Comments (22)

【方案调研】搜广推场景下的特征工程

导读：在搜推广领域中，特征工程对业务效果具有很大的影响，并且占据了算法工程师很多精力。正如业界经典的话所说“Garbage in, garbage out”，数据和特征决定了机器学习算法的上限，模型、算法的选择和优化只是在不断逼近这个上限，而特征工程就是数据与算法之间的桥梁。特征工程的前提是收集足够多的数据，使用数据学习知识，从大量的原始数据中提取关键信息并表示为模型所需要的形式。本文调研了阿里、腾讯、美团等公司在CVR模型、CTR模型工作中特征工程的相关做法和方案，主要分为4部分内容：

- 为什么要精做特征工程
- 什么是好的特征工程
- 各厂搜广推场景下的特征工程经验
- 特征工程的常见问题及处理

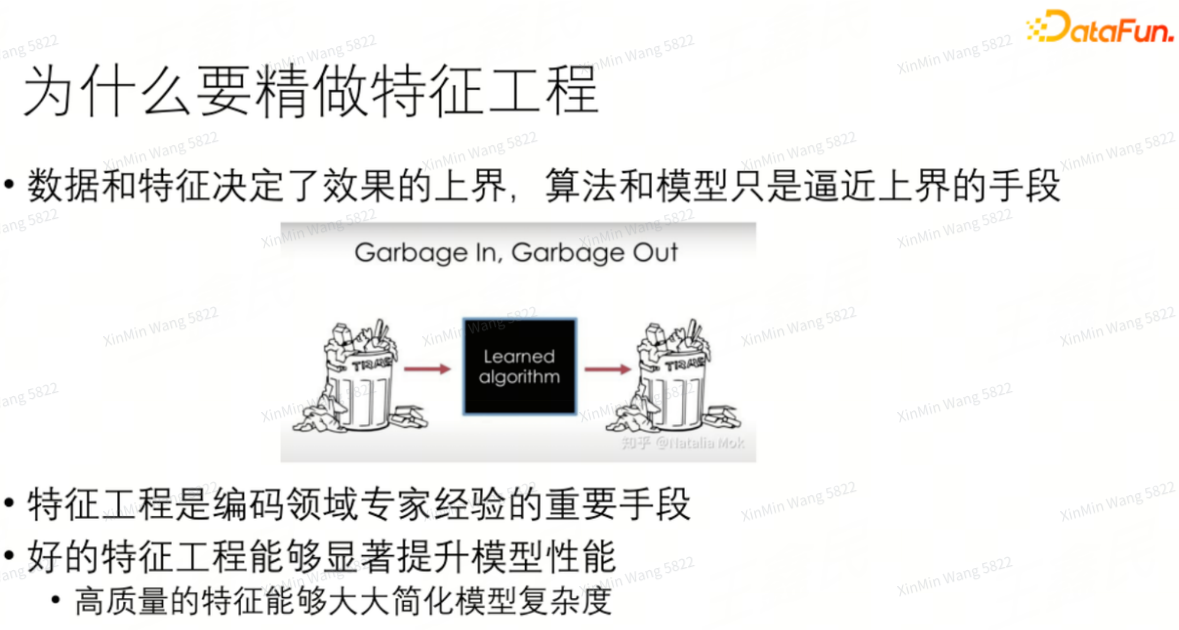
1. 为什么要精做特征工程

1.1 精做特征工程的原因

特征工程就是对原始数据进行一系列工程处理，将其提炼为特征，作为输入供算法和模型使用。在完整的机器学习流水线中，特征工程不仅要考虑特征的处理，同时也需要考虑样本的处理，即选择哪部分样本输入模型，供模型训练使用。

推荐算法中的特征工程一文中详细说明了为什么要细致化地进行特征工程的工作，主要原因也是机器学习领域的经典共识**“Garbage In, Garbage Out”**，数据和特征决定了模型机器学习算法的上线，模型和算法只是不断去逼近这个上限，只有收集足够多的数据，构造足够完善的特征工程，并从中提取关键信息，才能构造一个优秀的模型。**特征工程处于机器学习流水线的上游位置，处理结果的好坏关系到后续模型的效果。**

许多大厂的经验表明，**特征工程往往占据算法工作80～90%的时间，特征工程是最能显著提升模型性能的工作，同时，也可以大大简化模型复杂度，降低模型的维护成本，高质量的特征在简单的线性模型上也能表现出不错的效果。**



对于传音CVR模型当前现状：

1. **离线实验中，特征工程初版优化可以带来AUC&打分准确性的提升**

就是对原始数据进行一系列工程处理，将…

XinMin Wang Jan 29, 2023

这里的“一系列工程处理提炼为特征的工作”，即“通过一系列变换映射到新的向量空间”，并使算法模型能够在新的向量空间更好的学习数据中表达的信息/规律。

2. **随着业务的发展&CVR模型性能提升需求，特征工程优化工作势在必行。**特征工程决定效果上限，后期效果类/转化类广告主不断加入，同时也面临细粒度转化成本达标问题，当前特征工程可能无法满足CVR模型长期迭代更新（保证转化成本）的需求，因此CVR模型特征工程优化非常有必要。

1.2 关于特征工程的3个误区

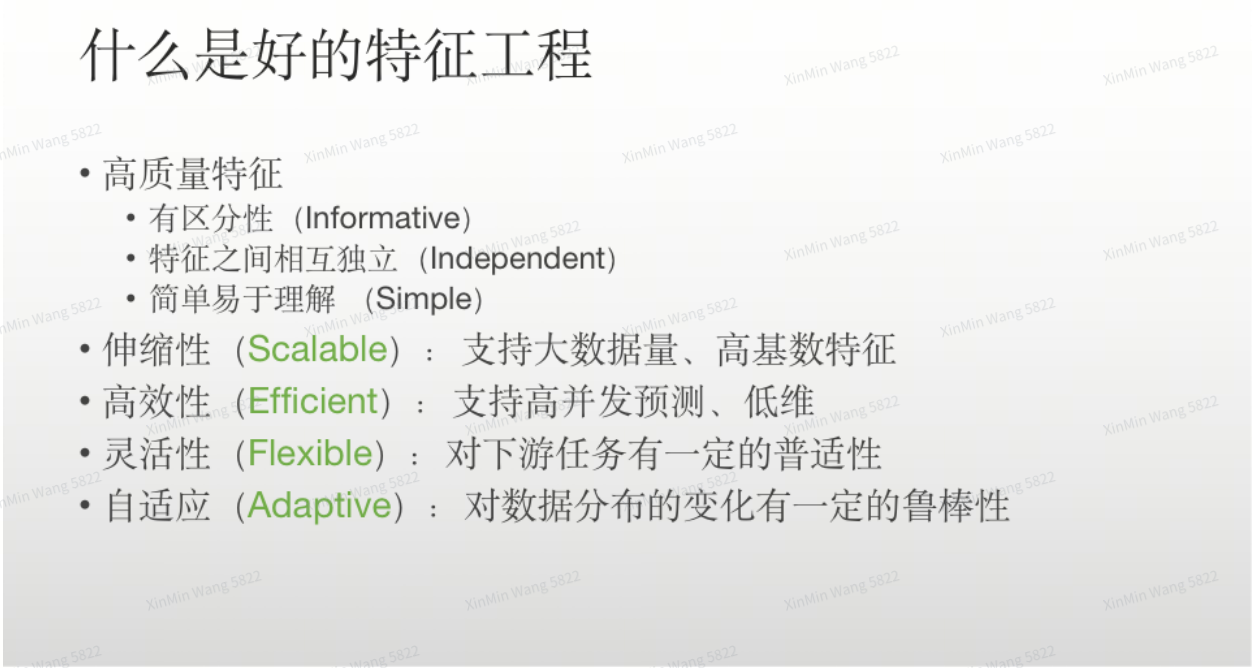
误区点	说明
1. **深度学习不需要精细化的特征工程**	1. cv、nlp场景中end-to-end的学习方式，使得在这些领域手工做特征工程的重要性大大降低 2. 搜广推场景下的特征数据主要以关系型结构组织和存储，关系型数据上的特征生成和变换操作主要有两大类型：基于行的特征变换（row-based）和基于列的特征变换（column-based） 3. 深度学习模型在一定程度上可以学习到row-based的特征变，但是很难学习到column-based的特征变换 4. 特征工程可以高效提炼有价值的特征，降低模型学习的难度。
2. **有了AutoFE工具就不需要手工做特征工程**	1. AutoFE的研究尚处于初级阶段，在使用效率上的问题还没有得到很好的解决 2. 特征工程非常依赖于数据科学家的业务知识、直觉和经验，通常带有一定的创造性和艺术性
3. **特征工程没有技术含量**	1. 算法迭代更新速度太快，总会有效率更高、效果更好的模型被提出 2. 特征工程的沉淀是会使我们成为业务领域的专家，对业务贡献无可替代的价值

但是很难学习到column-based的特征变换

XinMin Wang Jan 29, 2023

深度模型一次只能接受一个小批次的样本，无法建模到全局的统计聚合信息

2. 什么是好的特征工程



特征是对某个行为过程相关信息的抽象表达，构建特征工程的原则是：尽可能地让特征工程抽取出一组特征，能够保留用户行为过程和场景中所有“有用”信息，并摒弃“冗余”信息（**有用，避免冗余，简单易于理解**）。

高质量的特征应具**有区分度，相互独立，简单可解释性强**等特点，对**所有的机器学习任务都适用**。

在搜广推领域，我们对特征工程有一些其他要求。首先特征工程的伸缩性要强，支持高基数特征，支持大数据场景下的任务；其次特征的设计还要让模型在线预测的时候支持高

并发度，预测效率高；并且特征的设计要具有灵活性，一个好的特征工程应该适用于多个模型任务；最后也是最重要的，特征工程要对数据分布变化有一定的鲁棒性，因为在真实的场景中，数据分布变化在某种程度上来说是难以避免的，比如电商场景下经常会有些大促活动，这些活动的举办就会对数据分布产生影响，能够适应这种变化的特征才称之为一个好特征。

3. 各厂搜广推场景下的特征工程

引言：特征工程涵盖的内容和工作量庞大，各厂在不同场景下特征工程部分进行了大量工作及相关探索，本章节主要聚焦各厂在特征工程部分“如何设计此部分架构，特征构造时可以选择哪些特征”进行调研总结（即从特征通用性&业务场景特性出发选取特征），以期为传音广告场景下的CTR&CVR模型特征工程设计中特征构造的工作提供指导。

3.1 总述

结合对阿里、腾讯、美团等大厂特征工程方案的调研，使用按照特征主体划分的特征工程分类体系构造特征适合传音广告场景的cvr模型特征工程，条理清晰且便于尤其迭代。从特征主体划分，特征主要分为 User 特征、Item 特征、Scene 特征、交叉特征。总结如下（**加粗表示传音广告场景下规划的特征**）：

```
ssp_id, national_id, state_id, city_id,code_seat_id,
code_seat_type, advertiser_id, ad_plan_id, ad_group_id,
ad_creative_id, brand, ~~media_id~~, app_id, brand_model, model,
~~os_type_ver_lang~~, os_type, os_version, language, promotion_type,
~~image_area~~, image_width, image_height, adv_seat_type, ~~screen_area,~~
screen_width, screen_height, ~~adver_csid, group_csid~~,
net_connect_type, cate1_gp
```

特征主体	主要类型	特征
User特征	1. 用户标签 2. 用户属性 a. 地点信息 b. 设备信息 3. 用户行为 4. 用户关系	• 用户标签&属性 ◦ 基础信息：**性别、年龄、消费水平、兴趣爱好、语言**、职业、学历、薪资、颜值、感情状况 ◦ 设备信息：**手机品牌信息（brand、model、series）、手机价格、手机使用天数、运营商（operator）**、手机使用频次（活跃情况）、时长、使用时间段 ◦ **地点信息：国家、省份、城市**、城市等级 ◦ 产品使用习惯：频次（活跃情况）、时长、使用时间段、用户价值、用户生命周期、用户等级 • 用户行为：用户在产品上的各种操作。电商场景下如浏览、点击、播放、购买、搜索、收藏、点赞、转发、加购物车、甚至滑动、在某个位置的停留时长、快进等等一切操作行为 • 用户关系：**通过用户关系建立关系图，生成Embedding**
Item特征	1. Item标签 2. Item属性	• 标签&属性基础信息 ◦ **物品ID（advertiser id、广告计划id、广告组id、广告创意id）** ◦ Item包含的信息：**广告主gp分类**、包体大小等

如何设计此部分架构，如何选择特征进行…

XinMin Wang Jan 29, 2023 (edited)

此部分内容只针对有哪些可供参考添加的特征进行说明，即从特征主体的角度总结相对完善的特征体系；关于不同分布类型的特征、样本构造、特征处理、特征重要性评估等内容将在后续部分详细说明。

地点信息 设备信息

XinMin Wang Jan 31, 2023

地点信息和设备信息作为 user 特征构造还是作为上下文特征需要实验测试模型效果，拓量后非传音系用户的此类信息在用户宽表中获取可能较为困难，需要从 oss 日志上报的逻辑获取

设备：手机品牌信息、手机使用频次（活…

XinMin Wang Jan 29, 2023

在传音广告场景下，user 是 gaid 即手机 id，与电商场景下的用户 id 有所区别，因此设备信息此处属于 user 特征，若在电商场景中，设备信息属于 scene/上下文特征。

地点信息：国家、城市、城市等级

XinMin Wang Jan 30, 2023 (edited)

传音广告场景中，根据数据产出逻辑，地点场景信息适合作为 user 特征

产品使用习惯：频次（活跃情况）、时长…

XinMin Wang Jan 30, 2023

		◦ 物品关系：**item embedding（app embedding）** • item国家定向
Scene/上下文特征	1. 网络场景 2. 时间场景 3. 应用（app）场景	• 网络场景：**网络环境**、上网时长、上网时段 • 时间场景：**当前时间、是否是周天、月份、季节、小时、**是否是节假日 • 应用（app）场景：**应用场景（app_id、code_seat_jd、code_seat_type）** 推荐场景中，context 特征通常是客户端带的信息，在用户授权的前提下可以直接获取，比如请求时间、用户手机品牌、手机型号、操作系统、当前网络状态（3g/4g/wifi）、用户渠道等实时属性特征以及之间的 cross 特征。
交叉特征	1. 用户&物品交叉特征 2. 用户属性&物品交叉特征 3. 用户&物品属性交叉特征 4. 用户属性&物品属性交叉特征 5. 物品&场景交叉特征	• **统计特征：用户物品交叉统计特征、用户属性物品交叉统计特征、用户物品属性交叉统计特征、用户属性物品属性交叉统计特征** ◦ 用户交叉统计特征诸如“该用户在该游戏上的历史点击率”、“该用户在该媒体上的历史点击数”等，描述每个用户对每个广告的交叉特征；用户属性标签统计交叉特征诸如“该职业用户在该游戏上的历史点击率”、“该年龄段用户在该广告模板上的历史点击率”等，描述一个用户属性标签下所有用户对每个广告的交叉特征 • **序列特征：用户点击广告主序列、用户点击广告创意序列、用户点击计划序列、用户点击广告类型序列** ◦ 针对每个序列先对其 Embedding 化，然后分别进行 Max-Poolling 和 DIN-Attention，并最终进行 concat，作为序列建模层的输出。 • **物品&场景交叉统计特征：反映商品热度** ◦ **item不同周期的ctr、clk_uv、cvr等** ◦ **item当前场景/维度下的ctr、clk_uv、cvr等**

3.2 【阿里】搜广推场景的特征工程

推荐算法中的特征工程

阿里在搜广推场景下为了应对高基数特征、大数据样本、在线推理的实时性等问题，使用了大量**统计特征**。

通过对用户或商品某些特征在不同的行为类型、不同时间周期、不同的标签上分别统计正负样本数量得到统计特征。这些统计量经过特征变换，包括特征缩放、分箱和统计编码后作为最终特征向量的一部分。

手机 gaid 中产品使用习惯相当于设备使用习惯

生成Embedding

XinMin Wang Jan 29, 2023 (edited)
使用 Graph Embedding 的方法生成用户和物品的 Embedding

XinMin Wang Jan 30, 2023
传音场景中 user embedding 是基于 pkg 打开序列、pkg 安装序列和 pkg 卸载序列生成

Scene/上下文特征

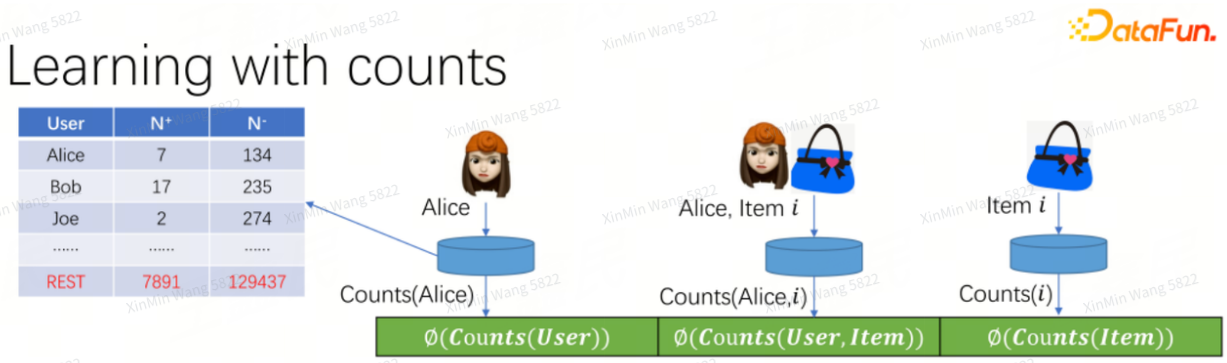
XinMin Wang Jan 29, 2023 (edited)
描述行为产生的场景信息。最常用的上下文信息是“时间”和通过 GPS、IP 地址获得的“地点”信息。根据场景的不同，上下文信息的范围极广，如推荐场景中，除了上面提到的时间和地点，还包括“当前所处页面”“季节”“月份”“是否节假日”“天气”“空气质量”“社会大事件”等等。

序列特征：用户交叉序列embedding特…

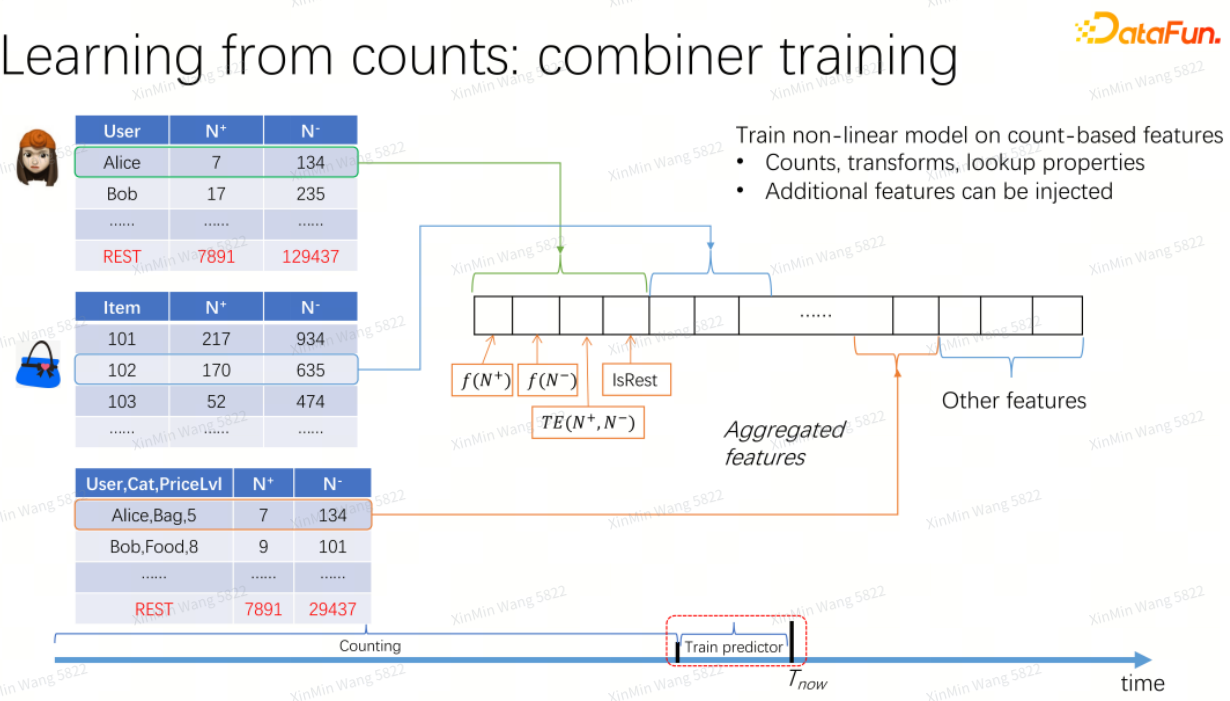
XinMin Wang Jan 30, 2023
序列特征是将用户的行为序列借鉴 NLP 文本序列的处理方法进行 embedding 放入模型中

这些统计量经过特征变换，包括特征缩放…

XinMin Wang Jan 29, 2023
这里推荐使用 Gauss Rank 来做特征缩放，因为 Gauss Rank 对特征的分布变化具有一定的鲁棒性，不易受到流量波动的影响。



在统计正负样本数量之前，要先对推荐的实体（如用户、物品、上下文）进行分箱，称为bin counting，如上图所示。在得到用户和商品的分箱统计特征后，我们还可以对这些特征进行交叉组合得到新的统计量，称为cross counting，这些统计量最后都会转化为特征。



那么如何能把所有的特征都考虑到，做到不重不漏呢？可以按照如下描述的结构化方法来枚举特征：

- **列存实体**。推荐场景的实体主要是用户、商品和上下文；广告场景的实体主要是用户、广告、搜索词和广告平台。
- **实体分箱**。可以针对用户或物品的自然属性来分箱，例如用户可以针对每一个用户画像做分箱，物品可以基于类目、价格等做分箱，得到单维度的特征，利用历史行为信息对单维度特征做统计、编码。
- **特征交叉**。在实体分箱的基础上做特征交叉，得到二阶、三阶或更高阶特征。

总结，阿里搜推广场景下的常用特征工程套路可以总结为一个词“Bin&Counting”，也就是先做bin，再做counting，中间结合cross counting。

3.3 【美团】搜索广告业务CTR&CVR模型的特征工程

深度学习在美团搜索广告排序的应用实践

美团搜索广告业务囊括了关键词搜索、频道筛选等业务，覆盖了美食、休娱、酒店、丽人、结婚、亲子等200多种应用场景，用户需求具有多样性。同时O2O模式下存在地理位置、时间等独特的限制。

结合上述场景，抽取了以下4大类特征：

特征类别	特征组成
用户特征	• 人口属性：用户年龄，性别，职业等。 • 行为特征：对商户/商圈/品类的偏好（实时、历史），外卖偏好，活跃度等。 • 建模特征：基于用户的行为序列建模产生的特征等。
商户特征	• 属性特征：品类，城市，商圈，品牌，价格，促销，星级，评论等。 • 统计特征：不同维度/时间粒度的统计特征等。 • 图像特征：类别，建模特征等。 • 业务特征：酒店房型等。
Query特征	• 分词，意图，与商户相似度，业务特征等。
上下文特征	• 时间，距离，地理位置，请求品类，竞争情况等。 • 广告曝光位次。

特征工程在不同场景下的特性

场景	特性
用户的消费场景	• “**附近**”请求：美团和大众点评App中，大部分用户发起请求为“附近”请求，即寻找附近的美食、酒店、休闲娱乐场所等。因此给用户返回就近的商户可以起到事半功倍的效果。“**请求到商户的距离**”特征可以很好地刻画这一需求。 • “**指定区域（商圈）**”请求：寻找指定区域的商户，这个区域的属性可作为该流量的信息表征。 • “**位置**”请求：用户搜索词为某个位置，比如“五道口”，和指定区域类似，识别位置坐标，计算商户到该坐标的距离。 • “**家/公司**”：用户部分的消费场所为“家”或“公司”，比如寻找“家”附近的美食，在“公司”附近点餐等，根据用户画像得到的用户“家”和“公司”的位置来识别这种场景。
多品类	针对美食、酒店、休娱、丽人、结婚、亲子等众多品类的消费习惯以及服务方式，将数据拆分成三大部分，包括**美食、酒店、综合**（休娱、丽人、结婚、亲子等）。其中**美食**表达用户的餐饮需求，**酒店**表达用户的旅游及住宿需求，**综合**表达用户的其他生活需求。
用户的行为轨迹	实验中发现用户的实时行为对表达用户需求起到很重要的作用。比如用户想找个餐馆聚餐，先筛选了美食，发现附近有火锅、韩餐、日料等店，大家对火锅比较感兴趣，又去搜索特定火锅等等。用户点击过的商户、品类、位置，以及行为序列等都对用户下一刻的决策起到很大作用。

用户的消费场景

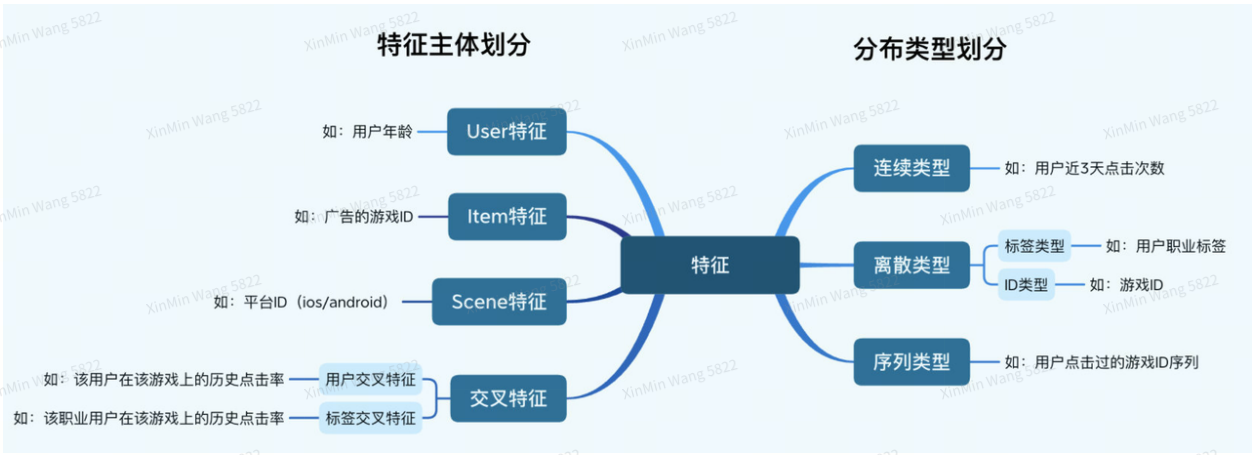
XinMin Wang Jan 29, 2023

距离&位置属性特征在美团消费场景下至关重要

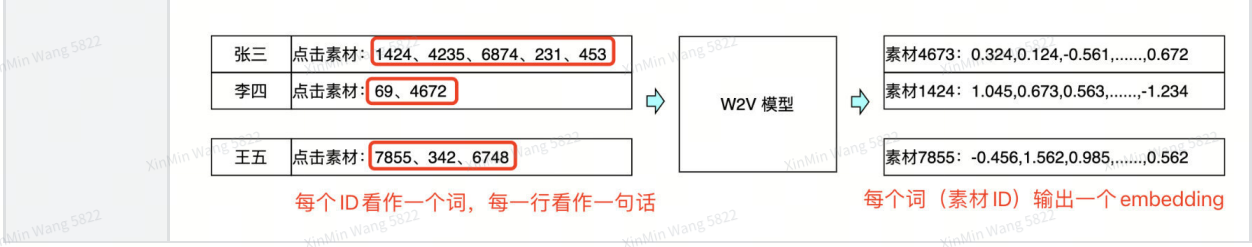
3.4 【腾讯】计算广告&微视推荐系统特征工程

一文彻底搞懂 CTR 建模

腾讯计算广告场景中的特征工程同样遵从按照特征主体分类的方式构建，对于特征工程的构造细节没有阐述，特征侧带来模型auc提升的主要工作包括**构造人工交叉特征、引入序列特征和添加素材特征**3方面：



特征优化	主要内容
人工交叉特征	1. 人工交叉可以高效提炼有价值的交叉特征，降低模型学习的难度。模型自动交叉不需要人工参与，可以自动根据数据集进行交叉，但是对于复杂的交叉关系需要极其大量的数据进行训练，在数据稀疏场景下并不现实。 2. 腾讯游戏广告场景下主要是用了 30+个“用户交叉特征”和 80+个“用户属性标签交叉特征”，用户交叉特征诸如“该用户在该游戏上的历史点击率”、“该用户在该媒体上的历史点击数”等，描述每个用户对每个广告的交叉特征；用户属性标签交叉特征诸如“该职业用户在该游戏上的历史点击率”、“该年龄段用户在该广告模板上的历史点击率”等，描述一个用户属性标签下所有用户对每个广告的交叉特征。
引入序列特征	1. 引入6个序列特征：用户点击游戏序列、用户点击游戏品类序列、用户点击创意序列、用户点击广告序列、用户点击计划序列、用户点击广告类型序列 2. 针对每个序列先对其 Embedding 化，然后分别进行 Max-Poolling 和 DIN-Attention，并最终进行 concat，作为序列建模层的输出。
添加素材特征	基于用户对素材的点击序列，构造素材 ID 的 Embedding作为特征加入模型。



基于用户对素材的点击序列（看作一个st…

鑫民 XinMin Wang Jan 29, 2023

用户刷到广告后，与用户体验相关性最强的就是素材。我们也加入了一些素材的 Embedding 特征，我们尝试将单个用户近期点击过的素材 ID 看作一个集合，同处在一个集合中的素材 ID 看作是一次“共现”。这个关系与 NLP 中，一个句子中词的共现是类似的，只不过素材 ID 的共现中没有前后顺序的联系。所以我们建立了一个简单的 W2V 模型，根据素材 ID 之间的共现关系，学习每个素材 ID 的 Embedding，并将该 Embedding 加入到 pCTR 模型中，给模型的 AUC 稳定地带来了约 0.0003 的提升。

Expand

浅谈微视推荐系统中的特征工程

微视场景属于富媒体形态的短视频平台，此场景下的特征工程主要做的是“如何通过视频内容特征以及用户属性和行为数据，来精准预测用户对短视频的喜好的”。特征工程主要组成包括4部分：user特征、item特征、context特征和session特征。此场景属于视频推荐场景，目标函数中包括点击、播放时长、完播率等，与广告场景的ctr&cvr建模目标有所区别。

item在不同维度下的rank特征

鑫民 XinMin Wang Jan 29, 2023

在构造效率、排名这类型统计特征时，我们需要防止特征穿越问题，一定要使用 yesterday 计算出来的数据进行模型训练。这是因为模型在线上 inference 时是拿不到当天的效率特征的（实时效率除外）。

3.5 【飞猪】搜索广告场景下特征工程

飞猪日常业务中使用的特征体系，CVR模型特征工程主要由4个部分组成：item侧特征、user侧特征、query侧特征和上下文context特征。

特征类别	特征组成	注意事项
item特征	1. item静态特征：价格、目的地、类目特征 2. item行为特征：反映商品热度，item不同周期的ctr、clk_uv、cvr等 3. item在不同维度下的rank特征：item在不同query下的排名特征、item在不同用户购买力下的排名特征、item在不同用户年龄阶段的排名特征、item在不同lbs下的排名特征	**谨防特征穿越** 在构造效率、排名这类型统计特征时，我们需要防止特征穿越问题，一定要使用yesterday计算出来的数据进行模型训练。这是因为模型在线上inference时是拿不到当天的效率特征的（实时效率除外）
user特征	1. user静态特征：性别、年龄、购买力等 2. user偏好特征：根据用户历史行为序列计算不同周期内对不同目的地的偏好、不同类目的偏好、不同poi的偏好等 3. user实时特征：用户当前的lbs信息、当前时间信息、用户实时的点击序列、用户实时的购买序列	使用实时特征注意**线上线下特征一致性问题**。使用实时特征前先埋点。
context特征	1. match特征：match type、match score、召回的trigger特征 2. 粗排特征：粗排打分 3. user上下文：用户当前lbs信息与商品lbs的距离 4. item上下文：入口trigger item与目标item的相似性特征、item价格在排序集合中的排名特征、item距离在排序集合中的排名特征	最大可能利用未曝光的数据，构造丰富的context特征。
query特征	1. query基础特征：query type、query vector 2. query相关性特征	**query归一化**对排序稳定性有非常重要的影响，

特征处理/特征提取工作

XinMin Wang Jan 30, 2023 (edited)

即通过特征处理和变换，去除原始数据中的杂质和冗余

生成高质量特征

XinMin Wang Jan 30, 2023 (edited)

第二节中介绍过，高质量的特征应具有区分度，相互独立，简单可解释性强等特点，对所有的机器学习任务都适用。

4. 特征工程的常见问题及处理

引言：上节从业务理解层面介绍了各厂特征工程特征构造部分的工作，并规划了传音广告场景下的特征构造。广告场景下的原始数据类型主要有**数值型、离散型和序列型**。由于诸如数据分布不均、异常值、数据稀疏等问题，直接使用原始数据进行模型训练往往会造成巨大的误差。**本节将从技术角度上说明特征工程中的特征处理/特征提取工作，如何生成高质量特征，供模型训练使用。**

4.1 传音广告场景下数据处理的主要问题

传音广告场景中，CVR模型特征工程优化初版方案的组内评估中，我们讨论了如下潜在风险和修改意见：

1. 对数值特征的归一化、分桶&异常值处理问题
2. 特征数据分布问题
3. 特征缺失率
4. 特征有效性/重要性评估问题
5. 避免特征重复&冗余共线性问题
6. 序列特征在multi_tower的使用原理问题
7. item行为特征在不同媒体/不同维度数据量级差异问题

问题补充：

8. **新建广告创意但是素材不变导致广告创意粒度特征为空/统计值偏差 - 按照广告素材id构建groupby特征**

序列型特征：针对每个序列先对其 Embe…

XinMin Wang Jan 30, 2023

pooling、CNN、RNN、GNN、attention、transformer、bert

9. **行业特征的重要性，需要加上**
10. 用户对广告行业的交叉特征 老板重点关注

4.2 常见数据问题&特征变换方案

针对特征工程中原始数据存在的问题，本节将按照数据类型对特征的常见变换操作进行总结。

结合各厂经验，特征工程中原始数据的处理（特征变换）适合按照特征/数据的分布类型，即数值型、离散（类别）型和序列型，选择合适的解决方案。常见的特征变换操作如下：

- 数值型特征：分桶、缩放、缺失值处理、平滑
- 离散型特征：onehot、multihot（multi-hot编码原理）、哈希分桶，打分排名编码
- 序列型特征：针对每个序列先对其 Embedding 化，然后分别进行 Max-Poolling 和 DIN-Attention，并最终进行 concat，作为序列建模层的输出。

<table>
<tr><th>数据类型</th><th>常见问题/处理</th><th>处理方法</th></tr>
<tr><td>数值型</td><td>简单特征预处理</td><td>有些特征的特点是分布偏态十分严重，且极差极大。如“用户近 90 天游戏付费金额”，大多数用户该特征都是 0，但是极少量用户的付费金额可能达到数十万。这种情况如果直接让深度模型学习，会导致模型很难收敛，就算收敛了模型效果也大打折扣。对于这种情况需要进行特征预处理，通常业内常采用“归一化”、“标准化”、“截断”
<table>
<tr><th>预处理方法</th><th>公式</th></tr>
<tr><td>归一化</td><td>$x' = \frac{x - min(x)}{max(x) - min(x)}$</td></tr>
<tr><td>标准化</td><td>$x' = \frac{x - \overline{x}}{\sigma}$</td></tr>
<tr><td>截断</td><td>$x' = \begin{cases} k_2 \ (x > k_2) \\ x \ (k_1 \le x \le k_2) \\ k_1 \ (x < k_1) \end{cases}$</td></tr>
</table>
</td></tr>
<tr><td></td><td>特征缩放 - 处理归一化&数据分布问题</td><td>面对各指标数据分布&数量级的差异，需要进行特征归一化（Normalization），以便进行分析&模型训练。

特征缩放就是把一些取值范围较大的特征缩放到较小的范围内。如果不做特征缩放，较大值的特征会支配梯度更新的方向，导致梯度更新在误差超平面上不断震荡，模型学习效率变低。另外，一些基于距离度量的算法，如KNN，K-means等也很大程度上会受到是否进行特征缩放的影响。不做特征缩放，取值范围较大特征会支配距离函数的计算，导致其他特征失去作用。在实际应用中通过梯度下降法求解的模型通常是需要归一化的，包括线性回归、逻辑回归、支持向量机、神经网络模等模型。但对于决策树模型并不适用，因为决策树在进行节点分裂是主要依据数据集D关于特征的信息增益，而信息增益与特征是否归一化是无关的。

常用的特征缩放方法包括Min-Max、Z-score、Log-based、L2-normalize、Gauss-Rank等。

Gauss Rank是推荐系统中效果比较好的一个特征变换操作。首先我们对数据的统计值做一个排序，从小到大或从大到小都可以，得到数据的Rank值，然后将Rank值缩放到(-1,1)区间，最后调用erfinv逆误差函数，就可以将变换后的Rank值分布调整为高斯分布。深度学习模型偏好高斯分布的数据输入，这也是为什么深度学习中经常使用的一个操作是Batch Normalization。</td></tr>
</table>

特征缩放 - 处理归一化&数据分布问题

XinMin Wang Jan 30, 2023 (edited)

归一化应用的关键在于如何根据不同场景选择最适合的方法。

思考题 1：如何量化短视频的流行度（假设以播放次数来衡量）？

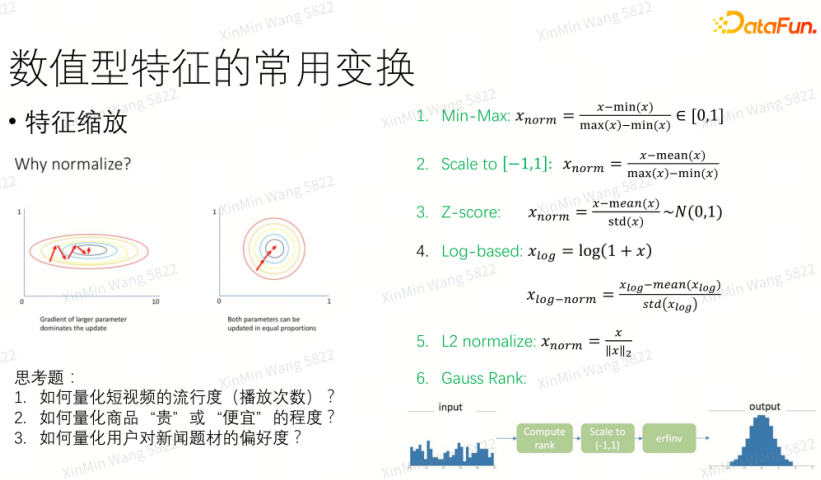
参考答案：短视频的播放次数在整个样本空间中服从幂律分布，即长尾分布，少量的热门视频播放次数会很高，大量的长尾视频播放次数都相对较少。这个时候最好采用 Log-based 变换，即先对播放次数取 log，取完 log 之后的值做 Z-score 标准化处理，最终得到的值分布比较均匀。如果不做 log 变换直接用 Z-score 处理，会导致大部分特征值被压缩到一个非常窄的区域。

思考题 2：如何量化商品“贵”或“便宜”的程度?

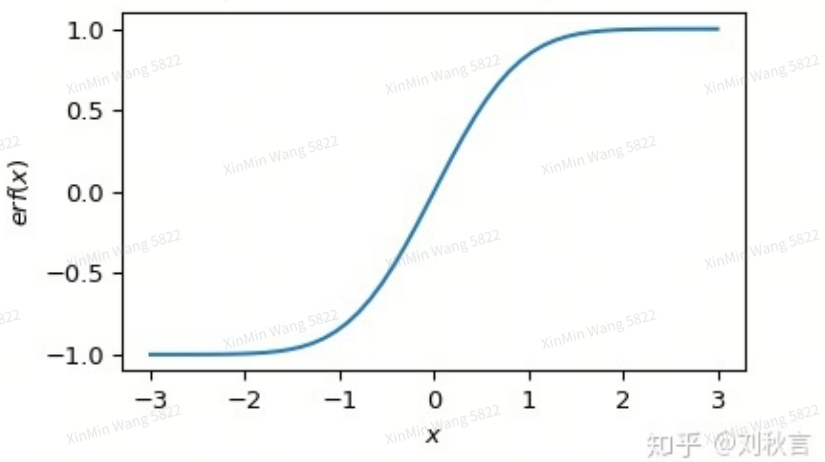
参考答案：首先商品的价格不能量化商品“贵”或“便宜”的程度，因为不同品类的商品价格区间本来差异就很大。比如，1000 块钱买到一部手机，顾客感觉很便宜；但同样 1000 块钱买一只鼠标，顾客就会觉得这个商品的定价很贵。因此，量化商品“贵”或者“便宜”的程度时就必须要考虑商品的类目，这里推荐的做法是做 Z-score 标准化。要注意的是 Z-score 的均值和标准差的计算都要限制在同类商品集合内，而不是对整个数据集，并且最好采用叶子类目，即最细一层的类目，但如果叶子类目的商品种类太少，回溯一层也是可以的。

思考题 3：如何量化用户对新闻题材的偏好度（假设以阅读次数来衡量）？

参考答案：不同用户的活跃度是不同的，有些高活跃用户可能对多个题材的阅读量都比较大，而一些低活跃用户对可能只对某几个题材有中等的阅读量。我们不能因为高活跃度的用户对某题材




```python
from scipy.special import erfinv
def scale_rankgauss(x, epsilon=1e-6):
    '''rankgauss'''
    x = x.argsort().argsort() # rank
    x = (x/x.max()-0.5)*2 # scale
    x = np.clip(x, -1+epsilon, 1-epsilon)
    x = erfinv(x)
    return x
```



1. 归一化和标准化只能解决分布方差大的问题，不能解决偏态严重的问题
2. 截断的方法损失了大量信息，损失了这部分特殊用户的区分度

为了解决这些问题，我们针对部分需要进行预处理的特征，采用如下公式进行转换：

$$x' = sign(x) \cdot log(|x| + 1)$$

这样做的目的是尽可能不损失区分度的情况下，对该特征进行压缩，避免模型无法收敛的情况。由于对于不同日期的数据，不论其分布如何，都统一采取同样的预处理方式，所以不会出现模型更新后，分布不稳定的情况。经过多次离线实验，采用该变换公式解决收敛性和稳定性问题的同时，并未导致 AUC 降低，真是一个好办法！

异常值问题	Robust scaling是阿里搜广推场景业务中使用的异常值处理方式 数据中存在异常值（上图中红色区域）时，用Z-score、Min-max这类特征缩放算法都可能会把转化后的特征值压缩到一个非常窄的区间内，从而使这些特征失去区分度。这里介绍一种新的特征变化方法：Robust scaling，其中median(x)是x的中位数；IQR为四分差，等于样本中75%分位点的值减去25%分位点的值。经过Robust scaling变换，如上图最后一列数据所示，数据中较小的值依然有一定的区分性。然而，对于这种存在异常值的数据，最好的处理方法还是提前将异常值识别出来，然后对其做删除或替换操作。

的阅读量大于低活跃度用户对相同题材的的阅读量，就得出高活跃度用户对这种类型的偏好度大于低活跃度用户对同类型题材的偏好度，这是因为低活跃度用户的虽然阅读量较少，但却几乎把有限精力全部贡献给了该类型的题材，高活跃度的用户虽然阅读量较大，但却对多种题材“雨露均沾”。所以建议对这种问题先按照用户分组，组内再做 Min-max 归一化。

Expand

分箱处理(Binning)

XinMin Wang Jan 29, 2023

思考题 1：如何衡量用户的购买力？如何给用户的购买力划分档位？

	Original	Standardization	Max-Min Scaler	Rubost Scaler
2	6.9314183	-0.2244971	0.0000003	0.8283487
3	2.6674115	-0.2244979	0.0000001	0.0690181
4	7.7248183	-0.2244970	0.0000003	0.9696367
5	5.7388433	-0.2244973	0.0000002	0.6159760
6	0.8965615	-0.2244982	0.0000000	-0.2463333
7	4.5147618	-0.2244975	0.0000002	0.3979926
8	2.9934144	-0.2244978	0.0000001	0.1270724
9	4.8708377	-0.2244975	0.0000002	0.4614023
10	4.2797819	-0.2244976	0.0000002	0.3561476
11	1.0085616	-0.2244982	0.0000000	-0.2263885
12	5.5166580	-0.2244974	0.0000002	0.5764094
13	1.1171326	-0.2244981	0.0000000	-0.2070542
14	0.4069897	-0.2244983	0.0000000	-0.3335159
15	5.0536949	-0.2244975	0.0000002	0.4939654
16	8.4068370	-0.2244969	0.0000003	1.0910900
17	8.9588050	-0.2244968	0.0000003	1.1893840
18	0.9543401	-0.2244982	0.0000000	-0.2360442
19	94750.5292279	-0.2079018	0.0037104	16872.6857158
20	2051.2433203	-0.2241390	0.0000803	364.8776314
21	25536631.9371928	4.2485000	1.0000000	4547540.7645023

缺失值	实际问题中经常会遇到特征缺失的情形，关于缺失值处理的方式， 总结有以下几种： • 不处理（这是针对xgboost等树模型），有些模型有处理缺失的机制，所以可以不处理 • 如果缺失的太多，可以考虑删除该列 • 插值补全（均值，中位数，众数，建模预测，多重插补等），需要视具体情况进行选择； • 分箱处理，将缺失作为一种信息（作为一个箱）进行编码输入模型让其进行学习，比如用户性别缺失，可以直接将未知作为一种类别进行处理。
分箱处理 (Binning)	分箱就是将连续的特征离散化，以某种方式将特征值映射到几个箱(bin)中。比如预测电商场景下用户的点击率，其中有一个特征是时间特征，即一天24小时，但并不是说时间值越大点击率就会越高。通过历史数据显示，深夜事件段用户点击率都比较低，午饭和晚饭后用户比较活跃，点击率较高，依靠这些经验将一天划分为不同时间段再输入给模型可以增加模型的表达能力。 **为什么要做特征分箱?** • **引入非线性变换，增强模型性能**。因为原始值和目标值之间可能并不存在线性关系，所以直接使用模型预测起不到很好的效果。 • **增强模型可解释性**。通过分箱可以得到一个分段函数，模型可解释性更强。 • **对异常值不敏感，防止过拟合。**异常值最终也会被分到一个箱里面，不会影响其他箱内正常特征值，分箱的在一定程度上也可以防止过拟合。 • 最重要的是分箱之后我们还**可以对不同的箱做进一步的统计和特征组合**。比如按年龄段分箱后对不同年龄段的人群做一个CTR统计。 **分箱有无监督和有监督两种方法**。无监督方法包括固定宽度分箱、分位数分箱、对数转换并取整等，实际中应用较多，有监督的方法应用较少。 
数据平滑	常用的行为次数与曝光次数比值类的特征，由于数据的稀疏性，这种计算方式得到的统计量通常具有较大的偏差，需要做平滑处理，比如广告

背景：电商场景下用户的购买力是一种很好的属性，可以反映用户的消费倾向，用户是倾向于高质量消费还是高性价比消费。购买力是长期的稳定的用户画像，与用户近期的消费金额无关。

参考答案: 首先要划分商品的价格档位，根据商品类目分组，组内按价格升降排序，利用等宽分箱方法得到价格档位。然后根据用户的历史消费行为把其购买过的商品的价格档位聚合到用户身上，注意同一个用户对不同类商品的购买力也是不同的，比如有的用户愿意花高价购买电子产品，对其他种类的商品的购买力一般，因此可以针对每个类目计算购买力。

思考题 2：经纬度如何分箱?

参考答案：经纬度是一个整体，不能把其拆开成多个独立变量来单独做分箱，而是要把这些变量当做一个整体来考虑。经常用到的一个方法是 GeoHash，简单来说就是把地图划分成一个二维网格，不同的网格有唯一的 hash 编码，代表不同的区域。

Expand

数据平滑

XinMin Wang Jan 30, 2023 (edited)

结合归一化方法共同解决上节中问题 7. item 行为特征在不同媒体/不同维度数据量级差异问题

点击率常用的贝叶斯平滑技术。在推荐场景中，也会用到很多统计类特征、比率特征。如果直接使用，比如由于不同 item 的下发量是不同的，这会让推荐偏向热门的类目，使得越推越窄，无法发现用户的个体差异，也不利于多样性的探索。我们可以把曝光量进行分段，同一个曝光量级的指标进行比较，也可以用该 item 所属类目统计量的平均值进行平滑处理。对于离群值较多的数据，我们会使用更加健壮的处理方法，比如使用中位数而不是均值，基于分位数而不是方差。而在短视频业务上较短或较长的视频在播放完成度上存在天然的差距，我们按视频本身长度离散，观看时长做分位数处理，同时做威尔逊置信区间平滑，使得各视频时长段播放完成度相对可比，避免出现打分因视频长度严重倾斜的情况。以及短视频 app 的投稿数量大，对于长尾的视频和类目都是需要做平滑处理的。下面介绍两种较为常用的平滑技术。

- 贝叶斯平滑

电商领域中经常需要计算或预测一些转化率指标，比如 CTR。这些转化率可以是模型的预测值，也可以作为模型的特征使用。以商品点击率预测为例，CTR 的值等于点击量除以曝光量。理想情况下，例如某个广告点击量是 10000 次，转化量是 100 次，那转化率就是 1%。但有时，例如某个广告点击量是 2 次，转化量是 1 次，这样算来转化率为 50%。但此时这个指标在数学上是无效的。因为大数定律告诉我们，在试验不变的条件下，重复试验多次，随机事件的频率近似于它的概率。后者点击量只有 2 次，不满足“重复试验多次”的条件。如果对于一个新上线的商品，其曝光为 0，点击量也为 0，此时这件商品的 CTR 应该设为 0 还是赋一个初始值？初始值设 0 是可以的，但不太合理。当 CTR 作为特征使用时，表示这个商品完全没有点击，不太符合日常推断，通常是赋一个大于 0 的初始值。

以上两个问题可以使用平滑技术来解决。贝叶斯平滑的思想是给 CTR 预设一个经验初始值，再通过当前的点击量和曝光量来修正这个初始值。如果某商品的点击量和曝光量都是 0，那么该商品的 CTR 就是这个经验初始值；如果商品 A 和商品 B 的曝光量差别很大，那么可以通过这个经验初始值来修正。贝叶斯平滑就是确定这个经验值的过程。贝叶斯平滑是基于贝叶斯统计推断的，因此经验值计算的过程依赖于数据的分布情况。对于一件商品或一条广告，对于某次曝光，用户要么点击，要么没点击，这符合二项分布。因此对于点击率类的贝叶斯平滑，都可以基于以下假设：对于某件商品或广告，其是否被点击是一个伯努利分布。伯努利分布的共轭分布就是 Beta 分布，也就是说，点击率服从 Beta 分布。而所有的数据有一个自身的点击率分布，这个分布可以用不同的 beta 分布来拟合。beta 分布可以看做是对点击率的一个先验知识，我们可以根据观测来修改我们的先验，所以贝叶斯平滑就是估计 Beta 分布中的参数 α 和 β，其中 C 和 I 是点击次数和曝光量。实际应用时根据历史数据得到的 α 和 β 可以帮助确定平滑参数的大致范围，防止设置参数时偏离过大。

$$r = \frac{C + \alpha}{I + \alpha + \beta}$$

点击率预估（CTR）

在点击率（Click Through Rate，CTR）预估任务中，点击率 R 的计算如下：

$$R = \frac{C}{I}$$

其中 C 是点击量（Click），I 是曝光量（Impression）。

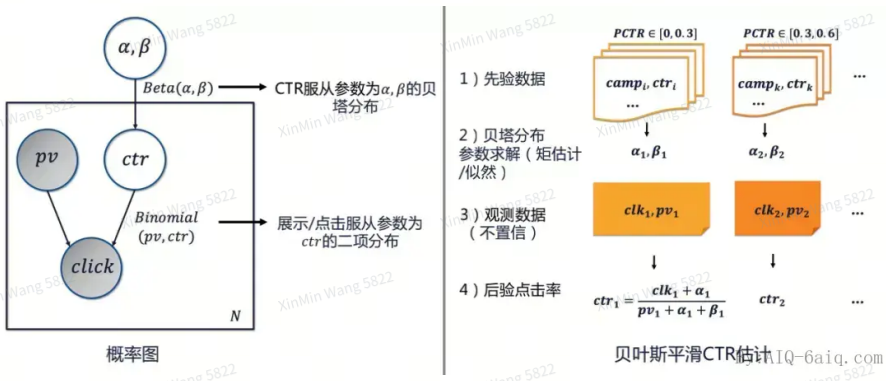
在真实场景中，如果直接用该 CTR 来进行排序则会有一个严重的问题，即 **新内容很难获得曝光**，曝光不足会导致 CTR 不准，甚至点击率为 0，以此算出来的 CTR 不能反映该内容的真实情况。

因此 贝叶斯平滑 被引入该场景，以下先给出带贝叶斯平滑的 CTR 计算公式：

$$\hat{R} = \frac{C + \alpha}{I + \alpha + \beta}$$
$$\alpha = (\frac{\overline{R}(1-\overline{R})}{S^2} - 1)\overline{R}$$
$$\beta = (\frac{\overline{R}(1-\overline{R})}{S^2} - 1)(1-\overline{R})$$

其中 $\overline{R}$、S^2 分别为点击率的均值、方差。

以该方式计算得到的结果是：**每个新内容刚开始时都会得到一个接近平均水平的初始值，然后在不断获得曝光后不断地调节 CTR 计算以接近自己的真实水平。**



参考：

转化率之贝叶斯平滑

贝叶斯平滑

- 威尔逊区间平滑

在现实生活中我们会接触到很多评分系统，如豆瓣书评、YouTube 影评，在这些评分中有 1 个共同问题是每个 item 的评分人数是不同的，比如 10000 个人打了 90 分似乎比只有 10 个人打了 90 分更能被相信该 item 是 90 分的。威尔逊区间法常用来解决此类问题，是一种基于二项分布的计算方法，综合考虑评论数与好评率，平滑样本量对评价的影响，我们画像兴趣分上也用到了威尔逊区间平滑。

假设 u 表示正例数（好评），n 表示实例总数（评论总数），那么好评率 p 就等于 u/n。p 越大，表示这个 item 的好评比例越高，越应该排在前面。但是，p 的可信性，取决于有多少人，如果样本太小，p 就不可信。我们已知 p 是二项分布中某个事件的发生概率，因此我们可以计算出 p 的置信区间。置信区间实际就是进行可信度的修正，弥补样本量过小的影响。如果样本多，就说明比较可信，不需要很大的修正，所以置信区间会比较窄，下限值会比较大；如果样本少，就说明不一定可信，必须进行较大的修正，所以置信区间会比较宽，下限值会比较小。威尔逊区间就是一个很好的修正公式，在小样本上也具有很强的鲁棒性。

在下面的公式中， p 表示样本的好评率，n 表示样本的大小，z 表示对应某个置信水平的 z 统计量，是一个常数。一般情况下，在 95%的置信水平下，z 统计量的值为 1.96。可以看到，当 n 的值足够大时，这个下限值会趋向 p 。如果 n 非常小，这个下限值会远小于 p，起到了降低好评率的作用，使得该 item 的打分变低、排名下降。

$$\frac{p + \frac{1}{2n}z^2_{1-\frac{\alpha}{2}} \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n} + \frac{z^2_{1-\frac{\alpha}{2}}}{4n^2}}}{1 + \frac{1}{n}z^2_{1-\frac{\alpha}{2}}}$$

p —— 概率，即点击的概率，也就是 CTR；

n —— 样本总数，即曝光数；

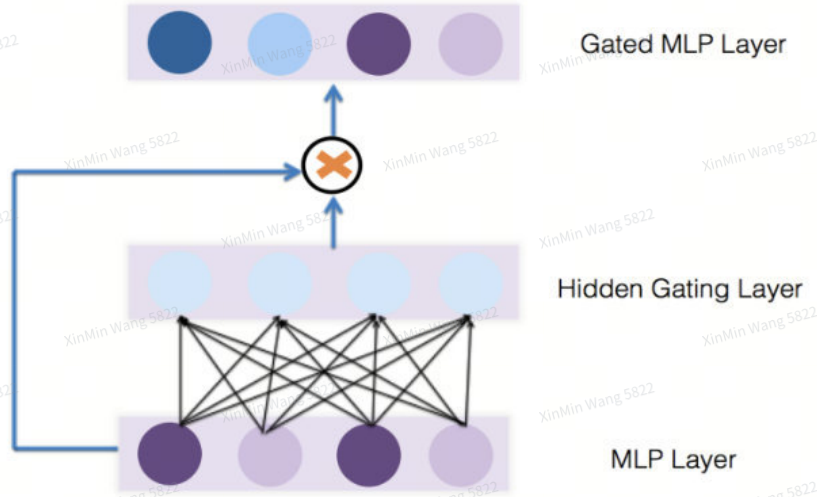
Z —— 在正态分布里， z=1.96 ，为95% 置信度

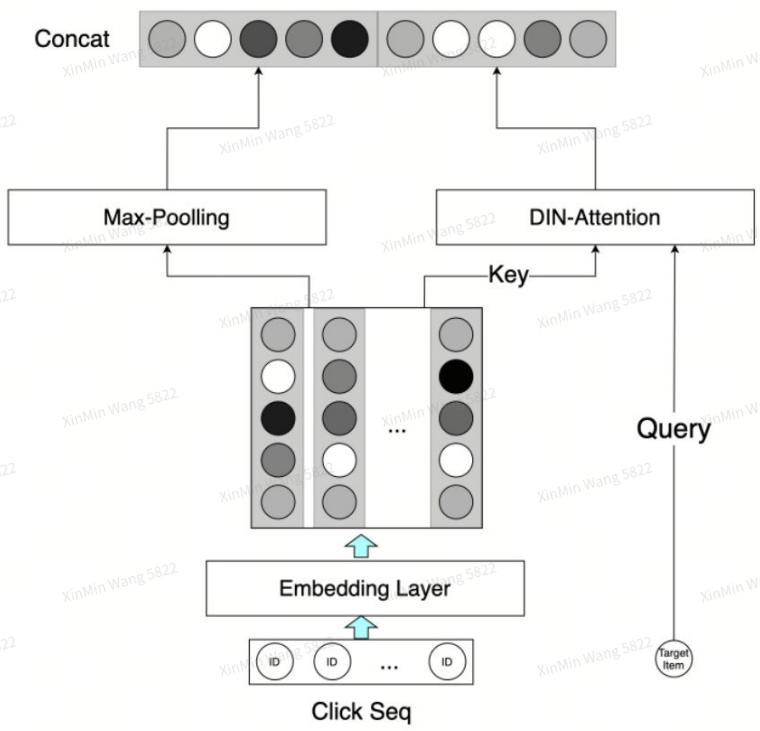
参考：

> 通常处理特征共线的方法是通过大量离线…

XinMin Wang Jan 30, 2023

常规做法

		王多鱼：广告点击率CTR修正-Wilson CTR
	特征冗余共线	在特征中不可避免的会出现特征之间的共线性，比如“用户近 24h 的点击数”和“用户近 72h 的点击数”，又比如“用户在该媒体上的历史点击率”和“用户近 3 天历史点击率”，不可避免的会有较高的共线性。高度共线的特征会使模型很容易在局部最小值处提前收敛，而且可能会导致模型参数量增加，但没有增加有效的信息，从而出现过拟合现象。 **通常处理特征共线的方法是通过大量离线实验，不断地删除可能导致共线的特征（可通过相关系数分析），并观察 AUC 的变化。从而在保证 AUC 不降低的前提下，剔除掉共线性最强的特征**。但是这样做仍然存在问题： 1. 既然剔除了特征，就不可避免的造成信息损失，降低了模型效果的天花板 2. 加入 A、B、C 三个特征共线，那应该剔除他们三个中的哪个呢？如果仅以剔除后的 AUC 来衡量，可能会在不同天的数据集上得出不同的答案（第一天剔除 A 最好，第二天剔除 C 最高）。即共线特征的重要性，可能是变化的，直接剔除特征无法反应这种变化 针对这种情况，腾讯借鉴了《GateNet:Gating-Enhanced Deep Network for Click-Through Rate Prediction》中的方案，在模型内部添加一个 Gate 层，本质上是为每个特征增加了一个“可学习的权重”。 读者可能会问：既然深度学习模型本身的学习机制就可以让无关特征的影响降低，从而起到变量筛选的作用，为什么还要单独一个 Gate Layer 来给变量赋权呢？答案就是：“假设空间太大”，一个全连接层的参数量与前后两层节点数的乘积成正比，参数量很大，不容易学习。相比之下 Gate Layer 中的参数量就少了很多，假设空间更小，更容易学习。 考虑到实际上在共线性的情况下，一个特征的权重不仅与自身有关，还与其他特征有关。所以采用 Bit-Wise Feature Embedding Gate 结构： 
离散（类别）型	分箱处理	类别型的特征有时候也是需要做分箱的，尤其是存在高基数特征时，不做分箱处理会导致高基数特征相对于低基数特征处于支配地位，并且容易引入噪音，导致模型过拟合。**类别型特征的分箱方法通常有以下3种：** - **基于业务理解**。例如对userID分箱时可以根据职业划分，也可以根据年龄段来划分。 - **基于特征的频次合并低频长尾部分(Back off)。** - **基于决策树模型。**
	统计编码	统计编码就是找到一个与类别本身以及目标变量相关的统计量来代替该类别特征，把类别特征转化为一个小巧、密集的实数型特征向量。 - **Count Encoding，统计某类别型特征发生的频次**。一般需要做特征变换后才能输入给模型，建议的特征变换操作包括Gauss Rank、Binning。

		• **Target Encoding，统计某类别特征的目标转化率**。如目标是点击就统计点击率，目标是成交就统计购买率。同时目标转化率需要考虑置信度问题，比如10次浏览有5次点击和1000次浏览500次点击置信度是不一样的，所以对于小的点击次数我们需要用全局的点击率做一个平滑处理。 • **Odds Ratio，可以用来度量用户对某一类目商品的偏好程度相对于其他类目是什么样的水平**。如上图所示，Alice对Bag类别的偏好程度相当于对其他类别偏好程度的0.7906。 • **Weight of Evidence，度量某类别特征不同分箱下取值与目标的相关程度**。值为正表示正相关，值为负表示负相关。 
独热编码	独热编码通常用于处理类别间不具有大小关系的特征，每个特征取值对应一维特征，能够处理缺失值，在一定程度上也起到了扩充特征的作用。但是当类别的数量很多时，特征空间会变得非常大，需要进行降维或分箱。	
序列型	腾讯对序列特征的处理方案	针对每个序列先对其 Embedding 化，然后分别进行 Max-Poolling 和 DIN-Attention，并最终进行 concat，作为序列建模层的输出。 
特征Embedding	特征embedding化的方案	对于每一个特征，都先将其 embedding 化，然后乘上该特征的 value 或 weight（对于连续特征是 value，对于离散标签特征是标签权重 weight）。最后可以将其 concat 到一起，或单独通过其他网络结构。

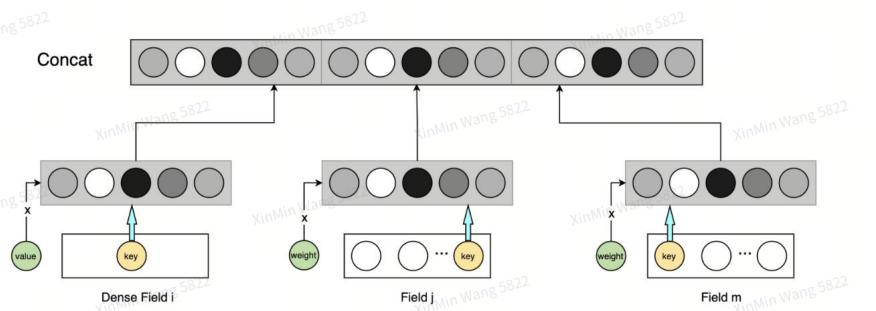
其他

鑫民 XinMin Wang Jan 30, 2023

双路并行

- wide 主要用作学习样本中特征的共现性，产生的结果是和用户有过直接行为的 item，通过少量的交叉特征转换来补充 deep 的弱点。wide 侧用于记忆，适合输入组合特征，用于记住那些已经存在过的特征组合。总结就是高维稀疏+常用交叉特征。
- deep 部分是一个前馈神经网络。通常情况下这部分特征使用的是用户行为特征，用于学习历史数据中不存在的特征组合。deep 侧用于泛化，适合输入非组合特征，包括离散特征和连续特征，用于泛化那些未曾出现过或者低频的特征组合。总结就是：deep 全用。

Expand



4.3 特征工程和模型的结合

- LR：LR的本质是评分卡模型，适合高维稀疏特征，对于离散型特征：类别多的使用哈希分桶（哈希大小为2～5），类别少的直接onthot；对于连续特征：对于分布基本均匀的使用log，不是很均匀的使用等频分桶。
- GBDT：GBDT的基础决策树，GBDT是适合处理高维稀疏特征，对于离散型特征：类别多的特征基本不用，类别少的直接labelencoder处理；对于连续特征：使用等频分桶或者直接使用原值。
- Embedding+MLP：本质是DNN，使用方式同LR
- 其他

5. 总结

本文作为【CVR模型特征工程优化】项目的方案调研，介绍了：

- 搜广推场景“为什么要精做特征工程”，
- “什么是好的特征工程”
- 结合各厂在不同场景的特征工程构造经验，设计了相对优化且适应当前数据开发的传音广告场景CVR模型特征工程特征体系
- 对特征工程中常见的原始数据处理问题，结合传音广告场景的数据处理相关问题，提出数据缩放、异常值处理、分桶、特征冗余共线性、数据平滑等相关问题的技术方案

“Garbage In, Garbage Out”特征工程处于机器学习流水线的上游位置，处理结果的好坏关系到后续模型的效果。特征工程不仅与模型算法相关，与实际问题更是强相关的，针对不同场景，特征工程所用的方法可能相差较大。在实际的工程应用中，需要深入理解数据和业务逻辑以及模型特点，才能更好地进行特征工程。

6. Q&A

Q1：手动做特征工程相比于XGBoost、LightGBM等可以自动分箱的方法有什么优点和缺点？

A1：XGBoost这类树模型是有监督分箱中基于决策树的方法。这类基于GBDT的模型很难处理高维度的特征，当特征维度比较高的时候，它的运行效率是比较低的。我们可以通过这类模型精选一些特征，然后再把这些特征输入给其他模型，如深度神经网络，LR等，相当于利用GBDT作为一个特征工程来初步提取特征。

Q2：特征工程这类经验性知识怎么获取和学习？

A2：学习特征工程的知识还是要深入理解业务。比如在某个场景下，我们认为年龄是一个比较重要的特征，但是现阶段用户画像里没有这个特征，那么我们首先要想办法通过一些现有特征，如用户的购买行为将年龄值预测出来。虽然预测出的年龄值与真实年龄值存在差异，但用户的真实年龄并不重要，重要的是他在平台上购物倾向所体现出的“购物年龄”，我们后续只会针对他的“购物年龄”做一个预测。此外，和有经验的人或团队交流也是可以收获很多特征工程经验的。

Q3：增加item侧行为特征是否会加重马太效应？

A3： 在实际应用中，增加item行为特征对模型AUC的提升起着至关重要的作用。只要注意特征穿越问题，item的效率特征能很好地体现商品的热度特征，进而得到较优的排序结果。

Q4： 统计类需要经过额外处理再作为模型的输入吗？

A4： 排序模型的输入是经过embedding化的特征，所以，统计类特征可以经过等频或者等宽离散化，经过一个embedding层得到输入特征。

7. 参考资料

1. 推荐算法中的特征工程
2. 深度学习在美团搜索广告排序的应用实践
3. 一文彻底搞懂 CTR 建模
4. 浅谈微视推荐系统中的特征工程
5. 王子一：飞猪稀疏高客单场景下的CVR优化实践
6. 推荐广告算法中的特征
7. 刀功：谈推荐系统特征工程中的几个高级技巧
8. 如何在工业界优化点击率预估:（三）特征

✅ **搜广推场景的特征工程介绍到此结束，由于本人水平有限，编写时间紧张，文中难免会出现一些错误或疏漏，恳请大家批评指正。在实际算法模型应用中，还涉及特征选择、特征重要性评估、样本采样优化等问题，后续将对此类问题及解决方案一一阐述。**