

评估方法与评价 离线评估及线上测试

1. 离线评估方法与在线评估指标

① holdout 分组 70% . 30%

② k-fold cross validation

③ leave-one validation

④ Bootstrap: 对比较少的重抽样，从而有效回随机抽样

2. 离线评估的指标

① Accuracy = $\frac{n_{\text{correct}}}{n_{\text{total}}}$

当样本很不平衡时，accuracy 失效

② Precision & Recall

Precision: 分类正确的正样本占判定为正样本比例

Recall: 分类正确的正样本占全部正样本比例

F₁: 综合 Precision & Recall:

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

实际应用：

在排序模型中，通常没有一个确定的阈值把预测结果直接判定为正样本或负样本，而是采用 Top N 排序结果的精确率 (Precision@N) 和召回率 (Recall@N) 来衡量排序模型的性能，即认为模型排序的 Top N 的结果就是模型判定的正样本，然后计算 Precision@N 和 Recall@N。

③ Root Mean Square Error (RMSE)

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}}$$

若 outliers, 则 RMSE 很差

④ Mean Absolute Percent Error (MAPE)

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\%$$

对每个的误差进行归一化，降低了个别样本异常的绝对误差的影响

④ log-Loss

$$\text{Log Loss} = -\frac{1}{N} \sum_{i=1}^N [y_i \log p_i + (1-y_i) \log (1-p_i)]$$

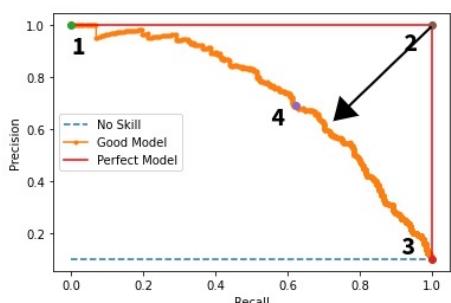
y_i : 第 i 个样本 x_i 的真实类别

p_i : 预测第 i 个样本 x_i 是正样本的概率 (Sigmoid / softmax)

N : 样本总数

2. 直接评估推荐系统的准确指标

① Precision-Recall Curve (P-R 曲线)



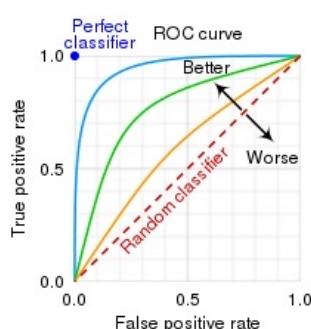
意义：在某一阈值下，模型将大于该阈值的结果判定为正样本，将小于该阈值的结果判定为负样本时，命中结果对应的召回率即精确率。

AUC: 描 P-R 曲线下面积大小, AUC 越大，模型效果越好！

② ROC curve

FPR: FP/N 假阳性率 (错判为正类的)

TPR: TP/N 真阳性率 (对的分成对的)



③ 平均精度值 (Mean Average Precision)

表 7-3 计算 precision@N 示例

推荐序列	$N=1$	$N=2$	$N=3$	$N=4$	$N=5$	$N=6$
真实标签	1	0	0	1	1	1
precision@N	1/1	1/2	1/3	2/4	3/5	4/6

AP 的计算只取正样本处的 precision 进行平均，即 $AP = (1/1 + 2/4 + 3/5 + 4/6)/4 = 0.6917$ 。那么什么是 mAP 呢？

每个用户算一个 AP，所有用户的 AP 平均就是 mAP

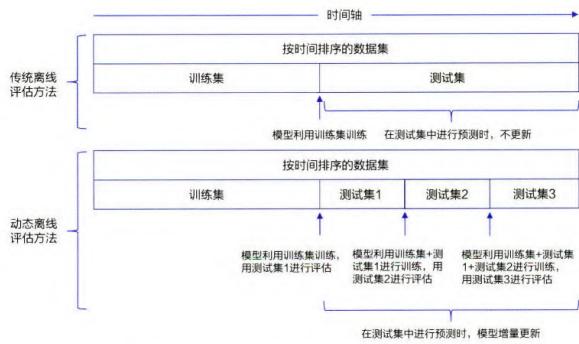
离线评估的目的在于快速定位问题，快速排除不可行的思路，为线上“评估”找到合适的候选者。

3. Replay

离线评估的重点是让离线评估的结果能尽量接近线上结果。

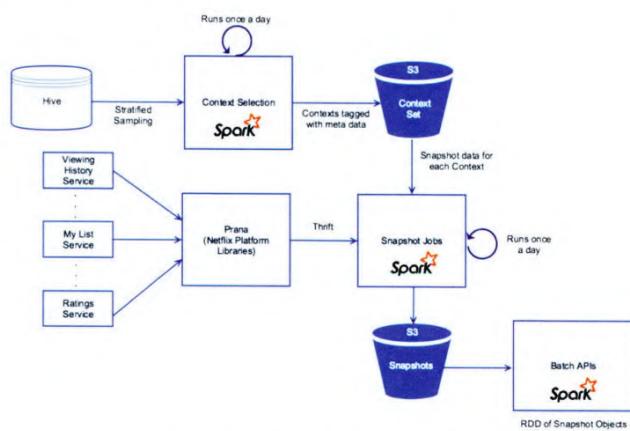
离线环境应接近线上环境。

动态离线评估方法



毫无疑问，动态评估的过程更接近真实的线上环境，评测结果也更接近客观情况。如果模型更新的频率持续加快，快到接收到样本后就更新。整个动态评估的过程也变成逐一样本回放的精准线上仿真过程，这就是经典的仿真式离线评估方法——Replay。

Netflix 的离线评估数据架构



要点：样本中不能有任何的未来信息，遵循历史数据穿越。

4. A/B 测试

将用户随机分成实验组和对照组，对实验组的用户施以新模型，对对照组用户施以旧模型，比较实验组和对照组在各维度上评估指标上的差异。

为什么 A/B 测试无法被替代？

1. 商业评估是完全依赖数据有偏现象的影响
2. 商业评估是完全还原线上的工程环境
3. 线上系统和某些商业指标在评估中无法计算

A/B 测试的分层和分流机制

1. 层与层之间的流量：正交

2. 同层之间的流量：互斥

以图 7-6 为例，在 X 层的实验中，流量被随机平均分为 X_1 （蓝色）和 X_2 （白色）两部分。在 Y 层的实验中， X_1 和 X_2 的流量应该被随机且均匀地分配给 Y 层的两个桶 Y_1 和 Y_2 。如果 Y_1 和 Y_2 的 X 层流量分配不均匀，那么 Y 层的样本将是偏的，Y 层的实验结果将被 X 层的实验影响，无法客观地反映 Y 层实验组和对照组变量的影响。所以穿过 X_1 层和 X_2 的流量都应被随机打散，均匀分布在 Y_1 组和 Y_2 组中。

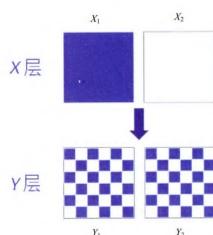


图 7-6 层与层之间的流量“正交”示例

A/B 测试的主要评估指标

表 7-4 各类推荐模型的线上 A/B 测试的主要评估指标

推荐系统类别	线上 A/B 测试评估指标
电商类推荐模型	点击率、转化率、客单价（用户平均消费金额）
新闻类推荐模型	留存率（x 日后仍活跃的用户数 / x 日前的用户数）、平均停留时长、平均点击个数
视频类推荐模型	播放完成率（播放时长 / 视频时长）、平均播放时长、播放总时长

5. Interleaving — 快速线上评估方法

传统 A/B test 存在的问题

A、B 两组之间的重度消费者份额小、不平衡，可能也对评估结果有影响。
不区分 A、B 组，而是把不同的被测对象同时提供给消费者，最后根据
实验者喜好得出评估结果的方法就是 Interleaving！

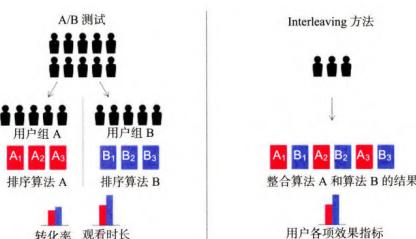
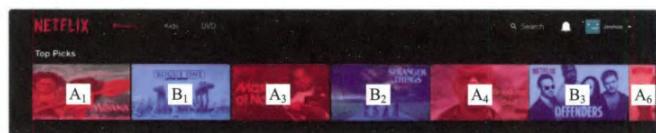


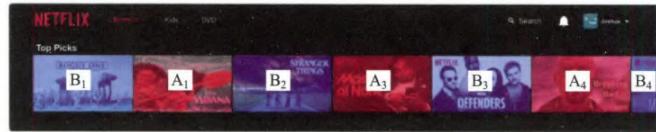
图 7-8 传统 A/B 测试和 Interleaving 方法的比较



如果算法 A 选择了第一个位置



如果算法 B 选择了第一个位置



Interleaving 算法能取代 A/B test 吗?
· 从灵敏度和正确性两方面验证

Interleaving 方法 缺点

1. 性能差
2. 只能验证“A 和 B 哪个好”，对于深层的数据分析，无法验证。

总结

