Report

The given problem is to predict the CTR using the historical data. This is a case of predictive analysis of historic data.

Out of the four data files (Bids, Impressions, Clicks, Conversions), I choose to work on impression and clicks files. My main focus was to predict the CTR and this data is available in the impressions and clicks files.

I started by working on imp.20131019.txt, clk.20131019.txt files and made the following observations.

- Conversions $\subset$ Clicks $\subset$ Impressions.
- I've considered the log type (1 = impression, 2 = click) as the class labels. In this case the classes were highly imbalanced (3500:1).

Before moving onto dealing with the class imbalance problem and model selection, I needed to modify the dataset (Data cleaning and data transformation).

1. Since most of the data is made up of ID's (Hash values, IP Address), we cannot find the mean, variance or correlation between data. As the first step I counted the number of unique values in each field.
   By finding the number of unique values I could remove some fields (fields where number of unique values is approximately equal to the number of records and fields with a single unique value) from the data.
   Initially there were 24 fields after the after 13 features were left.
2. Next I converted the hash values to numbers. I did this by assigning a number to each hash value/ IP address.
3. The dataset has only numbers where each number is a category (categorical variables).
4. Then I've performed descriptive analysis (variance and correlated features) and found out that the features 'City – Ip address' , 'Domain – URL' are highly correlated . As the features 'City' and 'Domain' are more descriptive I've removed the Ipaddress and URL features.
5. Changed the log type to 2 based on the data from clicks file.

The following are the final features used for training the model.

City, Ad Exchange, Domain, Ad Slot Id, Ad Slot Width, Ad Slot Height, Ad Slot Visibility, Ad slot floor price, Creative Id, Paying price.

The following are some of the methods to solve this problem.

- Clustering
- Logistic regression
- Ensemble learning
- Naïve Bayes
- Linear regression

I've chose work with Naïve Bayes for the following reasons:

1. Firstly, CTR = clicks/impressions i.e. probability of the ad being clicked.
2. The data consists of discrete values and class labels, it is not possible to predict CTR(continuous values).
3. As CTR can be thought of as probability, and Naïve Bayes predicts the probability of a record belonging to a particular class based on the given features I chose Naïve Bayes classifier.

The problem of unbalanced class can be solved by upsampling the data. I didn't perform upsampling as it would add dummy records that may result in inaccurate probabilities.

I've run the classifier with a 75-25 split.(75% training data, 25% Validation data). Here the accuracy of classification is not an indicator of the model's performance.

In the jupyter notebook you find can the values of accuracy, precision, recall. The accuracy for 20131019 dataset is given below.

Accuracy = 99.94

Recall = 99.94

Precision = 100

The above mentions metrics cannot access this model due to highly imbalanced classes. The performance of this model can accessed by comparing the overall CTR of the data with the predicted CTR.

I've run the classifier on four files separately and the results are given below.

Dataset	CTR	Predicted CTR(validation)	Predicted CTR(total)
20131019	0.000363	0.005709	0.001568
20131020	0.0003033	0.001882	0.0006508
20131021	0.0005997	0.0008306	0.000868
20131022	0.000562	0.000781	0.000662