075 | 推荐系统评测之一：传统线下评测

2018-03-26 洪亮劼

AI技术内参 进入课程 >

讲述：初明明
时长 06:32 大小 3.00M

上周我们讨论了 EE 算法，介绍了 UCB（Upper Confidence Bound）算法和“汤普森采样”（Thompson Sampling）。

这周，我们回归到一个更加传统的话题，那就是**如何评测推荐系统**。这个话题非常重要，牵涉到如何持续对一个推荐系统进行评价，从而能够提高推荐系统的精度。

今天，我们先来看一看**推荐系统的线下评测**。

基于评分的线下评测

在过去 10 年里，随着 Netflix 大奖赛的举行，很多研究人员和工程人员往往把推荐系统的模型学习简化为对用户评分的一种估计。同时，在模型上面来说，对用户物品评分矩阵进行

分解成为了一种主流的方法。

在这样的场景下，如何对模型进行评测呢？

一种简单且直观的办法，就是**衡量评分的准确性**，换句话说，也就是看我们预测的评分和真实评分之间有多大的差距。

那么，有哪些方法可以用来衡量两个数值之间的差异呢？

在机器学习中，一个经常使用的测度叫“**均方差**”（Mean Square Error），或**MSE**。有时候，我们也会使用它的根号后的结果，叫作“**方差**”（Rooted Mean Square Error），或**RMSE**。

MSE 是这么定义的。首先，如果我们知道一个用户 i 和物品 j 的真实评分，假设叫 Y_{ij} ，那么我们的一个估计值是 Z_{ij}，MSE 计算的就是 Y_{ij} 和 Z_{ij} 的差值，然后取平方。平方以后的结果肯定是一个正数，也就是说这样做的好处是整个计算不会出现负数，我们的估计值比真实值小了或者大了，MSE 都可以处理。当我们对于每一个用户和物品都计算了这个差值以后，再对所有的差值的平方取一个平均值，就得到了我们所要的 MSE。

从计算上来讲，RMSE 就是在 MSE 的基础上再取一个根号。我们在很多实际应用中，往往使用 RMSE 来汇报模型的评测结果。同时，RMSE 也经常用在大多数的学术论文中，但这个评测有没有什么问题呢？

答案是，RMSE 其实存在很多问题。

首先，我们从刚才描述的计算过程中就可以看到，RMSE 需要取一个平均值。这是对所有用户在所有物品上评分误差的估计的平均。那么，如果一个用户在数据集里面对很多物品进行了评分，这个用户误差的贡献就会在最后的平均值里占很大一部分。也就是说，最后的差值大部分都用于描述这些评分比较多的用户了。

上述情况存在一个弊端，如果我们得到了一个比较好的 RMSE 数值，往往很可能是牺牲了大部分用户的评分结果，而对于少部分的高频用户的评分结果有提高。说得更加直白一些，**那就是 RMSE 小的模型，并不代表整个推荐系统的质量得到了提高。**这是 RMSE 很容易带来困惑的地方。

RMSE 的另外一个问题就是，这个指标并没有反应真实的应用场景。什么意思呢？真实的应用场景，我们往往是从一大堆物品中，选择出一个物品，然后进行交互。在这样的流程下，物品单独的评分其实并不是最重要的。更进一步说，就算一个推荐系统能够比较准确地预测评分，也不能证明这个推荐系统能够在真实的场景中表现优异。

基于排序的线下评测

当研究人员意识到 RMSE 的问题后，不少人就开始回归到问题的本质，**究竟什么样的评测更能反应推荐系统在真实场景中的表现呢?**

很多人很自然地就想到了搜索。

我们来回忆一下，搜索的结果是根据某个查询关键字，然后搜索引擎返回一系列文档结果。在这样的场景中，如果我们来和推荐进行对比，就会发现，这里面最大的区别仅仅是有没有一个查询关键词。

所以，我们其实**可以把搜索的一些指标“移植”到推荐中来使用。**比如，我们在搜索中讲过的**基于二元相关度的指标。**下面简单回顾一下这个指标。

什么叫“二元相关度”呢？简单说来，就是指针对某一个查询关键字而言，整个测试集里的每一个文档都有一个要么“相关”，要么“不相关”的标签。在这样的情况下，不存在百分比的相关度。而每个文档针对不同的关键字，有不同的相关信息。假定某个系统针对某个关键字，从测试数据集中提取一定量的文档而不是返回所有文档，我们就可以根据这个提取的文档子集来定义一系列的指标。

有两个定义在“二元相关度”上的指标，成为了很多其他重要指标的基石。一个叫“**精度**”（Precision），也就是说，在提取了的文档中，究竟有多少是相关的。另一个叫“**召回**”（Recall），也就是说， 在所有相关的文档中，有多少是提取出来了的。

“精度”和“召回”的相同点在于，分子都是“既被提取出来又相关的文档数目”。这两个指标的不同之处则是他们的分母。“精度”的分母是所有提取了的文档数目，而“召回”的分母则是所有相关的文档数目。如果我们返回所有的文档，“精度”和“召回”都将成为 1（也就是说，在这样的情况下是没有意义的）。因此，我们注意到，这两个指标其实都假定，提取的文档数目相比于全集而言是相对比较小的子集。

我们其实就**可以利用“精度”和“召回”来评测推荐系统。**

然而，这有一个问题，那就是对于搜索而言，相关度大多数时候是通过人工标注的，但这个对于推荐系统来说是不可能的。因为推荐的结果对于每个人来说都是不一样的，所以，没法针对所有人来进行统一的人工标注。

一种折中的办法，就是使用用户的回馈信息。在这里，因为我们需要二元信息，所以可以使用像用户的点击信息或者购买信息来作为二元的相关度。也就是说，如果用户点击了某个物品，那我们就认为是相关的，反之则是不相关。

顺着这个思路下去，其实我们就可以计算类似于**NDCG**等更加复杂的指标，只不过我们需要自己去定义相关信息。

利用排序的思路来评测推荐系统，已经成为了目前推荐系统线下评测的一个标准指标。

小结

今天我为你讲了如何评测推荐系统的好坏，今天的重点是线下评测的两类指标。

一起来回顾下要点：第一，我们聊了聊非常通用的 RMSE 的评测方法，并且指出这类方法的缺陷；第二，我们介绍了怎么把搜索里的评测方法给移植到推荐中。

最后，给你留一个思考题，基于排序的评测有什么致命的问题吗？

欢迎你给我留言，和我一起讨论。

上一篇　074 | 推荐的Exploit和Explore算法之三：汤普森采样算法

下一篇　076 | 推荐系统评测之二：线上评测

精选留言 (2)

写留言

ninenight
2018-11-13

利用ndcg在做线下评测的时候怎么具体操作呢，是先标注点击相关吗，但是线下评测的时候还没上到线上呢，也不知道点击数据，这个时候怎么在线下评测呢，比如我要现在线下评测下效果，然后再上线，这个时候怎么能评测效果呢，请指教

林彦
2018-03-27

基于排序的推荐系统会一直推荐用户有交互行为的物品，发掘新物品和保持多样性的能力会降低