Proposal for the Starbucks's Capstone Challenge

by Ferenc Török

Introduction

The Starbucks Corporation is an American multinational coffeehouse chain. The company was founded in 1972 and today operates in more than 30000 locations in 77 countries worldwide. The dataset given was created by simulation and mimics purchasing decisions and how those decisions are influenced by promotional offers in a 30 day period. Each person in the simulation has some hidden traits that influence their purchasing patterns. In the simulation people produce various events, including receiving offers, opening offers, and making purchases.

Of course in real life, Starbucks has a wide variety of products to offer, however for the sake of simplicity the simulation includes only one product. There are however 10 types of promotions about this one product, such as 'BOGO' (buy-one-get-one) or even a simple advertisement. The dataset contains records of some basic personal data of the costumers, the transactions with time-stamped data (transactions, when did the costumer receive an offer, when did he view it etc.) and the details of the offers.

With the amount of data gathered in recent times and with the advance of Machine Learning techniques and Artificial Intelligence a new era in advertising raises. This is the era of personal advertisement where based on data about costumers one is able to send relevant and only the most relevant promotions and advertisement to them. This stimulates purchasing and hence the company is able to gain larger profit.

Problem Statement

The aim of personal advertisement is to send each costumer promotions which are the most probable to make them satisfied. Also it is important not to send them promotions which will most probably not interest them since these might even have a negative effect on the consumption. Also in some situations the people fulfill some promotions without even noticing that they existed. These are also situations to avoid since in this situation the company gave these people some discount although it was not necessary.

In this respect the aim of the project is to classify offers given a costumer and its purchasing history into one of the following categories:

- Will not even be viewed
- Will be viewed
- Will be viewed and completed
- Will be completed without viewing it

After an offer is classified into the groups above one can decide to send or not to send that particular offer to the costumer. A reasonable choice would be to send someone an offer if it falls into the categories 2 or 3 and not to send if it falls into 1 and 4.

Datasets and inputs

The dataset is stored in 3 '.json' files: 'profile.json', 'portfolio.json' and 'transcript.json'.

profile.json

The profile.json file contains personal data about costumers in the following fields (17000 costumers):

- gender: (categorical) M, F, O or null
- age: (numeric) missing value encoded as 118
- id: (string/hash)
- became_member_on: (date) format YYYYMMDD
- income: (numeric)

portfolio.json

The portfolio.json file contains the data of the offers sent during the test period in fields (10 offers):

- reward: (numeric) money awarded for the amount spent
- channels: (list) web, email, mobile, social
- difficulty: (numeric) money required to be spent to receive reward
- duration: (numeric) time for offer to be open, in days
- offer_type: (string) BOGO, discount, informational
- id: (string/hash)

Transcript.json

The transcript.json file contains timestamped data about transaction and offers in the following fields ((306648 events):

- person: (string/hash)
- event: (string) offer received, offer viewed, transaction, offer completed
- value: (dictionary) different values depending on event type
  offer id: (string/hash) not associated with any ”transaction”
  amount: (numeric) money spent in ”transaction”
  reward: (numeric) money gained from ”offer completed”
- time: (numeric) hours after start of test

Solution Proposal

To solve the classification problem of the Problem Statement we propose to use the following solution:

We are only going to use data about offers which did not have an expiration date after the end of the 30 day period, because the reaction to these offers could not be measured accurately. Hence these transcript points are discarded. Other than that, all received offers are going to be examined and are going to be classified into the four outcome groups stated in the Problem Statement section.

During constructing the feature vector for every outcome (outcome of received offer) we are going to fabricate a feature vector that contains the following information:

- personal information of the costumer
- average consumption of the costumer until receiving the offer
- some statistics about how the costumer reacted to previous offers until receiving the offers

The key of the proposed method for feature engineering in this respect is that outside of the personal data the consumption data is only used until the time the costumer received the offer. This allows to simulate the real word situation for which we would like to use the trained model in the first place: based on the consumption history of the costumer, try to predict how he/she will react to an offer.

The big advantage of this method is that the trained model will be able to predict the reaction of costumers with a wide variety of lengths of recorded history. It will even be able to predict the reaction of old costumers as well as new ones about whom we do not yet have any record. (We do realize however that for this specific task, to offer products to totally new costumers, other methods could be more accurate.)

We have chosen a deep neural network as model and we are going to train it on the data engineered according to the guidelines proposed above.

Benchmark model

To be able to measure the performance of the deep neural network we are going to compare it with our benchmark model which is chosen to be a k-nearest neighbor (kNN) classifier.

Performance measure

As the task is multiclass classification, accuracy is the suitable performance measure. (False positive, False negative etc. is not defined for the multiclass case.) The accuracy is hence going to be calculated with the usual formula:

$$Accuracy = \frac{\sum_{i=1}^{N} \mathbb{I}\left(\hat{y}(x) = y(x)\right)}{N}$$

where $\hat{y}(x)$ and $y(x)$ are the predicted and real class respectively and $\mathbb{I}(statement)$ is the indicator function which is 1 if *statement* is true and 0 otherwise.

Project Design Outline

- The data has already been examined. The distributions of the interesting features is checked, some statistics of the data are plotted.
- The features are going to be engineered and the anomalies in the dataset which were discovered during data exploration are cured.
- The data is split into train, validation and test sets
- The models are trained
- The results of the models are examined and compared.
- Conclusion is derived