特征工程｜特征设计、特征可用性评估

原创 Thinkgamer 搜索与推荐Wiki 2020-07-17

收录于话题

#特征工程 92 #Thinkgamer 11 #精品小系列内容 25

特征设计是整个特征工程中最消耗时间的一步，但其却也是十分重要的一步。不同领域的特征在设计上也会有所区别（当然也存在相同的地方），这时候需要结合领域知识进行思考和总结，制定适合自己业务的特征，然后可以进行相关的特征可用性评估，决定开发哪些基础的特征。

业务理解

在具体实践中，我们使用的特征通常分为四大维度：

- 用户维度特征
- 事物维度特征
- 类别维度特征
- 组合属性特征

对于用户维度的特征，其可以划分为：

- 客观属性：即事物本身的属性，不因人的看法而改变
- 主观属性：和客观属性相对，可以理解为人对事物的看法

用户维度的客观属性在各个平台的使用方式和制定方式几乎都是一样的，因为这些属性伴随事物本身存在，比如电商平台中商品的价格、品牌、类别等，不会说因为换了一个电商平台，白色鞋子就会变成黑色的。但其主观属性在不同平台内理解是不一样的，比如用户的偏好属性，每个平台都会有自己偏好的定义和理解，这时候就要结合具体的业务进行判断和分析。比如用户对类别的偏好在电商平台中可以根据用户在类别下的行为统计进行挖掘，得到对用户对类别的不同偏好程度，但是在保险类平台，由于保险个数少，类别少，这时候其实并不需要关心用户对哪个类别的保险产品感兴趣，只需要根据用户的身体健康状态去匹配相应的保险产品即可。

事物维度的属性也可以分为客观属性和主观属性，其客观属性的使用方式是相同的，但其主观属性则会有区别，比如同一篇文章会因为其侧重点有所不同，在不同的内容平台上就会将其分为不同的类别。

类别维度特征泛指所有类别、比如电商平台中的品类、店铺、品牌等，内容平台的标签、作者等，视频平台的演员、导演等。在具体使用过程中可以根据类别做汇聚，计作该类别下的特征。比如视频平台中计算用户对某部电影的喜好的时候，类别特征可以是用户对该电影导演的下的观看电影次数、点赞电影次数、评论电影次数等。

对于组合属性特征相对来讲就比较灵活了，可以做一切我们认为有关系的组合属性特征，比如用户与类别交叉，用户与事物交叉，性别与类别交叉等。同样，组合属性特征在不同的平台使用方式也是不一样的，除了一些简单的数值型特征，也可以做很多特征处理的工作，而不同的特征处理方式带来的效果也是不一样的。比如说对于电商平台，构建用户与事物之间的交互维度特征无非就是各种行为次数，但是像小说平台，除了次数这些特征之外，我们还需要考虑的是阅读的深度，阅读的连续性等，需要结合小说内容的形态进行特征的理解和设计。

再比如一些短视频平台，在构建特征的时候要考虑事物的热度特征和传播性特征，比如我们经常刷到时下的一个热度事件或者热度任务等，但是对于电商平台，热度特征和传播性特征其实就没有那么重视，同样在短视频平台中用户的社交关系特征也比较明显，比如我经常刷到我认识的人（但存在我没有关注的人）的视频，可能想让大家有共同交流的机会，这就要去问负责相应业务的人了，但是在一些其他内容平台上，这种社交关系特征就会被弱化很多。

笔者另外一个感触比较深的是，数据中隐含的一些特征往往需要进行深入的分析才能得到。比如不同人群的偏好倾向、使用平台的习惯等、特征人群对于特定类别的事物点击率有高低之分等。这些特征在个性化推送、用户冷启动、feed流推荐等模块都会产生很大的影响。

因此在设计我们建模所需要特征的时候，一定要结合业务和最终的目标导向制定最适合的特征，同时也需要算法工作者对数据有极强的敏感性，且具备基本的数据分析能力，便于设计特征和后期模型的分析。

特征可用性评估

在确定好使用哪些特征之后，并不是全都开发，而是需要先进行初步的特征可用性评估，可以分成几个方面来考虑：

- 获取这批数据的难易程度，比方说有的数据非常隐私，这批数据获得的难度就很大，因为在获取之后可能会带来用户的投诉和一些法律上的责任。
- 其次就是这批数据的覆盖率。比方说要构造某个年龄的特征，那么这些用户中具有年龄特征的比例是多少就是一个关键的指标。如果覆盖率低，那么最后做出的特征可以影响的用户数量就会有限制。如果覆盖率高，那么年龄特征做得好，对最后的模型训练结果都会有一个明显的提升。
- 再就是这批数据的准确率，因为从网上或者其他地方获取的数据，会由于各种各样的因素（用户的因素，数据上报的因素）导致数据不能够完整的反映真实的情况。这个时候就需要事先对这批数据的准确性作出评估。

因此设计完特征不要着急进行开发，进行初步的判断则可以避免开发过程中踏入各种坑而无法自拔。

> 点击【阅读原文】发现更多精彩！

The end

▼ 往期精彩回顾 ▼

特征工程｜数据的分类、特征工程的定义、意义和应用