1. 推荐系统评测

评价一个推荐系统的优劣，怎么样的推荐系统是我们所需要的，目前还没有一个统一的标准，但是推荐系统终将被应用到实践之中，即在商业项目中投入使用，因此用户的满意度是推荐系统最佳的评价，但客户满意度需要我们投入到商业应用中去，其中消耗的人力物力较大，并且具有很大的风险性，而且用户满意度是无法量化到用数值区评价，常规的推荐系统的评价流程为：离线实验（offline experiment）、用户调查（user study）和在线实验（online experiment），测试的范围从小到大，逐步规避调风险。

1）离线实验

离线实验通常采用数据集进行测试、常用于开发阶段：

a）获得用户数据，例如用户的购买记录、播放记录等，需要收集测试数据集合

b）将数据集按照一定的规律分成训练集和测试集；

c）在训练集上训练推荐模型并利用测试集进行预测；

d）利用事先定义的离线指标评测算法在测试集上的预测结果。

e）对于推荐算法进行改进

离线实验的优点是：

不需要用户参与成本低

利用离线数据验证、时间短、效率高

可以无人值守，时间地点方式灵活

离线实验的缺点是：

受到数据集合的局限性、适用范围的影响，例如一个数据集中没有包含某用户的历史行为，则无法评价对该用户的推荐结果，一旦数据集不准确就无法为用户准确预测

对于评价结果缺乏主观性，无法考虑到人的主观思想变化，缺乏灵活性

难以找到离线评价指标和在线真实反馈(如搜索率、点击率、转化率、点击次数、时长、购买客单价、购买商 品类别等)之间的关联关系

对于用户的变化无法做到及时响应，例如人突然转变风格等

2）用户调查

用户调查需要真实的用户在待测推荐系统上完成任务，并且观察和记录用户的行为以及回答一些问题，分析用户的行为和答案，了解测试系统的性能。

用户调查的优点是：

可以获得用户主观感受的指标，对于用户的主管想法进行记录和考究

测评范围小，造成风险的影响范围较小，风险易控制

用户调查的缺点是：

对于测试用户招募消耗巨大；

无法组织完全覆盖的测试用户，统计意义不足；

3）在线实验

通过离线实验和用户调查测评之后，可以将系统上线做AB测试，即将新的推荐系统与旧的推荐系统共同投入使用进行测试，通过两个推荐系统的对比进行测评，并且减少风险。

在线实验通过一定的规则将用户随机分成几组，对不同组的用户采用不同的推荐算法，然后通过统计不同组的评测指标，比较不同算法之间的优劣，其的核心思想是:

a) 控制变量多方案共同运行

b) 比较各个方案测试结果

c) 以某种规则进行修改优化或者是淘汰

在推荐系统的评价中，唯一变量就是推荐算法的选择。

1. 评测算法的选择

由于条件的限制，本次采用离线测评，即根据推荐系统在离线数据集合中的表现，基于一些离线评估指标来衡量推荐系统的效果，相比较在线系统，离线系统更加经济、方便，利用验证数据集来测试推荐系统的效果。但是评价效果有限，无法和真实用户表现来模拟。

本论文采用的评估环节有两个：拆分数据集、选择离线评估指标。

a）数据集拆分

在机器学习过程中，通常需要将数据集拆分为训练数据集、验证数据集，测试数据集合。它们的功能分别如下。

训练数据集（Train Dataset）：用来进行机器学习模型的构造。

验证数据集（Validation Dataset）：辅助模型构造，评估模型。

测试数据集（Test Dataset）：评估训练完成的模型各项指标。

本论文推荐系统采用协同过滤推荐算法，暂不涉及机器学习应用，因此将数据集合随机分为训练数据集合（Train DataSet）、测试数据集合（Test DataSet）

b）离线评测指标的选择

在分类模型评判的指标中，常见有1.混淆矩阵 2. ROC曲线 3. AUC面积 进行评判。

1. 准确率评价指标

混淆矩阵（也称误差矩阵，Confusion Matrix）：混淆矩阵是评判模型结果的指标，属于评估模型的一种，多用于判断分类器（Classifier）的优劣，在预测分类实验中将真实结果真假、预测结果真假分不到2x2的表格中进行分析，在表格里可以清晰的看到每个类别识别正确的数量和识别错误的数量。

|  |  |  |
| --- | --- | --- |
| 真实结果 | 预测结果 | |
|  | 真 | 假 |
| 真 | True Positive | True Negative |
| 假 | False Positive | False Negative |

混淆矩阵的指标：

TP（True Positive）：将真值预测为真，真正类，正确预测

TN（True Negative）: 将真值预测为假，伪假类，错误预测

FP（False Positive）: 将假值预测为真，伪真类，错误预测

FN（False Negative）：将假值预测为假，真假类，正确预测

根据对基础指标的归类得到了二级指标进行结果的衡量：

准确率（Accuracy）—— 针对整个模型：预测正确

精确率（Precision）

灵敏度（Sensitivity）：就是召回率（Recall）

特异度（Specificity）

例如：

精确率：

召回率

准确率

AUC

ROC

FPR

TPR

Fscore

1. 评分预测评价指标

平均绝对误差

均方误差

均方根误差

预测评分关联指标

皮尔逊积距相关

斯皮尔曼等级相关

肯德尔等级相关

预测排序准确性

评价排序分