## Time Series

***DAY 1***

What is Time Series:

Data that changes with time or dependent on time or indexed on time is time series data.

Applications of time series are:

1. Prediction of the stock market.
2. Predicting traffic condition on any road with respect to time

Major areas of study or aspects in Time Series:

1. Understand the data: Analyzing the data of any company and answering questions asked by the client. Analyzing how data is behaving with respect to time.
2. Forecasting:Making future predictions based on the past data.

Major difference between Machine Learning and Time Series

* In ML we deal with two types of variables. Y, dependent variables and x, independent variables. ML work upon y vs x

Eg: Given height we try to predict weight of a person

* In time series we have only x in the given data. Y is related to time. At every step we try to find a relation between y vs time.

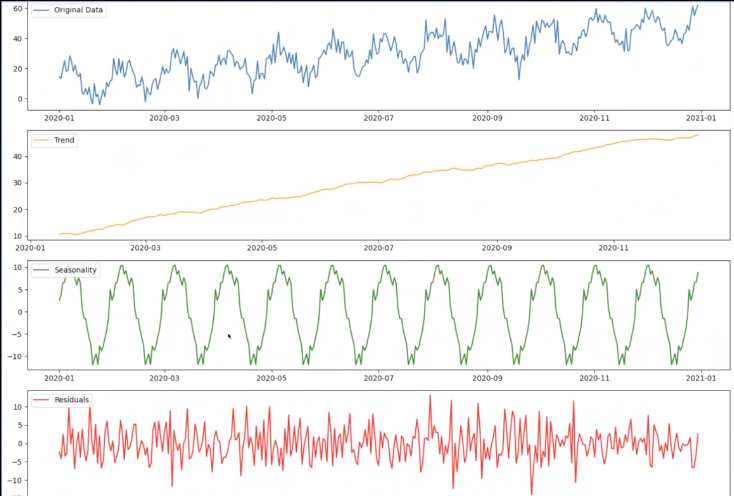
Eg: We try to predict weight of a person based on the age

How to understand the data in time series:

There are 3 major components for data in time series: Trend, Seasonality and Noise.

* Trend: Behavior or pattern of data on a larger time frame
* Seasonality: Behavior of data in a specific time frame. Repeating pattern in data during a regular interval of time.
* Noise: Abnormal behavior in data. It is also known as Residuals or Errors.

Actual data always comprises Trend, Seasonality and Noise.



Scale for trend is always equal to the scale of original data and the scale of seasonality and noise will be less than the scale of original data.

Processes for Forecasting:

1. AR - Auto Regression
2. MA - Moving Average
3. AR + MA = ARMA
4. ARIMA - Auto Regressive Integrated Moving Average
5. Exponential Techniques

Basic techniques for Forecasting:

1. Taking simple average of the data to make future prediction
2. Consider the value of the date close to the prediction date

Stationarity: If distribution of data across time is not changing then we can say its Stationary data. The forecasting techniques need the data to be stationary.

To check if data is stationary or not, we can use the Augmented Dickey Fuller (ADF) Test. If data is not stationary then we can try to convert it to stationary or use exponential techniques.

***DAY 2***

Simple ways to check stationarity of data:

1. Take two random time intervals and compare the means. If means are different then they are non stationary.
2. Global mean vs local mean. If they are in sync then data is stationary
3. Plot the graph
4. ADF

Augmented Dickey Fuller Test:

For ADF we do Hypothesis testing.

Null Hypothesis (H0): Data is Non-Stationary

Alternate Hypothesis (H1): Data is Stationary

P-value < 0.05 -> we reject null hypothesis. So, if p-value < 0.05, Data is Stationary

P-value > 0.05 -> We accept the null hypothesis. So, p-value > 0.05, Data is non-stationary

If data is stationary, we use AR, MA, ARIMA, etc. If not we try to convert data to stationary.

Techniques to Converting data to stationary:

1. Instead of using the data directly, use the differences of the data.

For example,

* if we have data points as D1, D2, D3, D4, D5, D6, D7, D8, D9, D10. Then we take the differences as: y1 = D2-D1, y2 = D3-D2, y3 = D4-D3, y4 = D5-D4, y5 = D6-D5, y6 = D7-D6, ……… This is called data differentiation. This will make the data stationary.
* If data points are like, 1,2,1,2,1,2,1,2,1,2,1,2. Here once step difference will give 1, -1, 1, -1, 1, -1, …. But this is not going to work out. So we go for a two step difference, that difference of the alternate data points, this gives, 0,0,0,0…. Which will be stationary.
* If data points are like 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024.

Here we apply log, then we get 1,2,3,4,5,6,..... Then we do one step difference to make it stationary.

If you see data is rapidly growing, then we can go for logarithmic differences

Techniques for converting to stationary:

1. Yn - Yn-1, Yn - Yn-2, …. -> Lag Difference
2. logYn - logYn-1, logYn - logYn-2, …… -> Log Difference

***DAY 3:***

Date column should be the index column as it will help us in plotting the data for visualization.

**AR - Auto Regression:**

The data on any particular day is dependent on yesterdays data.

This is one level of dependency.

Dt → Dt-1 == AR1. This is also known as AR model with lag of 1

If data on a particular data is dependent on yesterday’s and day before yesterday’s data then its called two level of dependency

Dt → Dt-1 + Dt-2 == AR2. This is also known as AR model with lag of 2

Similarly we can have 3 level of dependency. == AR3. This is also known as AR model with lag of 3

In this way we can have AR1, AR2, AR3, ……., ARn.

To figure out which AR model is best we use ACF (Autocorrelation Function) or PACF (Partial Autocorrelation Function)

Correlation can be positive correlation, negative correlation or no correlation.

In regression there is a relation between y vs x but in time series there is no x. We have a relation between Yt (current time) vs Yt-1(previous time).

Yt vs Yt-1 = ACF1. Yt vs Yt-2 = ACF2, …., Yt vs Yt-100 = ACF100

**Mathematical Representation of AR models**:

AR(1) can be written as Yt = mYt-1 +C +Et

AR(2) can be represented as Yt = mYt-1 + mYt-2 + C +Et

**Auto Correlation Function** tries to figure out the escalation between current data vs previous data. Autocorrelation measures the relationship between a variable's current value and its past values.

But there is a problem in this; If Yt depends on Yt-1 and Yt-1 depends on Yt-2, then Yt indirectly depends on Yt-2. There was an indirect relation between Yt and Yt-2. This means Yt is indirectly dependent on every past data.To correct this, PCAF was designed.

In **Partial Autocorrelation** we remove the impact of the middle variable.

Let's say Yt depends on Yt-1 = Yt → Yt-1 and Yt-1 depends on Yt-2 = Yt-1 → Yt-2. Here Yt-1 is the middle variable, which should be deducted from Yt and Yt-2.

So we consider the relation between [Yt - Yt-1] and [Yt-2 - Yt-1]. After doing this if we find out that they are still related to each other then we get the PACF for these data points.

**MA model - Moving Average:**

Moving Average says current data only depends on external factors.

For example: Current body weight can depend on external factors such as food intake.

In Agriculture, the production depends on weather conditions (external factors) not the previous data.

Lets say Yt, Yt-1, Yt-2, ……, Yt-n is a time series. MA model says each data point depends on external factors such as Et, Et-1, Et-2, ……Et-n

MA(1) model can be written as Yt = mEt-1 + C +Et

MA(2) model can be written as Yt = m1Et-1 + m2Et-2 + C + Et

**ARMA:** It says current data depends on both previous data and external factors

Example, current body weight depends on yesterday's body weight and food intake

ARMA model can mathematically represented as:

ARMA(1) model: Yt = m1Yt-1 + p1Et-1 +C +Et

***DAY 4:***

AR model => Future date depends on past data

MA => Future data depends on external factors

ARMA => Future data depends on past + external factors

ACF => It captures the direct and indirect relationship between data points

PACF => It gets rid of all the indirect relationships and focuses on direct relationships only.

ARMA(p,q) => This means P number of components are considered in AR model and q number of components are considered in MA model.

Future data depends on P many past data and q many external factors

ARMA(1,1) => AR(1) + MA(1)

=> Yt = m1Yt-1 + p1Et-1 + C + Et

ARIMA : Autoregressive Integrated Moving Average

It processes the data ( it takes the difference factor) to make it stationary and then applies ARMA. ARIMA is not a different model. It helps us to convert non stationary data into stationary data by taking ‘d’ order difference.

ARIMA(p,q, d) here d is a different factor. WE try out hyper parameter tuning of Grid Search to figure out the appropriate value of d.

For AR, MA, ARMA cannot handle non stationary data but ARIMA can take non stationary data provided the difference factor is mentioned.

***DAY 5:***

Exponential model: It can be used for stationary and non stationary data. But when data is stationary, it gives a better performance compared to when the data is non stationary.

* Simple exponential: Used when data has no trend or seasonality
* Double exponential: Used when trend is there but no seasonality
* Triple exponential: Used when there is both trend and seasonality

Simple Exponential:

Consider data for 180 days, from Jan to June temperature data

D1, D2, ….., D180.

Y1, Y2, ……, Y180.

Ask: Predict the temperature for 1st July as we have data up to June 30.

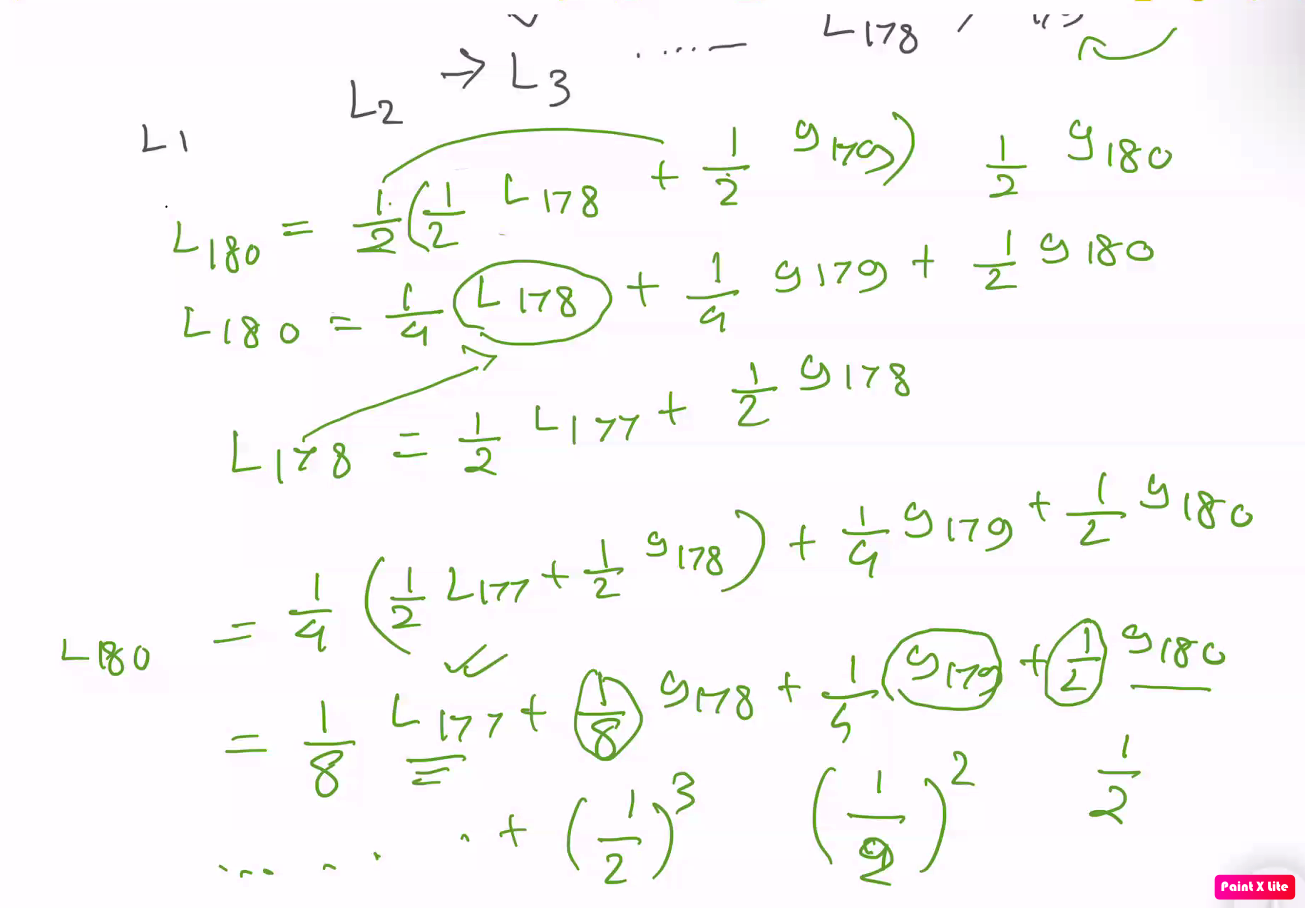
Solution options: Focus more on recent data and reduce the focus as you go backwards.

To predict 1st July temp, focus most on 30 June, then little less on 29 June, then a little more less on 28 June and so on. So, the focus is reduced exponentially as we go backwards.

Global average: Avg of all the data

Local average: Avg of nearby data. -> local avg of previous day + temp of the current day

On 30 June, the local avg (L180) = ½ (L179) + ½ (Y180)

Local avg for June 29 (L179) = ½ (L178) + ½ (Y179)

The prediction of the temp of July 1st will be the local average of June 30.

Simple exponential: All the future predictions will be the latest local average. And its only applicable for data that does not have any trend or seasonality.

Double Exponential: This model, data has trend but no seasonality. For this model, we need the local average and the trend (T1,T2, ……, Tn)

Local avg : Ln = ½ Ln-1 + ½ Yn

In double exp we add the trend to the local avg, now the local avg will be calculated as:

Ln = ½ (Ln-1 + Tn-1) + ½ Yn

Calculate trend with Tn = ½ Tn-1 + ½ (Ln - Ln-1)

To make the prediction for the next day it will be Ln + Tn

Triple Exponential Model: Data should have both trend and seasonality

Consider the data as price of AC for 10 years

Y1, Y2, ……., Y120

We can convert triple exponential data to double exp, we can get rid of seasonality. To remove the seasonality from the data, we subtract the seasonality from the original data value.

Y1 - S1, Y2 - S2, ……., Y120 - S120

Now the formula for local average will be the same as the double exp.

L120 = ½ (L119 + T119) + ½ (Y120 - S108) Seasonality is an annual thing so seasonality of previous year will be S120-12 = S108.

T120 = ½ (T119) - ½ (L120 - L119)

S120 = ½ (S120-12) + ½ (Y120 - L120)

= ½(S108) + ½ (Y120 - L120)

Given the data, we check all three models and see whichever model works the best.

***DAY 6:***

Simple exp: Ln = aLn-1 + (1-a)Yn ; a is a hyperparameter it's ½ in the above examples but it can be any value of our choice

Double exp: Ln = a(Ln-1 + Tn-1) + (1-a)Yn; a is a hyperparameter.