Offline Evaluation of Response Prediction in Online Advertising Auctions

Olivier Chapelle
Criteo
Palo Alto, CA
o.chapelle@criteo.com

ABSTRACT

Click-through rates and conversion rates are two core machine learning problems in online advertising. The evaluation of such systems is often based on traditional supervised learning metrics that ignore how the predictions are used. These predictions are in fact part of bidding systems in online advertising auctions. We present here an empirical evaluation of a metric that is specifically tailored for auctions in online advertising and show that it correlates better than standard metrics with A/B test results.

Categories and Subject Descriptors

H.3.4 [**Information Storage And Retrieval**]: Systems and Software—*Performance evaluation*

Keywords

Online advertising; auction; response prediction; metrics

1. INTRODUCTION

Online advertising is a major business for Internet companies, reaching $43 billion revenue in 2013. Its main forms are paid (or sponsored) search – where text ads appear alongside search results of a web search engine – or display advertising where advertisers pay publishers for placing graphical ads on their web pages. Second price auctions are now commonly used as a mechanism for selling advertisements opportunities on web pages [3].

Several payment options are offered to the advertisers in an online advertising market. If the goal of an advertising campaign is getting their message to the target audience (for instance in brand awareness campaigns) then paying per impression (CPM) with targeting constraints is normally the appropriate choice for the advertiser. However, many advertisers would prefer not to pay for an ad impression unless that impression leads the user to the advertiser's website. Performance dependent payment models, such as cost-per-click (CPC) and cost-per-conversion (CPA), were introduced to address this concern. In order to determine the winner of the auction, these bids need to be converted to an *expected* price per impression (eCPM).

The focus of this paper is evaluating the performance of a *bidder*, defined as an agent that takes the CPC or CPA that the advertiser is willing to pay and submits a CPM bid for the impression. The bidder may either be external to the marketplace, for instance a demand-side platform (DSP) bidding on a real-time bidding (RTB) exchange [9], or internal as in paid search.

The eCPM of a CPC or CPA bid will depend on the probability that the impression will lead to a click or a conversion event. Estimating these probabilities accurately is thus critical, both for paid search [7] and display advertising [1]. Offline evaluation of these probabilities is usually done as with any supervised learning problem, using standard metrics such as log likelihood or squared error. But these metrics are oblivious to how the probabilities will affect the bidder performance. We propose here an empirical study of a metric recently proposed in [5] that is specifically designed for evaluating a bidder.

The idea is to take into account the value of the highest competing bid on the exchange in the metric: if that bid is much lower (resp. much higher) than the eCPM, then a misprediction has hardly any influence because the bidder would have won (resp. lost) the auction anyway. The metric aims at estimating the bidder profit under a distribution of the highest competing bid. The reason for having such a distribution instead of using the observed bid value (i.e the cost of the impression) is to increase the robustness by considering that this bid could have been different.

We will first introduce in Section 2 the setting in which the bidder is to be evaluated and then discuss the *expected utility*, a metric designed for evaluating a bidder. An experimental evaluation of that metric, including correlation with A/B test results will be presented in Section 3, before the concluding remarks of Section 4.

2. BIDDER EVALUATION

2.1 Setting

The setting is as follows. A bidder competes for an impression opportunity and needs to submit a CPM bid for that impression in a second-price auction. It values a certain action (such as a click or a conversion) with a value v. The bidder predicts that if the ad is displayed, the probability of that action occurring ($a = 1$) is p. The value of an impression is thus vp and since the second-price auction is

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742566.

incentive compatible, the bidder bids $b = vp$ for the impression.[1] Whether the bidder wins the auction depends on the highest competing bid c. When $b > c$, the bidder wins and pays c; if the impression is followed by the action of interest, the bidder receives a reward of v.

The payoff for the auction can thus be written as:

$$\begin{cases} av - c & \text{if } pv > c; \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

2.2 Baseline metrics

We are given a set of historical events (x_i, a_i, v_i, c_i), where x_i are some context features and where lost bids have been filtered out. In other words, if $\bar{p}$ is the production function used to collect the data, then $\forall i,\ \bar{p}(x_i)v_i > c_i$. This last inequality would not hold if *exploration* data were included, i.e. impressions for which the bid was higher than the eCPM. We suppose in the rest of this paper that no such data was collected. A new prediction function p can be evaluated using standard metrics [8] such as,

Log likelihood

$$\sum_i a_i \log(p(x_i)) + (1 - a_i) \log(1 - p(x_i))$$

Squared error

$$\sum_i (a_i - p(x_i))^2$$

These metrics focus on the quality of the predictions but ignore the bidding system in which they are used. Another possibility is to replay the logs and to estimate the profit (1) that the bidder would generate with the new prediction function:

$$\sum_i (a_i v_i - c_i) \mathbb{I}(p(x_i)v_i > c_i),$$

where $\mathbb{I}$ is the indicator function. We call this metric the *utility*. The issue with that metric is that without any exploration data, overpredictions are not penalized. In particular, if $p \geq \bar{p}$, then $p(x_i)v_i \geq \bar{p}(x_i)v_i > c_i$ and the utility is the same for p and $\bar{p}$:

$$\sum_i (a_i v_i - c_i) \mathbb{I}(\bar{p}(x_i)v_i > c_i) = \sum_i a_i v_i - c_i = \sum_i (a_i v_i - c_i) \mathbb{I}(p(x_i)v_i > c_i). \quad (2)$$

2.3 Expected utility

Using exploration data with the utility would penalize overpredictions and fix the issue describe above. But collecting such data is expensive and might not be feasible. Another way to penalize overpredictions is to pretend that the highest competing bid may have been different. More precisely, given an observed second price c, let $\Pr(\tilde{c} \mid c)$ be the distribution capturing which other values the highest competing bid may have taken. The *expected utility* (EU) under this distribution is:

$$\text{EU} = \int_0^{pv} (av - \tilde{c}) \Pr(\tilde{c} \mid c) d\tilde{c} \quad (3)$$

We empirically study in this paper the suitability of (3) for the purpose of evaluating a bidder. In a closely related work, [5] analyzes the same expected utility as a *loss function* to train machine learning models. The idea of introducing a corrupting distribution to make a statistic or a loss function more reliable is also known in the machine learning literature as *noise injection* [2, 6].

Two limit distributions for $P(\tilde{c} \mid c)$ are noteworthy:

Dirac at the observed price c: this original utility (1).

Uniform : it is easy to show (see [5, Theorem 2]) that the expected utility is, up to an irrelevant additive term and multiplicative factor, equal to the squared error, weighted by the value: $v^2(a - p)^2$.

One of the compelling feature of the expected utility is thus that it includes two baseline metrics (utility and weighted squared error) as special cases. And it also well-calibrated in the sense that its expected value is maximized under perfect predictions, $p_i = \mathbb{E}(a_i \mid x_i)$ [5, Theorem 4]. This last property is a direct consequence of the truthfulness of a second-price auction.

2.4 Competing bid distribution

Given an observed second price c, which distribution $\Pr(\tilde{c} \mid c)$ should we choose to compute the expected utility (3)? A natural choice would be a log-normal distribution centered at c. Log-normal distributions have indeed been found to fit quite well observed winning bids in Yahoo's RightMedia exchange [4]. We have in fact experimented with this distribution and obtained satisfactory results. But this distribution does not include the special case of the uniform distribution discussed above: the log-normal does not converge to a uniform distribution as its variance goes to infinity.

The Gamma distribution on the other hand contains the two baselines as special cases. A Gamma distribution with $\alpha = \beta c + 1$ and free parameter β has the following density function:

$$\Pr(\tilde{c} \mid c, \beta) \propto \tilde{c}^{\beta c} \exp(-\beta \tilde{c}).$$

As $\beta \to 0$, $\Pr(\tilde{c} \mid c, \beta)$ convergences toward the uniform distribution. And as $\beta \to \infty$, the variance of the distribution, α/β^2, goes to 0, while the mean, α/β, converges toward c. In other words, it convergences towards a Dirac distribution at c.

With this choice of the highest competing bid distribution, the expected utility becomes:

$$av\gamma(\beta c + 1, \beta pv) - \frac{\beta c + 1}{\beta} \gamma(\beta c + 2, \beta pv),$$

where γ is the *incomplete gamma function*. We experiment in the next section with this metric for different values of β.

3. EMPIRICAL EVALUATION

For this work we collected traffic logs of Criteo, a global leader in performance display advertising, specialized in retargeting. Criteo acts as an intermediary between publishers and advertisers by paying publishers on a CPM basis and

[1] The value may be non-zero when the bidder loses the auction: the user may convert without seeing the ad; there might also be further impression opportunities for him. In these cases the bid should be lowered, but we ignore these possibilities in this paper.

gets paid by advertisers whenever there is a click on an ad (CPC) or when there is a conversion following that click (CPA).

We compare the performance of a CPC bidder offline and online (i.e. through an A/B test). A CPC bidder takes as input a bid request, that includes various features used to estimate the CTR of the display opportunity, as well as the value of a click. The bidder predicts the CTR with a prediction function and it multiplies this prediction by the value of the click to form the bid. The metrics (offline and online) were computed on a 4 days period during which the A/B test was live. The offline metric is the expected utility (3) computed on all the logs with various values of β, while the online metric is the profit on its own population, in other words the utility, $\sum_i a_i v_i - c_i$.

Instead of computing a correlation across different pairs of predictions functions, we fixed a pair of functions and computed a correlation across *publisher networks*. A good correlation means that if we estimate offline a prediction function to be better than the other one on a given network, then the corresponding bidder would indeed generate more profit on that network. A publisher network corresponds roughly to an RTB platform. Small networks were grouped together in order to ensure that each of them had at least 30M impressions. We ended up with 25 networks.

The difference in online metrics is much noisier than the difference in offline metrics; this is because the former is an unpaired difference while the latter is a paired one. In order to reduce the noise, we computed the correlations by taking into account the confidence intervals on the online results. This was done by randomly sampling the online profits according to their confidence intervals, computing the correlation with the offline metrics, and averaging the correlations over 100 trials. The experimental protocol is summarized in algorithm 1. Two correlation coefficients were computed, Pearson r (linear correlation) and Kendall τ (ranking correlation).

Algorithm 1 Computation of the correlation coefficients

Require: Two prediction models, an A/B test
1: Compute offline metrics for both models and each publisher network during the A/B test period.
2: Take the difference and normalize by the the total number of displays for that network.
3: **for** i=1 to 100 **do**
4: Resample the logs
5: Compute the bidder profit on each population.
6: Take the profit difference and normalize.
7: Compute the correlation between these online and offline metrics.
8: **end for**
9: Report the average correlation.

The correlation coefficients as a function of β are plotted in figure 1. The best correlation is achieved for $\beta = 10$. Note that this value corresponds to a rather large variance: for a cost[2] $c = 10^{-3}$ the variance is about 0.1.

The correlations for different offline metrics are reported in table 1. The weighted MSE is in fact a negative MSE in such a way that a positive difference means an improvement

[2]a cost for thousand impressions (CPM) is indeed of the order of $1

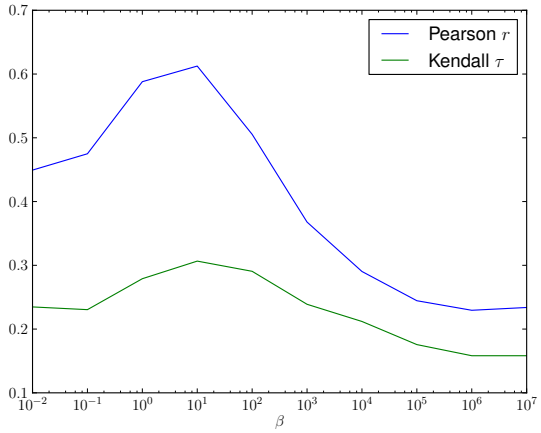

Figure 1: Correlation between A/B test results and the expected utility as a function of β. As $\beta \to 0$, the metric is equivalent to a weighted MSE, while as $\beta \to +\infty$, it is equivalent to the utility.

	Pearson r	Kendall τ
MSE	0.283 ±0.04	0.153 ±0.03
Weighted MSE	0.441 ±0.06	0.234 ±0.03
Utility	0.243 ±0.02	0.158 ±0.02
EU (β=10)	**0.608** ±0.03	**0.311** ±0.03

Table 1: Correlations coefficients and their standard deviation – as computed by algorithm 1 – between A/B test results and offline metrics.

and that the correlation coefficients are expected to be positive. As mentioned before, the columns weighted MSE and utility correspond to $\beta = 0$ and $\beta = \infty$ respectively. The correlation of the regular MSE is worse than the weighted MSE, presumably because the regular MSE does not take into account the values associated with each action: it is more costly to make a prediction error while bidding on an item with a large value than on one with a small value.

Finally a scatter plot of the offline / online differences is shown in figure 2. Even though the correlation between expected utility and online results is not that large (probably because of noise in the online measurements), it is still much better than the correlation with the weighted MSE.

4. CONCLUSION

For the purpose of evaluating the accuracy of a prediction model, the MSE seems well suited, but does not take into account how the predictions are used by the bidder. The regular utility on the other hand tries to estimate the bidder profit but may fail to correctly penalize mispredictions. The expected utility, a metric recently proposed in [5], captures best of both words. This paper demonstrated empirically that indeed the expected utility achieves better correlations with online results.

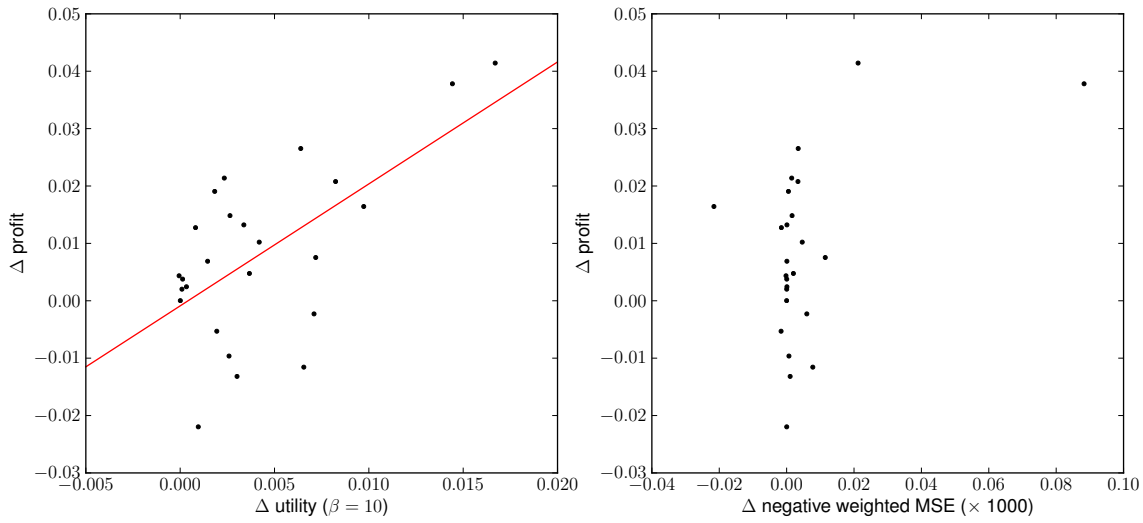

Figure 2: Scatter plots of offline vs online results, each point representing a publisher network. Left: expected utility, right: weighted MSE

5. REFERENCES

[1] O. Chapelle, E. Manavoglu, and R. Rosales. Simple and scalable response prediction for display advertising. *ACM Transactions on Intelligent Systems and Technology*, 5(4), 2014.

[2] O. Chapelle, J. Weston, L. Bottou, and V. Vapnik. Vicinal risk minimization. *Advances in neural information processing systems*, pages 416–422, 2001.

[3] B. Edelman, M. Ostrovsky, and M. Schwarz. Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *American Economic Review*, 97(1):242–259, 2007.

[4] A. Ghosh, P. McAfee, K. Papineni, and S. Vassilvitskii. Bidding for representative allocations for display advertising. *CoRR*, abs/0910.0880, 2009.

[5] P. Hummel and P. McAfee. Loss functions for predicted click through rates in auctions for online advertising. Unpublished, 2013.

[6] L. Maaten, M. Chen, S. Tyree, and K. Q. Weinberger. Learning with marginalized corrupted features. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 410–418, 2013.

[7] H. B. McMahan, G. Holt, D. Sculley, M. Young, D. Ebner, J. Grady, L. Nie, T. Phillips, E. Davydov, D. Golovin, S. Chikkerur, D. Liu, M. Wattenberg, A. M. Hrafnkelsson, T. Boulos, and J. Kubica. Ad click prediction: a view from the trenches. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2013.

[8] J. Yi, Y. Chen, J. Li, S. Sett, and T. W. Yan. Predictive model performance: Offline and online evaluations. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1294–1302. ACM, 2013.

[9] S. Yuan, J. Wang, and X. Zhao. Real-time bidding for online advertising: Measurement and analysis. In *Proceedings of the Seventh International Workshop on Data Mining for Online Advertising*, ADKDD '13, 2013.