COMPM041 Web Economics

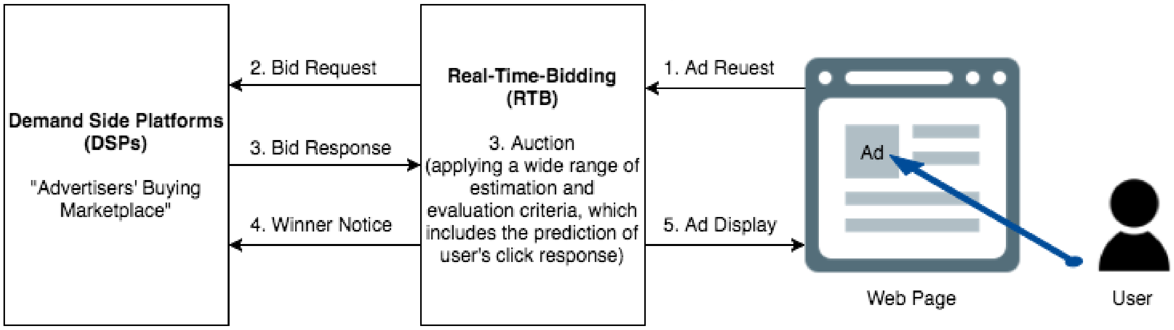
Zunran Guo

April 2016

1 Introduction

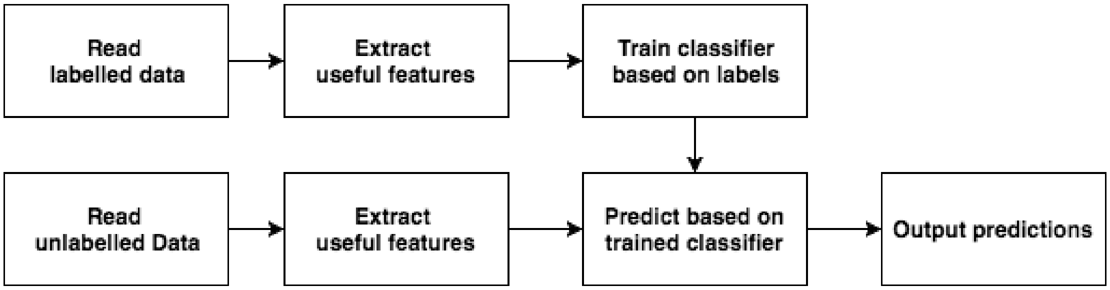
Recently, the computational advertising area has become one of the dominant sections in the online marketing space since the emergence of Real-Time Bidding (RTB) in 2009 [1]. Its mechanism is briefly illustrated in the figure below.

Figure 1: Real-Time-Bidding (RTB) Mechanism


Demand Side Platforms (DSPs)
"Advertisers' Buying Marketplace"
2. Bid Request
3. Bid Response
4. Winner Notice
Real-Time-Bidding (RTB)
3. Auction (applying a wide range of estimation and evaluation criteria, which includes the prediction of user's click response)
1. Ad Reuest
5. Ad Display
Ad
Web Page
User

In this project, the goal is to predict the probability of clicking the advertisement (ad) in RTB. In other words, given the information of the incoming bid request, the bid agent should estimate the probability of the user's click on the displayed ad. A supervised learning model is therefore constructed to predict the user's click response based on the training data. Figure 2 below states the processes of the project. Firstly, the features of the data are extracted to convert the data information as a readable vector. After that the labeled data as the training dataset is read to train the initial model with the various supervised learning methods and seek the best parameters. This classifier model is used to predict the class of the unlabeled data. Finally, in this project, the probability of clicking the advertisement will be estimated.

Figure 2: Training and Predicting Mechanism



Python is selected as the language in this project mainly for its quick prototyping and development. In addition, for data processing and classifier training, external libraries like scikit-learn, SciPy, and pandas provide a wide set of flexible machine learning and fast data frame processing tools, along with clean friendly parameter interface. Besides Python's compatibility with other develop environment, the data type in this project, pandas and NumPy (included in SciPy), could also be quickly injected into other popular libraries like Matplotlib.

2 Data Description

The data used in this project is sampled and reformulated from the iPinYou RTB dataset, which can be retrieved from UCL Web Economics Algorithm Challenge 2016, along with its data description [2]. As for the training data, each line contains the classification label, (1 for click, 0 for no click) followed by other ad features, delimited by tab, "\t". In the labeled training data set specifically, 2076 data points are labeled for click response, and 2845726 data points labeled for no-click response. Thus, given the sparsity of user's click response, the ratio of the number of click response to the number of the no-click response is really small, only about 0.0007.

Table 1: Data Description

Column	**Feature**	**Example**
0	Click	0 (only in data_train.txt)
1	Weekday	1
2	Hour	13
3	Timestamp	20130218134701883
4	Log Type	1
5	User ID	dF_5qwD1UDI
6	User Agent	windows_chrome
7	IP	119.163.222.*
8	Region	146
9	City	147
10	Ad Exchange	2
11	Domain	e80f4ec7f5bfbc9ca416a8c01cd1a049
12	URL	hz55b000008e5a94ac18823d6f275121
13	Anonymous URL ID	null
14	Ad slot ID	973726_9023493
15	Ad slot width	300
16	Ad slot height	250
17	Ad slot visibility	2
18	Ad slot format	1
19	Ad slot floor price	22
20	Creative ID	f80f4ec7f5bfbc9ca416a8c01cd1a049
21	Key Page URL	361e128affece850342293213691a043
22	Advertiser ID	3386
23	User Tags	10057,10063,10024,13800,13866,10110

Note that columns with * are hashed or modified before releasing.

2.1 Feature Extraction

There are 23 original features in both training and testing data. None of these features are used directly in training without preprocessing, like binarization and normalization.

Part of the features are dropped based on the assumption that they do not have meaningful (valuable) information that may help predict a user's click response. For example, a unique *User ID* could be generated randomly and assigned to each data point, while such unique ID does not have any valuable information that may help predict a user's click response. Even some of the features are not entirely unique, given its excessively large dimension, we also dropped those feature based on similar reasoning. These dropped features are *Timestamp*, *Log Type*, *User ID*, *IP*, *Domain* (15146 dimensions), *Anonymous URL ID* (715316 dimensions), and *Ad slot ID*.

As for the binarized features, such as *Weekday*, *Hour*, *User-Agent*, *Region*, *Ad Exchange*, *Ad slot width*, et al., their dimensions ranges widely from as small as 3 (*Ad Exchange*) as many as 370 (*City*). While the expanded dimensions of the data may increase the computation significantly, the result shows that it is worth to do as the AUC score is enhanced. Further details will be discussed in the later sections. Table 2: Feature Process and Extraction below shows each feature's name, the dimension of each feature, and the corresponding processing method, which will be discussed in details in the following.

Table 2: Feature Process and Extraction

Feature	Dimension	Method	Feature	Dimension	Method
Weekday	7	binarize	**Anonymous URL ID**	1	drop
Hour	24	binarize	**Ad slot ID**	N/A	drop
Timestamp	N/A	drop	**Ad slot width**	11	binarize
Log Type	1	drop	**Ad slot height**	6	binarize
User ID	N/A	drop	**Ad slot visibility**	4	binarize
User-Agent	36	binarize	**Ad slot format**	3	binarize
IP	N/A	drop	**Ad slot floor price**	1	binarize/normalize
Region	35	binarize	**Creative ID**	11	binarize
City	370	binarize	**Key Page URL**	2	binarize
Ad Exchange	3	binarize	**Advertiser ID**	1	drop
Domain	N/A	drop	**User Tags**	68	binarize
URL	N/A	drop	**Ad slot area***	11	generate; binarize

Note: *Ad slot area* is generated from the product of *Ad slot width* and *Ad slot height* and followed by binarization.

The most common processing method used in this data is binarization, which maps a vector with n distinct categories to n dimensions. For example, the attribute *Weekday*, which has 7 categories (7 days in 1 week), will be mapped to 7 dimensions, i.e. 7 attributes, each attribute representing one unique weekday. So if a specific weekday is Tuesday, of the 7 newly generated attributes, the second attribute becomes 1 and all the other 6 attributes become zeros. In addition, as mentioned, there are some features that are simply dropped. *Advertiser ID* for example, like *User ID* discussed above with its dimension being one, does not contain valuable information to predict user's click response. At the same time, since the features like timestamp and IP have duplicated information that can be extracted relatively easily from other features like *Weeky*, *Hour*, *Region*, and *City*, they are dropped in data processing as well. All the features are extracted as candidates before training, but when the model is trained, there are different combinations of the features to train the model to find the optimal set finally.

References

[1] W. Zhang, S. Yuan, J. Wang, X. Shen, *Real-Time Bidding Benchmarking with iPinYou Dataset*, July 25, 2014, http://arxiv.org/abs/1407.7073

[2] iPinYou, *iPinYou Global Bidding Algorithm Competition Data Description*, 2013, http://contest.ipinyou.com/data.shtml