推荐系统评估方法笔记

评价方法分为以下三种：

- Offline experiments，线下测试
- User studies，用户调查
- Online experiments，线上测试

我们使用的基本上是 Offline 方法。

实际上线时，需要综合多种评价标准（准确度，可信度，覆盖度等等）来衡量推荐系统的好坏，但是针对推荐算法，最重要的指标是推荐/预测的准确度（Accuracy）。

线下测试的基本准则：

- Hypothesis，测试前需要做出假定，比如：将使用的算法 A 比已有的算法 B 好
- Controlling variables，控制变量，即除假定要求的变量外，其他变量保持一致，比如测试算法好坏，要使用相同数据集
- Generalization power，得出的结论要有一定的可推广性，不限于特定的测试用例

线下测试的基本方法：

1. 选定数据集
   选择和应用相关的数据集。数据需要是无偏的（unbiased），通过随机抽样能满足要求。
2. 建立用户模型
   *
3. 模拟用户行为
   通过已有的历史数据，隐藏部分来让系统预测。比如从历史数据中获得用户 A 评价的一半 item，让系统推荐，然后和另一半比较。论文中常用的方法是指定一个 n，通过 n 个 item 推荐其他 item 或通过除 n 之外的所有 item 来推荐这 n 个 item。（这种方法貌似不适用于冷启动问题）
4. 评估

有必要的话，需要进行多次测试，再进行假设检验确认结论的可靠性。

准确度的评价方法：

对于一个元素是 user-item 对(u, i)的集合 T，实际评分为 r_{ui}，预测评分为 r_{ui}（估计值，带小角）。

Mean Absolute Error：

$$\text{MAE} = \frac{1}{|\mathcal{T}|} \sum_{(u,i)\in\mathcal{T}} |\hat{r}_{ui} - r_{ui}|$$

伪代码：

```
double mae(user_item_pairs, estimated_ratings, actual_ratings):
    error_sum = 0.0
    for pair in user_item_pairs:
        diff = estimated_ratings[pair] - actual_ratings[pair]
        squared_error_sum += abs(diff)
    mae = sum / user_item_pairs.size

    return mae
```

Root Mean Squared Error：

$$\text{RMSE} = \sqrt{\frac{1}{|\mathcal{T}|} \sum_{(u,i)\in\mathcal{T}} (\hat{r}_{ui} - r_{ui})^2}$$

伪代码：

```
double rmse(user_item_pairs, estimated_ratings, actual_ratings):
    squared_error_sum = 0.0
    for pair in user_item_pairs:
        diff = estimated_ratings[pair] - actual_ratings[pair]
        squared_error_sum += diff * diff
    rmse = sqrt(sum / user_item_pairs.size)

    return rmse
```

无论是通过 MAE 还是 RMSE 计算，最终的结果值越小证明结果越准确。但从公式可以看出，RMSE 通过平方扩大了偏离量，同样的两组结果用 RMSE 得出的差异值将比 MAE 更大。Netflix Prize 中使用的评价方法是 RMSE。

除了评价预测准确度的方法外，还有评价分类准确度（推荐的 item 对用户是否有用？没推荐的 item 真的对用户没用？）和排序准确度（对给定的 k 个 item，用户最可能的选择顺序和推荐产生的顺序一致性如何？）的方法，如果要用到再记之。

ref:

Chapter 8, Recommender Systems Handbook

Evaluating Recommender Systems: An evaluation framework to predict user satisfaction for recommender systems in an electronic programme guide context