Indian Institute of Technology Dharwad
CS214: Artificial Intelligence Laboratory

Lab 10: Autoregression

You are given a dataset containing details about new Covid-19 cases recorded in India on daily basis as a csv file (**daily_covid_cases.csv**). It shows the rolling 7-day average of newely confirmed cases starting from 30th-Jan-2020 to 2-Oct-2021. Rows are indexed with dates, first column represents the date and second column represents the new Covid-19 cases recorded that day. We will use this dataset to build an auto regression (AR) model..

1. Autocorrelation line plot with lagged values:

   (a) Create a line plot with the x-axis as index of the day and y-axis as the number of Covid-19 cases, as shown in Figure 1. Observe the first wave (around August-2020) and second wave (May-2021) of COVID-19 in India..

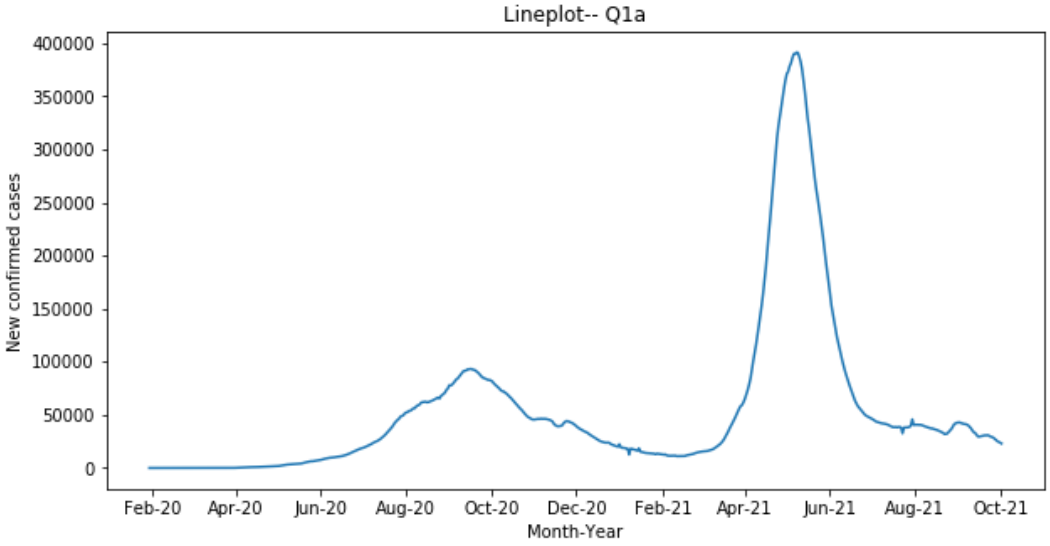

Figure 1: Line plot of new covid cases vs time

   (b) Generate another time sequence with one day lag to the given time sequence. Find the Pearson correlation (autocorrelation) coefficient between the generated one-day lag time sequence and the given time sequence.

   (c) Generate a scatter plot between given time sequence and one day lagged generated sequence in 1.(b). What do you infer regarding correlation? Does it match with the computed correlation coefficient in 1.(b)?

   (d) Plot a correlogram or Auto Correlation Function using python inbuilt function ***plot_acf*** up to 60 lag values. Observe the trend in the line plot with an increase in lagged values.

2. A general auto-regression (AR) model estimates the unknown data values as a linear combination of given lagged data values. For example, data value at $(t+1)$ instant, denoted by $x(t+1)$ can be estimated from its previous instance values, such as $x(t+1) = w_0 + w_1 * x(t) + w_2 * x(t-1) + ... + w_p * x(t-p+1)$. The coefficients $w_0, w_1, ...w_p$ can be estimated while training the auto-regression model on training dataset.

(a) Split the data into two parts. The initial 65% of the sequence for training data and the remaining 35% of the sequence as test data. (You may use slicing operation for the same to maintain the order of the sequence. Note that, you should not shuffle randomly.) This test set approximately covers the second wave of COVID-19. Plot the train and test sets. Generate an autoregression (AR) model using ***AutoReg()*** function from ***statsmodels*** library. This function generates an AR model with the specified training data and lagged values (given as its input). Use 5 lagged values as its input (p=5). Train/Fit the model onto the training dataset. Obtain the coefficients $(w_0, w_1, ..., w_p)$ from the trained AR model.

(b) Using these coefficients, predict the values (using the relation given above) for the test dataset. Note that, you have to make a 1-step ahead prediction each time. An example code snippet is given below (Code snippet to train AR model and predict using the coefficients).

i. Give a scatter plot between actual and predicted values.

ii. Give a line plot showing actual and predicted test values.

iii. Compute Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) between actual and predicted test data.

3. Generate AR models using ***AutoReg()*** function with lag values of 1, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55 and 60 days.

(a) Give a line plot showing actual and predicted test values with time.

(b) Compute the RMSE and MAPE between predicted and original test data values in each case. Give a bar chart showing RMSE on the y-axis and lagged values on the x-axis. Also, give a bar chart showing MAPE on the y-axis and lagged values on the x-axis. Infer the changes in RMSE and MAPE with changes in lagged values.

4. Compute the heuristic value for the optimal number of lags up to the condition on auto-correlation such that $\boldsymbol{abs(AutoCorrelation) > 2/sqrt(T)}$, where T is the number of observations in training data. Use it as input in *AutoReg()* function to predict the new COVID-19 cases on daily basis and compute the RMSE and MAPE value. Compare this result with that of Question 3.

Functions/Code Snippets:

```
from statsmodels.tsa.ar_model import AutoReg as AR

# Code snippet to train AR model and predict using the coefficients for lag=5.
# Train test split
train_size = int(len(df) * 0.65)
train, test = df.iloc[:train_size], df.iloc[train_size:]

### Train Autoregression model with lag=5
lag = 5
model = AR(train["new_cases"], lags=lag).fit()
# Get the coefficients of AR model
coef = model.params
print(f"AR(5) Model Coefficients: {coef}")
```

```
# To get the previous lag value from training data to predict the test value
history = list(train["new_cases"].values)
# List to hold the predictions, 1 step at a time
predictions = []
for t in range(len(test)):
    lag_values = history[-lag:]
    # Compute, x(t + 1) = w0 + w1  x(t) + w2  x(t1) + ... + wp  x(tp+1).
    yhat = coef[0] + sum(coef[i+1] * lag_values[lag-i-1] for i in range(lag))
    # Append predictions to compute RMSE later and show the predictions
    predictions.append(yhat)
    # Append actual test value to history, to be used in next step.
    history.append(test["new_cases"].iloc[t])
```

Note

Please upload the completed Jupyter Notebook. Copy all Jupyter Notebook and data files in a folder and rename the folder as ***your_rollnumber_Lab10***, then create the zip file with the name ***your_rollnumber_Lab10.zip***. Finally, upload the zip file on the Moodle for evaluation.