点击率预估校准调研报告

张航　　黄蔚

April 2017

点击率预估校准是为了解决降采样带来的 pctr 大于广告真实点击率的问题, 主要的评估指标是**偏差率**, 定义为

$$\left|\frac{\overline{pctr} - ctr}{ctr}\right|$$

即某个素材当天的平均 pctr 和 ctr 相对误差. 我们校准的目标是控制每个广告素材的偏差率小于 20%. 由于目前我们的模型更新周期是天，所以偏差主要分为两部分：

1. 训练误差，即模型在训练集上平均预估 ctr 与真实 ctr 的偏差；
2. ctr 波动，同一素材每天的点击率会有变化，特别是投放周期的开始和结束阶段, 因此我们用历史数据训练校准模型自然后会有偏差.

现在线上使用均值偏移校准方法，其表现基本稳定但存在两个问题:

1. 原生广告位偏差相对较大，测试期平均偏差率在 15% 左右，其中最大值超过了 20%
2. 存在较大的训练误差，且加入新特征后训练误差会变大（1.1 节从理论和数据上说明了这一点）.

我们调研了 Google 论文 [1] 提出的分段单调校准方法. 测试表明新方法：

1. 大幅降低了偏差率，原生广告平均偏差率由 15% 降到 9.9%, 视频关联位维持在 6% 左右的低位基本不变
2. 提升了校准方法的鲁棒性，原生广告位五天测试周期内偏差超过 20% 的素材比例由 28% 降至 9%，视频关联位维持在 8% 左右的低位
3. 理论分析和数据验证训练误差几乎为 0

下面对两种方法校准方法分别给出了误差分析。

1 误差分析

1.1 均值偏移校准

均值偏移校准使用校准前平均 pctr 和 ctr, 利用 Sigmoid 函数的反函数 Logit 函数计算模型 bias 所需的偏移量，从而校准 pctr 到接近 ctr. 这种方法的主要问题是训练误差受到 pctr 分布的影响，当 pctr 方差较大时预估的 pctr 容易偏大.

设校准前 pctr 为随机变量 $X\colon \Omega \to [0,1]$, 期望 μ, 方差 σ^2, 概率密度函数 $p(x)$, 已知真实点击率 $c \in (0,1)$, 校准后 pctr 为随机变量

$$\begin{aligned} Y &\coloneqq \operatorname{sigm}\left(\operatorname{logit}(X) - \operatorname{logit}(\mu) + \operatorname{logit}(c)\right) \\ &= \operatorname{sigm}\left(\log\left(\frac{X}{1-X}\cdot\frac{1-\mu}{\mu}\cdot\frac{c}{1-c}\right)\right) \\ &= \frac{1}{1+(\frac{1}{X}-1)(\frac{1}{1-\mu}-1)(\frac{1}{c}-1)} \\ &= \frac{c(1-\mu)X}{\mu(1-c)-(\mu-c)X} \end{aligned}$$

其中 $\operatorname{logit}(p) = \log\left(p/(1-p)\right)$, $\operatorname{sigm}(p) = 1/(1+\exp(-p))$. 因此训练集上误差

$$\begin{aligned} err_{sys} &\coloneqq \frac{\mathbf{E}(Y)-c}{c} \\ &= \frac{1}{c}\left(\int_0^1 \frac{c(1-\mu)x}{\mu(1-c)-(\mu-c)x}p(x)\mathrm{d}x\right) - \int_0^1 p(x)\mathrm{d}x \\ &= \int_0^1 \frac{(1-c)(x-\mu)}{\mu(1-c)-(\mu-c)x}p(x)\mathrm{d}x \\ &= \int_0^1 \frac{1}{\mu}\frac{1}{1-\frac{\mu-c}{\mu(1-c)}x}(x-\mu)p(x)\mathrm{d}x \\ &\approx \int_0^1 \frac{1}{\mu}\left(1+\frac{\mu-c}{\mu(1-c)}x\right)(x-\mu)p(x)\mathrm{d}x \\ &= \frac{1}{\mu^2}\frac{\mu-c}{1-c}\int_0^1 x(x-\mu)p(x)\mathrm{d}x \\ &= \frac{\sigma^2}{\mu^2}\frac{\mu-c}{1-c}, \end{aligned} \tag{1}$$

其中用了 Taylor 展开取第一项和

$$\int_0^1 (x-\mu)p(x)\mathrm{d}x = 0.$$

根据(1), 当校准前 pctr 均值确定时, 训练误差和 σ^2 正相关. 我们可以得出训练误差小的概率密度函数需要满足:

1. 方差较小, 大部分点分布在均值附近.

2. 右偏态, 右侧的尾部更长，分布的主体集中在左侧.

3. 右侧的长尾部分的最大值较小

下面用真实数据验证上面的结论. 我们取三个素材在 2017-04-05 上的校准前 pctr, 假设真实 ctr 都为 0.02. 由于它们校准前 pctr 均值都在 0.14 附近, 因此影响训练误差的只有 pctr 分布. 如图 1所示, 从上到下三个素材误差依次增大. 54000370540 的 pctr 基本集中在均值附近且比较对称, 右侧长尾部分几乎没有样本所以误差最小, 而 54000507885 在均值附近的对称性不如前者因此误差变大, 54000495399 的长尾部分样本最多导致其误差最大.

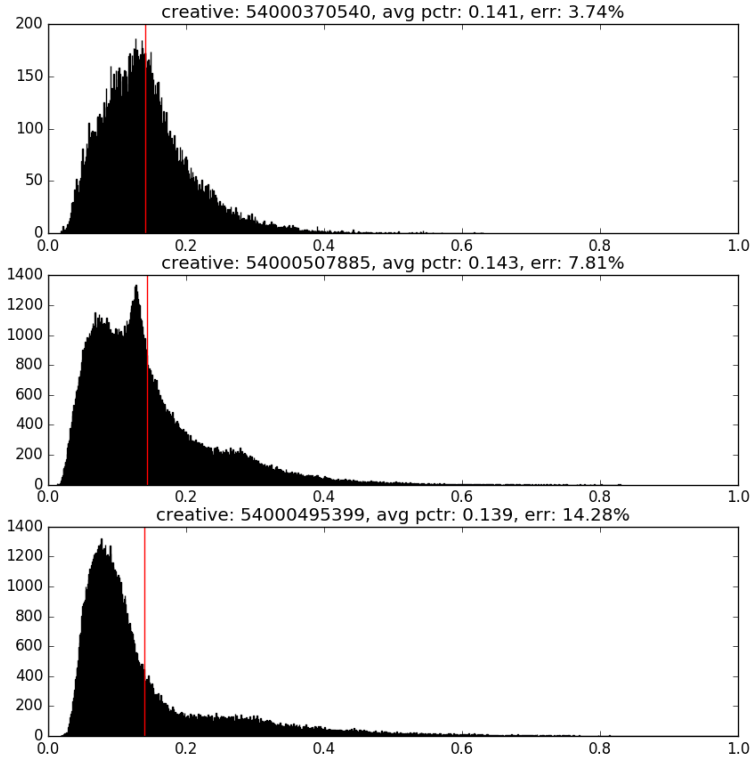

图 1: pctr 分布对系统误差的影响

1.2 分布形态归因

我们在素材 54000507885 上分析 pctr 出现不同分布的原因, 从图 2可以看出在加入用户观影偏好 (favorite entity) 特征之前, pctr 有两个明显的峰, 右侧的峰较小. 进一步地特征分析可知, 信息流广告位主要出现在三个频道: 热点、资讯和搞笑. 热点的流量最多约占 85%, 后两者共计 15%, 而资讯和搞笑的点击率远高于热点, 该素材在资讯搞笑频道下的点击率达到 4.2%, 热点下只有 1.3%, 因此 pctr 根据频道分成两部分, 第二个峰主要对应资讯搞笑的流量.

加入用户观影特征后, 由于该特征对流量的区分性较强, pctr 的分布范围明显增大, 部分流量的 pctr 增加, 另一部分减小. 因此可以得出**有效特征会使得 pctr 分化, 出现更多的高值和低值**, 结合上面的结论, 这会导致**训练误差变大**.

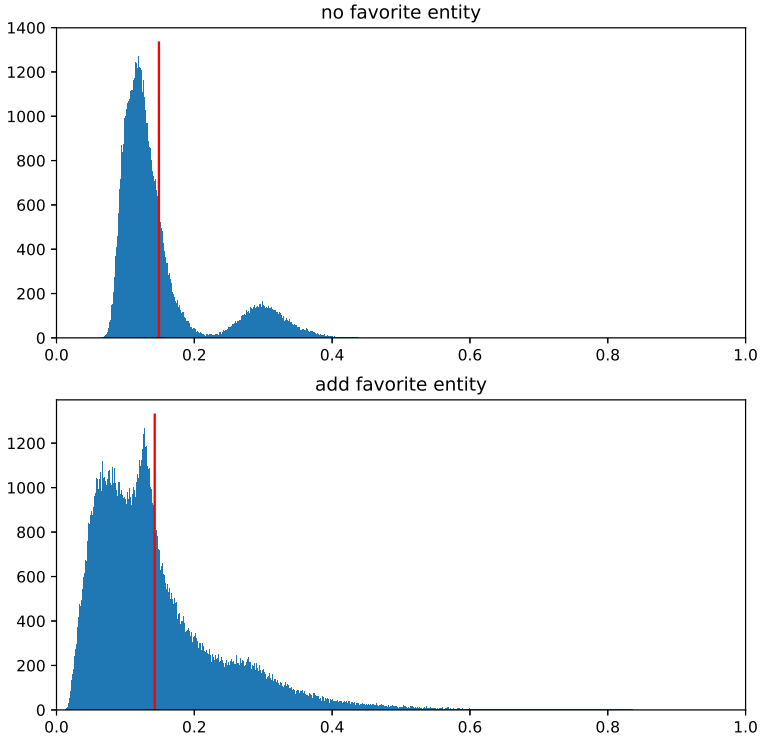

图 2: 素材 54000507885 加入观影偏好特征前后的 pctr 分布

1.3 分段单调校准

分段单调校准首先将训练数据按 pctr 排序之后切, 假设分成 n 个片段, 每片包含相同数量的样本, 第 i 片数据记为 $\mathcal{D}_i$, 我们可以得到 $\mathcal{D}_i$ 上的点击率 c_i, 每片上样本的 pctr 取平均得到 p_i, 满足 $p_1 < p_2 < \cdots < p_n$. 然后使用保序回归 (Isotonic Regression) 拟合点 (p_i, c_i), 即求解优化问题

$$\begin{aligned} &\min_{d_i} \sum_i w_i (d_i - c_i)^2 \\ &\text{s.t. } d_1 \le \cdots \le d_n. \end{aligned} \tag{2}$$

(2)的最优解 $\hat{d}_i$ 存在唯一, 如果取 $w_i = 1$, 再加上两端的默认点 $(p_0 = 0, c_0 = 0), (p_{n+1} = 1, c_{n+1} = c_{max})$ 我们可以得到校准函数

$$f(x) := \frac{\hat{d}_{i+1} - \hat{d}_i}{p_{i+1} - p_i}(x - p_i) + \hat{d}_i, \ p_i \le x < p_{i+1}.$$

如图 3所示, $f(x)$ 是单调分片线性函数.

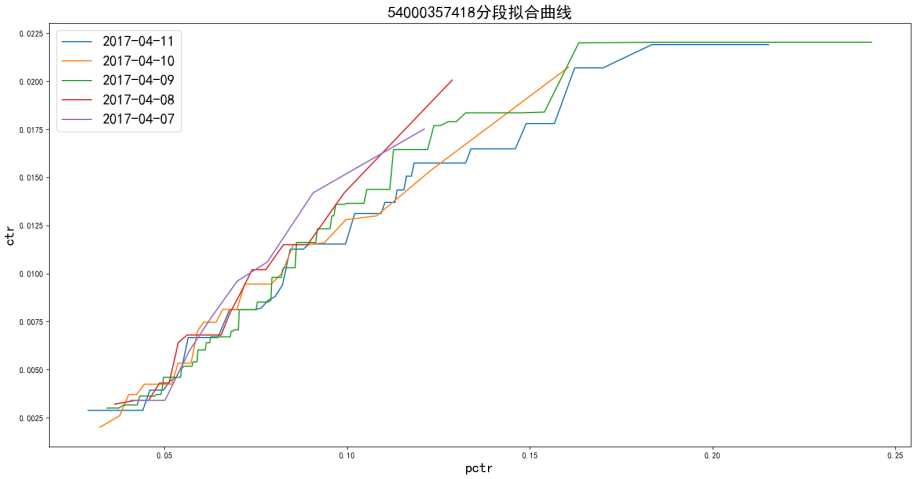

图 3: 单个素材在不同天的分段单调校准函数

因此校准后随机变量 $Y = f(X)$，我们有

$$\begin{aligned}\mathbf{E}(Y) &= \sum_i \int_{p_i}^{p_{i+1}} f(x)p(x)\mathrm{d}x \\ &= \sum_i \left(\frac{\hat{d}_{i+1} - \hat{d}_i}{p_{i+1} - p_i} \int_{p_i}^{p_{i+1}} (x - p_i)p(x)\mathrm{d}x + d_i \int_{p_i}^{p_{i+1}} p(x)\mathrm{d}x \right) \\ &\approx \frac{1}{n} \sum_i \frac{\hat{d}_{i+1} - \hat{d}_i}{2} + \frac{1}{n} \sum_i \hat{d}_i \\ &\approx \frac{1}{n} \sum_{i=1}^{n} \hat{d}_i\end{aligned}$$

由优化问题(2)的 KKT 条件可知 $\sum_i \hat{d}_i = \sum_i c_i$, 因此分段校准**训练误差近似为 0，且基本不受 pctr 分布的影响**.

1.4 测试结果

详细测试数据见表 1

	原生广告位		视频关联位	
日期	线上	分段	线上	分段
2017/4/14	12.20%	**8.90%**	**5.70%**	8.30%
2017/4/15	14.30%	**8.90%**	**5.30%**	6.40%
2017/4/16	15.00%	**5.30%**	**5.40%**	7.70%
2017/4/17	22.90%	**17.60%**	7.60%	**7.30%**
2017/4/18	13.70%	**8.90%**	6.10%	**5.00%**
2017/4/19	**8.30%**	8.60%	**6.60%**	8.60%
平均	15.60%	**9.90%**	**6.00%**	6.90%

表 1: 校准方法详细偏差率对比数据

2 后续优化

影响分段单调校准效果的主要因素是 CTR 波动，同一广告素材的点击率随时间变化会有较大幅度波动, 尤其在投放的初期和末期. 如图 4所示, 波动带来的误差在现有框架下较难解决，我们预计未来上线 online learning 后能有效感知 ctr 变化，调整模型, 减小偏差.

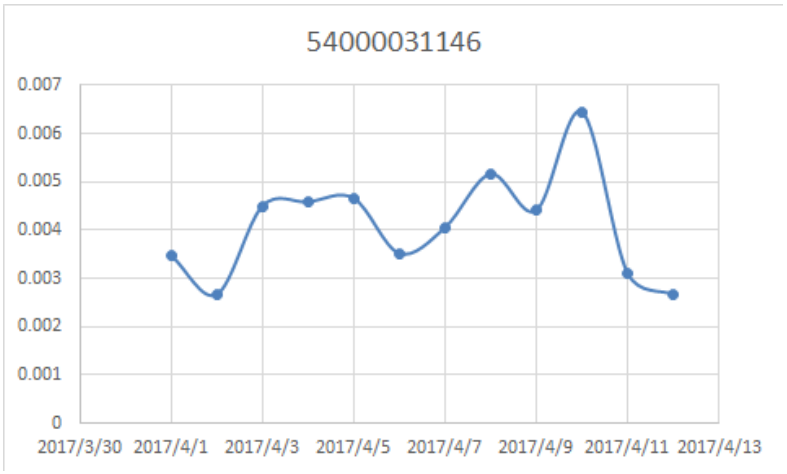

图 4: 素材在整个投放周期的 ctr 波动，可以看出投放前期 ctr 呈现上升趋势，末期下降明显

References

[1] H Brendan McMahan et al. “Ad click prediction: a view from the trenches”. In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2013, pp. 1222–1230.