Each team is expected to submit one copy of their project report, code, and data.

(1) The report should be a pdf file.

(2) Include your team team, team member and UID , the role of each team member in the first page of the report.

(3) The report should be include sections 1) Introduction (show the background, motivation, challenge, contribution) 2)related work 3)notation and problem statement 4)your proposed model 5)Experiments (Include at least one competitors) 6) Conclusion. (7) **Detailed** description how to run your code.

(4) Code

(5) Data: If you can demonstrate the performance of the model using part of the entire data, you can just use part of data. CCLE only allow submissions under 100 MB. If you want to submit the entire data, you may want to submit the entire data in person by appointment.

(6) Make sure the code is run-able under linux or windows. Make it clear. What is the operating system, What is the software to run your code.

(7) We will strictly follow the descriptions in your report to run your code. Do not hardcode anything (for example, path). Write your descriptions as detailed as possible.

(8) No late submissions or drafts will be accepted

Project Report

Team Name: Observer

Team Member:

1) XIANG ZHONG 204412666

2) Yang Pei 304434922

3) Hongbo Zhao 604426609

4) Qianwen Zhang 004401414

5) Jing Zhao 404426610

6) Zhe Sun 604435430

Role of each member:

Sun Zhe:

Doing surveys about the algorithms and features and Exploitinghow to use those state of the art algorithms/packages/modular

Xiang Zhong:

Data Preprocessing and figure out how to select good figures

Yang Pei and Hongbo Zhao:

Implementing the algorithms, and do evaluation of different algorithms.

Qianwen Zhang and Jing Zhao:

Writing reports and prepare the presentations.

1. Introduction:

Background and Motivation:

In online advertising, click-through rate (CTR) is a very important metric for evaluating ad performance. The meaning of the click rate of advertising is to measure basically how much people have clicked the advertisements during a certain period. Obviously, more click means more money return for the advertisers. Therefore, as a result, click prediction systems are essential and widely used for sponsored search and real-time bidding. With a known click rate, the advertisers can try to adjust their advertisement in a more proper way.

1. Related Work:
2. Problem Statement:

The objective of our project is to predict whether a mobile ads will be clicked or not, given some information (attributes) of the ads. The datasets we have are training data set, which contains ten days of information about the ads, and a testing data set, which contains one days of information about the ads.

The training data set is about 7 GB, which have nearly 20 million records, and testing data set is about 1.27 GB, which have nearly 5 million records for us to predict.

Inside the data set, there are some attributes are known, such as:

banner\_possite\_idsite\_domainsite\_categoryapp\_idapp\_domainapp\_categorydevice\_iddevice\_ipdevice\_modeldevice\_typedevice\_conn\_type.

However, there are also some attributes that we don’t know what are they.

And also, it should be noted that all the data or value of each attributes are categorical data, and all ads’ ids are unique in the dataset. Therefore, we don’t have any numerical data and cannot compute the click rate respected to each ads as many of the similar problem that used to predict click rate, which makes the problem becomes difficult and limited our model selection and etc.

Also, the number of values of the attributes varies from 4 to 10000, which are very different from each other.

The following is a short example of the dataset:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| id | click | C1 | C2 | site\_category |
| "1000009418151090000 | 0 | 1fbe01fe | f3845767 | 28905ebd |
| "10000169349117800000 | 0 | 1fbe01fe | f3845767 | 28905ebd |
| "10000371904215100000 | 1 | 1fbe01fe | f3845767 | 28905ebd |
| "10000640724480800000 | 1 | 1fbe01fe | f3845767 | 28905ebd |

Evaluation Metric:

In this problem, because what we can predict is the probability of whether ads will be click, we can use the AUC as the evaluation metric, which is also a well-defined metric to evaluate this kind of problems.

That means, bigger AUC, the better the performance of the prediction.

1. Attributes selection

Because the data set are so large, it is not possible for us to put all the attributes or the data for training and testing, we try to filter out some attributes that are not useful. Therefore, the main challenge here is how to find out those attributes that are not useful in classification in this problem.

The meaning of useful of an attribute is that, if some values have high positive correlation of clicks of ads, while some values have high negative correlation of clicks of ads, we should consider this kind of attribute is useful. Otherwise, the attribute is not useful and we should filter it out.

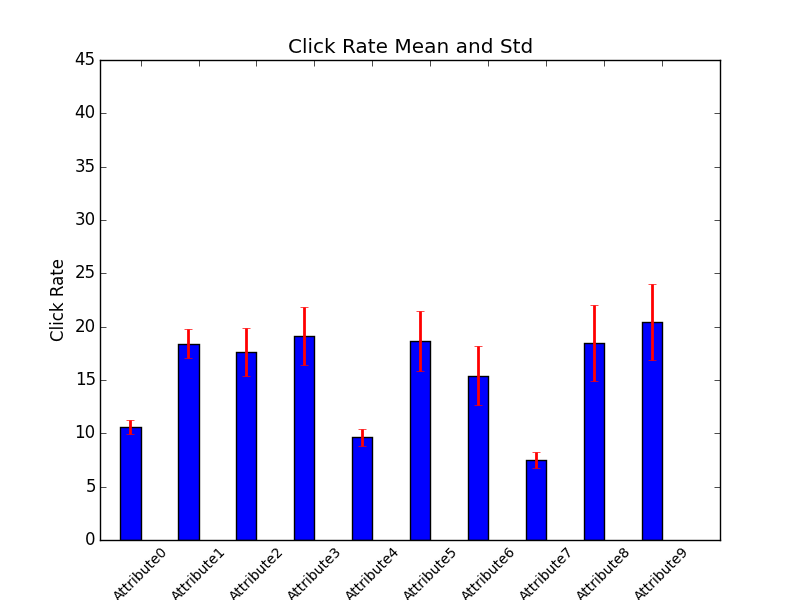
To find the correlation, we try to do some algorithm like frequent pattern mining, here we can think of it high variance pattern mining.

For a single attribute, we try to compute each value’s click rate (the meaning of click rate is how much click will appear if the value appears). If the variance is high, it means the attribute is useful since some values have high positive correlation of clicks of ads, while some values have high negative correlation of clicks of ads.

Beside just consider the relation between click and a single attribute, we also compute the relation between click and two attributes, for example, it makes sense that the ads’ site category combining the app’s category will have a strong relation with click of the ads.

However, because our limit of computing resource, we don’t consider the relation between click and more than 2 attributes, which I am sure it can produce more accurate results.

Here is the example of the variance and mean of the click rate of each attribute:

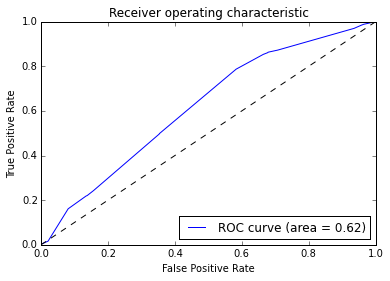


You can see that some attributes is good for classification, such as attribute 8 and attribute 9, while attribute 7 is bad.

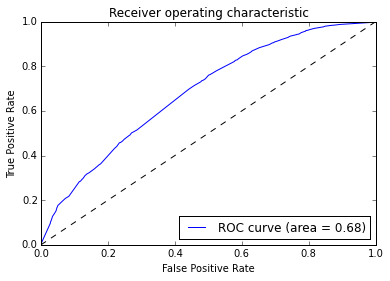
And the following is the improvement with and without attributes filtering.

(Using the ROC and AUC of logistic regression as examples)

Before filtering:



After filtering:



We can find that the performance of using attributes after filtering is better.

1. Proposed Model

4.1 Decision Tree

4.2 Naïve Bayesian Method

4.3 Logistic Regression

The logistic Regression algorithm is very useful in Classification if the given value of data set is numerical data. However, because the data set we have don’t have any numerical data. We have to find some way to transfer our categorical data into numerical data.

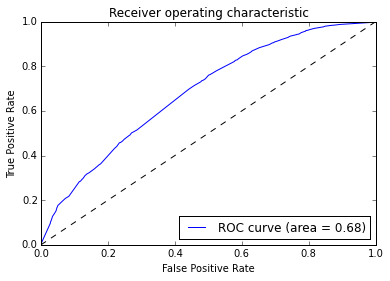
To do the transformation, we use the techniques of binary extraction and dimension increasing. That is, every dimension of the input to the logistic regression is a flag of whether a particular value of a particular attribute exists. For example, if the attribute app category has two values: game or tools, after the transformation, the input to the logistic regression will have two dimensions, one is whether the app is a game, and the other is whether the app is a tool. In this case, we are able to transfer the categorical data into numerical data, which can be able to apply to the logistic regression algorithm.

It should be noted that sometimes because of the number of values of all the attributes, the input will have too much dimension, which is hard for the algorithm running, that’s why the previous step attributes filtering is important. In addition, even though after filtering, the dimension is so high that other regression models such as SVM are difficult for our computing resources.

4.4 Neural Network

1. Experiments and Evaluation

Logistic Regression:



The following is a table summary the performance of different models:

1. Conclusion

6.1) Summary

6.2) Future Work

1. Details on how to run the demo code