Project1 移动推荐算法报告

一 设计题目及其目的：

在真实的业务场景下，我们往往需要对所有商品的一个子集构建个性化推荐模型。基于移动端用户对全网商品的行为数据构造面向生活服务行业的推荐模型。形式化定义 –U---用户集合 –I---商品全集 –P---商品子集。

第一部分是用户在商品全集上的移动端行为数据, 表名tianchi\_fresh\_comp\_train\_user\_2w，包含如下字段：



第二个部分是商品子集（P）,表名为tianchi\_fresh\_comp\_train\_item\_2w，包含如下字段：



目标是利用D来构造U中用户对P中商品的推荐（购买预测）模型，从而得到用户在接下来一天对商品子集购买行为的预测结果。采用经典的精确度(precision)、召回率(recall)和F1值作为评估指标。

二 设计内容：

总体算法流程分为三部分：

1.数据集分割和预处理：

对于线下模拟，我们用从11月18日到12月15日行为数据作为知识集，用来提取相关特征，然后利用12月16日真实的情况来给特征集打上label，如对于某个用户商品对，如果用户在12月16日真实购买了该商品，则在特征表的该UI对最后一个字段赋值1，这样我们就得到了一个训练集。然后我们再从11月18日到12月16日的行为数据中提取相关特征，得到一个12月17日的测试集。为了保证训练的模型具有稳定性，我们还需要一个验证集作对比，即从11月18日到12月17日提取特征，得到一个12月18日的验证集。如果用12月16日训练集得到模型，在测试集和验证集上得到的F1值差不多(相差在0.02%左右)，则说明这个模型是可靠的，同时也表明这些特征构建也是有效的。然后我们可以利用11月18日到12月17日的数据来构建特征，以12月18日的真实购买情况来打上label，得到一个线上的训练集，用该训练集训练出来的模型来预测12月19日的购买情况(以11.18~12.18作知识集提取特征)。

数据预处理包括:1.异常点剔除 2.缺失值填充

1. 特征构建：
2. 用户特征:

用户特征就只是针对用户来说的，反映的是用户整个购物习惯与购物规律，与具体哪件商品无关，比如用户是不是喜欢浏览购物网站、用户的购物频率等。

1. 商品特征:

商品特征反映的是商品本身的品质或者受欢迎程度如何，而与具体哪一个用户没有关系，同时商品特征也体现了商品的活动规律，即被用户购买的频率、最后有人购买的时间等。

1. 协同特征：

协同特征则是以<用户，商品>作为统计对象，是用来表现某个用户对某件商品的喜爱程度或是购买的可能性，这一部分特征直接与测试集数据对接，对预测结果起着决定性的作用。

1. 模型与调参

1. 模型选择

随机森林（Random Forest）

随机森林，是用随机的方式建立一个森林，森林里面有很多的决策树组成，随机森林的每一棵决策树之间是没有关联的。在得到森林之后，当有一个新的输入样本进入的时候，就让森林中的每一棵决策树分别进行一下判断，看看这个样本应该属于哪一类（对于分类算法），然后看看哪一类被选择最多，就预测这个样本为那一类。

在建立每一棵决策树的过程中，有两点需要注意：采样与完全分裂。首先是两个随机采样的过程，random forest对输入的数据要进行、列的采样。对于行采样，采用有放回的方式，也就是在采样得到的样本集合中，可能有重复的样本。假设输入样本为N个，那么采样的样本也为N个。这样使得在训练的时候，每一棵树的输入样本都不是全部的样本，使得相对不容易出现over-fitting。然后进行列采样，从M个feature中，选择m个(m << M)。之后就是对采样之后的数据使用完全分裂的方式建立出决策树，这样决策树的某一个叶子节点要么是无法继续分裂的，要么里面的所有样本的都是指向的同一个分类。

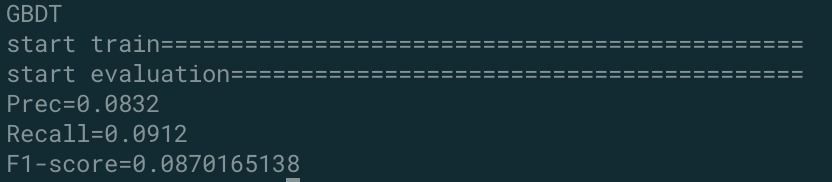
GBDT（Gradient Boosting Decision Tree）

迭代决策树在我理解更像是一种数据的串行处理，对于每一个训练样本，一棵决策树在进行处理后将其交给下一个决策树，其核心就在于每一棵树学的是之前所有树的结论和的残差，这个残差就是一个加预测值后能得真实值的累加量。

2.调参

1. 每个节点使用的特征数量
2. 树的深度
3. 最小叶子节点数目

三：结果



四：总结

在构建特征时要主要保证训练集和预测集上的特征分布一致，不然可能会造成线上线下偏差很大

通过调参使得准确率与召回率更加均衡。