给定查询和用户信息后预测广告点击率

Task2 描述

搜索广告是近年来互联网的主流营收来源之一。在搜索广告背后，一个关键技术就是点击率预测-----pCTR(predict the click-through rate)，由于搜索广告背后的经济模型（economic model ）需要 pCTR 的值来对广告排名及对点击定价。本次作业提供的训练实例源于腾讯搜索引擎的会话日志(sessions logs), soso.com，要求学员们精准预测测试实例中的广告点击率。

训练数据文件 TRAINING DATA FILE

训练数据文件是一个文本文件，里面的每一行都是一个训练实例（源于搜索会话日志消息）。 为了理解训练数据，下面先来看看搜索会话的描述。搜索会话是用户和搜索引擎间的交互，它由这几部分构成： 用户，用户发起的查询，一些搜索引擎返回并展示给用户的广告，用户点击过的 0 条或多条广告。为了更清楚地理解搜索会话，这里先介绍下术语：在一个会话中展示的广告数量被称为深度（depth），广告在展示列表中的序号称为广告的位置(position)。广告在展示时，会展示为一条短的文本，称之为标题（title），标题后跟着一条略长些的文本和一个 URL，分别叫做描述(description)和展示链接（display URL）。

我们将每个会话划分为多个实例。每个实例描述在一种特定设置

（比如：具有一定深度及位置值）下展示的一条广告。为了减少数据集的大小，我们利用一致的 user id, ad id, query 来整理实例。因此，每个实例至少包含如下信息：

UserID

AdID

Query

Depth

Position

Impression

搜索会话的数量，在搜索会话中广告（AdID）展示给了发起查询(query)的用户（UserID）。

Click

在上述展示中，用户（UserID）点击广告（AdID）的次数。

此外， 训练数据，验证数据及测试数据包含了更多的信息。原因是每条广告及每个用户拥有一些额外的属性。我们将一部分额外的属性包含进了训练实例，验证实例及测试实例中，并将其他属性放到了单独的数据文件中，这些数据文件可以利用实例中的 ids 来编排索引。如果想对这类数据文件了解更多，请参考 **ADDITIONAL DATA FILES** 部分。

最后，在包括了额外特征之后，每个训练实例是一行数据（如下），这行数据中的字段由 TAB 字符分割：

1. Click： 前文已描述。

2. DisplayURL：广告的一个属性。

该 URL 与广告的 title（标题）及 description（描述）一起展示，通常是广告落地页的短链(shortened url)。在数据文件中存放了该 URL 的 hash 值。

3. AdID: 前文已描述。

4. AdvertiserID ： 广告的属性。

一些广告商会持续优化其广告，因此相比其他的广告商，他们的广告标题和描述会更具魅力。

5. Depth：会话的属性，前文已描述。

6. Position： 会话中广告的属性，前文已描述。

7. QueryID： 查询的 id。

该id是从0开始的整数。它是数据文件'queryid_tokensid.txt'的key。

8.KeywordID : 广告的属性。

这是 'purchasedkeyword_tokensid.txt'的 key。

9.TitleID: 广告的属性。

这是 'titleid_tokensid.txt'的 key。

10.DescriptionID：广告的属性。

这是'descriptionid_tokensid.txt'的 key。

11. UserID

这是 'userid_profile.txt'的 key。当我们无法确定一个用户时，

UserID 为 0。

附加的数据文件 ADDITIONAL DATA FILES

这里还有前面提到过的 5 个附加的数据文件：

1. queryid_tokensid.txt

2. purchasedkeywordid_tokensid.txt

3. titleid_tokensid.txt

4. descriptionid_tokensid.txt

5. userid_profile.txt

前 4 个文件每一行将 id 映射为一个记号列表，在 query（查询）, keyword（关键字）, ad title（广告标题）及 ad description（广告描述）中都是如此。 在每一行中，TAB 字符将 id 及其他记号集分隔开。一个记号最基本可以是自然语言中的一个词。为了匿名，每个记号以 hash 后的值来表示。 字段以 ‘|’分割。

‘userid_profile.txt’ 文件的每一行由 UserID, Gender, 和 Age 组成，用 TAB 字符来分隔。注意，并非训练集和测试集中的每个 UserID 都会出现在‘userid_profile.txt’文件中。每个字段描述如下：

1. Gender:

'1' for male(男), '2' for female（女）, and '0' for unknown（未知）.

2. Age:

'1' for (0, 12], '2' for (12, 18], '3' for (18, 24], '4' for (24, 30], '5'

for (30, 40], and '6' for greater than 40（6 代表大于 40）.

TESTING DATASET（测试数据集）

除了广告展示及广告点击的数量不同外，测试数据集与训练数据集的格式一致。 广告展示及广告点击次数用于计算先验的点击率（empirical CTR）。 训练集的子集用于在 leaderboard 上对提交或更新的结果进行排名。测试集用于选举最终冠军。用于生成训练集的日志与之前生成训练集的日志相同。

EVALUATION（评估）

希望参与的团队以文本格式（text format）提交结果，要求结果文件中每行的顺序与下载的文件中行顺序一致，并且每行只有一个字段，即：预测出的点击率值（the predicted CTR）。在结果文件里，与验证数据集中相同的那些行将被用于打分，进而在 leaderboard 上排名（除了最后一天----2012 年 6 月 1 日）。 并且，结果文件中与测试集相同的那些行将被用于在 2012 年 6 月 1 日在 leaderboard 上排名并选取最终获胜者。预测集的表现好坏会通过 AUC 评价指标来打分。