Exploration / Exploitation

目　　录

第 1 章 Thompson Sampling

平衡探索 (Exploration) 和利用 (Exploitation) 是优化个性化排序中比较重要的问题，若个性化排序模型仅关注利用，将导致新的商品曝光不足，若仅关注探索 (Exploration)，将导致流量浪费问题。因此有效平衡好探索和利用，对个性化排序系统闭环健康流转至关重要。下面将首先描述汤普森采样算法用于解决 EE 问题的总体框架，在此基础上介绍该算法在给定上下文下的参数估计算法，最后给出参数估计中涉及的高斯分布方差估计推导。本文档讨论的算法来自于 Olivier Chapelle 等人的工作[1]。

1.1 算法框架

问题设定：给定环境 x (可选) 及动作集合 $\mathbb{A}$。选定动作 $a \in \mathbb{A}$ 后，系统获得反馈 r。目标是找到某个策略 (policy) 使得选择的动作累计反馈最大。假设历史观察集合 $\mathbb{D}$ 由三元组 (x_i, a_i, r_i) 组成，通过似然函数 $P(r|a, x, \theta)$ 建模，该函数的参数为 θ。给定参数 θ 的先验分布 $P(\theta)$，及历史观察集合 $\mathbb{D}$ 后，可以通过贝叶斯法则进行参数的后验估计，即

$$P(\theta|\mathbb{D}) \propto \prod P(r_i|a_i, x_i, \theta) P(\theta) \tag{1.1}$$

在真实环境中，反馈 r 为动作 a，环境 x 及未知真实参数 θ^* 的随机函数。理想场景下，我们希望选择最大化反馈期望的动作，即 $\max_a E(r|a, x, \theta^*)$.

由于真实的 θ^* 未知，若我们希望最大化即时反馈 (即利用)，则可以选择动作 a 最大化期望 $E(r|a, x) = \int E(r|a, x, \theta) P(\theta|\mathbb{D}) d\theta$。当进行探索/利用时，可以根据动作为最优动作的概率随机选择最优动作 a，即动作 a 被选择的概率如下

$$\int \mathbb{I}\left[E(r|a, x, \theta) = \max_{a'} E(r|a', x, \theta)\right] P(\theta|\mathbb{D}) d\theta, \tag{1.2}$$

其中 $\mathbb{I}$ 为标识函数，实际计算时无需显式计算上述积分：可以在每一轮根据算法 1 抽样随机参数 θ。

Algorithm 1 Thompson sampling

$\mathbb{D} = \Phi$
for $t = 1, \ldots, T$ **do**
  `Receive context` x_t
  `Draw` θ^t `according to` $P(\theta|\mathbb{D})$
  `Select` $a_t = \arg\max_a E_r(r|x_t, a, \theta^t)$
  `Observe reward` r_t
  $\mathbb{D} = \mathbb{D} \cup (x_t, a_t, r_t)$
end for

[1] An Empirical Evaluation of Thompson Sampling

在标准的 K-armed Bernonulli bandit 问题中，动作即为选择某个 arm。第 i 个 arm 的反馈服从均值为 θ_i^* 的 Bernoulli 分布。建模每个 arm 的反馈采用 Beta 分布，因为该分布为 Binomial 分布的共轭分布。算法 2 为 Thompson sampling 算法应用于 Bernoulli bandit 问题的实例化算法。

Algorithm 2 Thompson sampling for the Bernoulli bandit

Require: α, β prior parameters of a Beta distribution
  $S_i = 0$, $F_i = 0$, $\forall i$. {Success and failure counters}
  for $t = 1, \ldots, T$ **do**
    for $i = 1, \ldots, K$ **do**
      Draw θ_i according to $Beta(S_i + \alpha, F_i + \beta)$
    end for
    Draw arm $\hat{i} = \arg\max_i \theta_i$ and observe reward r
    if $r = 1$ **then**
      $S_{\hat{i}} = S_{\hat{i}} + 1$
    else
      $F_{\hat{i}} = F_{\hat{i}} + 1$
    end if
  end for

1.2 参数估计

现在考虑个性化搜索排序问题。给定用户发起商品查询，该问题即为选择最优的商品展现给用户。该匹配问题的关键事项为点击率或转化率预估：在给定场景 (user, query) 下选定的商品被用户点击或下单的概率。这时存在一个探索/利用的两难问题：为了学习商品的点击率或转化率，商品需要被展现，这可能导致短期收入的损失。

这里考虑采用标准的正则化逻辑回归算法预测点击率或转化率，样本的特征可以通过深度学习模型建模。该模型中，参数的后验分布由协方差为对角阵的高斯分布近似。如下是正则化逻辑回归算法参数估计算法

Algorithm 3 Regularized logistic regression with batch updates

Require: Regularization parameter $\lambda > 0$.
  $m_i = 0$, $q_i = \lambda$. {Each weight θ_i has an independent prior $\mathcal{N}(m_i, q_i^{-1})$}
  for $t = 1, \ldots, T$ **do**
    Get a new batch of training data $(\boldsymbol{x}_j, y_j)$, $j = 1, \ldots, n$.
    Find $\boldsymbol{\theta}$ as the minimizer of: $\frac{1}{2}\sum_{i=1}^{d} q_i(\theta_i - m_i)^2 + \sum_{j=1}^{n} \log(1 + \exp(-y_j \boldsymbol{\theta}^\top \boldsymbol{x}_j))$.
    $m_i = \theta_i$
    $q_i = q_i + \sum_{j=1}^{n} x_{ij}^2 p_j(1 - p_j)$, $p_j = (1 + \exp(-\boldsymbol{\theta}^\top \boldsymbol{x}_j))^{-1}$ {Laplace approximation}
  end for

上述算法中首先基于参数 θ 服从的给定高斯先验和当前批次的样本预估 θ，并在此基础上更新参数 θ 的高斯分布参数。上述算法中高斯分布的均值为 θ 较容易理解，方差估计的公式由 Laplace 近似推理得到，下一小节给出具体的方差估计推理过程。

1.3 方差估计

本小节给出算法 3 中方差估计计算等式 $q_i = q_i + \sum_{j=1}^{n} x_{ij}^2 p_j(1-p_j)$ 的具体推导。假设随机变量 θ 服从高斯分布，因此高斯分布的均值 μ 应该在概率密度函数取最大值的位置，即 $f'(\mu)=0$。假设 θ 服从均值为 μ，方差为 σ^2 的高斯分布，因此有：

$$f(\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-(\theta-\mu)^2}{2\sigma^2}\right] \tag{1.3}$$

$$\Rightarrow f(\mu) = \frac{1}{\sqrt{2\pi}\sigma} \tag{1.4}$$

$$\Rightarrow \ln f(\theta) = \ln f(\mu) - \frac{1}{2\sigma^2}(\theta-\mu)^2 \tag{1.5}$$

对 $\ln f(\theta)$ 在 μ 处取 Taylor 二阶展开式近似，可得：

$$\ln f(\theta) = \ln f(\mu) + \ln' f(\mu)(\theta-\mu) + \frac{1}{2}\ln f''(\mu)(\theta-\mu)^2 \tag{1.6}$$

$$= \ln f(\mu) + \frac{f'(\mu)}{f(\mu)}(\theta-\mu) + \frac{1}{2}\ln f''(\mu)(\theta-\mu)^2 \tag{1.7}$$

由于 $f'(\mu)=0$，因此有：

$$\ln f(\theta) = \ln f(\mu) + \frac{1}{2}\ln f''(\mu)(\theta-\mu)^2 \tag{1.8}$$

结合等式1.5和等式1.8有：

$$\ln f(\mu) - \frac{1}{2\sigma^2}(\theta-\mu)^2 = \ln f(\mu) + \frac{1}{2}\ln f''(\mu)(\theta-\mu)^2 \tag{1.9}$$

$$\Rightarrow \quad \frac{1}{\sigma^2} = -\ln f''(\mu) \tag{1.10}$$

考虑到 Thompson sample 中假设参数间彼此独立，假设采用 MAP 方法进行参数估计，同时参数服从均值为 μ，方差为 σ^2 的高斯先验分布，则有：

$$P(\boldsymbol{\theta}|\mathbb{D}) \propto P(\mathbb{D}|\boldsymbol{\theta})P(\boldsymbol{\theta}) \tag{1.11}$$

对上述两边同时取自然对数可得

$$\ln P(\boldsymbol{\theta}|\mathbb{D}) = \ln P(\mathbb{D}|\boldsymbol{\theta}) + \ln P(\boldsymbol{\theta}) \tag{1.12}$$

同时对二分类问题有：

$$\ln P(\mathbb{D}|\boldsymbol{\theta}) = \sum_{i=1}^{n}\left[P_i \ln \hat{P}_i + (1-P_i)\ln(1-\hat{P}_i)\right] \tag{1.13}$$

考虑参数 $\boldsymbol{\theta}$ 服从均值为 $\boldsymbol{\mu} \in \mathbb{R}^n$，协方差矩阵 Σ 为对角阵的高斯分布

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 & 0 \\ 0 & \sigma_2^2 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \sigma_{n-1}^2 & 0 \\ 0 & 0 & \dots & 0 & \sigma_n^2 \end{bmatrix} \tag{1.14}$$

因此其概率密度函数形式为：

$$P(\boldsymbol{\theta};\boldsymbol{\mu},\Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu})^\top\Sigma^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})\right) \tag{1.15}$$

因此有：

$$\ln P(\boldsymbol{\theta};\boldsymbol{\mu},\Sigma) = -\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu})^\top\Sigma^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu}) + \text{const} \tag{1.16}$$

结合算法 3，因此有：

$$\ln P(\boldsymbol{\theta}|\mathbb{D}) = \underbrace{\sum_{i=1}^{n}\left[P_i \ln \hat{P}_i + (1-P_i)\ln(1-\hat{P}_i)\right]}_{A} \underbrace{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu})^\top\Sigma^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})}_{B} + \text{const} \tag{1.17}$$

上述公式中

$$\hat{P}_i = \frac{1}{1+\exp(-\boldsymbol{\theta}^\top\boldsymbol{x}_i)} \tag{1.18}$$

对公式1.17中 A 部分求关于参数 $\boldsymbol{\theta}$ 的二阶导：

$$f(\boldsymbol{\theta}) = \sum_{i=1}^{n}\left[P_i \ln \hat{P}_i + (1-P_i)\ln(1-\hat{P}_i)\right] \tag{1.19}$$

$$\Rightarrow \frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n}\left[\frac{P_i}{\hat{P}_i}\frac{\partial \hat{P}_i}{\partial \boldsymbol{\theta}} + \frac{1-P_i}{1-\hat{P}_i}\frac{\partial(1-\hat{P}_i)}{\partial \boldsymbol{\theta}}\right] = \sum_{i=1}^{n}\left[\left(\frac{P_i}{\hat{P}_i} - \frac{1-P_i}{1-\hat{P}_i}\right)\frac{\partial \hat{P}_i}{\partial \boldsymbol{\theta}}\right] \tag{1.20}$$

$$= \sum_{i=1}^{n}\left[\left(\frac{P_i}{\hat{P}_i} - \frac{1-P_i}{1-\hat{P}_i}\right)\frac{\partial \hat{P}_i}{\partial \boldsymbol{\theta}}\right] = \sum_{i=1}^{n}\left[\frac{P_i-\hat{P}_i}{\hat{P}_i(1-\hat{P}_i)}\frac{\partial \hat{P}_i}{\partial \boldsymbol{\theta}}\right] \tag{1.21}$$

$$= \sum_{i=1}^{n}\left[(P_i-\hat{P}_i)\boldsymbol{x}_i\right] \tag{1.22}$$

$$\Rightarrow \frac{\partial^2 f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} = -\sum_{i=1}^{n}\left[\hat{P}_i(1-\hat{P}_i)\boldsymbol{x}_i\boldsymbol{x}_i^\top\right] \tag{1.23}$$

对公式1.17中 B 部分求关于 $\boldsymbol{\theta}$ 的二阶导:

$$g(\boldsymbol{\theta}) = -\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu})^\top\Sigma^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu}) \tag{1.24}$$

$$\Rightarrow \frac{\partial^2 g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} = -\Sigma^{-1} \tag{1.25}$$

因此有:

$$\frac{\partial^2 \ln P(\boldsymbol{\theta}|\mathbb{D})}{\partial \boldsymbol{\theta}^2} = -\Sigma^{-1} - \sum_{i=1}^{n}\left[\hat{P}_i(1-\hat{P}_i)\boldsymbol{x}_i\boldsymbol{x}_i^\top\right] \tag{1.26}$$

由于 Σ 为正定对角阵，因此有:

$$\Sigma^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \dots & 0 & 0 \\ 0 & \frac{1}{\sigma_2^2} & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sigma_{n-1}^2} & 0 \\ 0 & 0 & \dots & 0 & \frac{1}{\sigma_n^2} \end{bmatrix} \tag{1.27}$$

其中 $\frac{1}{\sigma_i}$ 即为算法 3 中的精度 q_i，结合等式1.10可知:

$$q_i = \frac{1}{\sigma_i^2} = -\frac{\partial^2 \ln P(\boldsymbol{\theta}|\mathbb{D})}{\partial \theta_i} = q_i + \sum_{j=1}^{n}\left[x_{ij}^2\hat{P}_i(1-\hat{P}_i)\right] \tag{1.28}$$