G5 End Term Report

So far in the mid-term report we designed a RL based algorithm based on the actor critic model in which we fed up the NIFTY50 data and calculated the reward of the portfolio based upon the action space created by us.

Now for the End Term out of the ARIMA and the LSTM I chose the ARIMA model to work on as this is quite common and the resources are widely available.

ARIMA Model

Autoregressive Integrated Moving Averages

So basically, ARIMA model is a statistical analysis model that uses time series data to either better understand the data set or to predict future trends.

This is termed as the regressive as it predicts the future values based upon the past ones! Assuming that the changing variables are dependent on the others.

So, basically the ARIMA model is compromised of the three parts viz;

AR : Automatic regressive or the auto regressive. This refers to a model that shows a changing variable that regresses on its own lagged, or prior, values.

MA : The moving averages which brings the dependency between an observation and a residual error from a moving average model applied to lagged observations.

I : The middle one which is the integration of the both. This basically works on the differences in the raw observations in order to make it stationary.

To understand the ARIMA in the common layman terms let’s take the example the of the number of the light bulbs sold or the manufactured!

L sub t or the number of light bulbs that is going to be created this month is going to be equal to some coefficient beta sub 0 this is just a constant. That's the constant beta sub 1 is a different coefficient and then it gets interesting which is L sub t minus 1 which is the number of light bulbs created last month so this is the autoregressive bit which is if we were to just stop here then this would be a AR 1 model; because it basically means that how many light bulbs is needed to be created this month is a function of how many light bulbs were created last month. But of course we also have this ma one bit which is this part here which says that not only is it a function of the number of light bulbs created last year. It's also a function of this coefficient this V sub one is a coefficient and

it's a function of epsilon sub T minus one which is the error from the previous time period from last month.

$L_t = \beta_0 + \beta_1 L_{t-1} + \varnothing_1 \varepsilon_{t-1} + \varepsilon_t$

This basically says that last month the prediction about how many light bulbs to. Whether the prediction to this was positive or negative this error is epsilon sub t minus one is how much it was off by in the previous period. So the error is incorporated here that how much it was wronged by into the new prediction for this month and there's a little bit distinction to be made. Here this would be the process itself and the epsilon sub T is there at the end so this is the error from this month.

$\hat{L}_t = \beta_0 + \beta_1 L_{t-1} + \varnothing_1 \varepsilon_{t-1}$

L sub T hat as we remember in statistics or time series L sub T hat or anything hat is the pur predicted value whose real process is given by the thing above but of course the real process is unknown because if it was predicted the error from this month then with all the information the perfect prediction can be made which is obviously not true.

So the predicted value for light bulbs created this month is going to be equal to this coefficient beta sub naught plus the coefficient beta sub 1 L sub t minus 1. The exact number of light bulbs needed last month is present because last month is in the past and the past knowledge is with us plus V sub 1 epsilon sub T minus 1. We also have access to the error from last month because it's past knowledge but we don't have access to the error from this month because it hasn't happened yet so our prediction would basically be given by this function right here.

The differenced series is the change between consecutive observations in the original series, and can be written as $y'_t = y_t - y_t{-}1$.

A closely related model allows the differences to have a non-zero mean. Then

$y_t - y_t{-}1 = c + \varepsilon_t$ or $y_t = c + y_t{-}1 + \varepsilon_t$.

There are various types of the differencing too like the seasonal, second order.

A moving average model do not uses the past values of the forecast variable in a regression rather it uses the errors in the values that is the differences in the predicted and the real values.

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$, where ε_t is white noise.

The ARIMA model is a stochastic process defined by three parameters, p, d, and q, where p stands for the Auto-Regressive AR(p) process, d is the integration (needed for the

transformation into a stationary stochastic process), and q is the Moving Average MA(q) process.

In ARIMA model the standard notation is ARIMA with p, d, and q, where integer values substitute for the parameters to indicate the type of ARIMA model used. These parameters are:

p: this is the number of lag observations in the model.

d: the number of times the raw observations are differenced, that is for the removal of the seasonality!

q: the size or the order of the moving average window.

The main part of the ARIMA model combines AR and MA polynomials into a complex polynomial, as seen in below. The ARIMA (p, d, q) model is applied to all the data points of the total cost data.

The value of the ARIMA parameters (p, d, q) for AR and MA can be obtained from the behaviour of the Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF). These functions help to estimate the parameters that can be used to forecast data using the ARIMA model.

Stationary Series

A Stationary series is one whose statistical properties such as mean, variance, covariance, and standard deviation do not vary with time, or these stats properties are not a function of time. That is, stationarity in Time Series also means series without having any Trend or Seasonal components. I the series is stationary then the prediction of the trend by the statistical models is quite easier and also up to the mark.

There are various types of the stationarity like the season, trend and the strict one each being stationary in different terms.

Here in order to remove the stationarity from the data to be used in the ARIMA model I will be using the Dickey Fuller Test which is a type of the unit root test.

A unit root is a feature of some stochastic processes that can cause problems in statistical inference involving time series models. In simple terms, the unit root is non-stationary but does not always have a trend component.

Dickey-Fuller Test

It uses an autoregressive model and optimizes an information criterion across multiple different lag values. A Dickey-Fuller test is a unit root test that tests the null hypothesis that $\alpha=1$ in the following model equation. alpha is the coefficient of the first lag on Y.

Null Hypothesis (H0): alpha=1

where,

y(t-1) = lag 1 of time series

delta Y(t-1) = first difference of the series at the time (t-1).

It has a similar null hypothesis as the unit root test. That is, the coefficient of Y(t-1) is 1, implying the presence of a unit root. If not rejected, the series is taken to be non-stationary. After the ARIMA model we have calculated the root mean squared error and the mean absolute error value.