Predicting Commercial Promoted Contents Will Be Clicked By User

Name: Gary(Xinran) Guo Email: garyguo@stanford.edu

PREDICTING
As e-commerce, social media grows rapidly, Advertisements are everywhere in human daily activities. Clickstream data becomes incredibly large and it contains tremendous treasures that we can learn for people's interests, hobbies, and personalities. Our goal is to predict which pieces of recommended content users will be likely to click on. The idea is to rank the recommendations served to a specific user in a specific context. Each context (i.e. a set of recommendations) is likely to click on.

DATA
Public data set from Ourbain.com, the web's leading content discovery platform delivers these moments or advertisements while we surf our favorite sites. User Clickstream: the log of users visiting documents. Page Document: The detail metadata of document include, sources, publisher, topics, entities, Promoted Content: The detail metadata of advertisement, that each ad belongs to a campaign run by an advertiser

MODELS
Regularized Click Probability: only focus on ad itself, suppose not to consider any other related features. Calculate the click probability of each single ad via full training set. Test data click probability will ranked based on table.

RESULTS
Sample Output and Feature and Model prediction comparison

DISCUSSION
• This project is application based. So the major task is to applying different machine learning models to a practical problem and make prediction. Features selection and model comparison will be primary focus of project.
• Doing exploratory data analysis (EDA) is one of the most important tasks for this project. The cumulative volume of data sets exceed 20GB. Knowing feature distributions helps significantly on features selection.
• After EDA, finding that the average page view of single user is just 2.835, it indicates that user-based recommendation profile is not suitable for click-prediction, instead, document categories and topics have uniform distribution among data.
• In Regularized Click Probability model, adding a regularized term REG that penalizes ads with small amounts of data, therefore making it prefer an ad with large amounts of training data and a reliable probability. It improves score by 2%
• More advance features is not guaranteed to make prediction better. Comparing the result of basic feature, and advance features, The prediction score is actually decreasing when apply more unrelated features.

FUTURE
• Do more exploratory data analysis to discover more features distributions in huge amount of data set, such as calculate mean, sum, standard deviation.
• Join various tables, and select different features to do prediction and compare result, test error.
• Apply Keras, deep Learning library for Theano and TensorFlow, high-level neural networks library. To deeply dig more into user click stream (page view).
• Changing the parameters of the model in logistic-regression algorithm and compare results.

REFERENCES