Multivariate Modeling

DATS 6450

Term Project

Obtain some real experimental (not simulated) time series dataset from a public data base. The dataset must have at least 500 samples with multiple features. Develop a linear model representation of the data using Python. You can use either the codes your developed in class or you can use any python packages.

SPECIFIES

The formal report must be typed and should contain the following sections:

1- **Abstract.**
2- **Introduction**. An overview of the multivariate modeling process and an outline of the report.
3- Description of the dataset. Describe the independent variable(s) and dependent variable:
    a. Plot of the dependent variable versus time.
    b. ACF of the dependent variable.
    c. Correlation Matrix with seaborn heatmap and pearson's correlation coefficient.
    d. Preprocessing procedures: Clean the dataset (no missing data or NAN)
    e. Split the dataset into train set (80%) and test set (20%).
    f. All the performance tests must be taken on the test set.

4- **Stationarity**: Check for a need to make the dependent variable stationary. If the dependent variable is not stationary, you need to use the techniques discussed in class to make it stationary. You need to make sure that ADF-test is not passed with 95% confidence.
5- **Time series Decomposition**: Find the best decomposition of the time series which you are working with. Refer to the lecture notes for different type of time series decomposition techniques.
6- **Holt-Winters method:** Using the Holt-Winters method try to find the best fit using the train dataset and make a prediction using the test set.
7- **Feature selection:** You need to have a section in your report that explains how the feature selection was performed. You must explain that which feature was eliminated and why.
8- Develop the **multiple linear regression** model that represent the dataset. Check the accuracy of the developed model.
    a. You need to include the complete regression analysis into your report.
    b. Hypothesis tests like F-test, t-test
    c. AIC, BIC, RMSE, R-squared and Adjusted R-squared
    d. ACF of residuals.
    e. Q-value
    f. Variance and mean of the residuals.

9- **ARMA model** order determination: Develop an ARMA model that represent the dataset.
    a. Preliminary model development procedures and results. (ARMA model order determination). Pick at least two orders using GPAC table.
    b. Should include discussion of the autocorrelation function and the GPAC. Include a plot of the autocorrelation function and the GPAC table within this section).

c. Include the GPAC table in your report and highlight the estimated order.

10- Estimate ARMA model parameters using the **Levenberg Marquardt algorithm**. Display the parameter estimates, the standard deviation of the parameter estimates and confidence intervals.

11- **Model Selection**: You need to derive at least 2 ARMA model and pick the best one. Make sure to include
    a. Diagnostic analysis using chi-square test.
    b. Pick the final model and simplify by :
        i. looking at the confidence interval of the estimated parameters. Make sure zero is not included inside the confidence interval.
    c. Check for zero/pole cancellation by looking the roots of the numerator and denominator.
    d. Display the estimated variance of the error and the estimated covariance of the estimated parameters.
    e. Is the derived model biased or this is an unbiased estimator?
    f. Check the variance of the residual errors.

12- **Final Model selection**: There should be a complete description of why your final model was chosen over another model (i.e. regression versus ARMA).

13- **Predictions**: Make sure to add the plot of the predicted values versus the true value (test set) and write down your observations.

14- **Summary and conclusion:** You should state any limitation of the final model and suggestions for other types of models that might improve performance. You need to compare the performance of various models developed for your dataset (Holt-Winters, Multiple Linear regression, ARMA) and come up with the best model for the given dataset.

15- A **separate appendix** should contain documented python codes that you developed for this project.

16- The **soft copy of your python programs** needs to be submitted to verify the results in the report. Make sure to include the dataset in your submission. Make sure to run your code before submission. If the python code generates an error message, 50% of the term project points will be forfeited.

17- Include a **readme.txt** file (if applicable) that explains how to run your python code.

18- The final report is due by **11:59pm Wednesday, April 22nd** .

19- The final presentation is due by **6pm Wednesday, April 22nd** . You will be given 10-15 minutes to present you term project to the class. The presentation weighs 20% of the term project.

Upload the **final report (as a single pdf**) plus **the .py file(s)** through BB by the due date.