International Journal of New Technology and Research (IJNTR)
ISSN:2454-4116, Volume-4, Issue-12, December 2018 Pages 24-26

Principal Component Analysis Of TF-IDF In Click Through Rate Prediction

Ankita Pal

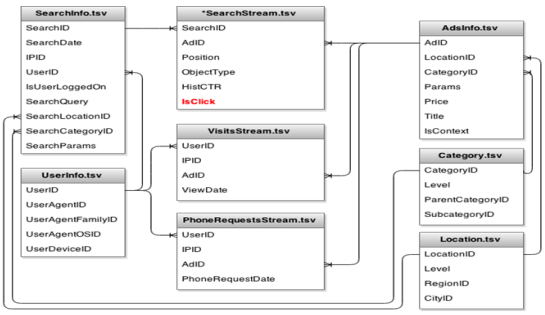
***Abstract*— This paper presents a model to predict the probability whether a user will click on a particular advertisement or not. The dataset used is that of Avito.ru provided as a part of the Kaggle competition- "Avito Contextual Ads Prediction". Here Principal Component Analysis on the Search and Query features is used, some extra count variables are made for integrate the categorical variables. Lastly logistic regression, SVM and gradient boosting algorithms are applied for classification into click or no-clicks.**

***Index Terms*— Avito Contextual Ads Prediction, Click Through Rate, Principal Component Analysis**

I. Introduction

Advertising is a major source of revenue for the companies which deals with search engines. These companies have a pay-per-click policy for charging the advertisers. As per the Pay-per-click policy, an advertiser has to pay for only those ads which have some clicks on them. They estimate the best possible location for placing ads so as to maximize revenue by using Click-through Rate (CTR). Hence, Click-through Rate prediction is very crucial to Sponsored Search.

In this paper, we have a training data set of 10.2GB and a test set of 531MB. For model formation we have taken small subsets of these data. The database has 8 tables as shown:



trainSearchStream.tsv and testSearchStream.tsv are two main data tables used in this paper of predictive model. They have a column ObjectType which shows which type of advertisement. Avito.ru supports 3 types of advertisements:

1. Regular: Free ads which go down with time
2. Highlighted: Advertisers pay a fixed amount for these ads. The ad stays on the top for a fixed amount of time then is shifted down.
3. Contextual: Advertisers pay per click. Therefore, it is necessary to predict the probability of click, so as to estimate the revenue from the ads.

In this paper we predict the probability of click of only the Contextual ads, which corresponds to ObjectType 3.

To reach our desired solution, we use the concept of Principal Component Analysis, a dimensionality reduction technique, on Term Frequency-Inverse Document Frequency weighted Document-Term matrix. Since, the Document –Term matrix is very large and very sparse, we need to reduce its dimensions to be able to use it in classification techniques. We also created some count variables of some categorical fields to be passed into the classification algorithm .We used various classification techniques and compare the results of various resultant models.

II. Methodology

A. *Data Pre-processing:*

1) *Data cleaning:* We have two text features –SearchQuery in SearchInfo and Title in AdsInfo. To integrate them into our predictive model, we need to prepare their Document Term Matrices (DTM) using the Term Frequency- Inverse Document Frequency weighting (TF-IDF). We make two corpuses, with each row of the two columns as documents of the respective corpuses. As a part of cleaning and pre-processing step we translate these corpuses from Russian to English and the remove punctuations, digits, stopwords, whitespaces and convert everything to lower case and also stem them.

2) *Feature engineering :*We make a DTM of the resulting 2 corpuses separately, using the TF-IDF weighting.

$$Wi,j = tfi,j \times log(\frac{N}{dfi})$$

Here, Wi,j =weight of i^{th} word in j^{th} document

tfi,j =number of occurrences of i^{th} word in j^{th} document

N= total number of documents

dfi=number if documents containing i^{th} word

We see the matrix is very large but very sparse. We need to reduce the dimensionality of the matrix , so we used Principal Component Analysis to reduce the number of dimensions and selected 2 each of the most relevant dimensions to represent the SearchQuery in SearchInfo and Title in AdsInfo.

In Principal Component Analysis,we decompose the DTM into Singular Value Decomposition (SVD) .

$$\begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix}_{m\times n} = \begin{bmatrix} u_{11} & \cdots & u_{1r} \\ \vdots & \ddots & \vdots \\ u_{m1} & \cdots & u_{mr} \end{bmatrix}_{m\times r} \times \begin{bmatrix} s_{11} & \cdots & s_{1r} \\ \vdots & \ddots & \vdots \\ s_{r1} & \cdots & s_{rr} \end{bmatrix}_{r\times r} \times \begin{bmatrix} v_{11} & \cdots & v_{1n} \\ \vdots & \ddots & \vdots \\ v_{r1} & \cdots & v_{rn} \end{bmatrix}^{T}_{r\times n}$$

Ankita Pal,Gurgaon, India

To get the principal components, we multiply:

$$B = V \times L^{\frac{1}{2}}$$
$$F = X \times B$$

Here F contains the principal components.
Another feature engineering step carried out was that of making count columns of LocationID, CategoryID and UserDeviceID to be passed into the classification algorithm.

B. Classification Algorithms:

1) *Logistic Regression:* It is one of the mostly widely used classification methods in CTR prediction. Binary Logistic Regression is a special type of regression where binary response variable is related to a set of independent variables, which can be discrete and/or continuous. In logistic regression the outcome is be 0 or 1. We apply the sigmoid function to the results of the linear regression model to get the dichotomous outcome.

$$f(x) = \frac{1}{1 + e^{\theta^T x}}$$

Here: f(x) =outcome variable
θ^T=matrix of weights
x =input vector

2) *Support Vector Machine (SVM)*: Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used classification as well as regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well.

3) *Gradient Boosting:* Gradient Boosting Machines(GBMs) are the learning procedure that consecutively fits new models to provide a more accurate estimate of the response variable. The idea behind this algorithm is to construct the new weak learners to be maximally correlated with the negative gradient of the loss function, associated with the whole ensemble. The loss functions applied can be arbitrary, but to give a better intuition, if the error function is the classic squared-error loss, the learning procedure would result in consecutive error-fitting.

C. Evaluation metrics:

The evaluation metric used was the log loss function. It is a classification loss function

$$\log loss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} log(p_{ij})$$

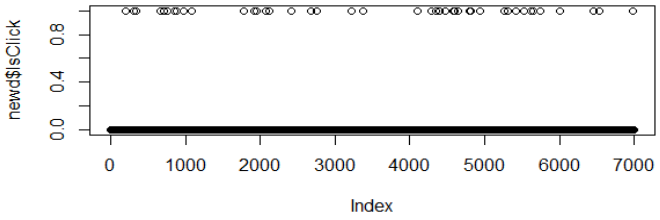
Here: N=number of samples
y_ij=is a binary indicator of whether or not label j is the correct classification for instance i
p_ij =is the model probability of assigning label j
When there are only 2 classes i.e. it is a Binary Classification then:

$$\log loss = -\frac{1}{N}\sum_{i=1}^{N} y_i log(p_i) + (1 - y_i) log(1 - p_i)$$

Here: N=number of samples
y_i =is a binary indicator of whether ad is clicked or not
p_i =is the model probability of the advertisement being clicked

Log-loss measures the accuracy of a classifier. It is used when the model outputs a probability for each class, rather than just the most likely class. Log loss measures the uncertainty of the probabilities of your model by comparing them to the true labels.

III. RESULTS



Plot of Clicks in first 7000 records with ObjectType 3

From figure we notice that the number of no-clicks is far more than clicks. Thus, we expect the click through rate to be very small numerically. The probability of click lies between 0 and 1 for all records in all models. We are more interested in how much close the probability is close to actual: 0 for not being clicked, 1 for being clicked.
We compare the log loss we get from the 3 models we made using the 3 classification models:

S.No	CLASSIFICATION ALGORITHM	LOG LOSS	Rank
With the text features reduced by PCA of TF-IDF weighted matrix, without count features.			
1	Logistic Regression	0.0331649	2
2	Support Vector Machine(SVM)	0.0871754	3
3	Gradient Boosting	0.0322150	1
With the text features reduced by PCA of TF-IDF weighted matrix, with count features.			
1	Logistic Regression	0.0326645	2
2	Support Vector Machine(SVM)	0.1252218	3
3	Gradient Boosting	0.0322914	1

IV. CONCLUSION

We observed the performance of various classification models:

- With the text features reduced by PCA of TF-IDF weighted matrix, without count features.
- With the text features reduced by PCA of TF-IDF weighted matrix, with count features.

We conclude that the best model is with Gradient Boosting, followed by Logistic Regression.
We notice that SVM is not an appropriate classifier.
The reasons for the good log loss scores for Gradient

Boosting is that it is an ensemble method and increases the accuracy by learning from weak classifiers, and does not need feature engineering. By feature engineering we further tune up the model for better predictions.

The main challenge was the large dataset provided as a part of the competition, which could not be handled very efficiently, given the restricted resources. So we took a subset of 10000 records for the project. The DTM for larger records could not be stored in memory. Another challenge was that of conversion of Russian data to English, which was done using Google Translate.

REFERENCES

[1] Steffen Rendle. Scaling factorization machines to relational data. In Proceedings of the VLDB Endowment, volume 6, pages 337–348. VLDB Endowment, 2013.
[2] 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.
[3] Zhipeng Fang, Kun Yue, Jixian Zhang, Dehai Zhang, and Weiyi Liu. Predicting click-through rates of new advertisements based on the bayesian network. Mathematical Problems in Engineering,2014, 2014.
[4] Norwegian Computing Center, P.B. 335 Blindern, N 0314 Oslo 3 (Norway) and Research Group for Chemometrics, Institute of Chemistry, Umed University, S 901 87 Umeci (Sweden). Principal Component Analysis
[5] AlexeyNatekin and AloisKnoll Gradient boosting machines a tutorial.2013 available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3885826/

Ankita Pal B.Tech in Mathematics and Computing from Delhi Technological University(2013-2017). General Secretary of Society of Industrial and Applied Mathematics for the period of 2016-2017. Currently working in Axtria India Pvt. Ltd.