User和Item的特征提取方式

挖掘最有价值的信息辅助模型训练

讲师: 王老师

开篇

本讲我们主要讨论一些特征和特征工程的事情.

1、item特征
2、user特征
3、交叉特征
4、Embedding特征
5、特征工程

item特征

item 一般指一个需要打分的个体.
搜索中, 是等待被搜索的个体;
推荐中, 是等待被推荐的个体.

但是在实际的业务中, 却往往不是那么简单.
同一个 item 在具有不同属性的时候, 到底是不是同一个 item?

item特征

在 item 中, 我们都用过什么特征?

编码方式

自然数编码、独热编码、count编码（替代类别特征）、目标编码

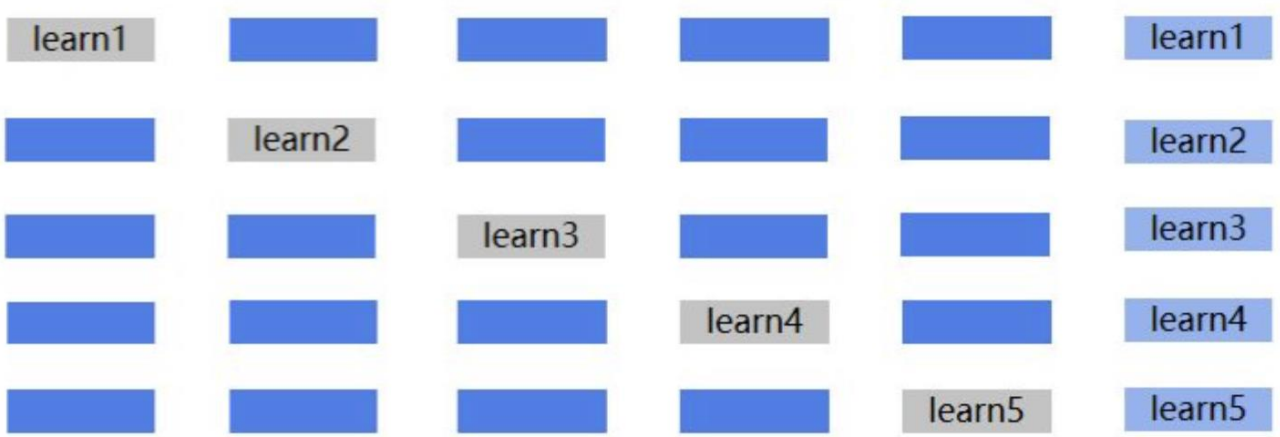
统计方式

count、nunique（宽度）、ratio（偏好）

item特征

目标编码（target encoding）

只要和目标相关的都可以进行进行编码，分类和回归均可构造，推荐广告中ctr和cvr特征。具体两种构建方式

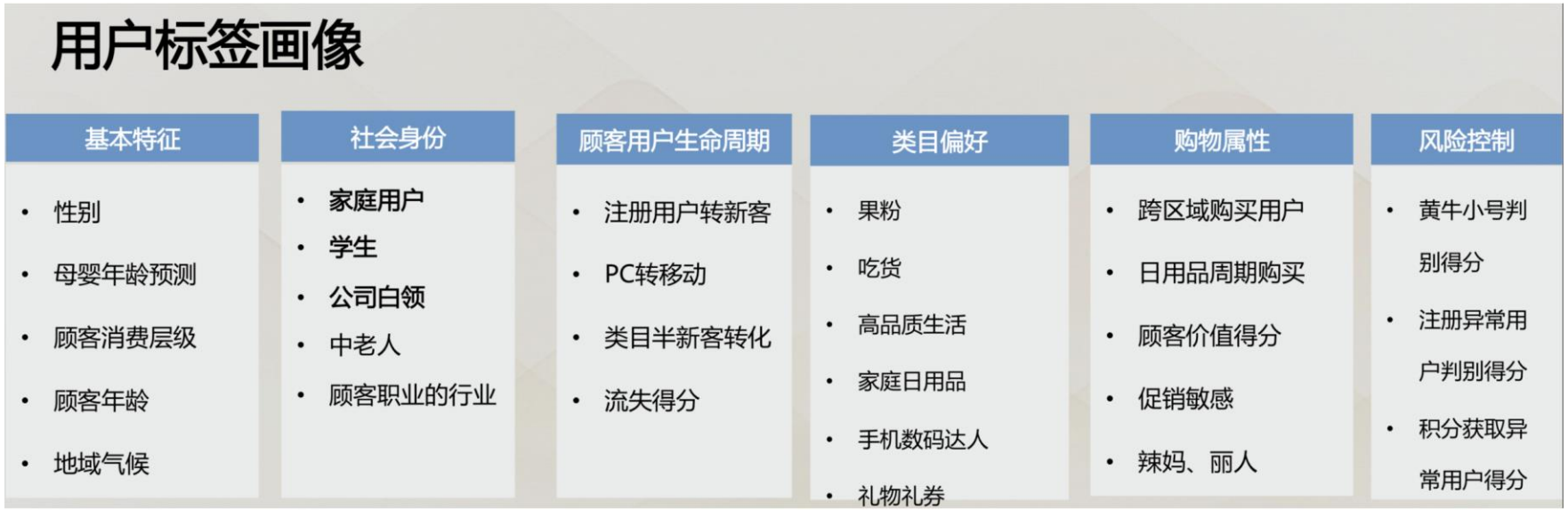


离散特征和连续特征

- ⊙ 为什么会有这个区分?
- ⊙ 离散特征和连续特征的特点
- ⊙ 离散和连续的相互转化

用户特征-用户画像

如果想到要对用户建模, 我们一般都能想到什么特征?



用户特征-特征挖掘

如何捕捉到真正有用的用户特征, 其实是和业务以及建模问题强相关的.

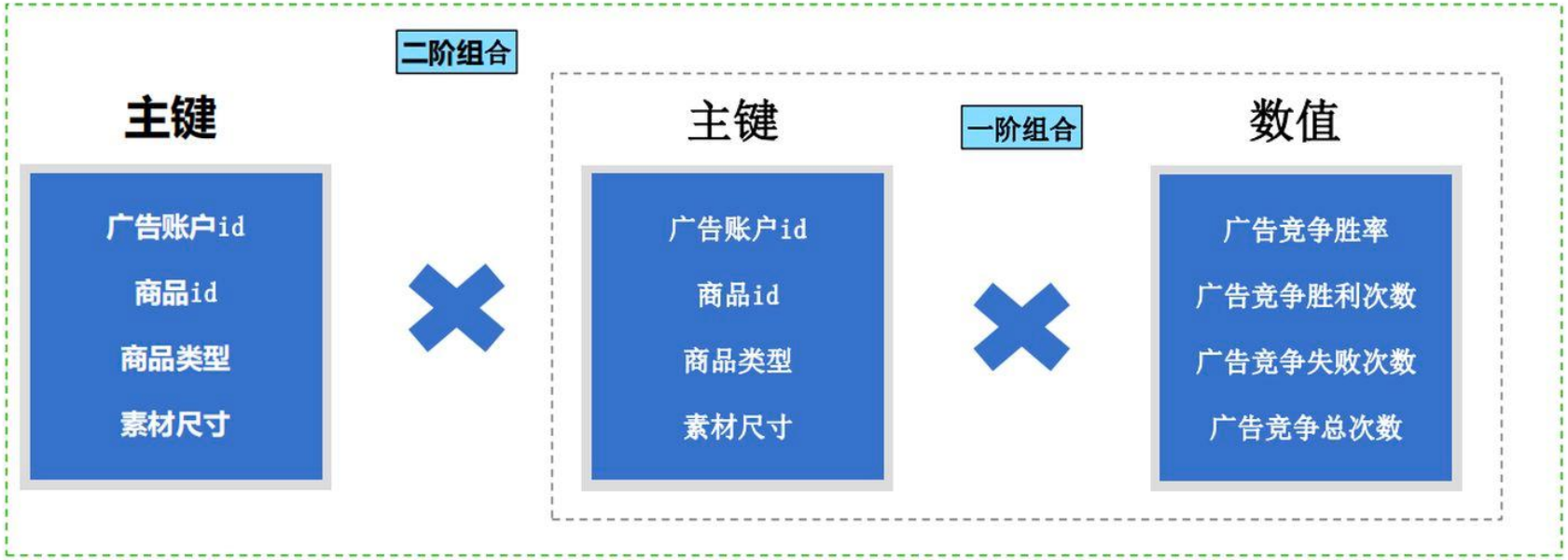
案例学习

用户停留时长对于广告建模. 如何挖掘日志和埋点,
是所有算法工程师面临的问题.

交叉特征

交叉方式：类别+类别，类别+连续，连续+连续

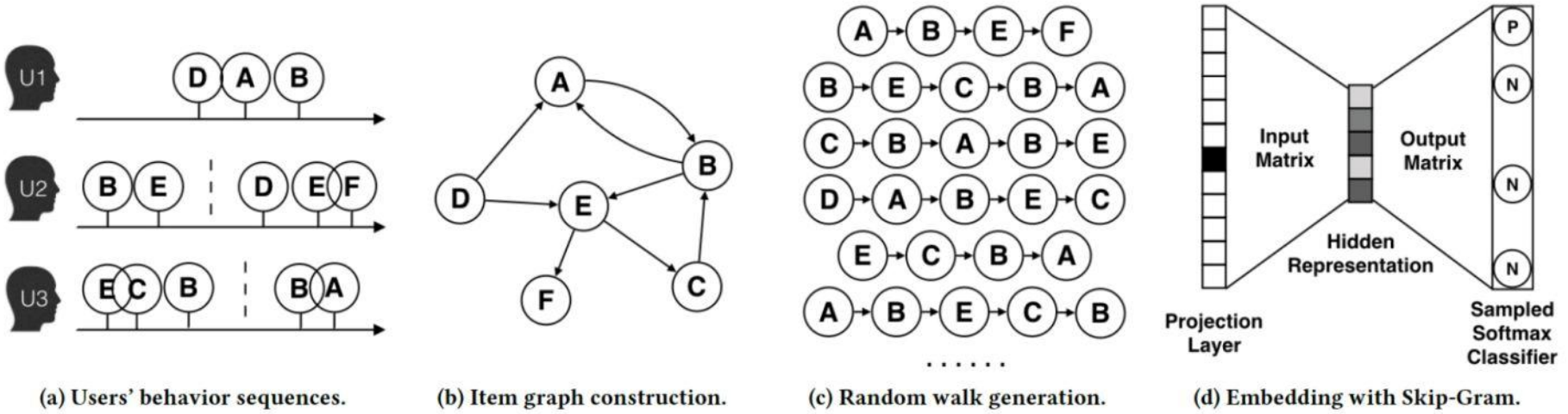
业务组合：用户侧+商品侧，用户侧+用户侧，商品侧+商品侧



Embedding特征

什么是Embedding?

我相信大家, 很多人听说 embedding, 和我一样, 都是从 word2vec 听说的. 那么, 究竟什么才是 embedding?



(a) Users' behavior sequences. (b) Item graph construction. (c) Random walk generation. (d) Embedding with Skip-Gram.

Distributed Representation

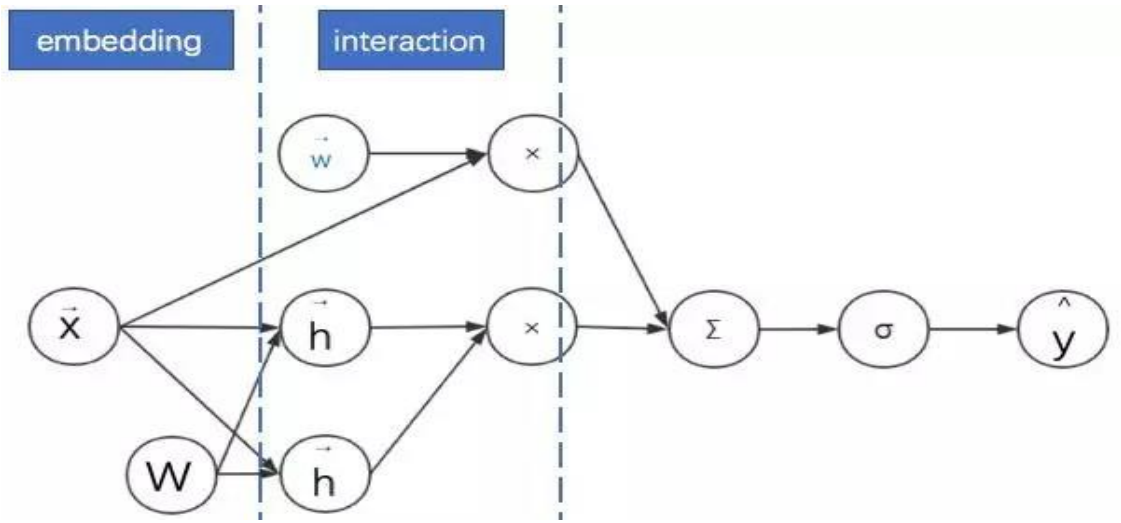
OneHot 无法考虑到不同维度的关系.
◎ 如国王-男 = 女王-女
◎ 如 1 和 2 都是数字

Embedding 是一个将离散变量转为连续向量表示的一个方式，具体怎么转化?

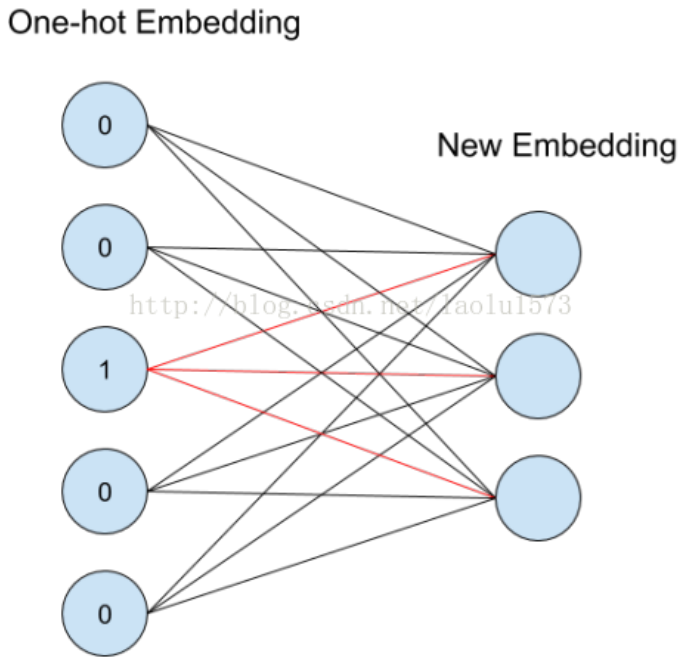
从神经网络的角度看 FM

$$y = \sigma(< w, x > + < W \cdot x, W \cdot x >)$$

- FM 首先是对离散特征进行嵌入，也叫做 embedding
- 之后通过对嵌入后的稠密向量进行内积来进行二阶特征组合
- 最后和线性部分结合



从神经网络的角度看 FM

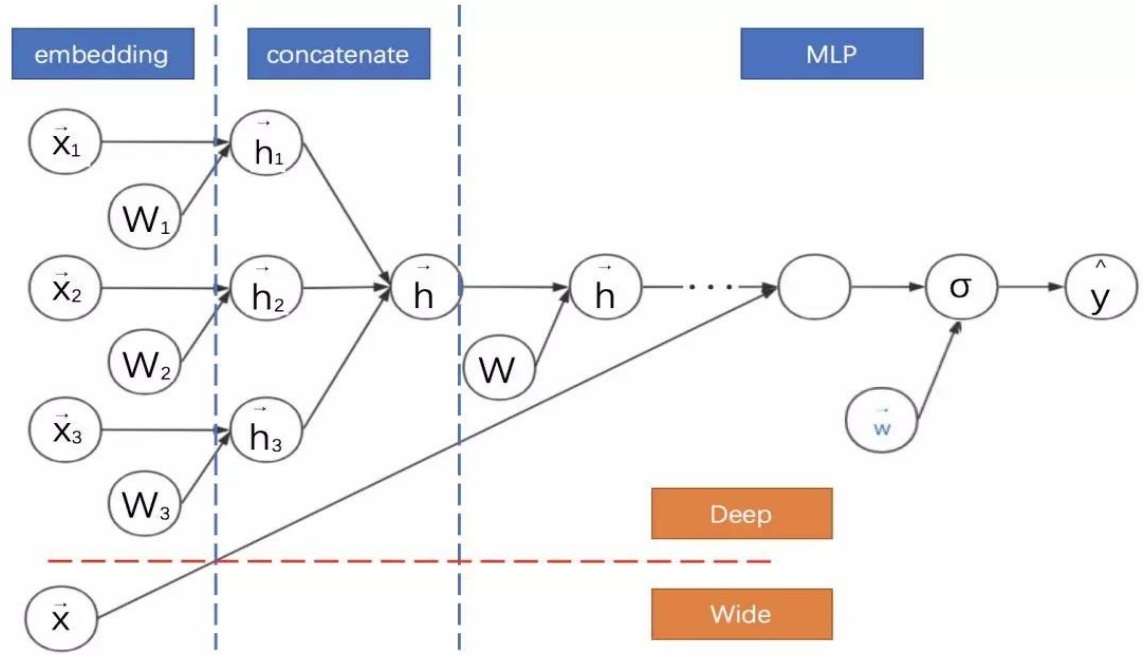


$$[0 \quad 0 \quad 0 \quad 1 \quad 0] \times \begin{bmatrix} 17 & 24 & 1 \\ 23 & 5 & 7 \\ 4 & 6 & 13 \\ 10 & 12 & 19 \\ 11 & 18 & 25 \end{bmatrix} = [10 \quad 12 \quad 19]$$

Wide and Deep

Wide and Deep 由谷歌提出, 采用神经网络联合训练的思路, 对 神经网络进行并联.

- Deep 部分是 MLP, 而且是 dense 特征的 MLP.
- Wide 部分是直接的 LR.
- 如果 Wide 部分采用了 FM, 就变成了 DeepFM.



Wide and Deep Revisit

为什么 Wide and Deep 是好的?

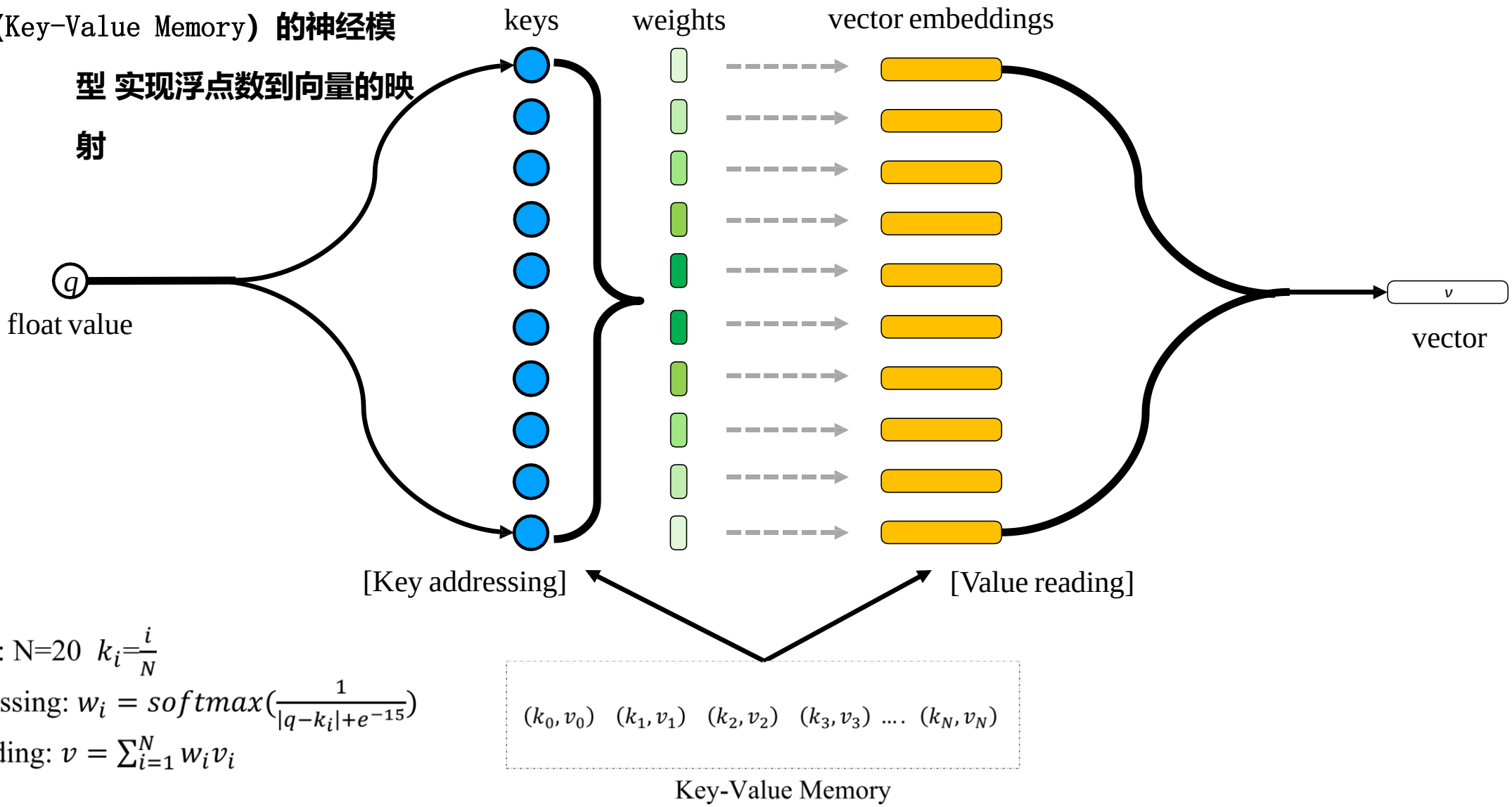
- 分开学习 wide 和 deep 部分
- 同时获得 **记忆性** 和 **泛华性** 的信息
- 模型简单效果好, 易于扩展

什么是Embedding?

- 从数学上看，是映射 $f : X \to Y$
- 从神经网络的角度看，是层与层之间的矩阵
- 从特征的角度看，是从一套特征映射到另一种表示方式

Key-Value Memory

键值存储（Key-Value Memory）的神经模型 实现浮点数到向量的映射



Parameter: N=20 $k_i = \frac{i}{N}$

Key addressing: $w_i = softmax(\frac{1}{|q-k_i|+e^{-15}})$

Value reading: $v = \sum_{i=1}^{N} w_i v_i$

怎么得到Embedding

- 使用 word2vec 预训练, node2vec
- 使用 FM 预训练
- 深度学习的 supervised learning

推荐系统中主要目的:

- 在 embedding 空间中查找最近邻，这可以很好的用于根据用户的兴趣来进行推荐。
- 作为监督性学习任务的输入。

特征工程

为什么要进行特征工程?

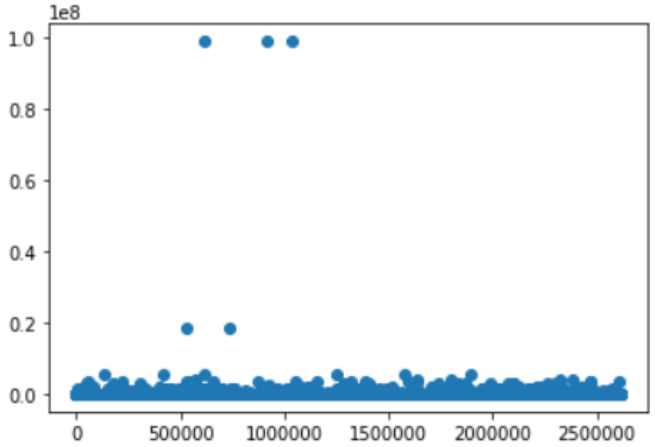
- 简单模型 + 复杂特征
- 复杂模型 + 简单特征

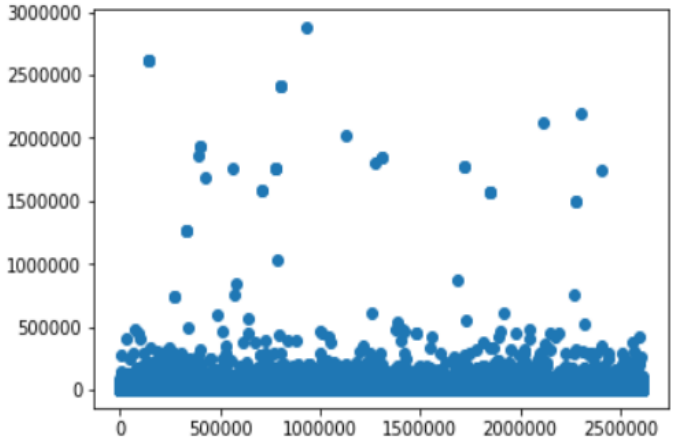
特征工程-线性模型

- 二阶交叉和高阶交叉
- 单变量的非线性变换
- 特征预处理和归一化 (梯度)

特征工程-预处理

离群点处理





处理方法

当作缺失值进行处理

删掉离群点所在样本

使用统计值进行填充

特征工程-预处理

缺失值处理

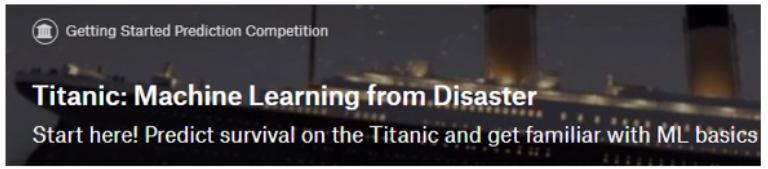
真正（意义）缺失?

否

有特定的业务含义!

- **填充max(fea)+1/min(fea)-1**

是



填充?

- **各种填充方案**
- **不填充,设为np.nan**
- **对比效果选择**

特征工程-特征选择

过滤法

相关系数 + 卡方检验 + 互信息

封装法

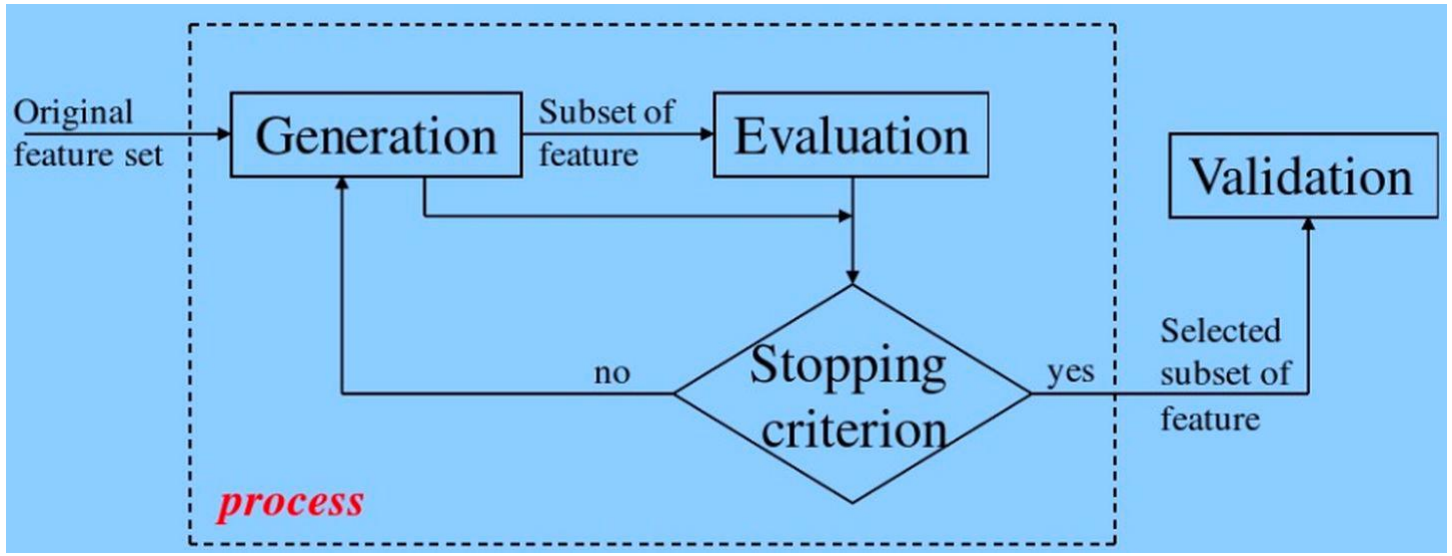
前向搜索 + 后向搜索

嵌入法

基于学习模型的特征排序

特征工程-特征选择

实际工作 覆盖率 + 信息增益（比） + xgboost



特征工程-特征降维

- PCA
- SVD
- LDA

```
from sklearn.decomposition import NMF, PCA, TruncatedSVD
decom_feature = []

for i, decom_func in enumerate([TruncatedSVD, NMF, PCA]):
    x = data[feats].values
    decom = decom_func(n_components=8, random_state=1024)
    decom_x = decom.fit_transform(x)
    decom_feas = pd.DataFrame(decom_x)
    decom_feas.columns = ['v_{}'.format(i) for i in range(i*8,i*8+8)]
    decom_feature += ['v_{}'.format(i) for i in range(i*8,i*8+8)]

    data = pd.concat([data, decom_feas], axis=1)
```

特征工程-离散化

怎么进行特征离散化?

一般有手动分桶和自动分桶两种方法.

- 手动分桶: 统计每个组的情况
- 自动分桶: GBDT+LR

特征工程-GBDT+LR

先在样本集上训练一个 GBDT 的树模型, 然后使用这个树模型对特征进行编码, 将原始特征 x对应的叶子节点按照 01 编码, 作为新的特征, 叠加到 LR 模型里再训练一个 LR 模型.

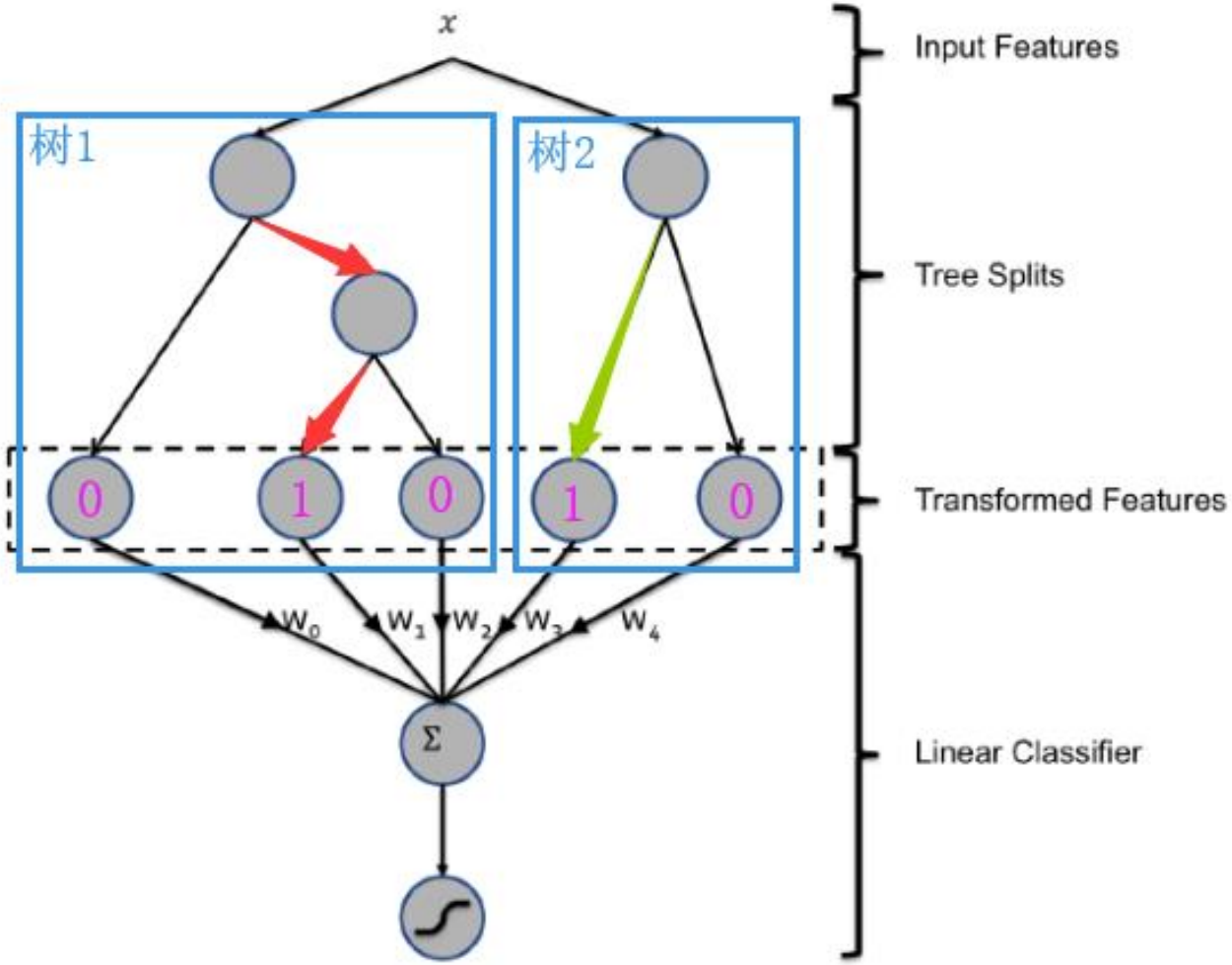
为什么这样做是有效的?

因为 GBDT 是在函数空间对残差进行连续的逼近, 精度很高, 但是容易过拟合; 在进行裁剪后, 利用叶子节点编码, 有效的把连续特征离散化, 因此适合 LR.

GBDT+LR

2004 年Facebook 在论文 Practical Lessons from Predicting Clicks on Ads at Facebook 中提出的 GBDT + LR 模型给出了一个可行的解决方案。

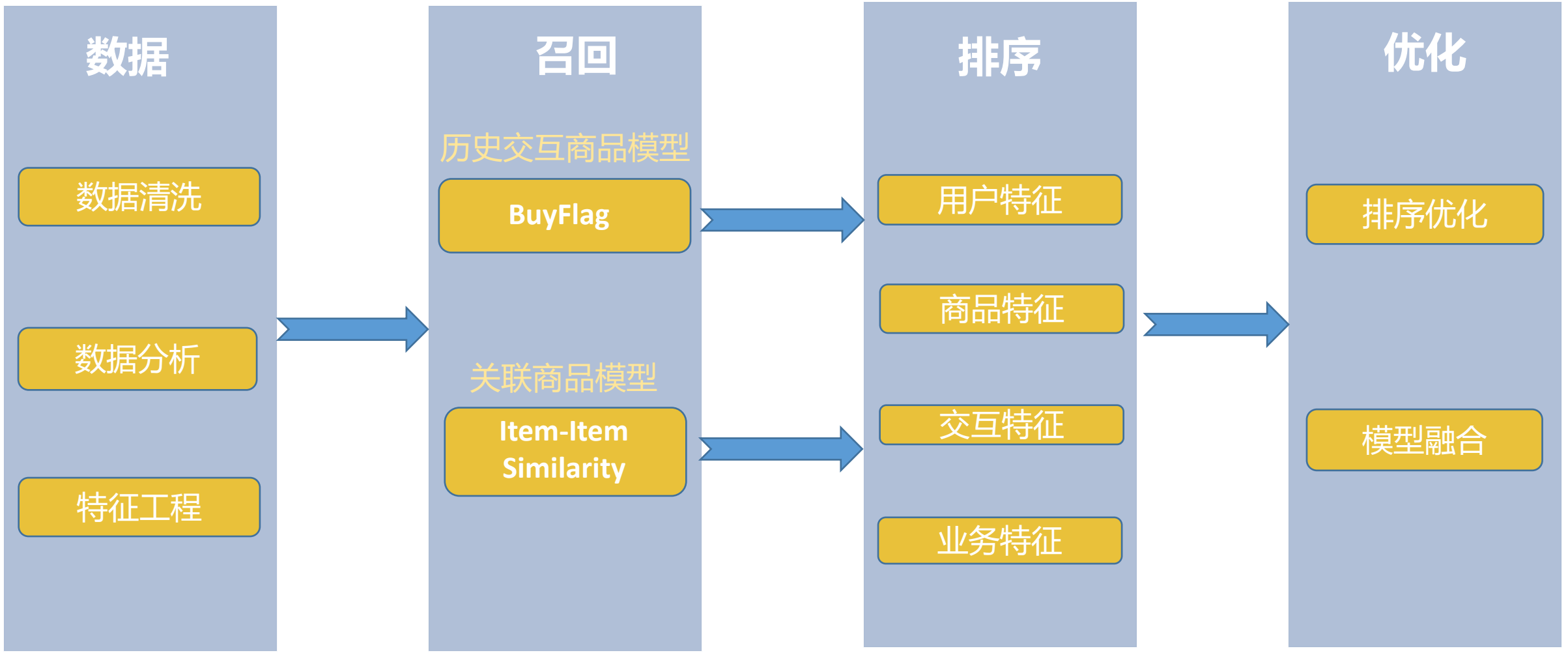
- GBDT构建特征，LR预估CTR
- 深度决定特征交叉阶数
- 特征工程模型化，模型的输入可以是原始的特征向量，实现端到端训练。



特征实战


2019
安泰杯
跨境电商智能算法大赛
AliExpress消费行为预测
联合主办
上海交通大学安泰经济与管理学院
阿里巴巴集团AliExpress
天池平台
立即报名

特征实战

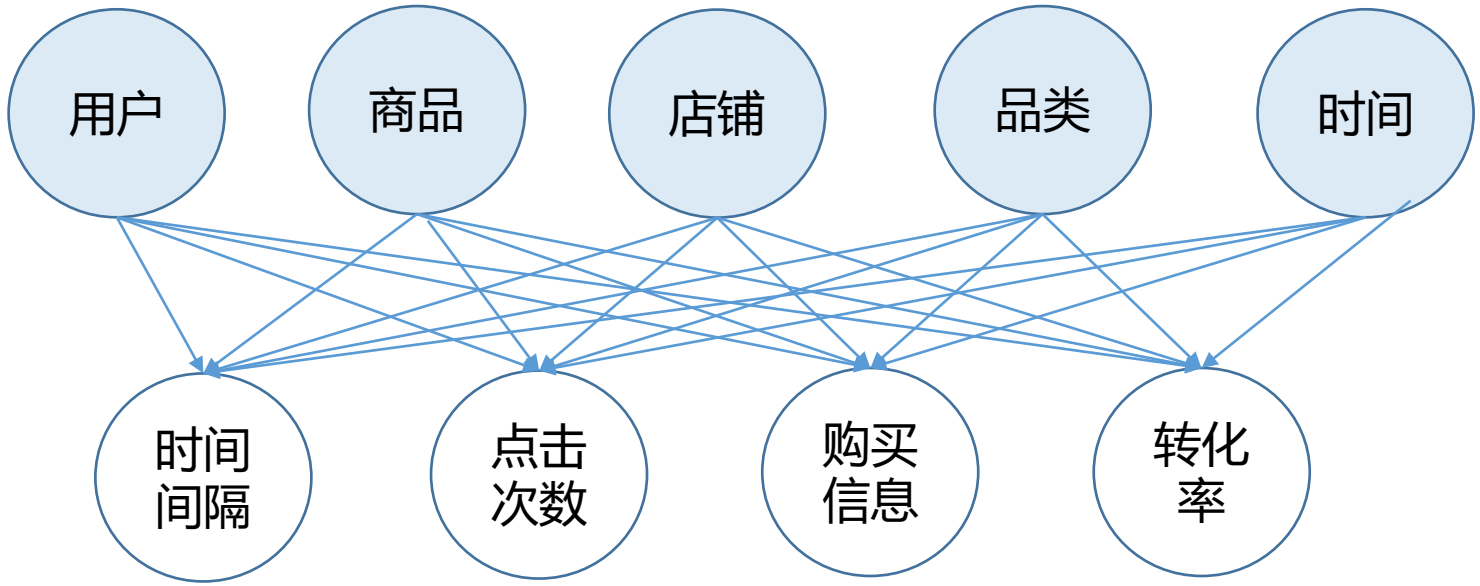

数据
数据清洗
数据分析
特征工程
召回
历史交互商品模型
BuyFlag
关联商品模型
Item-Item Similarity
排序
用户特征
商品特征
交互特征
业务特征
优化
排序优化
模型融合

特征实战

根据零售行业的人货场概念，赛题提供了关于用户行为日志的常见字段可分为如下部分：

用户：用户标识、用户国籍
商品：商品标识、店铺、品类、价格
场景：点击时间、访问排序、购买标记

特征实战



最大值　最小值　方差　均值　比率　排序

特征实战

通过对以上维度的交叉统计，形成高阶特征群，提取出购物决策的关键信息

用户画像

- 用户活跃度
- 用户品类偏好
- 用户店铺偏好

商品画像

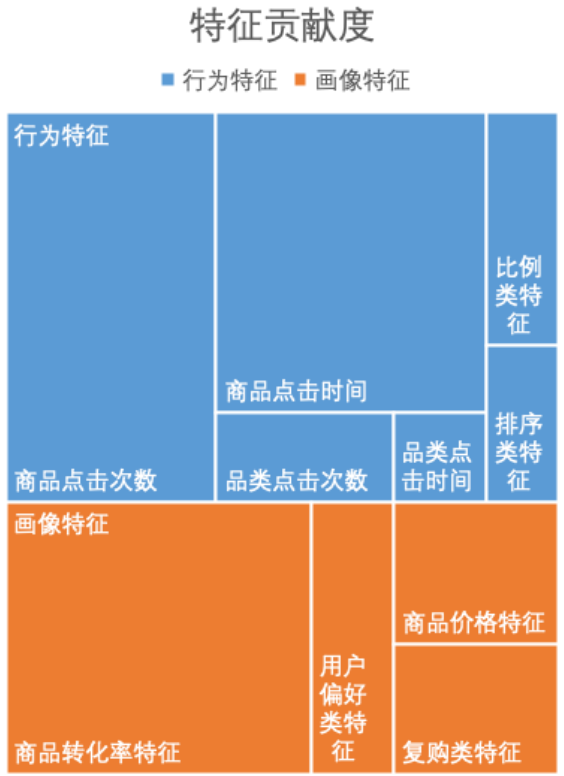
- 商品销量
- 商品转化率
- 商品热度

行为明细

- 用户点击次数
- 用户点击频率

业务知识

- 商品上架时间
- 促销节点
- 复购产品
- ....



本周作业

1、实战安泰杯竞赛召回模块
要求：关联召回+embedding召回

2、实战安泰杯仅是特征模块
要求：原有基础上添加CTR相关特征、embedding相关特征等

参考：https://github.com/RainFung/Tianchi-AntaiCup-International-E-commerce-Artificial-Intelligence-Challenge

感谢聆听 | Q&A