COMPGW02/M041: the Web Economics Project

Jun Wang, Emine Yilmaz
{j.wang,e.yilmaz}@cs.ucl.ac.uk
TA: Bin Zou, Sam Windels, and Ying Wen
b.zou@cs.ucl.ac.uk, sam.windels.16@ucl.ac.uk,y.wen@cs.ucl.ac.uk

1. SUMMARY

In this assignment, you are required to work on an online advertising problem. You will help advertisers to form a bidding strategy in order to place their ads online in a real-time bidding system. You are required to train a bidding strategy based on an impression training set. The aim of this project is to help you understand some basic concepts and write a computer program in real-time bidding based display advertising. As you will be evaluated both as a *group* as well as *individually*, part of the assignment is to train a model of your choice independently. The performance of the model trained by your team, which is either a combination of the individually developed models or the best performing individually developed model, will be (mainly) evaluated on the Click-through Rate achieved on a provided test set. In order for you to properly evaluate the performance of each of your models before that, a benchmark click-through rate on the validation set will be provided. Before the final submission, you are also given the opportunity to hand in the preliminary result of your team's model on the test set, which allows you to compare the performance to that of your peers. Should you have any question about the assignment, feel free to contact the TAs: Bin Zou (email), Sam Windels (email), and Ying Wen (email).

2. RULES

You are expected to work in a team of 3 people and you can sign in the team through this google sheet, by the end of this week (the 24th Feb). You will write a joint group report, and an individual report for each team member.

2.1 Submissions

- Preliminary submission: 23:55 17/03/2017. Please upload your submission through Moodle as a zip file, and rename the file as that: Group_xx.zip (e.g: Group_10.zip).

  For the preliminary submission, only include a bid price generated using your team's model applied on the testing set in the zip file.

- Final submission: 23:55 14/04/2017. Both the *Group Report* and *Individual Report* will be submitted together. Please upload your submission through Moodle as a zip file, and rename the file as that: Group_xx.zip (e.g: Group_10.zip). For the final submission, the file should include the reports (both the group and individuals) in PDF format (including the names of all members of your team, the reports should be named as group_xx_report.pdf and your_name_report.pdf), and your team's predictions for the provided test set. Please upload your code to GitHub and provide a link to it in the reports (again for both reports).

2.2 The Group and Individual Reports

Individually, you are required to submit an *Individual Report* (one per person). At the same time, as a group, you are required to submit one *Group Report* (one per group). Each of them will have 15 marks, and in total there are 30 marks. The rest 70 marks will come from the written exam in May.

The two reports will have different set of questions, respectively, as given in Section 3.2. The reports shall summarise what you (or your group) did and how you addressed the posed questions as described in the problem description.

You will use the ACM Proceedings style LATEX files for writing your reports. You can find out more about these styles as well as LATEX templates on the ACM website. Note that to make your report looks better, you could simply add the following lines before `\begin{document}` to remove the copyright notice:

```
\makeatletter
\def\@copyrightspace{\relax}
\makeatother
```

For the *Individual Report*, your submission should be 3-5 pages, with maximum 1 additional page for references. For the *Group Report*, your submission should be 5-7 pages, with maximum 1 additional page for references. Images and tables also count against the maximum. Rules written here trump all the non-formatting instructions in the ACM style files. Your report should address all the problems in Section 3.2.

2.3 Q&A and Office Hours of TAs

The office hours below are for this project only. For the office hours of the module itself, please refer to the moodle page.

- Time: 12:30-13:30 28/02/2017 and 14/03/2017.
- Location: Hub Level 1, 66-72 Gower Street.

3. THE PROJECT

In display and mobile advertising, the most significant evolution in recent years is the employment of the so-called Real-Time Bidding (RTB) mechanism to buy and sell ads.

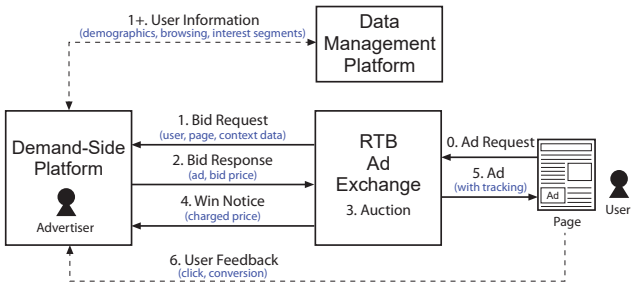

Figure 1: A brief illustration of the interactions between user, ad exchange and the advertiser's DSP bidding agent.

RTB essentially facilitates buying an individual ad impression in real time, automatically triggered by a user's visit. Although other types of auctions, such as the first price auction, are also popular, RTB exchanges typically employ the second price auction model. In this type of auction, the bidder winning the auction pays the price of the second highest offer. A summary of the process is provided in Figure 1. In this assignment you will be training your own bidding strategy models on a given impression data set, as described in detail next.

3.1 Data Description

The dataset includes a training set, a validation set, and a testing set. The dataset for the task can be downloaded from https://goo.gl/yIBqKx.

This data comes in CSV format, the first line in the file containing the header formatted as described in Table 1. As the testing set is used for final evaluation purposes, it does not contain the three fields: 'bidprice', 'payprice' and 'click'. Note that all numbers related to money (e.g., bid price, paying price and floor price) use the currency of RMB and the unit of Chinese fen × 1000, corresponding to the commonly adopted cost-per-mille (CPM) pricing model.

You can develop your model on the training data and use the validation set to compare the different models and correct for overfitting.

3.2 Problem Description

In this assignment, you are required to train a bidding strategy based on the impressions in the training set. This objective has been divided into the following five sub-problems, and you are required to solve and report these problems in either the group or individual report, as indicated below. The concept of a bidding strategy has been visualised in Figure 2.

3.2.1 *Problem 1: Data Exploration*

- Individual report, 5/30 marks

You should do some exploratory data analysis and show your understanding of this dataset. You can start from basic statistical information, such as num Imps, num Clicks, Cost, CTR, avg CPM, eCPC. You can also include further analysis on user feedback, bidding etc. A good example can be found in the paper [1].

3.2.2 *Problem 2: Basic Bidding Strategies*

Table 1: Fields in dataset

Field	Example	Supplement
click	1	1 if clicked, 0 if not.
weekday	1	
hour	12	
bidid	fdfe...b8b21	
logtype	1	
userid	u_Vh1OPkFv3q5CFdR	
useragent	windows_ie	
IP	180.107.112.*	
region	80	
city	85	
adexchange	2	
domain	trq...Mi	
url	d48a...4efeb	
urlid	as3d...34frg	
slotid	2147...813	
slotwidth	300	
slotheight	250	
slotvisibility	SecondView	
slotformat	Fixed	
slotprice	0	
creative	hd2v...jhs72	
bidprice	399	
payprice	322	Paid price after win the bidding.
keypage	sasd...47hsd	
advertiser	2345	
usertag	123,5678,3456	Contains multi-values, ',' as segmentation.

- Group report, 5/30 marks

Evaluate the following strategies on the validation set and discuss your results. Use the performance metrics as defined in Section 3.3.

- Constant bidding (Const). Bid a constant value for all the bid requests. The parameter is the specific constant bid price. You coud find a way to obtain an optimal constant value from the training set.
- Random bidding (Rand). Randomly choose a bid value in a given range. The parameter is the upper bound of the random bidding range.

3.2.3 *Problem 3: Linear Bidding Strategy*

- Group report, 5/40 marks

Apply CTR estimation to create a linear bidding strategy. The bid value is linearly proportional to the pCTR. The formula can be generally written as $bid = base\_bid \times pCTR/avgCTR$, where the tuning parameter $base\_bid$ is the bid price for the average CTR cases. Optimise the $base\_bid$ and the CTR estimation and evaluate the performance of your model using the metrics as defined in Section 3.3. Discuss your results.

3.2.4 *Problem 4: Your Best Bidding Strategy*

- Individual report, 10/40 marks

As the linear model may not be the best model, you are asked to experiment further in order to find a more optimal

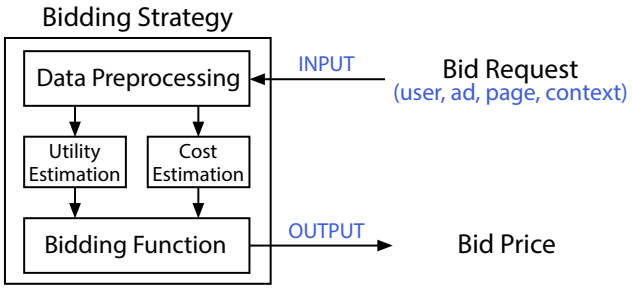

Figure 2: A bidding strategy can be abstracted as a function mapping from the given bid request (in a high dimensional feature space) to a bid price (a non-negative real or integer number).

model. The model you developed can be tested over the validation set. and you should explain your approach and discuss your results in the individual report. This does include, but is not limited to, reporting the performance metrics as defined in Section 3.3 of your solution and the parameter turning on the validation set.

Note that your bidding strategy described here will be part of the combined model (either pick up the best or "cleverly" combined) for the group, see problem 5 below, which will be tested over the test dataset. Therefore, it is advised, although not mandatory, that you try different models than your other team members in this section.

Some directions: A non-linear bidding strategy (e.g. ORTB) [2], or using other CTR estimators.

3.2.5 *Problem 5: The Combined One and Submission*

- Group report, 5/40 marks

Compare the individual bidding strategies developed in problem 4 and design a combined strategy. Your report should include em the discussion of the way you combine them and the performance of your combined model on the validation set.

Next, pick up the best strategy, this can be one of the individual models or the combined models and make your predictions on the given test set. (There are also the predictions that you are allowed to hand in preliminary.)

3.3 Evaluation

Alongside the final report, you are required to submit the bid prices for the testing set, generated using your group's best performing model. Performance will be evaluated based on the following metrics, within a limited budget of 25000 CNY fen.

- Click-Through Rate (this is the main metric you should focus on)
- Conversions (Number of Clicked Bids)
- CVR(Conversion Rates)
- Spend(Total Money Paid)
- Average CPM(Cost Per Mille)
- Average CPC(Cost Per Click)

The bidding prices on the testing set (not the training or validation set!), should be saved as a CSV file named 'testing_bidding_price.csv' with following format:

```
bidid,bidprice
xxxxxxxxxx,13.0
yyyyyyyyyy,23.0
....
```

During the evaluation, we (our script) will go through the file and only consider the bids until the accumulated cost is higher than or equal the allowed budget 25000 CNY fen.

3.4 Format of your reports

Your reports (both the individual report and the group report) should consist of the following sections: 1) the introduction section briefly explains the problem context and what you intend to do and summarise your results; 2) the related work section briefly summarise the related work of the specific problems you have found from the research literature. You could use https://scholar.google.co.uk/ to obtain the related research papers; 3) the approach and result section gives your answers to each of the specific questions and provide the evaluation results and your discussions. 4) the conclusion section concludes your report and point out the potential direction to improve your report (e.g., if you have time, you will do...). For the individual report, in the conclusion section, create a paragraph to explain your role in the group project and also comment on other members' contribution in the group. We will gather the information to understand how the collaboration has been done in the group.

4. REFERENCES

[1] Weinan Zhang, Shuai Yuan, Jun Wang, and Xuehua Shen. Real-time bidding benchmarking with ipinyou dataset. *arXiv preprint arXiv:1407.7073*, 2014.

[2] Weinan Zhang, Shuai Yuan, and Jun Wang. Optimal real-time bidding for display advertising. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1077–1086. ACM, 2014.