Click-through Rate Prediction of Online Advertisement

Zhou Li, Yingzi Xu, Cai Li

1 Description of Problem

- Predict the click-through rate (CTR) by carrying out numerical analysis (linear/logistic regression aiming at the entire data and support vector machine (SVM) aiming at a subsect of data) after preprocessing the raw data.

2 Main Finding

- We explore the R package 'ff' which provides an efficient way for the fast storage and access of big data. The key idea is to map a section of data each time while the whole data is stored on disk.
- The SVM method using linear kernel gives the best prediction on the testing data, with AUC=0.6657 compared to the naive mean benchmark which fixes the CTR as the mean value of all CTR's in the training data and gives AUC=0.5000. Some other benchmarks are also given, including the Ad ID benchmark which gives the best AUC (0.7196). Results of other methods are included in Table 1.

Method	Test AUC
Mean Benchmark	0.5000
User Id Benchmark	0.5922
Ad Id Benchmark	0.7196
Weighted Logistic Regression	0.5867
Unweighted Linear Regression	0.5967
Weighted Linear Regression	0.5970
Logistic Regression	0.5970
SVM (Linear Kernel)	0.6657

Table 1: Performance of different methods in terms of AUC on the testing data

3 Analysis

- As described in Figure 1, we first randomly split the training data into three subsets: training (50%), validation (25%) and testing (25%). The user information (gender and age) in the file 'userid_profile.txt' is merged and missing values are excluded.
- We also transform the raw data into another version with binary response in order to carry out SVM and logistic regression analysis. Specifically, for each instance, we split the raw data into (*Click*) positive responses '1' and (*Impression* - *Click*) negative responses '0' while other variables keep unchanged.
- Before introducing our models, we first describe the features used by our models here. The data set contains *AdID*, *UserID*, *Gender*, *QueryID*, *KeywordID* and *TitleID* as categorical features. It also contains *Depth*, *Position* and *Age* as continuous features. To simplify our analysis, we only take into account a subset of the features (discussed later) mentioned above, although it may be more reasonable to consider all features which give a complementary information if we want a more accurate prediction. Note that in our linear/logistic model, we do not consider *UserID* and *AdID* which are included in our

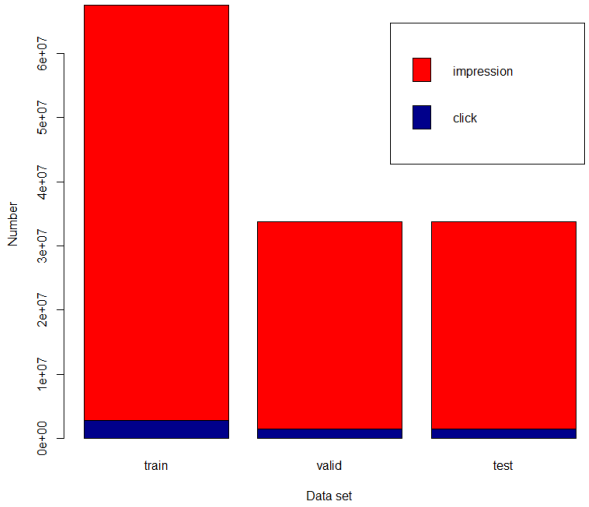

Figure 1: Number of clicks and impressions for training, validating and testing data

SVM model. This is due to the limitation of software because the `bigglm()` function in R cannot deal with such high dimensionality of predictors (as an illustration, there are 16,698,160 levels for *UserID* in our training data).

- Regression method:
  - The response is CTR = *Click* / *Impression* in the original data for linear model and the binary 0/1 response in the transformed data for logistic model.
  - To simplify the analysis, we only take into account a subset of the features as covariates in our model: *Depth*, *Position*, *Gender*, *Age*, *Relative Position*. Among them *Gender* is treated as categorical and all the others are treated as continuous. The *Relative Position*, which equals to *Position*/*Depth*, represents the relative order of an ad in the impression list.
  - We consider using *Impression* as weights when we are fitting the regression model, which makes sense because the CTR is more 'accurate' if more *Impression*s are included at one instance.
- SVM method:
  - The response for the SVM method is again the binary 0/1 response in the transformed data. Due to the complexity of the SVM method, only a very small subset (size of 5000) of the data is utilized for analysis.
  - The features we choose for SVM analysis include: *Depth*, *Position*, *Gender*, *Age*, *Relative Position*, *UserID* and *AdID*. Note that *UserID* is a categorical variable, we generate different dummy variables for each level of the *UserID*. That is, if one searching instance is from some user, we assign 1 to the corresponding dummy variable and 0 to the others. Similar transformation is done to *AdID*.
  - We consider the effect of *UserID* because we believe that different user may have a different tendency to click the ad. Although age and gender are given for each user, we feel that there may be other information 'hidden' behind the *UserID* (e.g. education and working status). Another reason is the average number of subjects who share the same *UserID* is greater than that for *AdID*, which indicates that *UserID* may provide more 'useful' information in our prediction.
  - The dummy variables mentioned above will dramatically increase the dimension of the features. In practice we use different combinations of features in our models. Specifically, we include (all

UserID, no *AdID*), (all *UserID*, top 25 *AdID*), (all *UserID*, all *AdID*) and (no *UserID*, all *AdID*) in our SVM model. We also compared linear kernel and radial basis kernel. The best result is from the linear kernel SVM including all *UserID* and top 25 *AdID*, which gives an AUC of 0.6657. Details are listed in Table 2.

Features			
UserID	*AdID*	Kernel	AUC
4489	0	Linear	0.6126
4489	0	Radial basis	0.5580
4489	25	Linear	0.6657
4489	25	Radial basis	0.5588
4489	4975	Linear	0.6339
4489	4975	Radial basis	0.5393
0	4975	Linear	0.5665
0	4975	Radial basis	0.4786

Table 2: SVM models and results

- Results are shown in Part 2.

4 Discussion

- Although we only take into account a few predictors in the regression model, the results provide a clear view of interpretation. Table 3 provides the parameter estimates for the unweighted linear regression model. We can easily draw some interesting conclusions from the results. For example:
  - people may not click through any ad when they see too many ads (see depth) at one instance;
  - elder people (see age) are more likely to click through some ad when using the search engine;
  - women (see gender level 2) prefer clicking through an ad online than men (see gender level 1).

	Coef	SE	P-Value
(Intercept)	0.1443	5×10^{-4}	0
depth	-0.0230	2×10^{-4}	0
position	0.0080	2×10^{-4}	0
factor(gender)1	-0.0032	2×10^{-4}	0
factor(gender)2	-0.0015	2×10^{-4}	0
age	0.0012	$0 \times 10^{+0}$	0

Table 3: Parameter estimates for the unweighted linear model

Nevertheless, these conclusions are based on our statistical models. Although they are statistically significant (actually under such a big sample size almost every effect is significant), they may not be practically significant. However, there is indeed evidence online to support our conclusions, see Older Online Users Are More Likely to Click On an Ad Than Those Young Whippersnappers, Women Click Facebook Ads More Than Men for more details.

- The implication of the small sample SVM analysis is, we can utilize some categorical ID features efficiently via SVM. This will surely increase the feature dimension as the cost, especially for our data, which has too many levels and hard to implement. And for those testing set, which do not have the same ID's as the training set, the performance level drops significantly. This is also supported by the result of our exploratory analysis. Thus, joining some discrete variable as a synthetic feature may help to improve the prediction performance at the cost of huge dimensionality. So is the case for the regression model.