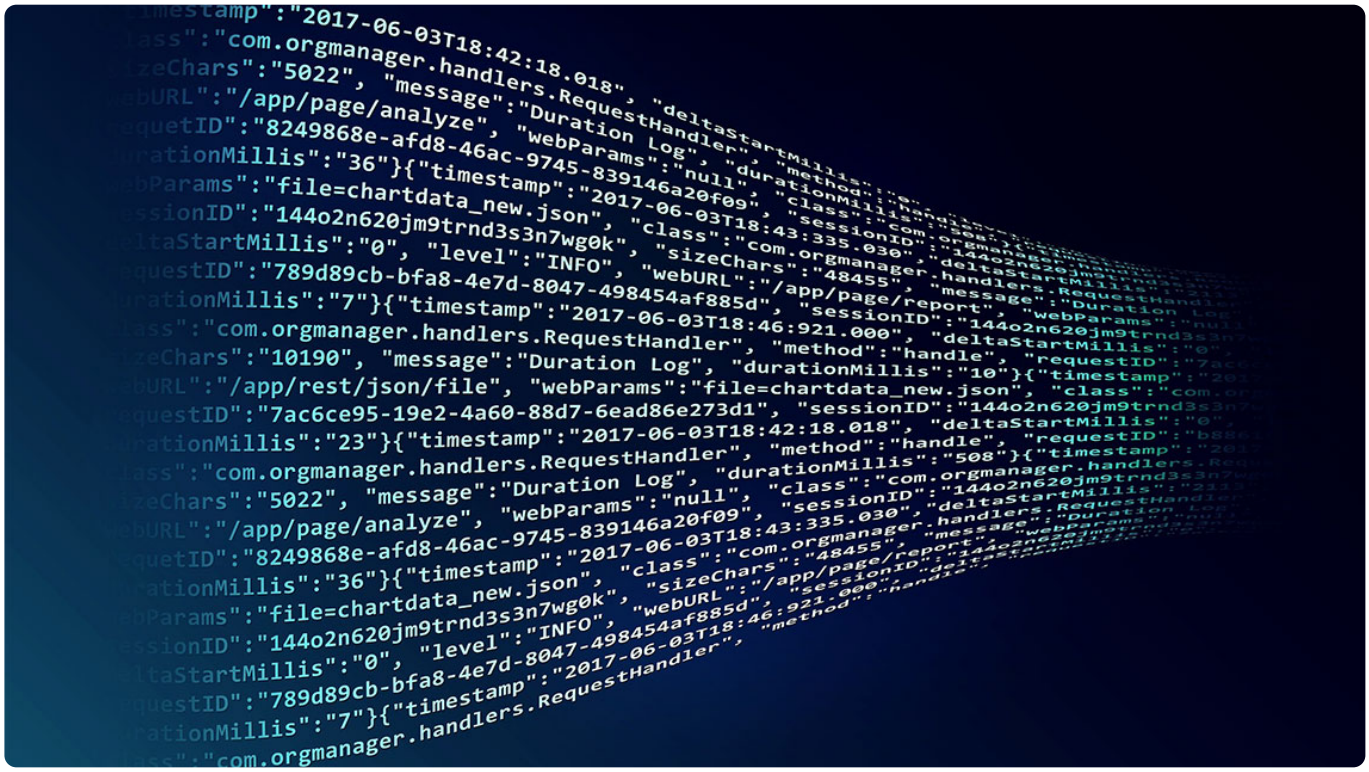
077 | 推荐系统评测之三：无偏差估计

2018-03-30 洪亮劼

AI技术内参

进入课程 >



讲述：初明明

时长 06:16　大小 8.63M

周三，我讲了推荐系统的线上评测，我们讨论了如何做在线评测，以及推荐系统在线评测可能遇到的一系列问题。

今天，我们来看一个比较高级的话题，那就是如何做到推荐系统的无偏差估计。

推荐系统的偏差性

在理解为什么需要无偏差估计之前，我们先来看一看现在系统为什么会存在偏差性，以及这种偏差性会对推荐系统的学习造成什么样的影响。

先假定我们讨论的是一个新闻推荐系统。为了方便，假设我们的系统中仅有两类文章，体育新闻和财经新闻。我们先来看一个极端情况，假设系统因为种种原因，只给用户推荐体育新

闻，会出现什么样的情况呢？

很明显，如果我们只给用户看体育新闻，那么，用户只可能在这些体育新闻里面选择点击，压根不知道财经新闻的存在。因为用户只在体育新闻里面通过点击浏览表达了喜好，任何一个机器学习系统，只能在这些新闻里面学习到用户的喜好。

具体来说，用户已点击的新闻可以认为是正例，没有点击的可以认为是负例。这样可以对用户在体育新闻上面的喜好加以建模。但是很明显，系统在用户对财经新闻喜好的学习上是一筹莫展的，因为我们根本没有任何用户对于财经新闻的回馈数据。

这就是系统的偏差性。在这样的设置下，系统存在的偏差是无法修正的。通俗地讲，我们压根没有给财经新闻任何机会，所以没有数据可以供系统学习。

其实还有更加严重的情况。因为我们当前的系统仅仅学习到了用户对于体育新闻的喜好，然后如果把这个系统部署到生产中，系统又会进一步选择体育新闻给用户看，如此循环下去。这其实就是说，系统的偏差性会随着机器学习系统而反复循环，甚至有逐渐放大的可能性。

当然在现实的应用中，我们不可能有这么极端的情况。但可能面临更加复杂的情况，比如，我们压根不知道系统存在什么样的偏差，或者说，我们该如何面对各种不同的偏差。

在有偏差的系统中，先通过数据学习得到模型，然后再部署到系统中去，这个流程其实严重阻碍了我们对用户真实喜好的检测。因此，这也是线下表现和线上表现不一致的一个原因。长期以来，偏差性都是困扰研究者的一个非常重要的问题。

无偏差估计

当我们知道系统有偏差以后，怎么来解决这个问题呢？

一个很容易想到的策略是，如果我们知道系统的某种偏差，那能不能在后面的评测过程中矫正这种偏差呢？

这就涉及“**矫正**”的思路。回到我们所说的体育新闻和财经新闻的例子。假设我们的系统在 80% 的情况下会显示体育新闻，20% 的情况下显示财经新闻。那么，当用户面对一篇体育新闻点击浏览，或者面对一篇财经新闻点击浏览，我们的系统该如何应对呢？

在我们已经提到过的传统评测手段中，例如计算 MAP 或者 NDCG 的时候，这两种点击是一样的。或者说，权重是一样的。然而，在这样的情况下，机器学习系统其实还是会更加偏重于学习到用户对于体育新闻的偏好，因为毕竟 80% 的情况下都是体育新闻。相对于财经新闻而言，这种情况就是处于劣势的，可能我们没有给财经新闻足够的机会。

所以，从矫正的角度来说，我们认为如果用户点击浏览了原本出现概率较低的文章，这个时候，我们反要给这类文章更大的权重。什么意思呢？也就是说，我们认为财经新闻出现的概率比较低，如果在这种情况下，用户点击浏览了财经新闻，那应该是真正的偏好。而相同的情况下，因为 80% 的新闻都是体育新闻，因此用户点击了其中的一篇也就不足为奇。

把这种思维放入到一种数学的表达中，也就是，我们希望用户的回馈按照出现的概率进行**反比矫正**，出现概率越大的物品，正样本权重越小；反之，出现概率越小的物品，正样本权重越大。具体来说，也就是正样本除以出现的概率，然后我们计算平均的加权点击率。这样加权平均后的结果，就是矫正后无偏差的点击率的计算结果。

很明显，无偏差估计是有一定假设的。首先，我们就需要假设收集的数据涵盖了整个数据集。什么意思？就是刚才我们说的极端情况，比如我们只显示体育新闻而压根一点都不显示财经新闻，这种情况是无法进行矫正的，因为在这种情况下，财经新闻的概率是 0。也就是说，无论什么类别的新闻，都需要有非零的概率出现。这是进行无偏差估计的一个基本假设和要求。

遗憾的是，虽然这个要求看似容易，但其实在现实中很难真正做到。

试想一个有百万文章量的新闻网站，要确保所有的新闻都有一定概率显示给用户是有挑战的。在实际的应用中，大量的新闻质量是呈指数下降的。也就是说，虽然有百万甚至更多的文章量，但是很有可能只有几百几千的文章相对比较有质量，而剩下的大量文章是低质量的文章。

然而，我们并不能完全确定哪些是低质量文章。如果我们真的需要做无偏差的估计，就需要针对所有的文章进行显示，也就是说，我们需要冒着给用户显示低质量文章的风险，显然这并不是很好的策略。

在如何收集数据这一方面，无偏差估计其实和我们之前提到过的 EE 策略又结合在了一起。也就是说，如何既能够让我们尽可能地把所有数据都呈现给用户，使得我们可以进行无偏差估计，又能够照顾到用户的体验，这是目前非常热门的研究领域。

小结

今天我为你重点讲了什么是系统的偏差以及如何处理偏差的思路。

一起来回顾下要点：第一，我们聊了聊在线系统的偏差出现的场景以及机器学习为什么会让这样的情况恶化；第二，我介绍了如何进行无偏差估计以及无偏差估计所需的条件。

最后，给你留一个思考题，假如一个系统，你不知道每一种新闻出现的概率，你该如何做无偏差估计呢？

欢迎你给我留言，和我一起讨论。

上一篇　076 | 推荐系统评测之二：线上评测

下一篇　078 | 现代推荐架构剖析之一：基于线下离线计算的推荐架构

精选留言 (1)

写留言

林彦

2018-04-05

简化的来想，如果我不知道每一种新闻出现的概率，假定所有新闻初始出现的概率是相等，用每种新闻的数量占新闻的总数量的比例来作为初始概率来进行无偏差估计。随着真实数据的搜集，再去调整这个概率。

展开