ST 810 Homework 1 (Due Sept. 30)

Cai Li, Zhou Li, Yingzi Xu

September 29, 2013

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=0.7560 compared to the naive method which fixes the CTR as the mean value of all CTR's in the training data and gives AUC=0.5000. Results of other methods are also included in Table 1.

Method	Test AUC
Naive (Mean Value)	0.5000
Unweighted Linear Regression	0.5967
Weighted Linear Regression	0.5970
Logistic Regression	0.5970
Weighted Logistic Regression	0.5867
SVM (Linear Kernel)	0.7560

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

3 Analysis

- As descreibed 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.
- 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.

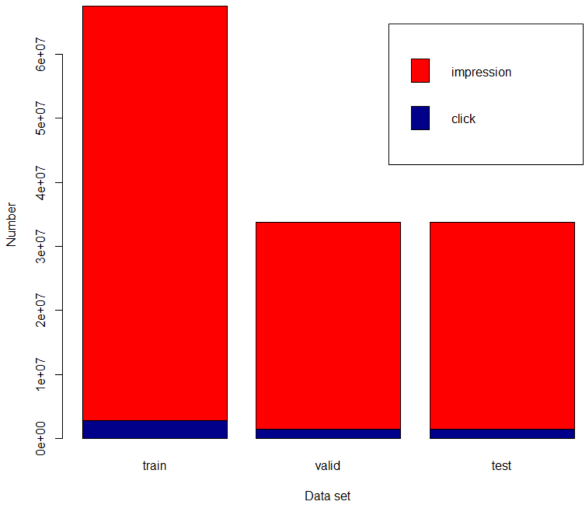

Figure 1: Number of clicks and impressions for training, validating and testing data

- 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 *Depth*/*Position*, 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 3000) of the training data is utilized for analysis.
  - The features we choose for SVM analysis include: *Depth*, *Position*, *Relative Position* and *UserID*. 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.
  - The dummy variables mentioned above will dramatically increase the dimension of the features, thus in practice we select the top 13 levels, which capture 10.73% (322/3000) records in our training data. The linear kernel and regularization parameter $c = 500$ is used for the SVM model.

4 Discussion

- Although we only take into account a few predictors in the regression model, the results provide a clear view of interpretation. Table 2 provides the parameter estimates for the unweighted linear regression model. We can easily draw some interesting conclusions from the result. 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;
  - men (see gender level 1) prefer clicking though an ad online than women (see gender level 2).

	Coef	SE	p
(Intercept)	0.1443	5e-4	0
depth	-0.0230	2e-4	0
posit	0.0080	2e-4	0
factor(gender)1	-0.0032	2e-4	0
factor(gender)2	-0.0015	2e-4	0
age	0.0012	0e+00	0

Table 2: Parameter estimates for the unweighted linear model

- 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. 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.