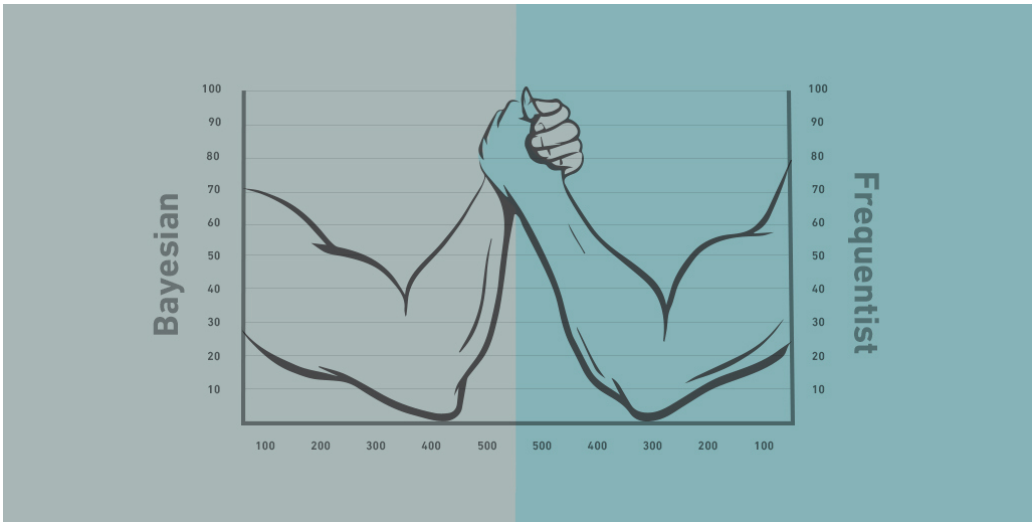

Multi-Armed Bandit: Thompson Sampling

冯伟
Hulu推荐算法

知识回顾

假设我们开了一家叫Surprise Me的饭馆

- 客人来了不用点餐，由算法从N道菜中选择一道菜推荐给客人
- 每道菜都有一定的失败概率：以1-p的概率不好吃，以p的概率做得好吃
- 算法的目标是让满意的客人越多越好。

在之前的MAB中我们介绍了Upper Confidence Bound (UCB) 算法来解决这个问题：

- 基于目前的数据，估计出每道菜被接受的概率 $\tilde{p}$ 以及浮动范围 Δ
- 我们乐观的认为下一个应该推荐的菜应该是 $\tilde{p}+\Delta$ 最大的那个

UCB算法的缺点

但是UCB算法也有一定的缺点：

- 算法是确定性的，结果也是固定的，在模型更新前，推荐结果不会改变
- 无法融合先验知识，比如我们事先知道某些菜是比较好吃的
- UCB的实际效果不一定好于我们接下来要介绍的方法

MAB里也有Frequentist vs Bayesian？！

回顾我们的问题：一道菜概率p = θ 做的好吃，以概率p=(1- θ)做的不好吃。

Frequentist 和 Bayesian学派对于参数 θ p的看法是不一样的：

- Frequentist学派的看法

- 认为 θ 是一个客观存在的、固定的值，我们要做的就是通过多次试验来推测 θ 的值，既 $\tilde{\theta} = \sum_i reward_i / n$ ，当采集的样本无穷大时， $\tilde{\theta}$ 会趋近于真实的 θ
- 现实中采样样本不可能是无穷大的，因此Frequentist还会计算出一个置信区间 Δ ，也就有了UCB算法

- Bayesian（贝叶斯）学派的看法
  - 虽然 θ 是一个客观存在的、固定的值，但我们可以用一个概率分布来描述 θ 的不确定性。随着样本的增加，这个概率分布在真实 θ 附近的概率密度会越来越大。

UCB是Frequentist学派的一个经典，本节我们介绍一个Bayesian方法 — Thompson Sampling

Bernoulli MAB和Thompson Sampling

回顾我们的问题：一道菜以概率p = θ 做的好吃(reward=1)，以概率p=(1- θ)做的不好吃(reward=0)，这是一个典型的Bernoulli（伯努利)分布

$$p(reward|\theta) \sim Bernoulli(\theta)$$

Bayesian学派会用概率分布来描述 θ 不确定性：

$$p(\theta|reward) = \frac{p(reward|\theta)p(\theta)}{p(reward)} \propto p(reward|\theta)p(\theta) = Bernoulli(\theta)p(\theta)$$

$p(\theta)$ 的选取直接决定了 $Bernoulli(\theta)p(\theta)$ 的函数形式。在贝叶斯统计当中， $Bernoulli(\theta)$ 经常和 $Beta(\alpha, \beta)$ 分布一起使用（称为共轭分布）， $Bernoulli(\theta)Beta(\alpha, \beta)$ 会得到一个新的 $Beta$ 分布：

- 如果 $Bernoulli(\theta)$ 的结果为1，则会得到 $Beta(\alpha + 1, \beta)$
- 如果 $Bernoulli(\theta)$ 的结果为0，则会得到 $Beta(\alpha, \beta + 1)$

有了 θ 的不确定性，Bernoulli MAB的解决方案也就出来了 - Thompson Sampling:

- 步骤1: 用 $p(\theta|reward)$ 刻画每道菜好吃的概率，得到 $\{p(\theta_1|reward_1), \ldots, p(\theta_N|reward_N)\}$
- 步骤2: 对每道菜 $p(\theta_i|reward_i)$ 随机抽取一个样本 θ_i ，得到 $\{\theta_1, \ldots, \theta_N\}$
- 步骤3: 推荐 θ_i 最大的那道菜，得到 $reward_i$
- 步骤4: 更新 θ_i 的分布： $p(\theta|reward) = Beta(\alpha', \beta') \propto Bernoulli(\theta)Beta(\alpha, \beta)$
  - 如果 $reward_i = 1$ ，那么会得到 $Beta(\alpha + 1, \beta)$
  - 如果r $reward_i = 0$ ，那么会得到 $Beta(\alpha, \beta + 1)$

编辑于 2017-12-28

「真诚赞赏，手留余香」

赞赏

还没有人赞赏，快来当第一个赞赏的人吧！

机器学习 在线机器学习 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

推荐系统（五）——利用上下文信息

Joooor

论强化学习和概率推断的等价性：一种全新概率模型

机器之心 发表于机器之心

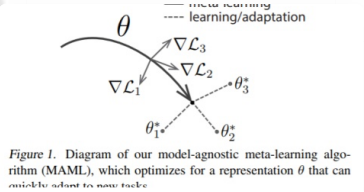

Figure 1. Diagram of our model-agnostic meta-learning algorithm (MAML), which optimizes for a representation θ that can

模型无关的元学习：learn to learn

刘芷宁

2 条评论 ⇌ 切换为时间排序

写下你的评论...

呜啦啦啦啦 4 个月前

推荐web.stanford.edu/~bvr/p...，里面补充了通用的TS算法～谢谢答主，写的很清楚了～

1

ansonwww 15 天前

感谢博主分享～ 关于frequentist vs bayesian的描述写得很清晰，受用了！

赞

▲ 赞同 41 ▼ 2 条评论 分享 收藏 ···