监督学习越来越准，我为什么要写bandit问题

冯伟
Hulu推荐算法

监督学习的典型场景

在涉猎bandit问题之前，监督学习是很好概括的：

- **步骤 1 刻画原始需求：** 给用户推荐一道菜，结果只有两个：用户喜欢或者不喜欢
- **步骤 2 映射成监督学习（二分类）问题**：给定特征向量x=(菜的类型：荤菜/素菜，顾客类型：性别、年龄性别，就餐时间：早/午/晚），预测顾客是否会接受这道菜，y=0或1
- **步骤 3 用历史数据训练模型**：选择常用的监督模型logistic regression/gdbt/神经网络，从大量的历史数据(x, y)中学习模型的参数，给定x，预测y越准越好，
- **步骤 4 部署上线做A/B Test**：观测线上效果。

准确率并不是监督学习的全部

然而，上面的抽象并不是完整的，我们先从一个例子开始。

(千篇一律的新闻) 现在的新闻客户端都使用了机器学习进行智能排序，你有没有跟我一样的体验：

- 某类型的新闻，你点击的越多，下次登录时就会看到的越多
- 看到的越多，点的机会也就越多
- 最后满眼的新闻都是千篇一律

借着这个例子，我们说说把现实问题映射到监督学习的过程中存在的坑：

- **历史数据的收集直接受到了算法影响，是有偏向性的**：算法推荐了一个新闻，用户才有机会给出反馈，系统才会收集到反馈。但是还有千千万万的新闻是没推荐出来的，用户不知道它们好不好吃，系统也没机会收集那些新闻的反馈
- **正反馈是很容易获得的，负反馈却需要自己去猜**：算法推荐了k个新闻，用户只点击其中的一个，这并不100%意味着是对剩下的新闻的否定：（1）这个新闻没有被用户注意到，（2）这个新闻用户也感兴趣喜欢，只是时间有限，这次没点

结论：既然历史数据是受算法影响的，用户又只提供了正反馈，那么根据历史数据训练就会不断强化自己去推荐已经推荐过的东西，使得模型陷入一个局部最优，潜在的好的东西迟迟得不到推荐。

为什么监督学习还能work

可是这么多年都是这么训练的，为什么也没见到大的问题?

- **特征工程时考虑到了泛化能力**：新闻到底属于财经类还是娱乐类、用户的年龄、性别是什么，这些特征都是普遍适用的。一个用户、一个新闻，即使没有推荐过，我们也能依据它们的特征判断的八九不离十。**但是，为了提高准确率，我们也会牺牲泛化能力，加入ID类特征**，包括用户ID和物品ID。
- **冷启动问题得到了足够的重视，弥补了特征泛化能力不足的问题**：一个新闻刚出现时，我们会有意识的采取手段确保他们能得到一定推荐。比如去看看新闻和用户已经点击过的新闻的相似性（基于内容去找关联）。

Bandit问题的核心

Bandit的研究总是需要回答3个核心问题：

- 如何预测点击率 p
  - Contextual Bandits使用了线性模型： $p = x^T\theta$
  - 我们当然也可以使用非线性模型，比如决策树、神经网络
- 如何衡量 p 的不确定性 Δ ，按照 $\tilde{p} \in [p-\Delta, p+\Delta]$ 对物品进行排序
  - UCB算法是Frequentist学派的代表，用置信区间来刻画
  - Thompson Sampling是Bayesian学派的代表，用概率分布来刻画

抓住了这个核心，我们看看之前的问题

- **冷启动有多冷**：一条新闻只被推荐过几次，它的不确定性 Δ 是很大的， Δ 很大表示这个新闻还很冷，按照 $\tilde{p} \in [p-\Delta, p+\Delta]$ 对物品进行排序是很有可能把新闻推荐出来的
- **算法和用户反馈的关系**：用户只会点击算法选中的新闻，
  - 利用已有历史信息(Exploitation)：推荐高质量的新闻，确保用户当前的体验，也就是 p 值较高的那些新闻
  - 勇于探索 (Exploration)：有些新闻才出来，或者用户以前没点击过，不确定性高，但如果推荐出来用户也有可能会喜欢，也就是 Δ 高的那些新闻
  - 如何平衡Exploitation和Exploration：万变不离其宗，我们是 $\tilde{p} \in [p-\Delta, p+\Delta]$ 对物品进行排序， $[p-\Delta, p+\Delta]$ 是一个区间，如何在这个区间取值反映了我们对Exploitation v.s. Exploration的偏好

工业界中的实践

微软在几个月前launch了Decision Service，感兴趣的读者可以看看本专栏的另一篇文章：

冯伟：解析微软云Azure Decision Service
zhuanlan.zhihu.com

欢迎订阅微信公众号 "零基础机器学习"，搜索微信号：ml-explained

编辑于 2018-01-17

「赞赏多一点，鼓励多一点」

赞赏

4 人已赞赏

在线机器学习 机器学习 强化学习 (Reinforcement Learning)

文章被以下专栏收录

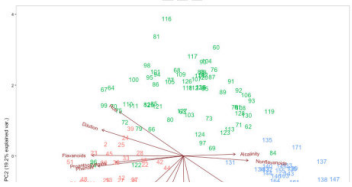
零基础机器学习 已关注

推荐阅读

深度学习算法索引（持续更新）

机器学习最近几年发展如同野兽出笼，网上的资料铺天盖地。学习之余建立一个索引，把最有潜力(不求最全)的机器学习算法、最好的教程、贴近工业界最前沿的开源代码收录其中。个人能力有限，希...

赵印 发表于计算广告学



怎样提升机器学习：特征工程的奇淫巧技

我爱机器学习

机器学习必知的10大算法

机器学习算法可以分为三大类：监督学习、无监督学习和强化学习。以下介绍 10 个关于监督学习和无监督学习的算法。监督学习可用于一个特定的数据集(训练集)具有某一属性(标签)，但是其他数据...

七月在线 发表于从零学AI

1 条评论 切换为时间排序

写下你的评论...

Jerry 1 年前

2个核心问题?

赞

▲ 赞同 69 ▼ 1 条评论 分享 收藏 ···