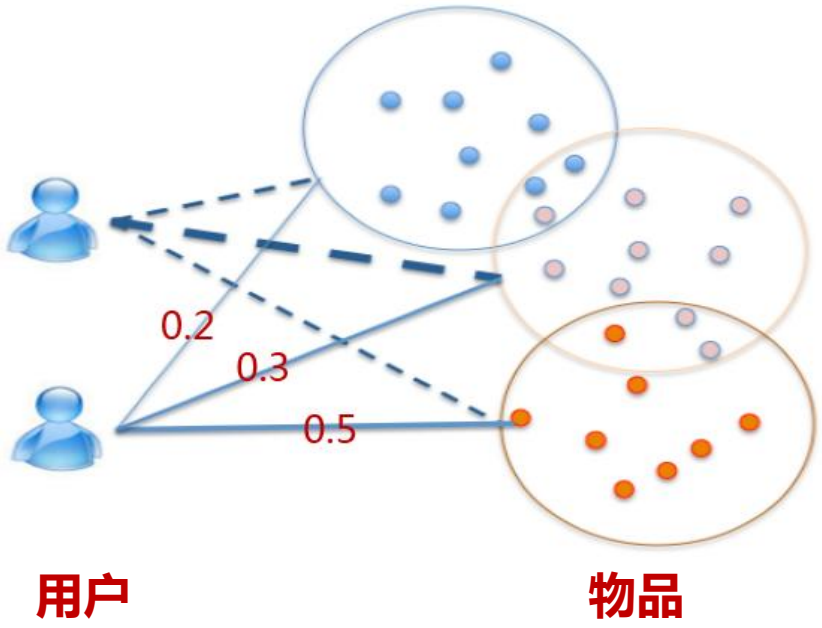
gibbs-sampling p(w|d):

0.053949	范冰冰
0.042701	苹果
0.037216	家具
0.031803	电影
0.018056	视频
0.014876	减肥
0.014846	佟大为
0.013595	近义词
0.012089	反义词
0.010294	电视剧

图 6 Peacock 文档语义推断示例 3：“苹果大尺度”

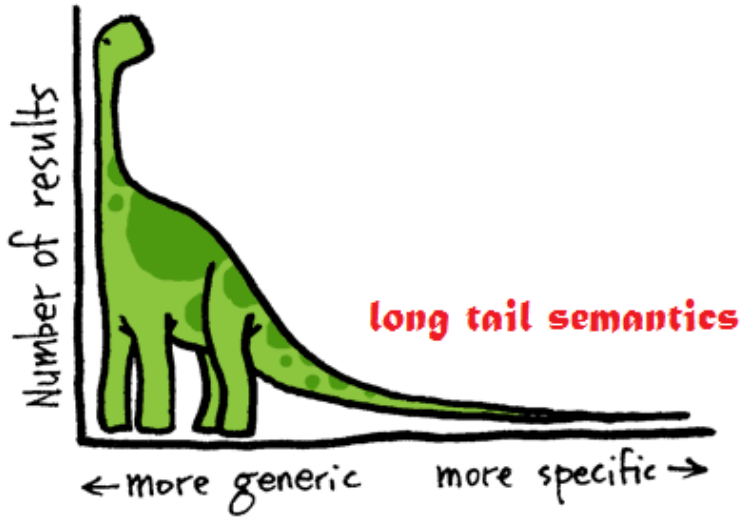
RecSys：Latent Factor Model

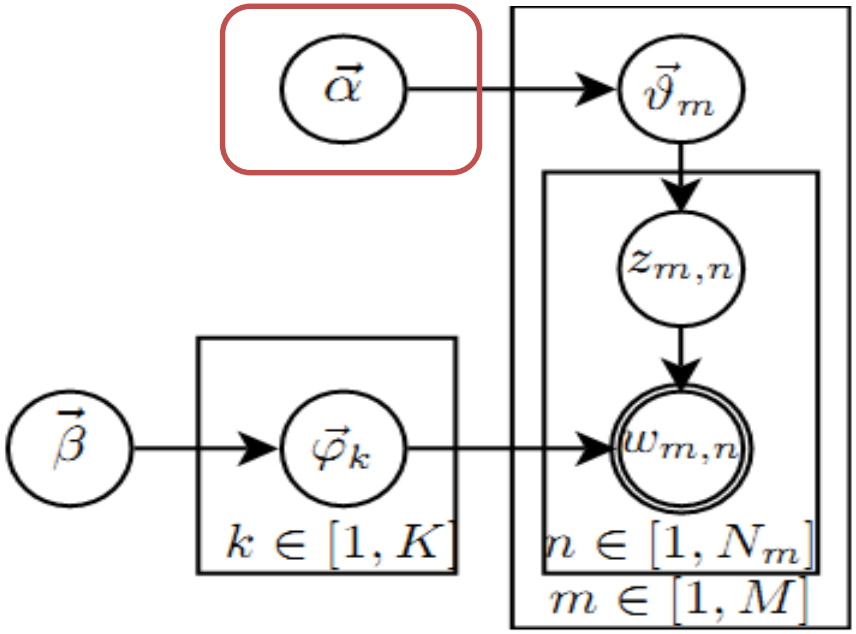




超参数的影响

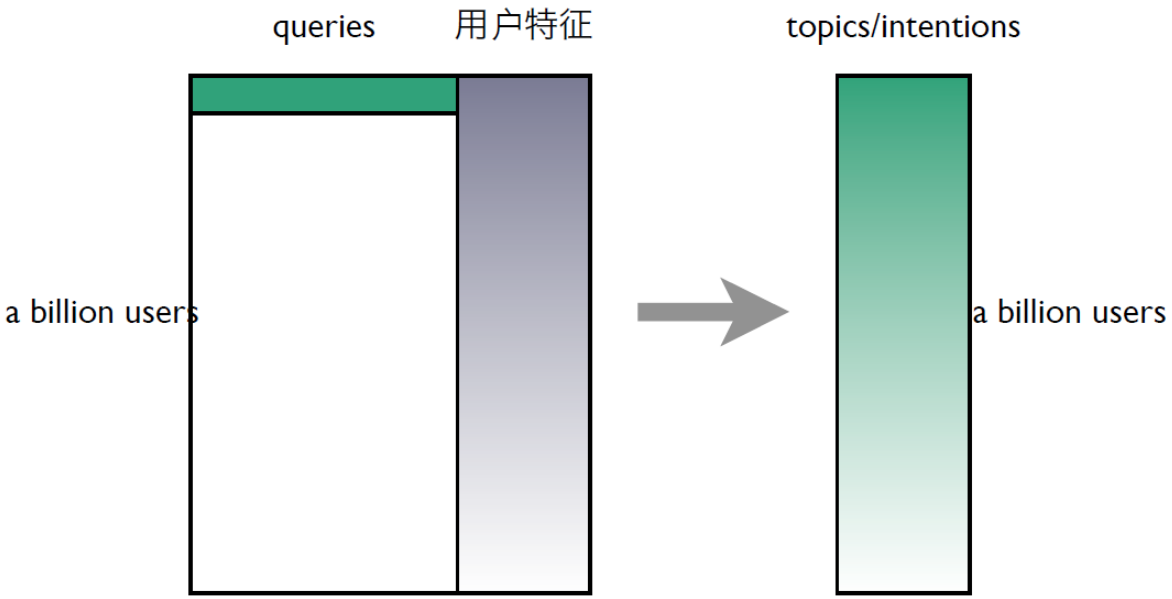
- 超参数 alpha 对模型质量有重要的影响
- 每轮迭代中，通过 MLE 估计优化 alpha

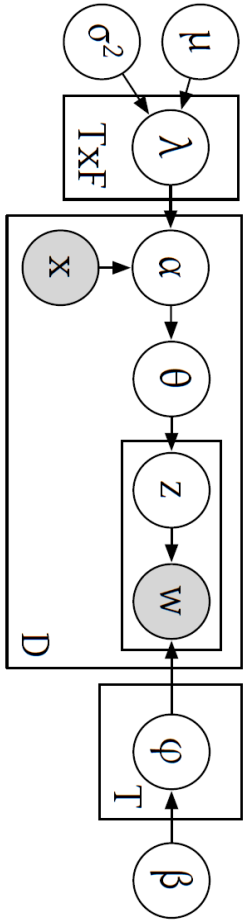




Dirichlet-Multinomial Regression

- 腾讯 20 亿用户，大部分用户都有高质量属性数据





DMR 年龄属性 例子

17 股票 投资 证券 黄金 理财 股 期货 股市 炒股 滚滚 金融 金 财源 财富 群 软件 盘 短线 股民

```
<default>   -3.4969567698702275
 age19-24   -1.108914651862329
 age25-30   -0.3828150512707217
 age31-40    0.12491698829717246
 ageUnk     -0.6510675869135051
 age41-inf   0.28808408849130157
 age1-18    -1.6687980266531255
```

49 同学 爱情 兄弟 开心 聊天 微笑 女人 缘分 美好 无聊 放弃 恋爱 说话 游戏 改名 男人 love 潮流 阳光

```
<default>   -2.6253915985636267
 age19-24    0.5921503327139054
 age25-30   -0.9256592908792939
 age31-40   -1.7998373628017386
 ageUnk     -0.2162909427852924
 age41-inf  -1.7028117456928522
 age1-18     1.5094800495869967
```

模型质量评价

- 内聚性
- 独特性

LDA 应用

- 文档（或用户）相关性计算
- $p(z|d)$（或 $p(w|d)$）作为特征
  - ✓ 分类器、LearningToRank、pCTR、pCVR 等
- 推荐系统

1	0	1	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0
0	1	1	0	1	0	0	0	0
0	1	1	2	0	0	0	0	0
0	1	0	0	1	0	0	0	0
0	1	0	0	1	0	0	0	0
0	0	1	1	0	0	0	0	0
0	1	0	0	0	0	0	0	1
0	0	0	0	0	1	1	1	0

$p(w|d)$

0.14	0.37	0.33	0.40	0.16	-0.03	-0.07	-0.10	-0.04
0.15	0.51	0.36	0.41	0.24	0.02	0.06	0.09	0.12
0.26	0.84	0.61	0.70	0.39	0.03	0.08	0.12	0.19
0.45	1.23	1.05	1.27	0.56	-0.07	-0.15	-0.21	-0.05
0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
0.22	0.55	0.51	0.63	0.24	-0.07	-0.14	-0.20	-0.11
0.10	0.53	0.23	0.21	0.27	0.14	0.31	0.44	0.42
-0.06	0.23	-0.14	-0.27	0.14	0.24	0.55	0.77	0.66

其他

- 模型相关
  - ✓ SVD、pLSA、LDA、word2vec、Rephil …
- 参数估计
  - ✓ VB、EP、Gibbs、SparseLDA、AliasLDA、LightLDA、超参数优化 …
- 并行训练系统
  - ✓ MR-LDA、PLDA、PLDA+、Yahoo_LDA、Peacock、LightLDA …

主要内容

1	什么是机器学习
2	监督学习：转化预估
3	非监督学习：文本主题建模
4	小结

小结

- 监督学习
  - ✓ 点击、转化预估：二分类
  - ✓ 层次文本分类器：多分类
  - ✓ 相关性：LearningToRank
  - ✓ 反作弊：二分类
  - ✓ Lookalike：PU-Learning（半监督）
- 非监督学习
  - ✓ 文本主题建模：TopicModeling
  - ✓ 推荐系统：MatrixFactorization
  - ✓ 用户、广告聚类：kMeans、MinHash