Machine Learning Engineer Nanodegree

Capstone Proposal

Haoran Zhu 06/21/2018

Domain Background

The central problem that this project is designed to accomplish is forecasting future sales. Forecasting sales are extremely essential for retailers to control cost. Depending on the type of retailer, the smaller the profit margin for a given market, the more critical on cost control. Optimizing stock, as the quantity of merchandise available on the premise of the store or warehouse, also means better utilization of shelf space and a greater variety of products to be sold at retail or online for maximizing sales.

Problem Statement

The sales forecasting for the project is based on one of the largest software firms in Russia, 1C Company. With the forecasting report, the company can make the precise decision on procurement to reduce cost and prevent overstocking. The data used for the forecasting is contributed by the Coursera course organizer who created the Kaggle competition "Predict Future Sales". The training data is the historical daily sales record. The sales record is individual entries that contain the date, price, count, product ID, category ID, and shop ID. The approach of solving the problem is to explore the sales data and find a correlation between sales to region and categories of items, then to determine how sale changes in terms of time. The data that contains strings, such as product name, category name, and shop name are all in the original language, Russian, and translation of the string to English may be required for better interpretation.

Datasets and Inputs

The time-series dataset consisting of daily sales data, kindly provided by one of the largest Russian software firms - 1C Company

File descriptions

- **sales_train.csv** - the training set. Daily historical data from January 2013 to October 2015.
- **test.csv** - the test set. You need to forecast the sales for these shops and products for November 2015.
- **sample_submission.csv** - a sample submission file in the correct format.
- **items.csv** - supplemental information about the items/products.
- **item_categories.csv** - supplemental information about the items categories.
- **shops.csv**- supplemental information about the shops.

Data fields

- **ID** - an Id that represents a (Shop, Item) tuple within the test set

- **shop_id** - unique identifier of a shop
- **item_id** - unique identifier of a product
- **item_category_id** - unique identifier of item category
- **item_cnt_day** - number of products sold. You are predicting a monthly amount of this measure
- **item_price** - current price of an item
- **date** - date in format dd/mm/yyyy
- **date_block_num** - a consecutive month number, used for convenience. January 2013 is 0, February 2013 is 1,..., October 2015 is 33
- **item_name** - name of item
- **shop_name** - name of shop
- **item_category_name** - name of item category

Solution Statement

For this problem, the sales contain sequential information of item price and the number of products sold and the target is the prediction of a monthly number of products sold. RNN could use sequential information, and the output being depended on the previous computations. Thus I choose LSTM(Long short-term memory), a special kind of RNN, to solve this problem.

Benchmark Model

In discussion board, a kaggler shared a solution "Playing in the Sandbox" which is using XGBoost. Its prediction score is 1.19015.

Evaluation Metrics

Predict the number of products sold in next month and upload the submission file to Kaggle. Kaggle will evaluate it with real sale number by root mean squared error (RMSE) and return a score (the smaller the better).

Project Design

1. Explore the dataset
2. Clear outliers
3. Reshape the dataset into training data
4. Normalization
5. Build LSTM model and train
6. Evaluate the performance
7. Try different architectures