Predict the click-through rate of ads given the query and user information.

TASK 2 DESCRIPTION

Search advertising has been one of the major revenue sources of the Internet industry for years. A key technology behind search advertising is to predict the click-through rate (pCTR) of ads, as the economic model behind search advertising requires pCTR values to rank ads and to price clicks. In this task, given the training instances derived from session logs of the Tencent proprietary search engine, soso.com, participants are expected to accurately predict the pCTR of ads in the testing instances.

TRAINING DATA FILE

The training data file is a text file, where each line is a training instance derived from search session log messages. To understand the training data, let us begin with a description of search sessions.

A search session refers to an interaction between a user and the search engine. It contains the following ingredients: the user, the query issued by the user, some ads returned by the search engine and thus impressed (displayed) to the user, and zero or more ads that were clicked by the user. For clarity, we introduce a terminology here. The number of ads impressed in a session is known as the ’depth’. The order of an ad in the impression list is known as the ‘position’ of that ad. An Ad, when impressed, would be displayed as a short text known as ’title’, followed by a slightly longer text known as the ’description’, and a URL (usually shortened to save screen space) known as ’display URL’.

We divide each session into multiple instances, where each instance describes an impressed ad under a certain setting (i.e., with certain depth and position values). We aggregate instances with the same user id, ad id, query, and setting in order to reduce the dataset size. Therefore, schematically, each instance contains at least the following information:

UserID
AdID
Query
Depth
Position
Impression

the number of search sessions in which the ad (AdID) was impressed by the user (UserID) who issued the query (Query).

Click

the number of times, among the above impressions, the user (UserID) clicked the ad (AdID).

Moreover, the training, validation and testing data contain more information than the above list, because each ad and each user have some additional properties. We include some of these properties into the training, validation and the testing instances, and put other properties in separate data files that can be indexed using ids in the instances. For more information about these data files, please refer to the section ADDITIONAL DATA FILES.

Finally, after including additional features, each training instance is a line consisting of fields delimited by the TAB character:

1. Click: as described in the above list.

2. DisplayURL: a property of the ad.

The URL is shown together with the title and description of an ad. It is usually the shortened landing page URL of the ad, but not always. In the data file, this URL is hashed for anonymity.

3. AdID: as described in the above list.

4. AdvertiserID: a property of the ad.

Some advertisers consistently optimize their ads, so the title and description of their ads are more attractive than those of others’ ads.

5. Depth: a property of the session, as described above.

6. Position: a property of an ad in a session, as described above.

7. QueryID: id of the query.

This id is a zero-based integer value. It is the key of the data file 'queryid_tokensid.txt'.

8. KeywordID: a property of ads.

This is the key of 'purchasedkeyword_tokensid.txt'.

9. TitleID: a property of ads.

This is the key of 'titleid_tokensid.txt'.

10. DescriptionID: a property of ads.

This is the key of 'descriptionid_tokensid.txt'.

11. UserID

This is the key of 'userid_profile.txt'. When we cannot identify the user, this field has a special value of 0.

ADDITIONAL DATA FILES

There are five additional data files, as mentioned in the above section:

1. queryid_tokensid.txt

2. purchasedkeywordid_tokensid.txt

3. titleid_tokensid.txt

4. descriptionid_tokensid.txt

5. userid_profile.txt

Each line of the first four files maps an id to a list of tokens, corresponding to the query, keyword, ad title, and ad description, respectively. In each line, a TAB character separates the id and the token set. A token can basically be a word in a natural language. For anonymity, each token is represented by its hash value. Tokens are delimited by the character ‘|’.

Each line of ‘userid_profile.txt’ is composed of UserID, Gender, and Age, delimited by the TAB character. Note that not every UserID in the training and the testing set will be present in ‘userid_profile.txt’. Each field is described below:

1. Gender:

'1' for male, '2' for female, and '0' for unknown.

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.

TESTING DATASET

The testing dataset shares the same format as the training dataset, except for the counts of ad impressions and ad clicks that are needed for computing the empirical CTR. A subset of the testing dataset is used to consistently rank submitted/updated results on the leaderboard. The testing dataset is used for picking the final winners.

The log for forming the training dataset corresponds to earlier time than that of the testing dataset.

EVALUATION

Teams are expected to submit their result file in text format, in which each line corresponds to a line in the downloaded file with the same order, and there is only one field in each line: the predicted CTR. In the result file, the lines corresponding to the lines from validation dataset will be used to score for the ranking on the leaderboard during the competition except the last day (June 1, 2012), and the lines corresponding to the lines from testing dataset will be used for the ranking on the leaderboard on the day of June 1, 2012, and for picking the final winners.

The performance of the prediction will be scored in terms of the AUC (for more details about AUC, please see ‘ROC graphs: Notes and practical considerations for researchers‘ by Tom Fawcett). For a detailed definition of the metric, please refer to the tab ‘Evalaution’.