Report Project 2 STAT 542

Yuhui Luo

11/7/2018

Pipeline Introduction

In this project,we built three time series models to predict 2 month ahead sales for different departments across different stores for walmart. In evalution of the model performance, we choose WMAE which gives more weight to the holiday season department sales compare to normal seasons. In total we have 30 WMAE(10 for each folds,10 folds in total with 3 models it's 30 WMAE).

For model trainning:first we get the sale data for specific department in a specific walmart sale, this part data is time series data for single store. Then we build seasonal naive model and other models based on each time series data and make prediction for corresponding test department for 2 month period. Lastly we loop over each department and get prediction for each test department.

Details of coding explanation is divided as three parts:

- Preprocessing
- Models
- Postprocessing

Preprocessing

All the preprocessing steps is built within `mypredict` function and being called on by **evaluationCode.R** and being looped over. Within `mypredict` , we first choose start_date and end_date for the designated period of training period,then we extract the corresponding 2 month ahead fold test data. From there, within `for (dept in test_depts) ... loop` , we extract store-level sales data from the `train_frame` , and call our model function and make predictions.

Models

In this project,we built 3 models in total:

- Naive model
- Time Series Trend model
- Hybrid model(tslm_basic for fold 1-7,stlf.svd for fold 8-10)

Naive Model

Naive model is being directly copied from instructor joshua's code. In his docs, naive model is the last corresponding observation from train set,so the test prediction equals excatlys as the previous train data point.

TSLM (Only on trend)

for the second column of `error.csv` , the method being used here is TSLM model,but we only built model on trend,no seasonal pattern were included here.

Hybrid model(tslm model for fold 1-7,stlf.svd for fold 8-10)

TSLM and STLF both are seasonal,trend linear regression model.But before linear model buildining, I conducted `SVD` (Single Value Decomposition) for train test to get a smoother train data and built model upon the transformed train set.

Between fold 1-7, we use TSLM, however after fold 7, we transformed our model to a stlf model, that's because the model learning of stlf requires at least 2 season, and therefore we can only start at fold 8.

Postprocessing

For the step of postprocess, it's more like undo preprocessing. In this part, `update_test()` is a function to update the test result to the test_frame, then `evaluationCode.R` will loop over all the saved test result and output the error for the ten folds.

Appendix

- Running Time: 2444.079
- System: 3.3 GHz,16 GB MacBook Pro

Fold	Model 1	Model 2	model 3
1	2078.726	3092.643	1967.499
2	2589.338	3143.451	1377.466
3	2253.936	2578.291	1385.451
4	2823.098	2194.317	1549.900
5	5156.012	5514.857	2310.403
6	4218.348	2931.628	1639.898
7	2269.904	2639.593	1686.314
8	2143.839	2794.890	1534.991
9	2221.145	2462.216	1381.363
10	2372.425	2128.114	1344.238
Averege	2812.677	2948.000	1617.752