Instacart推荐商品预测

张 晗

1 介绍

商品推荐功能是零售网站及App广泛使用的营销方式，系统通过记录分析用户偏好，预测用户较为感兴趣或较可能购买的商品并推荐给用户。准确的推荐预测会带来更好的购物体验，并且在一定程度上提高用户的商品购买率及点击率。这份报告通过在线超市购物应用程序Instacart提供的用户历史消费数据，预测在未来订购中用户最可能重复购买的商品。

2 数据集描述

Instacart的数据来自约20万用户的300万次订购记录，数据完整无缺失值。所有数据由数张相互关联的表构成，用户订单表、订单明细表将提供主要的订购相关信息。订单明细表的内容包括：订单ID、购买商品ID、添加到购物车的顺序、该商品是否属于再次订购，形式如下：

订单明细表			
订单号	商品编号	购买顺序	是否曾购买过
572	333	3	1
572	602	2	0

用户订单表的内容包括：用户ID、订单ID、用户下单顺序、在一周内的哪一天订购、在几点订购、本次订购距离前一次订购的间隔（按天计算），用户订单表的形式如下：

用户订单表					
用户号	订单号	序号	时间1	时间2	间隔
32	570	1	周一	10点	3天
32	571	2	周二	12点	2天
32	572	3	周六	20点	5天

另外三张表分别为商品明细列表、货架区域表以及商品所属部门表，后两者仅含名称与编号的对应，商品明细表的形式如下：

商品列表			
商品编号	商品名称	区域编号	部门编号
6	椰子汁	61	7
7	巧克力饼干	84	19
8	乌龙茶	127	7

总体上用户购买的商品较为分散，下图为用户订购总次数以及商品被购买总次数的密度分布，从中可以看到大多数用户只有30次以下的订购记录，而大多数商品总购买次数不超过20次，有大量商品只有5次以下的购买记录。分布有很长的尾部，导致有可能对于部分用户没有足够的历史数据判断其购物习惯和偏好。

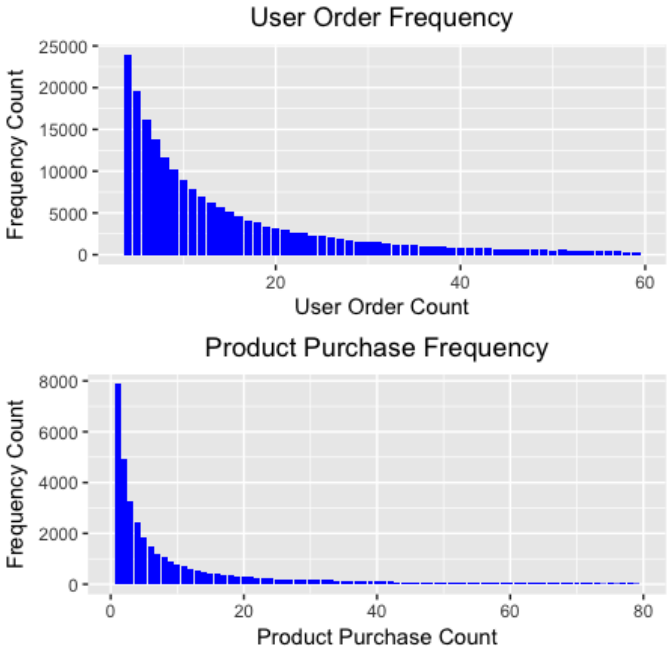

Fig 1: 用户订购次数与商品购买次数频率

在用户和商品特征方面，Instacart数据集除提供用户订购商品明细、下单顺序及下单时间之外，未提供太多与商品及用户相关的信息，所以

需要从现有数据中提炼构造出商品或用户独有的特征。下一节将介绍构造用户及商品特征变量的方法。

3 特征工程

由于预测目标为用户下次会购买哪些曾购买过的商品，特征变量应主要与商品和用户属性相关，主要考虑从以下四个方面构造特征变量：商品特征相关变量、用户特征相关变量、用户与商品交互作用特征变量、用户购物模式的频率统计。下面将对部分特征变量的意义进行说明：

1. 商品再订购率（商品特征/用户商品交互特征）：此特征可以作为商品单独特征，也可以作为用户商品交互特征。商品再订购率为商品在首次订购之后的购买次数与在首次购买后的可能被购买次数的比率。例如用户203和用户312对椰子汁的购买记录如下：

订单序号	1	2	3	4	5	6	7	8
用户203	0	1	0	0	1	无	无	无
用户312	1	0	0	1	0	0	0	0

如果作为物品特征，椰子汁再订购率为$\frac{2}{9}$。如果作为交互特征，用户203对椰子汁的再订购率为$\frac{1}{3}$，用户312的再订购率为$\frac{1}{6}$。再订购率在一定程度上反映了商品被多次订购的概率，部分商品可能因为质量问题或日常使用率很低导致难以被再次购买，再订购率可以反应这一特性。

2. 物品遗忘间隔（用户物品交互特征）：为用户对商品最后一次购买距离现在间隔与用户订单之间的最大间隔的差。如用户共有五次购物记录，其中最长的购物间隔为10天，最后一次购买樱桃的时间距今有30天，则物品遗忘间隔为30-10=20天。遗忘间隔越大，说明用户现在对商品的购买欲望越弱。

用户	商品	最近购买时间	最大购物间隔
375	气泡酒	14天前	10天
375	芒果汁	10天前	15天
375	矿泉水	7天前	6天

图二为被重复订购的商品及未被重复订购商品的遗忘间隔分布。两种情况下分布的形状较为类似，会被重复订购商品的遗忘间隔均值大约会相对缩短20天左右。

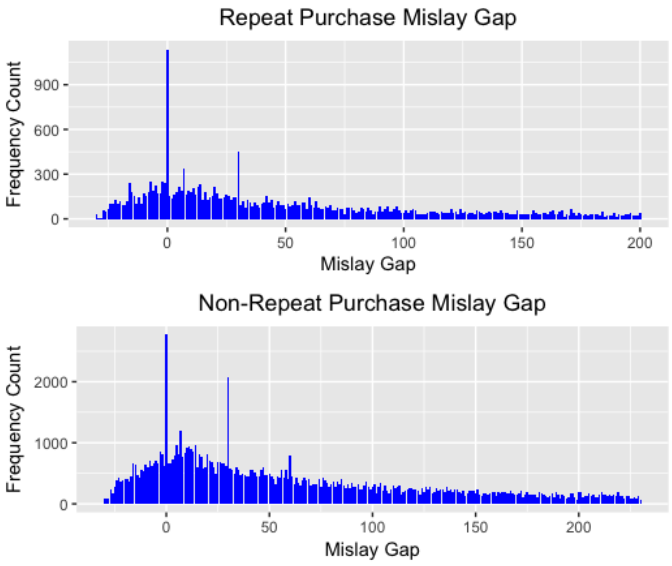

Fig 2: 商品遗忘间隔频率

3. 用户单品最大购买次数（用户特征）：用户特征变量。例如在该用户的所有购买记录中，购买芒果5次、巧克力饼干2次、乌龙茶3次，该用户的最大单品购买次数为5。单品购买次数越大，则该用户未来订单中含有重复商品的概率越大。

用户编号	商品	购买次数
761	芒果	5
761	巧克力饼干	2
761	乌龙茶	3

4. 用户平均订购间隔（用户特征）：例如某用户共进行四次订购，相邻两次订购的间隔为3天、10天、6天，则该用户的平均订购间隔为$\frac{19}{3}$。该间隔越短意味购物越频繁，同时也在一定程度上说明该用户更倾向于购买使用周期较短的商品，即更容易对商品进行重复购买。

5. 用户单次购买商品平均数量（用户商品交互特征）：某用户购物三次，购买的商品数为5个、2个、6个，则该用户的单次购物平均数量为$\frac{13}{3}$，单次购买的商品越多则越有可能出现重复购买的商品。

6. 用户前一次订购单是否包被重复购买的商品（频率统计）：根据Instacart提供数据的统计，如果前一次订单不包含重订购的商品，下一次订单同样也不包含重订购的商品的概率较小（小于35%）。

4 模型训练及误差评估

4.1 模型训练

目标是预测用户在下一次购物中会购买哪些以往购买过的商品，预测结果可以是空集，即该用户不会对任何商品进行重新购买；如果结果不是空集，则逐一预测用户对每一个曾经购买过商品的重新购买概率。例如用户375的购买记录为：

用户	订单号	下单顺序	商品
375	107	1	乌龙茶
375	107	1	芒果汁
375	107	1	巧克力饼干
375	129	2	玉米饼
375	129	2	芒果汁
375	129	2	樱桃
375	12	3	巧克力饼干
375	12	3	芒果汁

模型主要思路是将每个用户的消费记录按订单顺序归为一组，然后以用户编号和商品编号为主键，预测该用户对其购买过的每一个商品的再订购率，对应预测模型的形式为：

用户	商品	是否会重新购买
375	乌龙茶	是/否
375	芒果汁	是/否
375	巧克力饼干	是/否
375	樱桃	是/否
375	玉米饼	是/否

实际处理中随机抽取了10%的用户数据进行作为训练数据，将用户除最后一次订购外的数据作为历史购买记录，预测最后一次订购是否会购买以往的商品，使用最后一次实际购买记录来验证预测结果。因为绝大多数特征变量是数值型变量，且预测目标变量都为二元变量，可以考虑使用逻辑回归或者决策树。决策树的优点是可以较好抵抗极值、无关特征变量等不规则数据的影响。实际训练中使用的是R中的gbm[1, 2]模型，下面是1000次迭代后，生成模型的特征变量相对重要性对比：

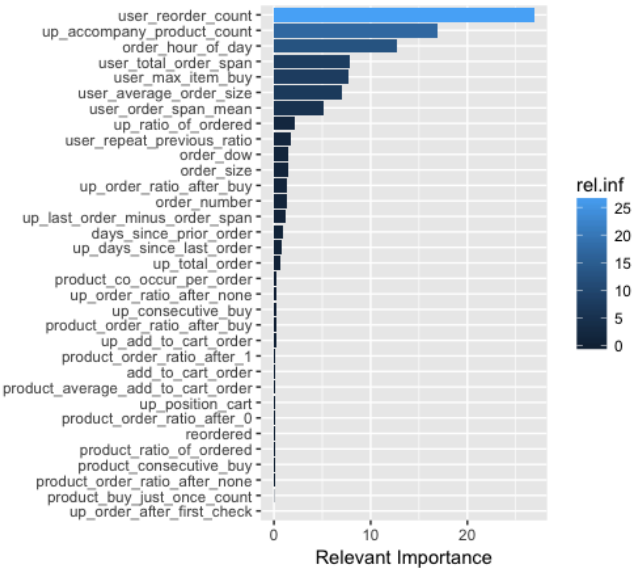

Fig 3: 特征变量相对重要性对比

图三显示了特征变量的相对重要性，其中对预测结果（是否重复购买）影响较大的变量包括：

- 用户购物的总时间跨度
- 用户重复订购率：例如用户用有7次订购，其中3次含有重复订购的商品，用户重复订购率为3/7
- 用户购买记录中重复订购的总次数
- 用户对特定商品的重订购次数
- 用户单品做大购买次数
- 用户对特定商品的购买总次数
- 用户单次购买商品平均数量
- 用户对特定商品的重复订购率
- 用户对特定商品在未购买/已购买情况下的购买率
- 用户对商品的遗忘间隔
- 最后一次购买商品距离现在的间隔
- 用户连续两次购买特定商品的次数

其中相对影响力最大的特征变量全部为用户特征变量，其次是用户与商品的交互特征变量，仅与商品相关特征变量对预测结果的影响力明显

弱于其他两种类型的特征变量。特征相对重要性排序不能绝对反映变量的重要程度，但在是否会重新购买以往商品的行为中，用户自身的特征以及用户自身对商品的感受起到了决定性的作用，而与商品独立于用户之外特性的关联较小。例如面包作为一个重复订购率较高的商品，但由于某用户购物频率很低或该用户对面包喜爱程度较低，依然会导致其对面包的购买率及重复购买率偏低。

4.2 误差评估

相比于用错误分类率作为预测精度的度量，使用F1-score能够更加全面的评估预测结果的准确性，F1-score通过结合预测的准确率和召回率来整体评估预测结果，其中准确率(Precision) 表示预测为购买的商品中实际被购买的比率，召回率(Recall)表示预测被购买的商品占所有将被购买商品的比率。F1-score的详细计算方式如下:

		实际购买 1	实际未购买 0
预测购买	1	true positive	false positive
预测未购买	0	false negative	true negative

$$Precision = \frac{True\ Positive}{True\ Positive+False\ Positive}$$

$$Recall = \frac{True\ Positive}{True\ Positive+False\ Negative}$$

$$F1 = \frac{2\times Precision\times Recall}{Precision+Recall}$$

使用gbm模型预测二元分布（bernoulli）的变量时，预测结果为商品被购买的概率，当通过预测概率决定商品是否被购买时，需要选定概率阀值，例如预测概率大于0.3的商品设定为将被购买。但因为误差的度量为F1-score，如果想要最小化F1-score，对于每个用户需要根据商品的预测概率结果设定不同的概率阀值。假定对于用户需要预测是否购买的商品为樱桃和芒果，下面将通过这些商品不同的两组预测概率来计算最大化相应F1-score 的阀值。

假设樱桃的预测购买概率为0.9，芒果的预测购买概率为0.3，下面分别计算只购买樱桃，只购买芒果及两者都被购买时的F1-score。

	只买樱桃		只买芒果		两者都买	
概率	F1	期望	F1	期望	F1	期望
0.63	1	0.63	0	0	0.66	0.42
0.03	0	0	1	0.3	0.66	0.02
0.27	0.67	0.18	0.67	0.18	1	0.27
0.07	0	0	0	0	0	0
总期望	0.81		0.21		0.71	

当樱桃和芒果的购买预测概率为0.9、0.3时，最大化F1-score的方式是只购买樱桃（F1期望为0.81），所以需要将概率阀值设置为大于0.3以过滤掉芒果。

	只买樱桃		只买芒果		两者都买	
概率	F1	期望	F1	期望	F1	期望
0.24	1	0.24	0	0	0.66	0.16
0.14	0	0	1	0.14	0.66	0.09
0.06	0.67	0.04	0.67	0.04	1	0.06
0.56	0	0	0	0	0	0
总期望	0.28		0.18		0.31	

当樱桃和乌龙茶的购买预测概率为0.3、0.2时，最大化F1-score的方式是同时购买樱桃和乌龙茶（F1期望为0.31），所以需要将概率阀值设置为小于0.2 以保证两者都能被选中。以上的方式将每种情况下的F1期望值都计算出来，可以精确每种组合下F1-score的期望，但计算量随用户购买的商品数量成指数型增长，即所有组合的数量为2^N，N为用户曾购买的总商品数。当N较大时，比如用户曾经购买的商品超过30个，则$2^{30} \approx 10^9$，此时对于单个用户的计算量已经过大。为了减少计算量，可以牺牲部分精度，通过物品被购买的概率，随机生成m个被商品购买列表，例如用户曾经购买过五个商品，随机生成的商品购买列表的形式为$[1,1,0,1,0]$，0代表未购买，1代表购买。可以根据用户曾购买的商品数量适当增加或减小m值。

假设筛选概率阀值低于购买概率最低的商品，此时所有商品都会被重复购买，计算此种情况下随机生成的被购买商品列表F1-score平均值（模拟F1-score的期望）；接着增加筛选概率以过滤掉购买率最低的商品，并依照上面的方法再次计算F1-score平均值，一旦F1-score均值比上一次低则停止计算。即每次去掉当下预测购买列表中购买率最低商品并计算F1-score的期望，当期望小于上一次结果时，则不再删除任何列表

中的商品并将此列表最为最终预测结果。该方法通过牺牲掉部分F1期望的精度，使计算效率达到了接近线性的效果。最终生成的预测结果的F1-score约为0.42。

5 结论

这篇报告主要阐述了用户特征、商品特征及其交互作用特征的构造，并通过生成的模型检验哪些变量能够对预测结果产生较大影响，经过这一过程的循环，获取对目标变量预测影响较大特征并改进预测精度。在误差评估的中，由于直接设定概率阀值筛选商品生成预测结果的方式并不适合F1-score误差度量，实际中使用了一种近似计算F1-score期望的方法来辅助设定筛选商品的概率阀值，使得对预测误差的评估能够更为合理准确。在对数据集进行训练时仅应用了决策树模型，如果在训练时能够使用多种模型（如逻辑回归或SVM）并对模型的预测结果进行适当融合，应该能够在一定程度上进一步降低误差并提高预测精度。

参考文献

[1] J.H. Friedman (2001) "Greedy Function Approximation: A Gradient Boosting Machine," *Annuals of Statistics* 29(5) : 1189 – 1232.

[2] G. Ridgeway (1999) "The State of Boosting," *Computing Science and Statistics* 31 : 172 – 181.