Click-Through Rate Prediction via Logistic Regression and Features Combination

Sen Lin

1 Introduction

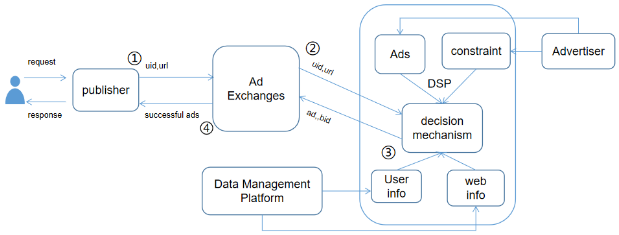
The rapid development of the Internet has fundamentally changed the Internet Adertising mode:the traditional form of adertising in which fixed images or texts are embedded in web pages is gradually transformed into a dynamic targeting mechanism based on web content and user characteristics [9]. Computational advertising is a sub-discipline that arises in this demand environment. It is an advertising mechanism based on the calculation of the best mathching advertisements for a given user and webcontent [1].

The Internet computation advertising industry chain evolved from the first three basic roles - Advertiser, Publisher, User, to later advertisers, advertising media, advertising exchange platforms, demand side platforms, data Management platform and users. Among them, advertisers hope to improve the user's ability to purchase goods or website registration by placing the most appropriate advertisements on the right users, so as to obtain the best publicity effect; users want to see useful advertising information instead of harassing information;Publishers can get the corresponding page profit by posting ads.Therefore, in the process of three-way interest interaction, the Click Through Rate (CTR) is an important core and balance point for the three parties to obtain benefits, and is an important link for accurately advertising, maximizing the interests of advertisers and users [5]. However, sparse, massive data makes accurate ad click-through rate prediction a very challenging part of the job.

CTR forecast is through advertising sites, users, and ads. The feature information of the party predicts the probability of occurrence of a click event for a specific advertisement by a specific user in a specific webpage environment. This paper proposes a logistic regression model based on multidimensional feature combination vector to predict the click rate of advertisements, model the click event of the advertisement, represent the feature vector parameters at different levels of hierarchy, and combine the feature vector set into the logistic regression calculation. In the model, it is thus possible to estimate the CTR more accurately with respect to the calculation of the one-dimensional feature vector. Finally, the feasibility of the algorithm and the improvement of the validity of the model are proved by experiments.

1.1 Advertising Mechanism

For advertisers, calculating CTR is a professional High, complex and difficult to implement algorithmic logic. Therefore, wide The advertisers urgently need the platform from the Demand-side Platforms(DSP) technical solutions [6]. DSP also provides advertising display services for multiple advertisers. It can be purchased directly from the Ad Server or from the real-time advertising exchange platform (AdExchange). If purchased directly from the ad server, the advertiser only needs to negotiate with the advertising media in advance to determine the cost of the ad display. In the real-time advertising communication platform, the DSP must obtain bidding opportunities through bidding, and the highest bidder will win the auction. When a user visits a particular web page, the various advertising campaigns of the advertisers should also be ready at the DSP. In such a specific environment, the core work that DSP has to accomplish is to combine the user, webpage, advertisement, promotional activities and other information to determine the best advertisement in the webpage display (the user is most likely to accept, most likely to click or even buy. Or a registered ad), and can give a reasonable bid price to get the opportunity to show the ad. The actual ad serving process is shown in Figure 1.
The specific process is as follows:



1. When a user sends a request to a web page, the web page Send the user's uid, url information to Ad Exchange;

2. Ad Exchange sends uid, url to DSP;

3. DSP is based on the received uid and url respectively To information, use the data management platform to query relevant user information and webpage information, and find matching advertising information according to the query information into the inventory library (the advertising information and the advertising constraint are pre-published by the advertiser on the DSP), and Decision-making advertising information and auction information are sent to Ad Exchange;

4. Ad Exchange obtains advertisements and auction information from multiparty DSPs and sorts them organically to generate successful auction advertisements and return them to the webpage in response to the user's request.

1.2 Formal Representation

Advertisers need to know the optimal price for each ad auction to maximize the benefits of the ad campaign. The optimal auction price for an advertisement depends on the value of the benefit that this advertisement can bring to the advertiser. The value of the advertiser's interest as a parameter of interest in the advertising campaign, usually expressed as the unit click fee CPC (Cost Per Click)[3].The optimal bid price will be determined by the expected cost of display. The performance of the impression cost (IC) is $IC = CPC * CTR$). In this case, the quality of the advertising promotion directly determines the accuracy of the CTR estimate, so the issue of optimal interest can be considered as the CTR estimation accuracy rate. For example, if the CTR is estimated to be higher, the bid for the auction will be higher than it should be, then the advertiser will waste the advertising campaign on the useless advertising display; otherwise, if the estimate is low, the bid price Will be low, advertisers will miss an ad showcase with good publicity.

In this link, the core module decision mechanism of DSP works. The key point is to calculate the click rate of the advertisement based on the user, webpage, and advertisement information, thereby obtaining the advertisement with the best click rate in the current environment and its specific click probability value. Model this process by formula as follows:

$$ad_{win} = f(u_i, p_i, Ad) \tag{1}$$

In the equation(1), the u_i refers to the speciific users,p_i represent the specific webpage.Ad represents a collection of advertisements in the DSP; ad_{win} on the left side of the equation represents the final decision. The best decision-making ad is equivalent to calculating the CTR of the highest eligible ad, so the problem can be Described as formula (2):

$$ad_{win} = max(CTR(u_i, p_i, ad_k))(k = 1, ..., n) \tag{2}$$

In the actual operating mechanism, user information, advertising information, and web page information constitute a huge amount of data with extremely low coupling. Faced with this data, the accuracy and time efficiency of ad click-through rate predictions pose a huge challenge to DSP.

2 Related Works

Since online advertisment has larget amount of profit, thereby, there numerous works about CTR prediction, This part will introudce some related work.

1. **Logistic Regression**.The early classification method is Logistic Regression(LR) [10].The advantage of LR is that it handles discretized features, and the model is very simple and easy to implement distributed computing.

The shortcomings of LR are also obvious. The features and features are independent in the model. For some features with crossover possibilities (such as: clothing type and gender, these two features are meaningful), a lot of labor is needed. Feature engineering is crossed. Although the model is simple, the manual work is a lot heavier. Moreover, LR needs to discretize features and normalize them, and boundary problems may occur in the process of discretization.

2. **Gradient Boost Decision Tree**.Gradient Boost Decision Tree(GBDT), the gradient lifting decision tree, is a nonlinear model with strong expression ability [3].The advantage of GBDT is that it handles continuous value features such as user history click rate and user history browsing times. And because of the tree splitting algorithm, it has a certain ability to combine features, and the model's ability to express is stronger than LR. GBDT is insensitive to the linear variation of the feature. It automatically selects the optimal splitting feature and the optimal splitting point of the feature according to the objective function. According to the number of splitting of the feature, the importance ordering of a feature can also be obtained. Therefore, the use of GBDT reduces the workload of manual feature engineering and performs feature screening.

3. **Factorization Machines and Factoriztion-field Machines**.Sine LR cannot handle cross-character automatically, Factor Machine(FM) and Factor field Machine(FFM) are aimed to handle this problem [11] [8].The advantage of FM is that it has the ability to handle quadratic crossover features, and it can achieve linear time complexity, and the model training is also very fast. The FFM is based on the FM, considering the characteristics of the field of feature intersection, but it also has no way to achieve linear time complexity, model training is an order of magnitude slower than FM, but the effect will be better than FM .

3 Methodology

since single feature are not able to work well during the model construction. we employ the FFM model to extract the combination feature. Then employ Logistics Regression to predict the target. for the optimization method, we select the SGD algorithm [4].

3.1 Logistic Regression

The application of logistic regression model for prediction is mainly divided into two steps.[2]. Firstly, using the training to construct a logistic regression model, the important feature vector parameter β in the logistic regression model is obtained through machine learning, and then the parameter set formula is applied to the test set to predict it. Consider the vector $x = (x_1, x_2, ..., x_k)$ of k independent variables, assuming that the conditional probability $P(Y =$

$1|X) = p$ is the probability of occurrence of an event based on the observation (ad click event), then the logic The regression model can be expressed as:

$$P(Y = 1|x) = f(x) = \frac{1}{1 + e^{-g(x)}} \tag{3}$$

in the formula, $g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k, \beta = (\beta_0, \beta_1, ..., \beta_k)^T$ is parameter of LR.
The conditional probability that the event does not occur is:

$$P(X = 0|x) = 1 - P(Y = 1|x) = 1 - \frac{1}{1 + e^{-g(x)}} = 1 + \frac{1}{1 + e^{g(x)}} \tag{4}$$

The goal is construct the logistic model is to compute the parameters of LR. $y = 1$ mean "click", $y = 0$ display "non-click". Suppose there are N obeserved values $Y_1, Y_2, Y_3, ..., Y_N$.Then randomly select n samples, labeled as $y_1, y_2, y_3, ..., y_n$. Suppose $P_i = P(y_i = 1|x_i)$ are conditional probability given the condition of $y_i = 1$, likewise, the conditional probability of $y_i = 0$ is $P(y_i = 0|x_i) = 1 - P_i$. Then, the probability of observed value is:

$$P(y_i) = p_i^{y_i}(1 - p_i)^{1-y_i} \tag{5}$$

Because the observations are independent of each other, their joint distribution can be expressed as the product of the marginal distributions:

$$l(\beta) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i} \tag{6}$$

Equation (6) is a likelihood function of n observations. The goal is to be able find the largest parameter estimate of this likelihood function.
The core of the maximum likelihood estimation is to find the parameter $\beta = (\beta_0, \beta_1, ..., \beta_k)$, so that the equation (6) takes the maximum value. Therefore, take the logarithm on both sides of equation (6):

$$L(\beta) = \sum_{i=1}^{n} [y_i ln(f(x_i)) + (1 - y_i) ln(1 - f(x_i))] \tag{7}$$

Solving the parameter orientation of the logistic regression model taking the maximum probability value The value β, you can calculate the value of CTR:

$$CTR = f(x, \beta) = \frac{1}{1 + e^{-\beta^T x}} \tag{8}$$

3.2 FM

FFM is effective method to dig the collaborative feature. The formula for FM model is:

$$f(x) = logsitics(linear(X) + \sum_{i=1}^{n} \sum_{j=i+1}^{n} w_{ij} x_i x_j) \tag{9}$$

The formula in the logistic function can be seen as two parts, one is the linear regression function we are familiar with, and the second part is the quadratic term. So FM handles the ability to cross between features.
But the quadratic weight w_{ij} requires us to store a variable of a two-dimensional matrix, and because the feature is massively discrete, the dimension of this two-dimensional matrix can be large. The author of FM uses the principle of matrix decomposition to decompose this weight matrix, ie $w_{ij} =< v_i, v_j >$.
Then the formula of FM becomes:

$$f(x) = logsitics(linear(X) + \sum_{i=1}^{n} \sum_{j=i+1}^{n} < v_i, v_j > x_i x_j) \tag{10}$$

4 Experiment Design

4.1 Dataset

The data we select is KDD cup 2012 track 2. The mission is to predict CTR from history click log of webpage. The task of the competition is to predict the click-through rate (CTR) of ads in a web search engine given its logs in the past. The dataset, which is provided by Tencent, includes a training set, a test set and files for additional information. The training set contains 155,750,158 instances that are derived from log messages of search sessions, where a search session refers to an interaction between an user and the search engine. During each session, the user can be impressed with multiple ads; then, the same ads under the same setting (such as position) from multiple sessions are aggregated to make an instance in the dataset. Each instance can be viewed as a vector (#click, #impression, DisplayURL, AdID, AdvertiserID, Depth, Position, QueryID, KeywordID, TitleID, DescriptionID, UserID), which means that under a specific setting, the user (UserID) had been impressed with the ad (AdID) for #impression times, and had clicked #click times of those. In addition to the instances, the dataset also contains token lists of query, keyword, title and description, where a token is a word represented by its hash value. The gender and segmented age information of each user is also provided in the dataset.

4.2 Preprocessing

For the category feature, we transfer them into numerical feature by one hot encoding. Besides, some discrete features have many values. For instance, we can utilize the cluster algorithm to cluster user id.
After processing the data, we divided the data into training data and validation data with ratio of 8:2. by handling this, we can employ 5 fold validation to train the model.

4.3 Evaluation Metrics

In computational advertising, AUC (Area Under Curve) is often used to count ROC (Receiver Operating Characteristics, Often used for pattern recognition, classifier performance presentation and performance evaluation) the area of the curve used to quantify the quality of the logistic regression model. The AUC is between 0 and 1, the larger the model indicates the better predictability of the model. After training the model, use the test set to enter the model and calculate the CTR. Predict the value and substitute it into the calculation AUC. The higher the AUC, the closer the predicted value is to the true value. Multi-dimensional feature combination ratio The model trained with one-dimensional feature combination has obvious AUC indicators. Rise.

4.4 Comparsion

To evaluate LR model, we compare other classifers such as Navie Bayes(NB), SVR and Ridge Regression(RR) [12] [2] [7].

5 Result

The best performance of each model on both validation AUC and test AUC can be shown as Table 1:

Model	Vadlidation AUC	Test AUC
LR	**0.8374**	**0.8023**
NB	0.8245	0.7932
RR	0.8137	0.7854
SVR	0.8203	0.7988

From the result of AUC we can find that LR outperform other models.

6 Conclusion

In the traditional model, the LR model is most powerful. This is the reason why LR are applied in CTR industry.Using Logistic Regression Models to Implement Advertising CTR Forecasting is Calculation Key research and hot issues in the field of advertising. According to the characteristics of the advertisement log information, multi-dimensional combination feature vector can be used to train a better predictive logistic regression model and a more accurate prediction of the click-through rate of the advertisement, thereby maximizing the business of the advertiser and the Internet user.

References

[1] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *The Journal of Machine Learning Research*, 14(1):3207–3260, 2013.

[2] Harris Drucker, Christopher JC Burges, Linda Kaufman, Alex J Smola, and Vladimir Vapnik. Support vector regression machines. In *Advances in neural information processing systems*, pages 155–161, 1997.

[3] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[4] Jerome H Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378, 2002.

[5] Saikat Guha, Bin Cheng, and Paul Francis. Challenges in measuring online advertising systems. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, pages 81–87. ACM, 2010.

[6] David Harrison. Advertisement targeting through embedded scripts in supply-side and demand-side platforms, June 20 2017. US Patent 9,686,596.

[7] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[8] Yuchin Juan, Yong Zhuang, Wei-Sheng Chin, and Chih-Jen Lin. Field-aware factorization machines for ctr prediction. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 43–50. ACM, 2016.

[9] H Brendan McMahan, Gary Holt, David Sculley, Michael Young, Dietmar Ebner, Julian Grady, Lan Nie, Todd Phillips, Eugene Davydov, Daniel Golovin, et al. Ad click prediction: a view from the trenches. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1222–1230. ACM, 2013.

[10] Nasser M Nasrabadi. Pattern recognition and machine learning. *Journal of electronic imaging*, 16(4):049901, 2007.

[11] Steffen Rendle. Factorization machines. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 995–1000. IEEE, 2010.

[12] Irina Rish et al. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46. IBM New York, 2001.