Analytics for Sequential Data
Homework 3

Instructions: Take your time to attempt this homework and do not be tempted to see the solutions till you give it an earnest attempt. Recall that homework will neither be collected nor graded. Solutions will be provided during the following week after this is posted.

Data: This is a forecasting homework for which two data sets are provided. One is `avocado.csv` and the other `temperature.csv`. View the files on excel first to determine which feature must be parsed as dates. Both files are located in Blackboard under "Datasets". In the avocado set, the most used columns are: AveragePrice, Total Volume, region, and Date. In the temperature set, we will only consider four cities: Atlanta, Boston, Houston, and San Diego. Keep only these four cities in the "region" column in the avocado set as well.

1. Merge the two datasets to create a new dataset which has Date, AveragePrice, Total Volume, and region from the avocado set (where the region column has only the 4 cities mentioned above) and the maximum, minimum, mean, and standard deviation of temperature for the week starting on the date in the Date column. Notice that the temperature is hourly data whereas the avocado set is weekly data.

2. Create a new data frame with only the Atlanta data sorted by date. Plot the time series of AveragePrice and also that of Total Volume. Do you see any trends or seasonality? Plot the ACF as well as PACF for both time series. Which lags are most important?

3. For the two time series with Atlanta data (and no other feature) create an exponential smoothing (ES) forecast and an ARIMA forecast. Do not use any seasonality. For ES, use additive trend. For the two time series check how much differencing would be optimal. Use that with $p = 2$ (why?) and $q = 1$. Use $T = 100$ to train. Then use a sliding window of 100 units each time to train the next. Obtain the usual metrics we used in our python codes. Notice which method did better.

4. For the Total Volume which was predicted less accurately in Atlanta than AveragePrice, let us add exogenous variables. Use the four weekly temperature statistics (max, min, mean, stdev) as exogenous features and predict using SARIMAX. First benchmark the AveragePrice with Gradient Boosting and then use XGB to forecast Total Volume. For Gradient Boosting and XGB select 3 lag features 1 week, 2 weeks, and 3 weeks ago (drop NaN) and use "month" of the date as a feature (along with the temperature statistics and the three lags). Also careful to not use any other feature. For Gradient Boosting use the same parameters we had in class. For XGB use 100 estimators, max depth of 3, min child weight 5.0, gamma of 5.0, and leave the other metrics what we had in class. Unlike SARIMAX which is retrained by sliding window, for GB and XGB we will simply use training data as dates of 2016-11-27 and prior, and 2016-11-28 and latter data to test. Obtain the performance of the model's predictions.

5. Create a new data frame with Boston's Total Volume of avocado. Then add columns of the Total Volume in Houston and Total Volume in San Diego. It is believed that it takes about 1 week for avocados to reach Boston from California and Mexico. So the Total Volume a week prior would be a good feature to use. Also, use lags of 1 week, 2 weeks, and 3 weeks as well as the Boston temperature statistics each week like we did before. Create three runs in XGB, (1) with the three lags, month, and the four temperature statistics (similar to what we did for Atlanta); (2) with only the three lags and month; (3) with the three lags, month, the four

temperature statistics, as well as the lagged Total Volume in San Diego and Houston. How do the models compare?