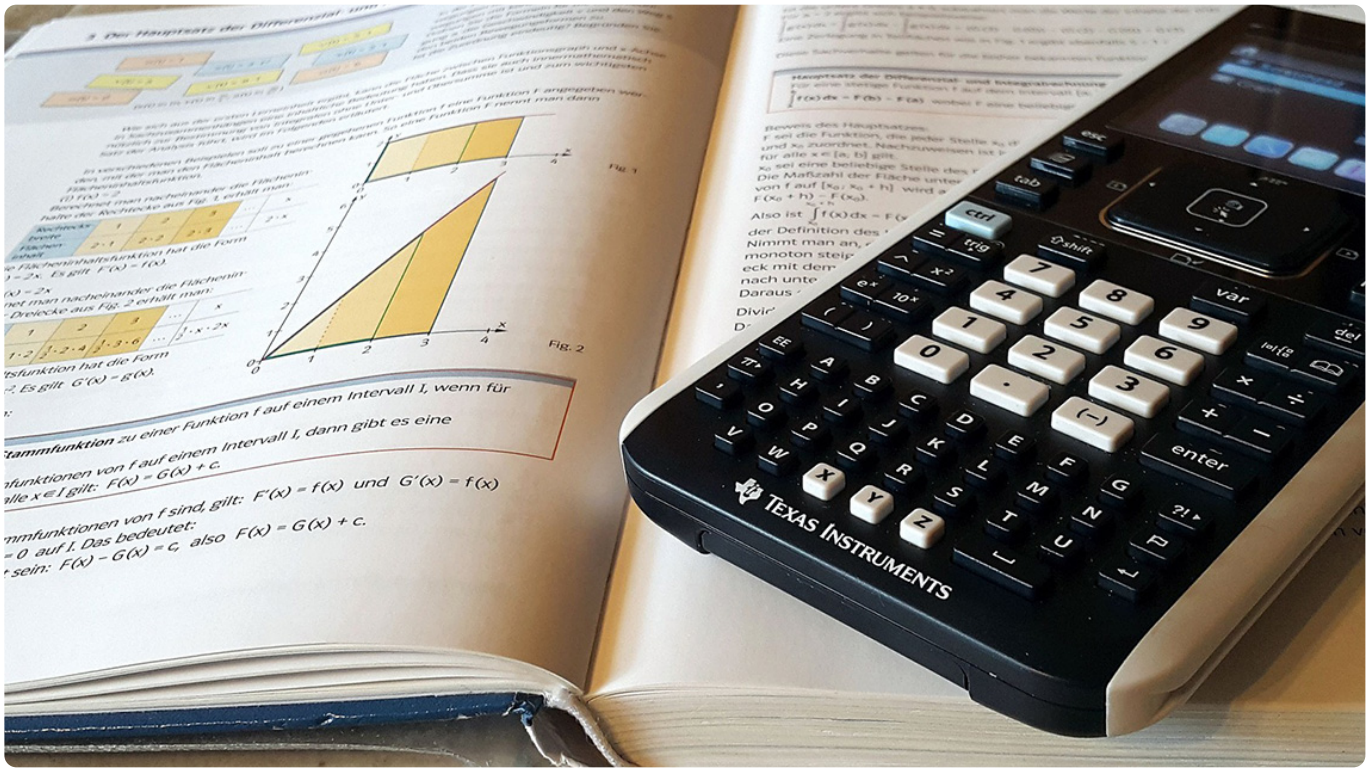
076 | 推荐系统评测之二：线上评测

2018-03-28 洪亮劼

AI技术内参

进入课程 >



讲述：初明明

时长 06:28　大小 2.97M

周一，我们聊了推荐系统的线下评测。线下评测是任何系统能够不断演化的最直接的要求。在线下的环境中，我们能够开发出系统的种种改进，并且希望能够通过这些线下评测的手段来选择下一个更好的版本。

今天，我们来讨论推荐系统的线上评测。任何系统在开发之后最终都要放到线上拿给用户使用。那么，在线上评测的时候需要注意什么呢？

线上评测的基础

推荐系统线上评测的基础和我们之前讲过的搜索系统有很多类似的地方。

线上评测的核心就是**在线可控实验**，有时候又称作是在线实验，或者叫作在线 A/B 实验。我们希望能够利用在线实验来对推荐系统的某一个部分进行检验，看是否对用户和系统的交互产生影响。

在线可控实验其实是建立“**因果联系**”（Causal Relationship）的重要工具，也可以说是唯一完全可靠的工具。这里面的基础是统计的**假设检验**。

具体来说，就是我们针对访问网站或者应用的人群，进行某种划分，一般情况下是平均随机划分。50% 的用户进入被划定的一个群组，叫作“控制组”（Control Bucket），而另外 50% 的用户进入另一个群组，叫作“对照组”（Treatment Bucket）。“控制组”和“对照组”的唯一区别在于所面对的系统。

假设我们有一个推荐系统，想对其中的某个部分进行改进。那么，可以保持住其他的部分，让这个希望得到改进的部分成为唯一的“**独立变量**”（Independent Variable），也就是在整个实验设置中的变量。这样，我们就希望通过在线实验以及假设检验的工具，来认定这个“独立变量”是否会带来系统性能上的提高或是降低。

这里面还有一个需要提前确定的，就是需要评测的指标，特别是用户指标，比如网站的点击率、搜索的数量等。这些指标我们称之为“**依赖变量**”（Dependent Variable）。说白了，**我们就是希望能够在“独立变量”和“依赖变量”之间通过假设检验建立联系。**今天在这里，我们就不重复假设检验的基础知识了，如果有兴趣深入了解，可以参考任意一本基础的统计教程。

对于在线可控实验，在概念上很容易理解，但在现实操作中有很多挑战。

首先，我们可以想一想，虽然理想状态下，我们可以把用户五五平分，进入“控制组”和“对照组”。然而，现实中，经过随机算法分流的用户群，在这两个群组中很可能并不呈现完全一样的状态。什么意思呢？举个通俗的例子，相比于“对照组”而言，“控制组”中可能存在更多的女性用户；或者是“对照组”中，可能存在更多来自北京的用户。

而在这样的情况下，“依赖变量”，比如网站点击率，在“控制组”和“对照组”的差别，就很难完全解释为“独立变量”之间的差别。也就是说，如果“控制组”下的点击率比“对照组”高，是因为我们更改系统的某部分的差别呢？还是因为这多出来的女性用户呢？还是女性用户和系统某些部分的交互产生一定复杂的综合结果导致的呢？这就比较难说清楚了。

当然，在现实中，如果说我们依然可以比较容易地通过算法来控制一两个额外的变量，使得在“控制组”和“对照组”里面这些变量的分布相当，那么，面对十几种重要变量（例如，年龄、性别、地域、收入层级等），要想完全做到两边的分布相当，难度很大。

另外一个难点是，即便能够做到对已知的变量通过随机算法，使得在两个群组中的分布相当，我们依然不能对当前还未知的变量进行上述操作。因此，现实中因为人群特性所带来的对结论的影响，是在线实验的难点之一。

在线实验的难点之二是，即便刨除刚才所提到的人群的差异以外，我们可能也很难在设想中的某个系统“控制组”和“对照组”中，确定唯一的“独立变量”。

在现代网站或者应用中，有很多服务、子系统、页面、模块在同时为整个网站服务。而这些服务、子系统、页面和模块，还有不同的前端系统和后端系统，很可能属于不同的产品以及工程团队。每个部分都希望能够做自己的可控实验，都希望自己改进的部分是唯一变化的“独立变量”。然而，我们从宏观的角度去看，如果每个部分都在做自己的实验，而我们做实验的基本单元依旧是每个用户的话，那这就很难保证用户之间的可比较性。

举个例子，如果用户 U_1，进入了首页的“控制组”，然后访问了购物篮推荐模块的“对照组”后离开了网站。而用户 U_2，直接访问了帮助页面的“对照组”，然后访问了购物篮推荐模块的“控制组”。那 U_1和 U_2两个用户最终产生的点击率的差别，就很难从他们访问网站页面的过程中得出结论。即便是在有大量数据的情况下，我们也很难真正去平衡用户在所有这些页面组别之间的关系。

实际上，如何能够有效地进行在线实验，包括实验设计、实验的评测等，都是非常前沿的研究课题。

推荐系统线上评测

在上次的分享中，我讲了过去比较流行的推荐系统线下评测的方法，比如利用 MSE 以及在此之上衍生出的 RMSE。然后，我又讲了如何从排序的角度来对一个推荐系统进行衡量。

到线上评测以后，很明显，RMSE 以及排序的一些相关指标，都不能完全反映用户是否对一个推荐系统产生好感。我这里讲几个通用的指标。

第一，用户的驻留或者停留时间（Dwell Time）。这个指标主要是衡量用户在某一个物品上的停留时间。用户总的停留时间往往和用户对网站的黏稠度有很强的关系。总体说来，如

果用户会长期反复访问网站，用户在网站平均驻留时间往往是比较多的。

第二，用户在相邻两次访问中的间隔时间，有时叫作“空缺时间”（Absence Time）。这个指标是越短越好。当用户反复访问你的网站，并且空缺时间越来越短，证明用户越来越依赖你网站提供的服务。

停留时间和空缺时间都是很好的推荐系统线上评测的指标。

小结

今天我为你讲了推荐系统评测的线上评测。

一起来回顾下要点：第一，我们聊了聊推荐系统的在线实验；第二，我们介绍了几个推荐系统线上评测的通用指标。

最后，给你留一个思考题，如何知道用户对于推荐的内容已经越来越不满意了呢？

欢迎你给我留言，和我一起讨论。

上一篇　075 | 推荐系统评测之一：传统线下评测

下一篇　077 | 推荐系统评测之三：无偏差估计

精选留言 (3)

写留言

兔子ORZ
2018-04-12

兴趣点降低，驻留时间，空缺时间两个指标会会在相对比率上降低。但是推荐系统实时响应会差一点。

damonhao
2018-04-06

观察依赖变量，如点击率。排除其他因素的干扰，如ue，内容等。做起来感觉很难。

林彦
2018-03-28

对于包含推荐内容的页面，(1)访问间隔时间变长，(2)每次访问的时间变短，(3)点击或其他可衡量的与推荐内容的互动次数变，互动时长变短。这些说明用户对推荐的内容越来越不敢兴趣。

展开