QCon 全球软件开发大会
INTERNATIONAL SOFTWARE DEVELOPMENT CONFERENCE

BEIJING 2018

深度学习在微博Feed流应用实践

刘博

新浪微博机器学习研发部关系流算法负责人

主办方 Geekbang. 极客邦科技 InfoQ

目录

微博Feed流产品介绍—排序场景

- 微博—社交媒体领跑者
  - DAU：1.72亿，MAU：3.92亿
  - 关注流基于关系链接用户与内容
- 信息获取方式
  - 主动获取（关注）
  - 被动获取（推荐）
- 内容形式
  - 博文/文章/图片/视频/问答/话题/...

微博Feed流特点介绍—排序原因

- 产品特点
  - 传播性强
  - 互动性好
- 存在问题
  - 信息过载
  - 信噪比低
- 排序目标
  - 提高用户的信息消费效率
  - 提升用户黏性

技术挑战

- 指标量化
  - 用户体验
  - 内容形式多样、非结构化
- 规模大
  - 用户和Feed内容数量大
  - 内容更新快，实时性要求高
  - 海量计算、超大规模模型优化

目录

CTR概要介绍

- CTR任务特点
  - 大量离散特征、高维稀疏
  - 特征关联性挖掘
- CTR预估常用算法
  - LR
  - GBDT
  - FM

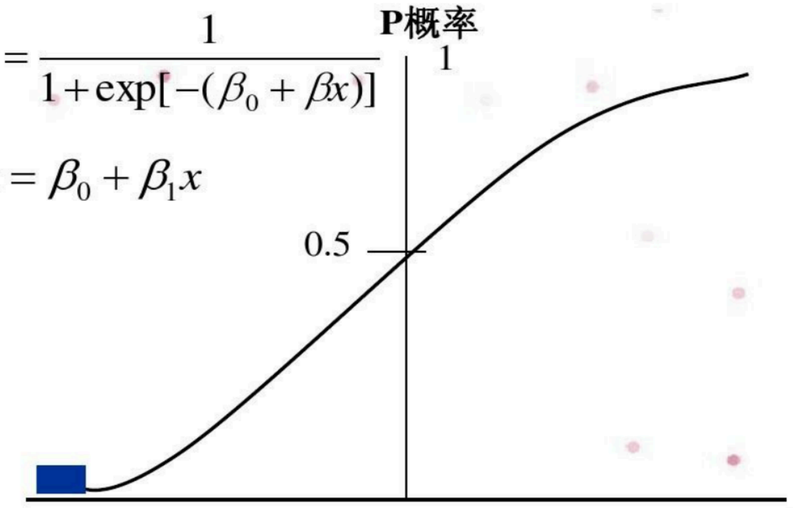
CTR一般流程

业务目标与模型选择

- 模型选择
  - 线性模型LR+特征工程
  - 排序基于pointwise的 learning to rank

$$p(y=1)=\frac{1}{1+\exp[-(\beta_0+\beta x)]}$$

$$z=\beta_0+\beta_1 x$$



- 模型优化目标
  - 互动（转发/评论/赞）
    点击（图片/视频/文章/链接等）
    阅读时长
  - 多目标预估

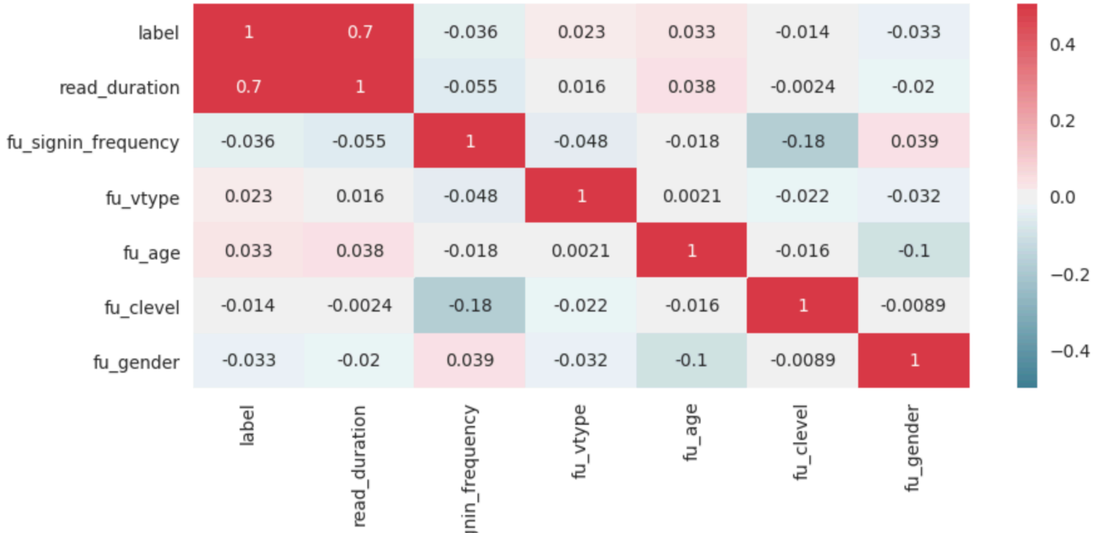
互动模型　点击模型　阅读模型

$$\text{Score}= m_{interact} * ictr + m_{click} * cctr + m_{read} * rctr$$

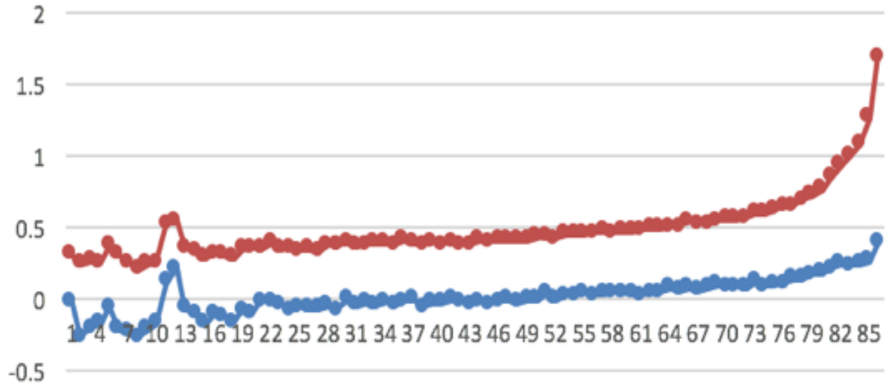
特征工程

- 特征工程非常重要
  - categorical特征
    - one-hot 表示
    - 假设检验方式
  - conitnues特征
    - 离散化/归一化处理
    - 相关系数评估
  - 特征组合
    - 手动组合——专家知识
    - GBDT+互信息——有效挖掘非线性特征及组合

皮尔逊相关系数特征评估



标签匹配度特征相关系数特征评估

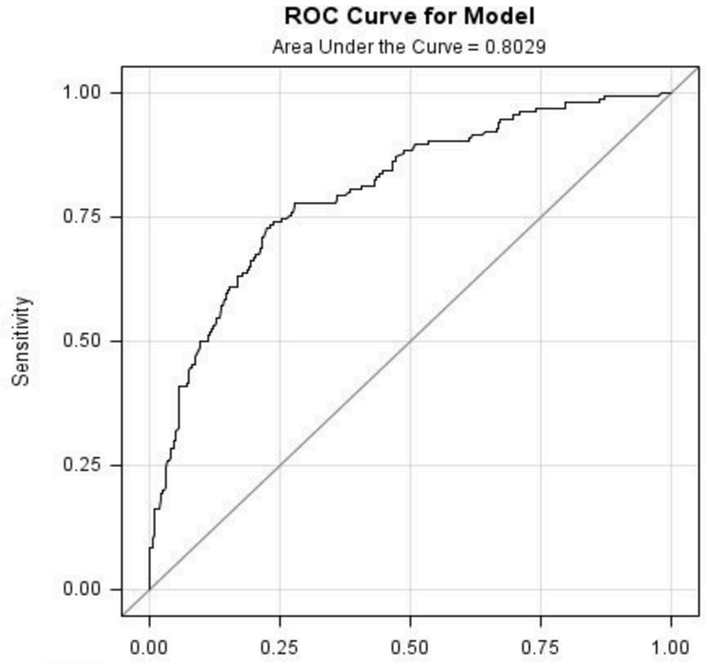


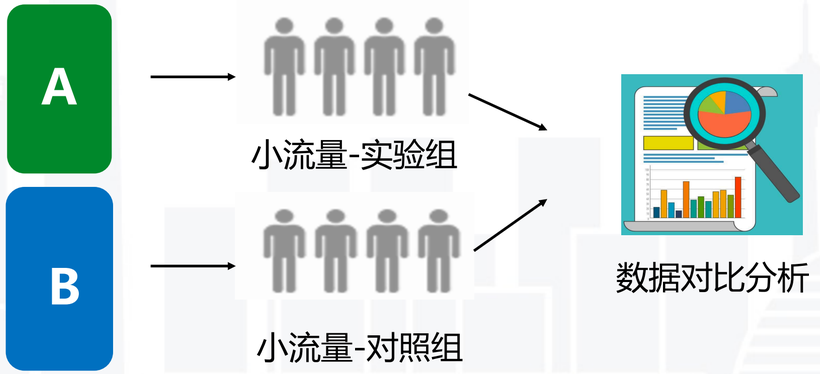
样本采集

- 存在问题
  - 头部效应
  - 正负样本比例严重失衡
  - 实时反馈类收集与在线存在差异性
- 解决方案
  - 对头部曝光进行降采样，长尾曝光上采样
  - 负样本进行下采样
  - 后端样本预采样

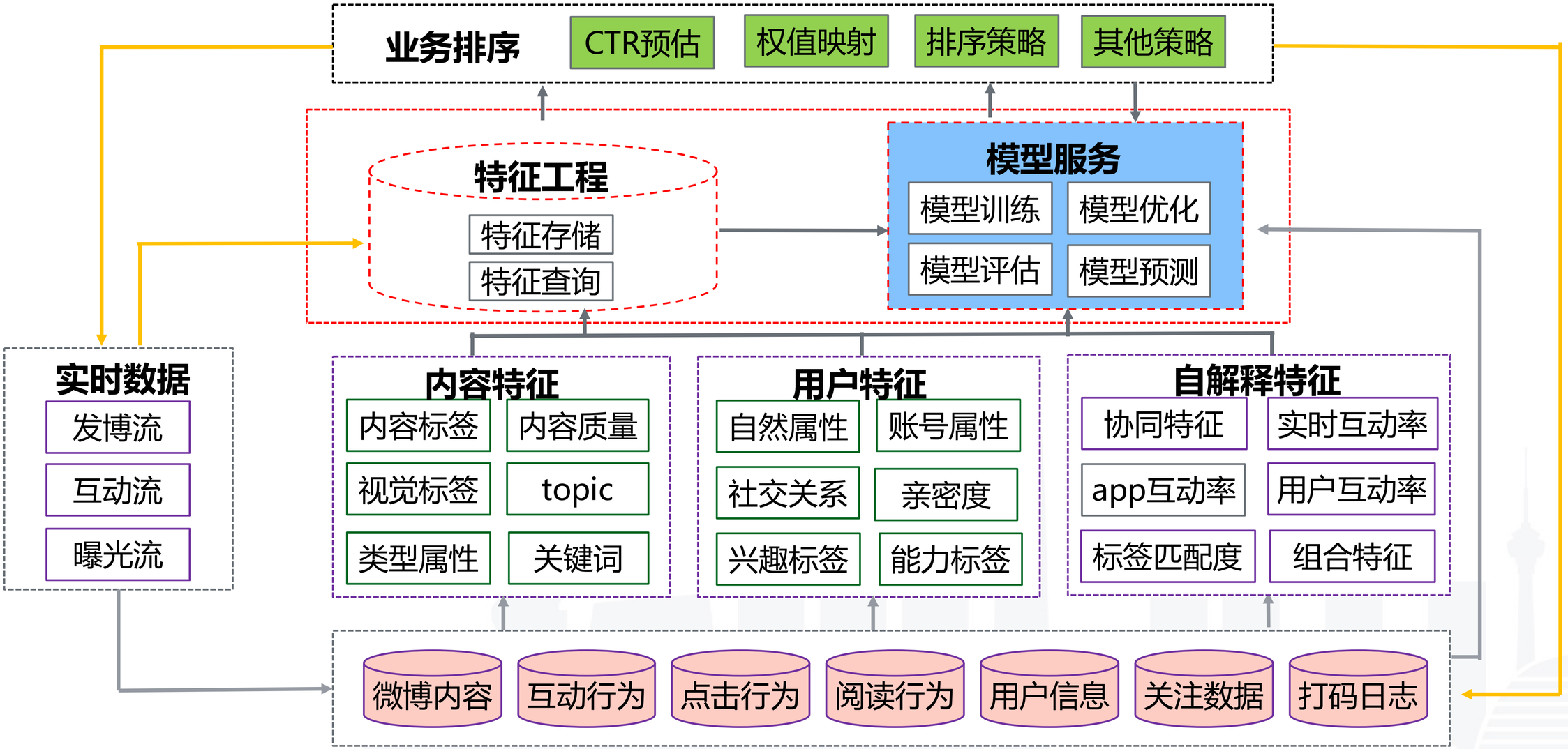
模型评估

- 离线评估
  - AUC / wAUC
  - 离线评估与线上效果正相关？
- 在线评估
  - A/B test测试
  - 分目标人群测试：地域、活跃度...





算法架构

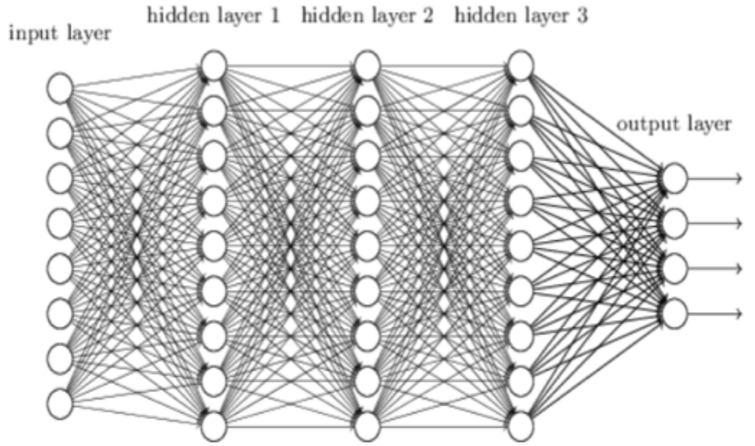

业务排序
CTR预估
权值映射
排序策略
其他策略
特征工程
特征存储
特征查询
模型服务
模型训练
模型优化
模型评估
模型预测
实时数据
发博流
互动流
曝光流
内容特征
内容标签
内容质量
视觉标签
topic
类型属性
关键词
用户特征
自然属性
账号属性
社交关系
亲密度
兴趣标签
能力标签
自解释特征
协同特征
实时互动率
app互动率
用户互动率
标签匹配度
组合特征
微博内容
互动行为
点击行为
阅读行为
用户信息
关注数据
打码日志

目录

为什么选择深度学习

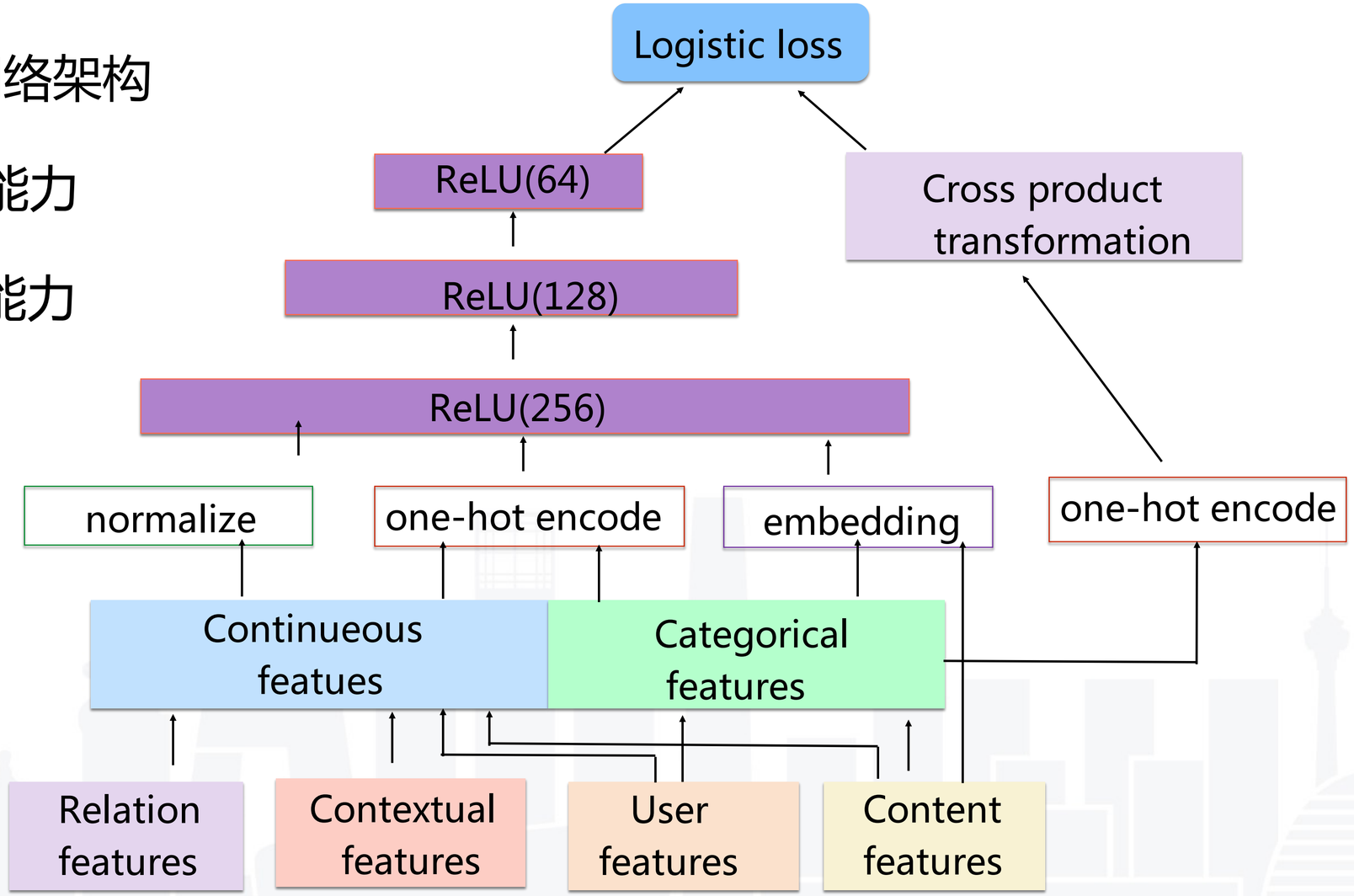
- 线性CTR模型
  - 优势：简单高效、可解释性强
  - 局限性：特征工程繁琐、无法表达高维抽象特征

- 深度学习模型（DNN based model）
  - 优势：
    - 表达能力强
    - 泛化能力强
    - 网络结构灵活



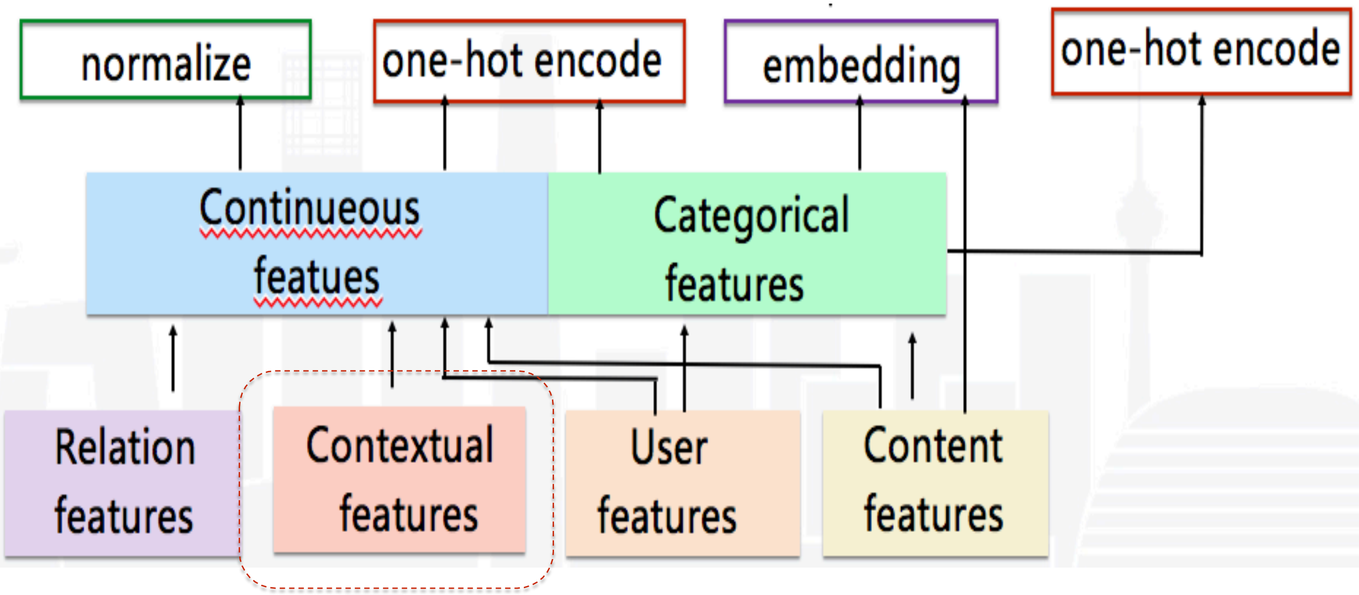
深度学习应用实践 —— wide & deep

- Wide & deep 网络架构
  - Deep—泛化能力
  - Wide—记忆能力



深度学习应用实践 —— wide & deep

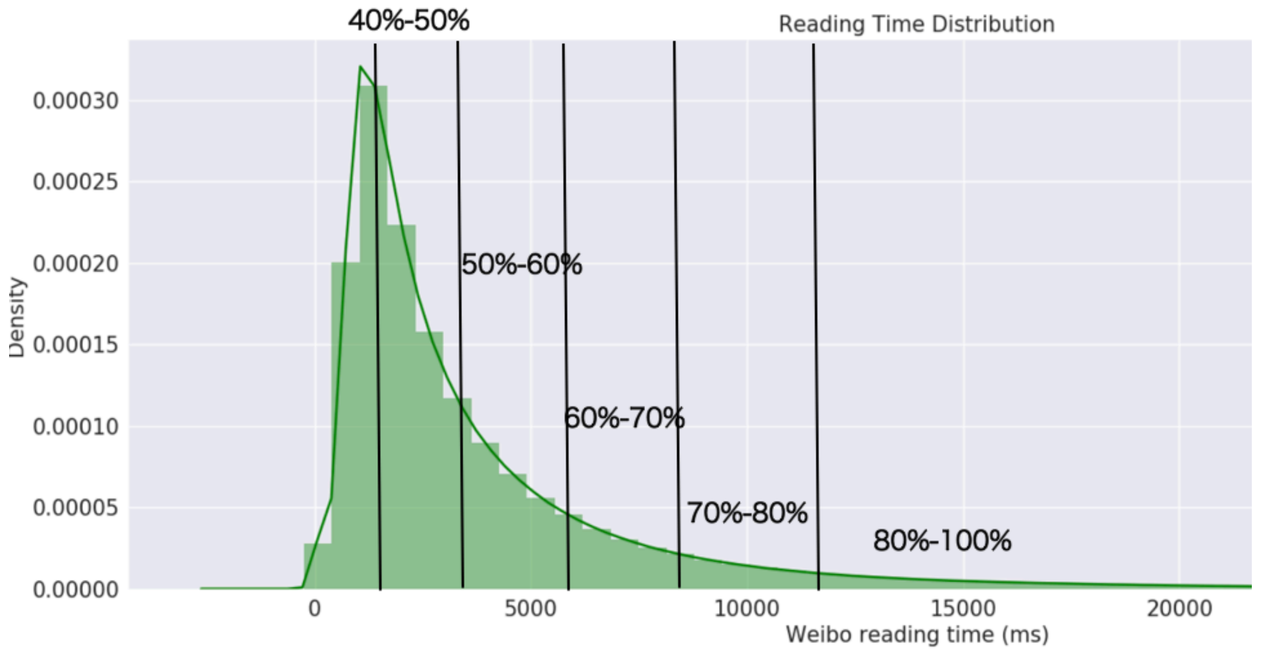
- 新增特征
  - Contextual featues：用户最近的平均阅读时长、用户最近的互动微博



- Deep部分依然需要特征工程
  - Wide输入
    - conitnues特征离散化 + 手动交叉特征
  - Deep输入
    - conitnues特征离散化 + 非连续特征embedding

深度学习应用实践 —— wide & deep

- 样本采样
  - Negative sampling：依据微博的平均阅读时间进行划分，将用户曝光但未阅读的微博作为负样本

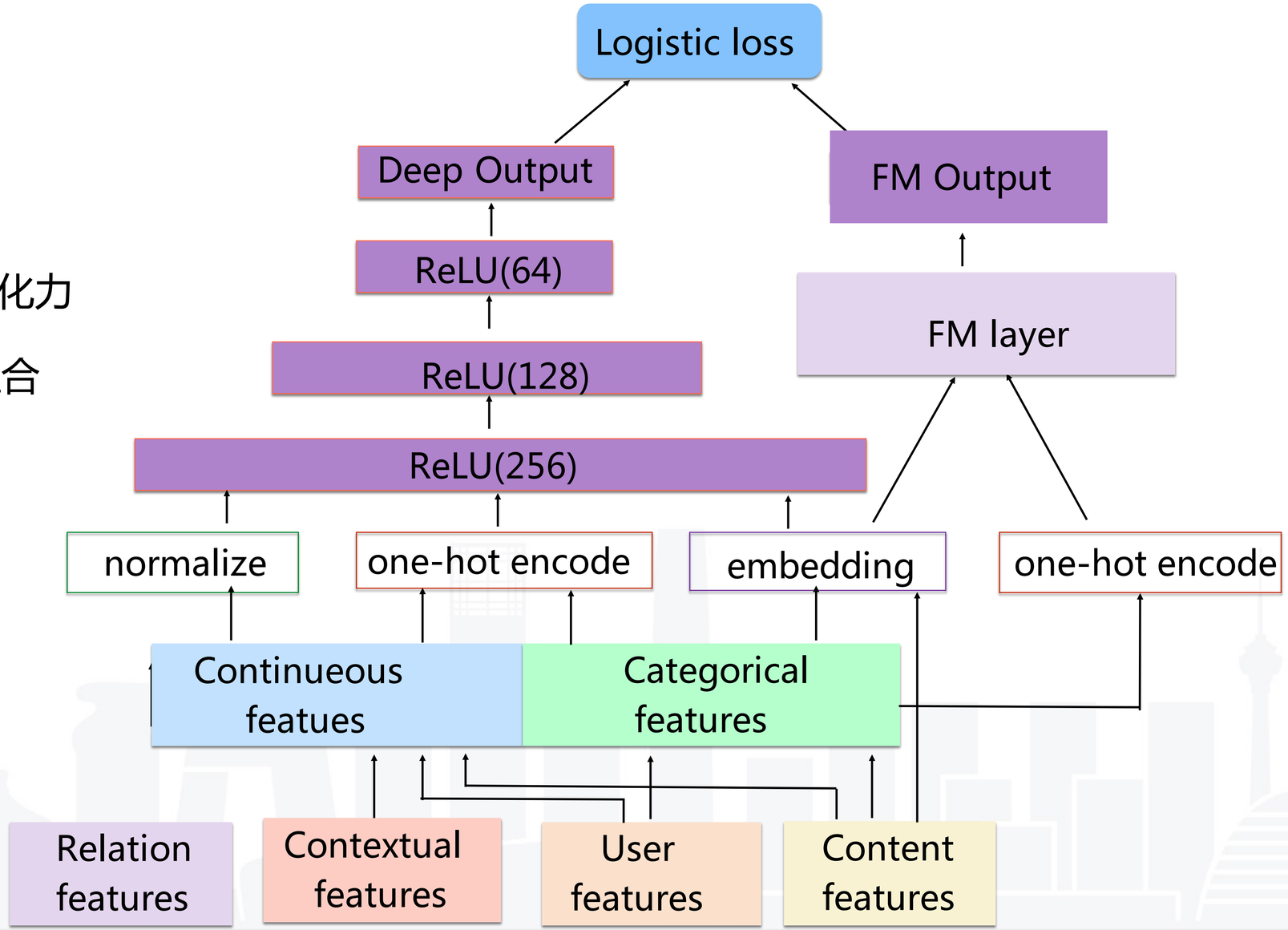


- 网络复杂度
  - 网络复杂度过高易导致过拟合
  - 网络深度达到一定数值AUC反而小幅降低

网络结构	logloss	wAUC
[1024,512,256]	0.049	0.743
[512,256,128]	0.043	0.753
[256,128,64]	0.039	0.761
[256,128]	0.045	0.749
[128,64]	0.053	0.741

深度学习应用实践 —— DeepFM

- DeepFM模型架构
  - End2End框架
  - Deep part — 泛化力
  - FM—低阶特征组合
- 优势
  - Deep和FM共享embedding层



总结展望

- 模型算法是手段
  - 业务和数据决定模型算法的应用场景
  - 模型算法殊途同归
  - 计算力和算法架构是保障

总结展望

- 未来工作
  - 多模态—更好的对非结构化内容进行表征
  - 用户行为序列embedding
  - 更多的融合网络结构适用于CTR预估场景

关注QCon微信公众号，
获得更多干货！

Thanks!

INTERNATIONAL SOFTWARE DEVELOPMENT CONFERENCE

主办方 Geekbang. 极客邦科技 InfoQ