

动态贝叶斯网络点击模型

The Dynamic Bayesian Network Click Model

Liam Huang*

2018 年 4 月 3 日

*Liamhuang0205@gmail.com

基础知识

目标问题

- Machine Learning Ranking 都是有监督学习.
- 对 query-document 相关性的人工标注成本高.
- 用户点击日志暗含了 query-document 相关性.
- 怎样利用用户点击日志, 获得 query-document 相关性?

相关概念

- Click Through Rate: 某种限制条件下的点击次数比上展现次数.
 - 全局下的: $\text{Global CTR} = \frac{\# \text{ clicks}}{\# \text{ shown docs}}$
 - 某 rank 位置上的: $\text{CTR}_r = \frac{\# \text{ clicks at rank } r}{\# \text{ shown docs at rank } r}$
- Examination: 用户对展现结果的观察行为.
- Perceived Relevance: 用户观察之后, 根据展现情况感知的相关性.
- Click: 用户的点击行为.
- Actural Relevance: 用户点击后, 根据页面实际内容判断的相关性.
- Position Bias: 由于展现位置带来的对 CTR 的影响.

机器学习模型三要素

- 模型结构 ← 模型是如何描述目标问题的？
- 目标函数 ← 怎样评判模型效果的优劣？
- 优化算法 ← 怎样求解目标函数上的优化问题？

全都是套路.

Impressive Approaches¹

Position Models

模型假设:

- Click = Examination + Perceived Relevance.
- $P(E = e | u, p) = P(E = e | p)$.
- $P(C = 1 | u, p, E = 1) = P(C = 1 | u, E = 1)$.

$$\begin{aligned} P(C = 1 | u, p) &= \sum_{e \in \{0,1\}} P(C = 1 | u, p, E = e) \cdot P(E = e | u, p) \\ &= \underbrace{P(C = 1 | u, E = 1)}_{\stackrel{\text{def}}{=} \alpha_u} \cdot \underbrace{P(E = 1 | p)}_{\stackrel{\text{def}}{=} \beta_p} \end{aligned}$$

- 假设: $\beta_1 \stackrel{\text{def}}{=} 1$,
- 则 α_u 表示当 u 位于 rank 1 时的 CTR.

¹Despite the impressive progress made so far, this goal has remained elusive. In this paper, we achieve this. —PRIMES is in P

COEC (Clicks Over Expected Clicks) 模型

继续假设, β_p 对于所有 query 和 session 是一致的, 则有:

$$\alpha_u \stackrel{\text{def}}{=} \frac{\sum_{i=1}^N c_i}{\sum_{i=1}^N \beta_{p_i}}.$$

问题: α_u 不仅包含了位置本身的信息 (position bias), 还包含了特定位置结果的平均相关性.

Examination 模型

解决: 在 N 个 session 中, 观察同一 URL 在不同位置的 CTR; 最大似然:

$$\alpha_u = \arg \max_{\alpha} \sum_{i=1}^N c_i \log(\alpha \beta_{p_i}) + (1 - c_i) \log(1 - \alpha \beta_{p_i}).$$

问题: 保证 $0 \leq \alpha \beta \leq 1$, 但无法保证 $0 \leq \alpha \leq 1$.

解决: 将 E 视作隐变量, 运用 Expectation-Maximization 算法解决.

问题: 原假设 $P(E = e | u, p) = P(E = e | p)$ 忽略了 URL 之间的相互作用.

Cascade Model

假设:

- Click = Examination + Perceived Relevance.
- 第 i 条结果的相关性: $P(A_i = 1) = \alpha_{u_i}$.
- 从第一条开始检查: $P(E_1 = 1) = 1$.
- 逐条检查: $P(E_i = 1 \mid E_{i-1} = 0) = 0$.
- 有点击后终止检查: $P(E_i = 1 \mid C_{i-1} = 1) = 0$.
- 无点击则继续检查: $P(E_i = 1 \mid C_{i-1} = 0) = 1$.

则有:

$$P(C_i = 1) = \alpha_{u_i} \prod_{j=1}^{i-1} (1 - \alpha_{u_j}).$$

问题:

- 点击了就满意了吗?
 - 无法处理多次点击的情况.
 - 无法区分看起来相关和实际上相关.
- 无法处理无点击的情况.

动态贝叶斯网络

模型

假设:

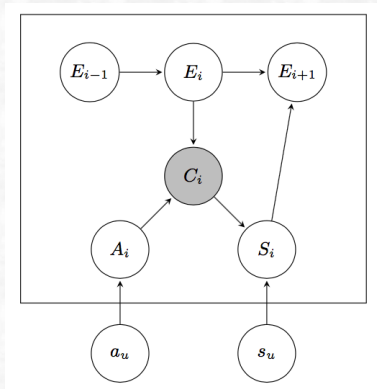
- Click = Examination + Perceived Relevance.
- 第 i 条结果的感知相关性: $P(A_i = 1) = a_{u_i}$.
- 从第一条开始检查: $P(E_1 = 1) = 1$.
- 逐条检查: $P(E_i = 1 \mid E_{i-1} = 0) = 0$.
- 有点击后实际相关性: $P(S_i = 1 \mid C_i = 1) = s_{u_i}$.
- 无点击则不被满足: $P(S_i = 0 \mid C_i = 0) = 1$.
- 被满足则不再检查: $P(E_{i+1} = 0 \mid S_i = 1) = 1$.
- 不满足时可能继续检查: $P(E_{i+1} = 1 \mid S_i = 0, E_i = 1) = \gamma$.

实际相关性:

$$\begin{aligned} r_{u_i} &\stackrel{\text{def}}{=} P(S_i = 1 \mid E_i = 1) \\ &= P(S_i = 1 \mid C_i = 1)P(C_i = 1 \mid E_i = 1) \\ &= s_{u_i} a_{u_i}. \end{aligned}$$

概率图

- 框内: session 级别变量.
- 框外: query 级别变量.
- 黑底: 非隐变量.
- 白底: 隐变量.



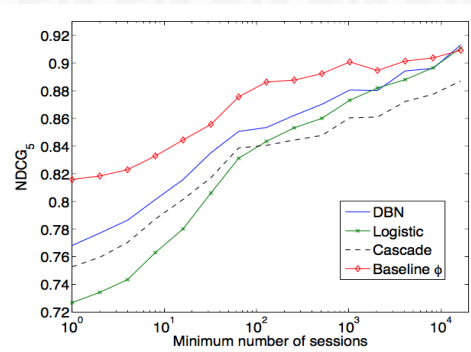
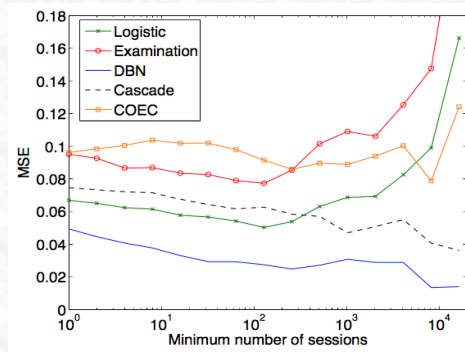
- 类似隐马模型.
- Expectation-Maximization 算法求解.
- γ 体现用户耐性, 可作为隐变量估计, 也可作为超参数统一配置.

实验结果

超参数 $\gamma \stackrel{\text{def}}{=} 0.9$.

CTR 准确率

作为 ranking 信号



A spiral-bound notebook with a white cover and a black spiral binding on the left side. The notebook is open to a blank white page. The letters "UGA" are printed in a large, bold, black, sans-serif font in the center of the page. The "U" is on the left, the "G" is in the middle, and the "A" is on the right. The letters are evenly spaced and have a consistent thickness. The background of the page is plain white, and the spiral binding is visible on the left edge.

UGA