逻辑回归预测CTR经典论文解读

通过上一篇文章（逻辑回归简介及实现）的介绍，我们对逻辑回归方法有了大致的了解。但对于点击率（Click-Through-Rate, CTR）预估问题而言，逻辑回归是怎么尝试解决的呢？为此，我阅读了一篇有关逻辑回归预测CTR的经典论文，并将阅读笔记整理在这里，一是为了加深自己对知识的理解和掌握，二是希望能够对有着相同疑惑的朋友起到一点点帮助。

论文信息

- 论文题目：Predicting Clicks: Estimating the Click-Through Rate for New Ads
- 发表于：Proceedings of the 16th international conference on World Wide Web，2007
- 引用次数：853（截至2020年04月22日）
- 论文地址：https://dl.acm.org/doi/pdf/10.1145/1242572.1242643

Motivation

这篇论文的题目表明了论文的目的——预测新广告的CTR。CTR我们知道是广告点击与广告展现的比率，那么为什么要预测CTR呢，而且为什么要预测新广告的CTR呢?

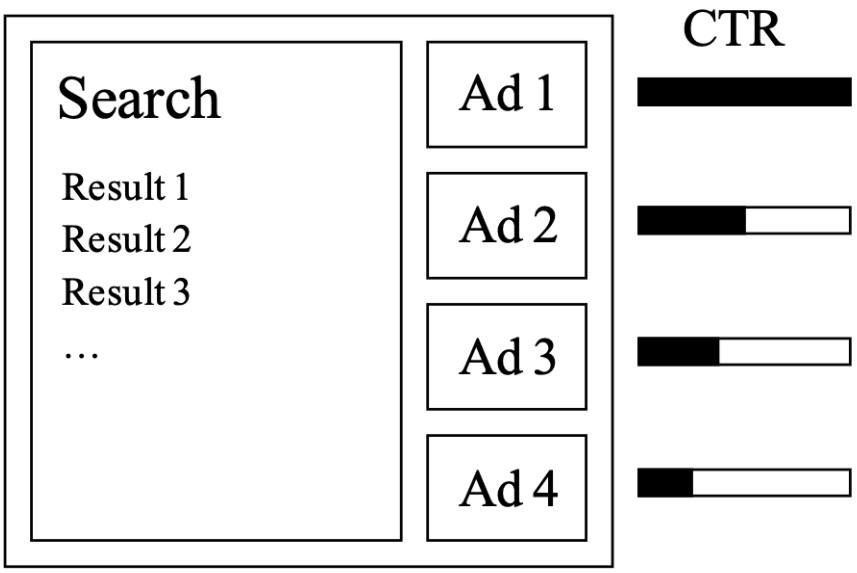
- 为什么需要预测CTR?

  目前投放广告是包括搜索引擎（Search Engine）在内的很多互联网公司重要盈利方式之一，比如以点击结算方式（Cost per Click，CPC）为例，**用户**每点击一次广告，**广告主**（推销产品的公司，比如可口可乐、百事可乐）都要付给**媒体**（负责投放广告的公司，比如谷歌、百度）一定的费用。此时，媒体或者本文主要关注的搜索引擎公司对某一个广告（ad）的期望收入可以写为：

$$E_{ad}[revenue] = p_{ad}(click) \cdot CPC_{ad}, \qquad (1)$$

  其中CPC_{ad}是规定的广告主为用户每点击一次广告付给媒体的费用，也就是点击单价，可以通过竞价的方式产生。$p_{ad}(click)$是用户点击该广告的概率，也就是点击率CTR。

  我们需要清楚的是，用户在使用搜索引擎时，搜索引擎需要根据用户的检索信息（query）来给出一组候选广告，并且广告的CTR与广告出现的位置有很大关系，其场景如下图所示：



这里先抛开广告位置对CTR的影响（下文会提到），搜索引擎需要在用户检索时给出一组候选广告的排序。在决定广告排序时，不能像返回用户检索结果$Result1, Result2, Result3$一样仅凭内容相关度大小（假设仅考虑广告放置在如上图所示的右侧情况），而是需要根据每个广告的期望收入$E_{ad}[revenue]$，即点击率$p_{ad}(click)$乘上点击单价CPC_{ad}，来得到最后的候选广告排序。而用户在每次使用搜索引擎时，我们可以认为此时的CPC_{ad}已经是固定的，搜索引擎只有准确的预测出用户点击广告的概率$p_{ad}(click)$或CTR，才能给出一个能够最大化期望收入的候选广告排序。

一个质量较好的广告排序，不仅能够最大化搜索引擎的广告收入，也能带动广告主的产品销售，同时用户满意度也会更高。而这其中的关键便是预测CTR。

- 为什么要预测新广告的CTR?

  为了最大化广告收入同时提升用户满意度，搜索引擎需要较为准确的预测出用户点击某一广告的概率（CTR）。虽然搜索引擎可以通过用户点击广告的历史记录来统计出以往某一广告的CTR，比如，一条广告在过去被展示了100次，用户点击次数为5次，那么其CTR可以估计为0.05，但是这样的估计是带有极大的方差（Variance）的，并且对于一个新广告而言，没有任何展示的历史记录，通过统计历史信息来预估CTR的方法变得更加困难。

  而且搜索广告市场近年来（论文时间2007年）不断发展，广告主每年每月甚至每天都在举办不同的广告推销活动，这给搜索引擎造成了大量的广告库存（Inventory），这些库存广告对于搜索引擎来说没有任何历史展示记录，也就是新广告。因此对于那些没有展示过的新广告或者展示次数不够多的广告，搜索引擎必须找到一个合理的方式来较为准确的预测其CTR，以提升广告收入和用户满意度。

另外需要说明的是，在这篇论文中，其所说的一个广告（ad）代表的是一条广告内容、一个广告主和一个关键词的组合。对于一条广告内容带有多个关键词的情况，这篇论文视其为多个广告。

Search advertising framework

当一个广告被展示在搜索结果页面的时候，它有一定的机率被用户看到，而且广告所处的位置越偏向页面下方，其被用户看到的概率也越低。在这篇论文中，作者做了一个简化，假定广告的点击率仅仅依赖于两个因素：

1. 广告被用户浏览到的概率
2. 在广告被用户浏览到的情况下，用户点击广告的概率

因此，作者将广告被点击的概率写为：

$$p(click|ad, pos) = p(click|ad, pos, seen)p(seen|ad, pos), \quad (2)$$

其中$p(seen|ad, pos)$表示的是一条广告（ad）在某一页面位置（pos）进行展示时被用户浏览到的概率，$p(click|ad, pos, seen)$表示的是在用户浏览到位置（pos）的广告（ad）后点击它的概率。当然这样的写法有一个前提，就是我们假设对于所有被点击的广告，其不被用户浏览到的概率为0，也就是只有在用户浏览到的情况下，该广告才有可能被点击。同时在这篇论文中，作者假设广告被浏览到与否，与广告本身的内容无关，而仅与放置的位置有关，并且假设在广告已经被用户浏览到的情况下，用户点击与否便和位置没有了关系，仅和广告本身内容有关，于是，广告被点击的概率可以简写为：

$$p(click|ad, pos) = p(click|ad, seen)p(seen|pos). \quad (3)$$

在这篇论文中，作者研究的CTR便是用户在看到广告的情况下点击广告的概率，即$p(click|ad, seen)$。而$p(seen|pos)$不是本文的研究范围，因为其被视为一个固定值，可以通过进行实验获得这一个值（比如把同一个广告放在不同的位置，统计所有用户中看到该广告的比例）。但是一旦较为准确的预测出$p(click|ad, seen)$，通过乘上一个系数$p(seen|pos)$，我们可以得到任何位置广告最后被点击的概率$p(click|ad, pos)$。

Data set

这篇论文收集了微软公司搜索引擎投放广告的一系列信息，来作为实验所用数据集。数据集中包含着来自1万个广告主的广告，且广告总数超过1百万条，其中每条广告都包含下面的信息：

- Landing page: 落地页，包括当用户点击时转向落地页的URL。
- Bid term("keywords"): 关键词，当用户输入关键词时，搜索引擎会投放相对应的广告。
- Title: 广告的标题。
- Body: 广告内容描述部分。
- Display URL: 广告中展示给用户的URL，比如广告内容里展示的“www.taobao.com”。
- Clicks: 广告自被投放以来，用户的点击数量。
- Views: 广告自被投放以来，用户浏览到的数量（通过实验获得）。

在这篇论文中所加的一个限制就是，作者研究的是预测新广告主的新广告CTR问题，也就是说，这类新广告既没有历史展示记录，也没有同一个广告主之前投放其他广告的记录。作者表明在以后的研究中再去考虑知道同一个广告主之前的广告投放信息，进而预测同一个广告主新广告的CTR。因此在划分数据集时，作者按照广告主划分数据，将70%、10%和20%的广告主分别划分到训练集、验证集和测试集中，并将隶属于同一个广告主的所有广告全部划分在同一个部分，比如同一个广告主的广告不会同时出现在训练集和验证集中，而只存在于三个部分中的任意一个。

同时，作者将数据集中那些浏览量低于100的广告过滤掉，因为这些广告浏览量过低，其统计出来的CTR方差大，但同时如果只选择那些浏览量过高的广告，又会使得模型更偏向此类广告，预测偏差大。还有就是作者对每个广告主最多随机选取1000个广告，防止不同广告主的广告数量相差过多。

Model

因为我们的目标是为了预测每个广告的CTR（一个0-1之间的实数值），在这篇论文中，将预估CTR问题看作是一个回归问题，并且采用形式如下的逻辑回归方法进行求解：

$$CTR = \frac{1}{1 + e^{-Z}}, Z = \sum_i w_i f_i(ad), \quad (4)$$

其中$f_i(ad)$代表的是广告ad的第i个特征，w_i代表的是相应的特征权重。在下文中，作者根据问题的具体情况构造了一系列的特征来求解问题，当然作者也尝试了梯度提升树方法，发现并没有逻辑回归的效果好。

同时在数据预处理阶段，为了提升模型的性能，作者对数据集进行特征标准化，为了防止离群点的影响，将所有超过5倍标准差的特征值截断为5倍标准差。在进行模型求解时，采用交叉熵为损失函数，但在后续实验结果中，采用模型预测值和测试集真实值之间的KL散度和均方误差进行展示（因为测试集的熵是固定的，而KL散度是熵和交叉熵的组合）。

在接下来的章节里，作者着重介绍了他们所构造的特征以及展示特征对模型性能带来的性能提升情况。

Estimating term CTR

在这节中，作者构造了一系列和广告关键词相关的特征，用来捕捉广告关键词对CTR的影响，进而提升模型的性能。

我们知道不同关键词的广告之间CTR相差很大，但是当我们训练好一个模型之后，往往希望它能够对带有同一个关键词的广告也能够给出相对合理的CTR预测。也就是说希望模型能够抓住那些带有同样关键词广告的共性，为此，作者构造了两类与关键词有关的特征。

- Term CTR

  作者首先关注的是带有相同关键词的其他广告的CTR，提出这个特征的思路是认为带有同样关键词的广告CTR之间有一定的联系，所以第一个特征具体计算方式是：

$$f_0(ad) = \frac{\alpha\overline{CTR} + N(ad_{term})CTR(ad_{term})}{\alpha + N(ad_{term})}, \qquad (5)$$

  其中$N(ad_{term})$代表的是带有同样关键词的广告数量，$CTR(ad_{term})$是这些带有同样关键词广告的平均CTR，而$\overline{CTR}$代表的是训练集上所有广告的平均CTR，加上这一项的目的是为了防止遇到一些少见的关键词，没有带有相同关键词的广告，而α可以视为平滑系数，在后续实验中其值被置为1。同时，作者将$N(ad_{term})$作为一个特征加入到逻辑回归模型中。

  因此，第一类特征包含两个特征，一是其他带有相同关键词广告的平均CTR，二是其他带有相同关键词的数量。

- Related Term CTR

  作者接着关注的是那些带有相关关键词广告的CTR。这里说的相关关键词，其定义如下：

$$\mathbf{R_{mn}(t)} = \left\{ \mathbf{ad:} \begin{array}{c} |ad_{term} \cap t| > 0 \; and \\ |t - ad_{term}| = m \; and \\ |ad_{term} - t| = n \end{array} \right. \qquad \mathbf{(6)}$$

其中$\mathbf{R_{mn}(t)}$指的是那些带有可以通过一定的编辑操作变成和关键词t一样的关键词的广告集合。这里说的编辑操作可以理解为词语之间的编辑距离，如果关键词ad_{term}定义为关键词t的相关关键词，那么将关键词t删除m个单词，同时将关键词ad_{term}删除n个关键词之后，二者会变为同样的关键词。

我其实不是很清楚上面给出的关于$\mathbf{R_{mn}(t)}$的定义公式，但根据后面作者给出的例子，我所理解的编辑操作是没有错的。比如，如果t为“red shoes”，那么带有关键词"buy red shoes"的广告会出现在集合$\mathbf{R_{01}}$中，同时带有关键词“shoes”的广告会出现在集合$\mathbf{R_{10}}$中，而带有关键词"blue shoes"的广告会出现在集合$\mathbf{R_{11}}$中。所以$\mathbf{R_{m0}}$中的广告，其关键词相比于t而言，少了m个词，而集合$\mathbf{R_{0n}}$中的广告，其关键词相比于t而言，是多了n个词。同时这里的m, n可以用$*$来代替，代表任意值。

当获取到广告集合$\mathbf{R_{mn}}$之后，那么相关广告的平均CTR可以通过如下的方式计算：

$$CTR_{mn}(term) = \frac{1}{|\mathbf{R_{mn}(term)}|} \sum_{x \in \mathbf{R_{mn}(term)}} CTR_x. \tag{7}$$

并且$CTR_{mn}(term)$使用平均值$\overline{CTR_{mn}}$使用与上一类特征同样的方式来进行平滑操作，同时也将相关广告的数量加入逻辑回归模型中作为一个特征:

$$v_{mn}(term) = |R_{mn}(term)|, m, n \in \{0, 1, 2, *\}. \tag{8}$$

因此，第二类特征也包含两个特征，一是带有相关关键词的广告平均CTR，二是带有相关关键词的广告数量。

在将这两类所构造的特征加入模型之后，相比于Baseline（仅仅使用$\overline{CTR}$一个特征，相当于不进行任何学习），其提升效果如下图所示：

Table 1: *Term* and *Related Term* Results

Features	*MSE (x 1e-3)*	*KL Divrg. (x 1e-2)*	*% Imprv.*
Baseline ($\overline{CTR}$)	4.79	4.03	-
Term CTR	4.37	3.50	13.28%
Related term CTRs	4.12	3.24	19.67%

如上图所示，在加入Term CTR特征集合后，KL散度提升了13.28%，在Term CTR的基础上，再加入Related term CTRs特征，模型KL散度一共提升了19.67%（其实根据表格中的KL散度结果计算出来的提升数值分别为13.15%和19.60%，但我认为表格中的数字是四舍五入后的结果，而表格中的提升数值是四舍五入之前计算得到的）。

所以在这一小节中，作者构造了两类特征，一是考虑包含相同关键词广告的CTR，二是考虑包含相关关键词广告的CTR。

Estimating ad quality

在上一节中，作者构造了两类和广告关键词有关的特征集合，但要知道预测CTR仅凭广告关键词是不够的，如下图所示，即使对那些带有相同关键词的广告来说，其CTR方差依然很大，比如带有关键词“digital cameras”的广告来说，CTR的最大值比均值高过3倍以上，带有“surgery”关键词的广告，出现的最大CTR比均值高过5倍以上：

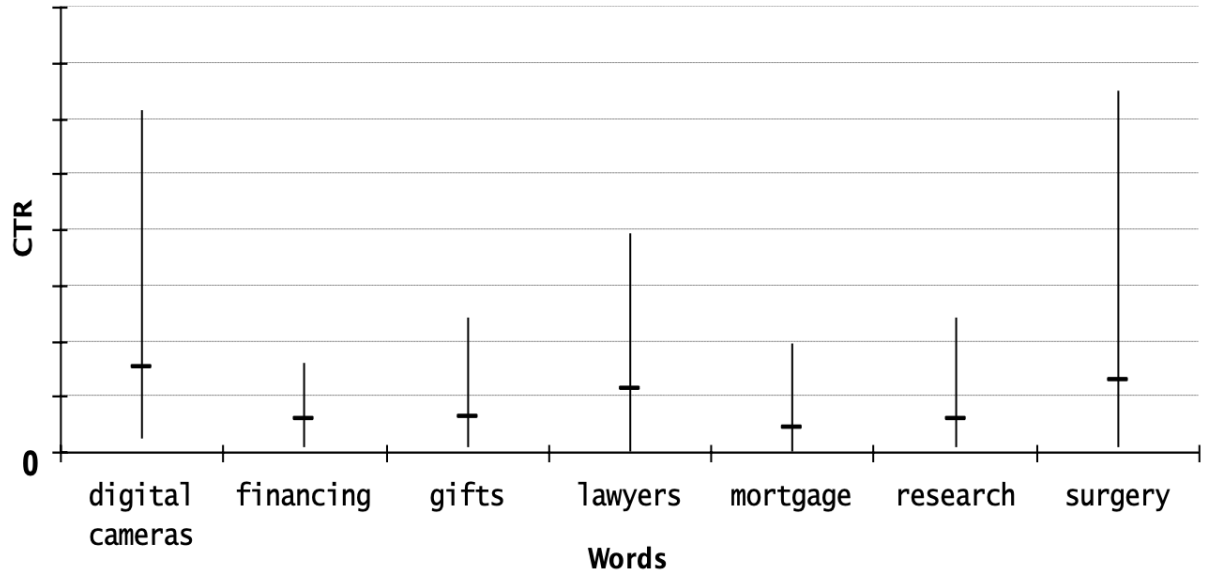


Figure 3. CTR variance across all ads for several keywords. Horizontal bars show average CTR; the bottom of the vertical bar is the minimum CTR, and the top is the maximum CTR.

因此在这一小节中，作者探讨了广告本身的质量对其CTR的影响，为此，构造了下面的5类特征：

- Appearance: 广告从美学设计上是否合理?
- Attention Capture: 广告是否足够吸人眼球?
- Reputation: 广告主知名度高吗？如果用户不熟悉广告主，那他们会觉得这个牌子是一个好牌子吗?
- Landing page quality: 虽然落地页是用户在点击广告之后才会出现，但是作者假设落地页的质量影响了用户对广告主品牌的看法，高质量的落地页也可能会引发用户多次点击。
- Relevance: 展示的广告和用户检索词汇是否相关?

作者从上述5个方向入手一共构造了81个新特征，来评价广告质量情况。比如，在Appearance类中，作者构造的特征有“广告标题中包含词汇的数量”、“广告中感叹号或其他标点符号的出现情况”等，在Reputation类中，构造的特征有“展示的域名地址是否以.com结尾？”、“域名的长度”和“域名的级数”，因为类似“books.com”的域名一般来说是好过“books.something.com”的，前者比较简短，词汇简单，这样的域名一般比较贵，能反映出广告主的实力。

此外，作者还统计了训练集中出现最多的10000个关键词，并对每个广告构造了一个10000维的二值向量，其中向量值为1代表相应的关键词出现在这个广告中，如果为0代表广告没有相应的关键词。论文这样设计特征的目的在于希望能够获取一些没有考虑到的信息，发现一些意想不到的重要特征。

如下图所示，在之前构造特征的基础上，加入有关广告质量的所有特征之后，性能提升了23.45%，如果只加入81个特征而不统计关键词是否出现的情况，性能会提升到20.72%，也就是说尽管作者认为设计的81个特征会十分重要，但其实在之前的基础上仅提升了不到1%，而后来设计的二值向量特征（unigrams）却达到了意想不到的效果。

Table 2: *Ad Quality* Results

Features	*MSE (x 1e-3)*	*KL Divrg. (x 1e-2)*	*% Imprv.*
Baseline ($\overline{CTR}$)	4.79	4.03	-
Related term CTRs	4.12	3.24	19.67%
+Ad Quality	4.00	3.09	23.45%
+Ad Quality without unigrams	4.10	3.20	20.72%

所以在这一节中，作者针对广告质量构造了两类特征，一是人工设计的81个特征来衡量广告本身质量，二是统计最频繁的10000个词汇在广告中是否出现。

Measuring order specificity

在考虑了关键词和广告本身质量对CTR的影响之后，作者认为广告主在生成一个广告订单的时候，该订单的指向是否明确也会对CTR产生影响。比如下图所示的广告订单：

```
Title:    Buy shoes now,
Text:     Shop at our discount shoe warehouse!
Url:      shoes.com
Terms:    {buy shoes, shoes, cheap shoes}.
```

当然广告主在输入如上图所示的订单时，作者认为会产生三个广告，因为上文提到过，本文所说的一个广告是一条广告内容、一个广告主和一个关键词的组合，上面的订单输入了三个关键词，因此会产生三个广告。我们看到，广告的Title是“Buy shoes now”明确指向想要买鞋的顾客，并且展示的Url是“shoes.com”，而且三个关键词都是围绕“shoes”的。接着我们来看下面展示的第二个广告订单：

```
Title:    Buy [term] now,
Text:     Shop at our discount warehouse!
Url:      store.com
Terms:    {shoes, TVs, grass, paint}.
```

相比于第一个广告订单来说，这个广告订单的指向性并没有特别明确，因为这个包含了四个不同类别的关键词，来应对用户不同的关键词输入，显然“store.com”是一个综合性的购物网站，并不是第一个订单展示的鞋品专卖店。直觉来说，第二个订单的CTR会低于第一个广告订单，因为其指向性模糊。

为此，作者使用朴素贝叶斯方法对所有关键词进行分类（共74个类别），然后统计广告订单中关键词类别的分布，以计算所得熵的大小来衡量订单的指向是否明确。同时作者还将订单中所出现的单词的数量（去重后）作为一个特征添加到模型中。

Table 3: *Order Specificity* results

Features	*MSE (x 1e-3)*	*KL Divrg. (x 1e-2)*	*% Imprv.*
Baseline ($\overline{CTR}$)	4.79	4.03	-
CTRs & Ad Quality	4.00	3.09	23.45%
+Order Specificity	3.75	2.86	28.97%

如上图所示，在添加了衡量订单指向是否明确的两个特征之后，模型在之前的基础上又提升了5%，相比于Baseline一共提升了28.97%。

External sources of data

除了上述构造的特征之外，作者还添加了一部分外部特征。作者并没有局限于广告本身数据的影响，还考虑到了关键词在网络中的词频，以及关键词在用户检索中的词频。作者添加这样的特征进来，我觉得是考虑到当人们说一个关键词越多，一般来说代表人们越需要相关的商品，进而影响广告的CTR。

对于第一个特征，关键词在网络中出现的频率，作者采用了一种比较简单的方式来统计，即在搜索引擎中检索相关词汇，统计页面数量来代表关键词在网络中的频率。对于第二个特征，作者在一个时长为三个月的搜索引擎用户检索日志上进行处理，获得关键词在用户检索中的词频。

Table 4: *Search Engine Data* results. *AQ* means the *Ad Quality* feature set, and *OB* means the *Order Specificity*.

Features	*MSE (x 1e-3)*	*KL Divrg. (x 1e-2)*	*% Imprv.*
Baseline ($\overline{CTR}$)	4.79	4.03	-
+Search Data	4.68	3.91	3.11%
CTRs & AQ & OS	3.75	2.86	28.97%
+Search Data	3.73	2.84	29.47%

如上图所示，在Baseline基础上，添加本节构造的两个特征（Search Data）之后，性能提升了3.11%，但如果在之前构造的特征基础上（CTRs&AQ&OS）添加，性能仅提升了0.5%，这代表着所构造的特征中会存在一些冗余特征。

总结

作者在设计了一系列特征之后，所构造的模型相比于Baseline而言，性能提升了29.47%。在论文的第10节中，作者依次分析了每类特征对模型提升的贡献程度，并估计了在每个广告前100次被浏览的过程中，本文所提模型相比于Baseline而言的提升程度。在论文的第11节，作者讨论了模型对业务的指导作用，比如根据模型告诉广告主如何改进他们的广告内容，同时还说明了未来的研究方向，比如构建一个能够时时更新的模型。

总体来说，仅从这篇论文来看，使用逻辑回归模型预测CTR的话，需要人为构造很多特征，这就需要了解数据集信息，清楚业务中什么样的特征对预测CTR有帮助。因为逻辑回归模型本身的优化迭代都是固定的，需要的是如何设计好特征。