LR原理与实践

# 逻辑回归模型

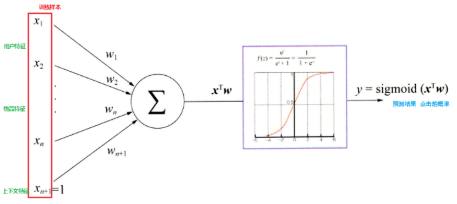
逻辑回归模型非常重要，在推荐领域里面，相比于传统的协同过滤，逻辑回归模型能够综合利用用户、物品、上下文等多种不同的特征生成较为“全面”的推荐结果，关于逻辑回归的更多细节， 可以参考下面给出的链接，这里只介绍比较重要的一些细节和在推荐中的应用。

逻辑回归是在线性回归的基础上加了一个 Sigmoid 函数（非线形）映射，使得逻辑回归成为了一个优秀的分类算法，学习逻辑回归模型，首先应该记住一句话：逻辑回归假设数据服从伯努利分布，通过极大化似然函数的方法，运用梯度下降来求解参数，来达到将数据二分类的目的。

相比于协同过滤和矩阵分解利用用户的物品“相似度”进行推荐，逻辑回归模型将问题看成了一个分类问题，通过预测正样本的概率对物品进行排序。这里的正样本可以是用户“点击”了某个商品或者“观看”了某个视频，均是推荐系统希望用户产生的“正反馈”行为， 因此逻辑回归模型将推荐问题转化成了一个点击率预估问题。而点击率预测就是一个典型的二分类，正好适合逻辑回归进行处理，那么逻辑回归是如何做推荐的呢？过程如下：

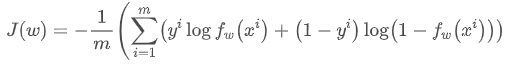
1. 将用户年龄、性别、物品属性、物品描述、当前时间、当前地点等特征转成数值型向量
2. 确定逻辑回归的优化目标，比如把点击率预测转换成二分类问题，这样就可以得到分类问题常用的损失作为目标，训练模型
3. 在预测的时候，将特征向量输入模型产生预测，得到用户“点击”物品的概率
4. 利用点击概率对候选物品排序，得到推荐列表

推断过程可以用下图来表示：



这里的关键就是每个特征的权重参数w，我们一般是使用梯度下降的方式，首先会先随机初始化参数w， 然后将特征向量（也就是我们上面数值化出来的特征）输入到模型，就会通过计算得到模型的预测概率，然后通过对目标函数求导得到每个的梯度w，然后进行更新w。

这里的目标函数长下面这样：



求导之后的方式长这样：



这样通过若干次迭代，就可以得到最终的了。

# LR优缺点

**优点**

* LR模型形式简单，可解释性好，从特征的权重可以看到不同的特征对最后结果的影响。
* 训练时便于并行化，在预测时只需要对特征进行线性加权，所以**性能比较好**，往往**适合处理海量id类特征**，用id类特征有一个很重要的好处，就是**防止信息损失**（相对于范化的 CTR 特征），对于头部资源会有更细致的描述。
* 资源占用小，尤其是内存。在实际的工程应用中只需要存储权重比较大的特征及特征对应的权重。
* 方便输出结果调整。逻辑回归可以很方便的得到最后的分类结果，因为输出的是每个样本的概率分数，我们可以很容易的对这些概率分数进行cutoff，也就是划分阈值(大于某个阈值的是一类，小于某个阈值的是一类)

**缺点**

* 表达能力不强，无法进行特征交叉，特征筛选等一系列“高级“操作（这些工作都得人工来干，这样就需要一定的经验，否则会走一些弯路），因此可能造成信息的损失
* 准确率并不是很高。因为这毕竟是一个线性模型加了个sigmoid，形式非常的简单(非常类似线性模型)，很难去拟合数据的真实分布
* 处理非线性数据较麻烦。逻辑回归在不引入其他方法的情况下，只能处理线性可分的数据， 如果想处理非线性，首先对连续特征的处理需要先进行**离散化**（离散化的目的是为了引入非线性），如上文所说，人工分桶的方式会引入多种问题。
* LR 需要进行**人工特征组合**，这就需要开发者有非常丰富的领域经验，才能不走弯路。这样的模型迁移起来比较困难，换一个领域又需要重新进行大量的特征工程。

所以如何**自动发现有效的特征**、**特征组合**，**弥补人工经验不足**，**缩短LR特征实验周期**，是亟需解决的问题，而GBDT模型，正好可以自动发现特征并进行有效组合。

LR模型有以下特点：

* 计算复杂度低；
* 易于并行化处理；
* 易于得到离散化目标值0或1，利用sigmoid函数将传统线性模型的输出值映射到(0,1)区间；
* 学习能力限于线性特征，需要提前进行大量的特征工程得到有效的特征及特征组合。

参考：

<https://mp.weixin.qq.com/s/XP5z_BEeFr6oJp9VmVJRqQ>