Machine Learning Engineer Nanodegree

Capstone Proposal

Vladimir Baranov December 18th, 2017

Proposal

Domain Background

Machine learning is the science of helping computers discover patterns and relationships in data instead of being manually programmed. It's a powerful tool for creating personalized and dynamic experiences, and it's already driving everything from Netflix recommendations to autonomous cars. One of the main goals of machine learning is to improve the quality of service provided by a company [1]. This leads to better user experience, customer satisfaction and retention resulting in higher sales. At the same time knowing the needs of a user helps a company to reduce operational costs [2].

Airbnb is an online marketplace where people list, discover, and book accomodations around the world. It has collected various data about users. This data about the usage patterns of its present user base can be utilized to predict patterns about its future users to provide them with customized suggestions to serve Airbnb's customers better.

This project is a Kaggle competition that was run by Airbnb from Nov 2015 to Feb 2016 [3].

Problem Statement

In this project we are going to train a classifier in order to predict which country will be the first destination of a new user given a dataset containig information of a user, his activities on airbnb, summary statistics of a user and destination countries.

Datasets and Inputs

The dataset was obtained from Kaggle [3]. It consists of 5 csv files.

train_users.csv - the training set of users containing 213451 rows of continious and categorical data

test_users.csv - the test set of users containing 62096 rows of continious and categorical data

- id: user id
- date_account_created: the date of account creation
- timestamp_first_active: timestamp of the first activity, note that it can be earlier than date_account_created or date_first_booking because a user can search before signing up
- date_first_booking: date of first booking
- gender
- age
- signup_method
- signup_flow: the page a user came to signup up from
- language: international language preference
- affiliate_channel: what kind of paid marketing
- affiliate_provider: where the marketing is e.g. google, craigslist, other
- first_affiliate_tracked: whats the first marketing the user interacted with before the signing up
- signup_app
- first_device_type
- first_browser
- country_destination: **target variable**

sessions.csv - web sessions log for users

- user_id: to be joined with the column 'id' in users table
- action
- action_type
- action_detail
- device_type
- secs_elapsed

countries.csv- summary statistics of destination countries in this dataset and their locations

age_gender_bkts.csv - summary statistics of users' age group, gender, country of destination

Solution Statement

To tackle this problem, we will use supervised machine learning. We will find up to five most probable solutions to this problem. We believe that the most suitable method to solve this problem is Logistic Regression, because it is computationally inexpensive baseline algorithm in most frameworks, which is widely used in applications [4]. But most inportantly, this method returns well calibrated probability predictions by default as it directly optimizes log-loss [5]. In case this this method will give poor results, the other option is to use the other methods such as SVM or Adaboost. However, these classifiers should be calibrated manually [5].

Benchmark Model

Taking into account that NDF is the most common class, we will use all-NDF benchmark. In other words, we will fill the fields to be predicted with 'NDF'. This submission yields 0.68411 NDCG (for details see below) score in the Kaggle leaderboard.

Evaluation Metrics

The evaluation metric for this competition is $NDCG_k$ (Normalized discounted cumulative gain) where k=5. NDCG is calculated as:

$$DCG_k = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{log_2(i+1)},$$

$$NDCG_k = \frac{DCG_k}{IDCG_k},$$

where rel_i is the relevance of the result at position i and $IDCG_k$ is the maximum possible (ideal) DCG for a given set of queries. All $NDCG$ calculations are relative values on the interval 0.0 to 1.0.

When we predicting the countries, the ground truth country is marked with relevance = 1, while the rest have relevance = 0. For example, if for a particular user the destination is 'FR', then the predictions become:

[FR] gives a $NDCG = \frac{2^1-1}{log_2(1+1)} = 1.0$

[US, FR] gives a $NDCG = \frac{2^0-1}{log_2(1+1)} + \frac{2^1-1}{log_2(2+1)} = \frac{1}{1.58496} = 0.6309$

Project Design

- Programming language : Python 2.7
- Libraries : Scikit-learn, Pandas
- Workflow:
    1. Preprocessing and data visualization.
        a. Drop date_first_booking since it contains 100% and 68% of missing data in test and train sets, respectively.
        b. Devide date_account_created and timestamp_first_active into three columns each - year, month, and day.
        c. Set age less than 15, more than 105 and >1999 to NaN. Set 1900<age<2000 as 2014-age, since in this case it means birth year.
        d. set NaN values in first_affiliate_tracked and in language to the most frequent value. The frequency of NaN here is less then 3% and this won't affect the data a lot.

e. set NaN values of age and gender columns to -1. The NaN frequency here is about 50%.
f. set NaN values of first_browser to 'missing'.
g. create visualisations for the data

2. Machine Learning
   a. Train the model using logistic regression, SVM, and adaboost. Calculate the scores and choose the best model.
   b. Tune the parameters using greed search.
   c. Train optimized model. Calculate score.
3. Results and Conclusions
   a. Add descriptions to the notebook.
   b. Report the results.

References

[1] Bernardo, Francisco, et al. "Interactive Machine Learning for End-User Innovation." American Association for Artificial Intelligence (2016).

[2] Lee, Wenke, et al. Journal of Computer Security **10**, 5 (2002).

[3] https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings.

[4] P. Harrington, Machine learning in action. Manning publication co. (2012).

[5] http://scikit-learn.org/stable/modules/calibration.html.