Walmart Store Sales Forecasting

Delong Meng

Nov 8th, 2021

Introduction

There has been great interest in the business field to predict sales, and machine learning technique is a great tool for this task. In this project, we are provided with historical weekly sales data from February 2010 to October 2012 for 45 Walmart stores located at different regions including 99 departments. In addition, we also have the information whether a particular week falls into a major event/holiday, such as Super Bowl, Labor Day, Thanksgiving, and Christmas. Our goal is to predict the future weekly sales for any given department at a given store based on all of the historical data.

Methods

Train/test splitting of the data

This dataset contains 421570 data points in total. We break down the dataset based on the dates. For example, as shown in Table 1, all of the 164115 data points between 2/2010 and 2/2011 (13 months) were used as the initial training set (train_ini). All of the remaining data spanning 20 months (from 3/2011 to 10/2012) were treated as the test set, and were divided into 10 folds, with each fold containing 2 months.

Table 1. Overview of the dataset

subset	size	date_from	date_to
train_ini	164115	2010-02-05	2011-02-25
fold_1	26559	2011-03-04	2011-04-29
fold_2	23543	2011-05-06	2011-06-24
fold_3	26386	2011-07-01	2011-08-26
fold_4	26581	2011-09-02	2011-10-28
fold_5	26948	2011-11-04	2011-12-30
fold_6	23796	2012-01-06	2012-02-24
fold_7	26739	2012-03-02	2012-04-27
fold_8	26575	2012-05-04	2012-06-29
fold_9	26599	2012-07-06	2012-08-31
fold_10	23729	2012-09-07	2012-10-26

Train/test process and performance evaluation

We performed the train/test process at a series of different time points. For example, we first went to the end of 2/2011, and used all the data prior to that time point, namely the initial training set, to train the model.

Instead of using the whole test set at once, we used all of the data points of 2 months immediately following the dates of the training set, in this case fold_1, as a test set to evaluate the initial model. Then we moved forward to next time point by two months, to the end of 4/2011, and updated our training data by including these months' data from fold_1. This time we fit a new model (using the same model building process though) based on the updated training data and used fold_2 as a new test set. We repeated this process 10 times until we came to the last data point, the end of 8/2012, used all data prior to that time point to build the model, and tested it using the last two months' data in fold_10.

By performing this training and testing process, we tested 10 slightly different models 10 times in total. In each time, we evaluate the corresponding model using the weighted mean absolute error (WMAE):

$$\text{WMAE} = \frac{1}{\sum w_i} \sum_{i=1}^{n} w_i |y_i - \hat{y}_i|$$

where w_i is the weight assigned to each data point (w = 5 for a holiday week and w = 1 for a regular week), n is the number of data points, and y_i and y_i ^{hat} represent the actual sales value and predicted value, respectively. By doing this, we obtained 10 WMAEs and were also able to calculate the average WMAEs as a metric to evaluate our modeling strategy.

Data pre-processing and modeling

In general, the weekly sales data have similar pattern in every year. For example, the Christmas week last year had much more sales compared to the weeks before and after, and this pattern still holds true this year. Thus, we extracted the week information from the date and then converted it to one-hot encodings. Note that there is a shift of the weeks in different years and can be fixed by subtracting 1 from weeks in 2010 so that the week number for the holidays among the years all matched.

In addition, different stores and departments have different sales. Specifically, we found that each department has a certain pattern of sales trend along the year across the stores. So we applied singular value decomposition (SVD) on each department's data in the training set, and used the first 8 components to represent a smoothed version of the original data.

Next, we treated specific store/department combination separately and built a specific linear regression model for the sales as response using year and week information as predictors. Note that week number were treated as a categorical variable as mentioned above and years treated as a continuous variable.

Results

Performance of the two models on the test data set

As shown below, the 10 test WMAEs and the average WMAE of the modeling strategy were summarized in Table 2.

Table 2. Test WMAEs

subset	WMAE
fold_1	1941.58
fold_2	1363.46
fold_3	1382.50
fold_4	1527.28
fold_5	2310.47
fold_6	1635.78
fold_7	1682.75
fold_8	1399.60
fold_9	1418.08
fold_10	1426.26
average	**1608.78**

Running time and computer system

Total running time of 10 cycles of modeling and testing were around 1.76 minutes. The computer system used in this project was: MacBook pro, 2.2 GHz, 16 GB memory.

Discussion

Forecasting is an important component of machine learning prediction applications in the business field because it can help people to allocate resources and prepare supplies in advance. In this Walmart store sales forecasting project, the overall performance of our modeling strategy was impressive. As expected, the seasonal and holiday information is critical for the sales. Note that although we didn't directly use the holiday information, it was already embedded in the week numbers. On the other hand, years reflect the overall trend and also contribute to the prediction. It is worth note that applying the SVD technique greatly improved the performance by reducing the noise and better extracting the true signal from the training data of a particular department. Different department has different patterns of sales throughout the year, and this is likely due to different needs on products of different departments in various seasons/holidays. In the future we can further improve the performance by fine tuning the prediction according to more precise information of the holidays, for example which particular day of the week the holiday falls into.

Acknowledgement

Dr. Feng Liang's instructions.

https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting