SOMAIYA VIDYAVIHAR UNIVERSITY
K J Somaiya College of Engineering

Somaiya TRUST

Unit 4:
Time series Analytics and Forecasting

Kaustubh Kulkarni
Assistant Professor,
Department of Computer Engineering,
KJSCE, SVU, Mumbai.

Definition

- Time series analysis is the endeavor of extracting meaningful summary and statistical information from points arranged in chronological order.
- It is done to diagnose past behavior as well as to predict future behavior.

Where to Find Time Series Data

- Prepared Data Sets
- The UCI Machine Learning Repository .
  - hourly air quality samples in an Italian city
  - Amazon file access logs
  - diabetes patients' records of activity, food, and blood glucose information
- The UEA and UCR Time Series Classification Repository
  - yoga movement classification dataset
  - Wine spectral analysis dataset
- Government time series data sets
  - The Open Government Data (OGD) Platform India is a single-point of access to datasets published by Ministries/Departments in India.
  - NOAA National Centers for Environmental Information, USA

Where to Find Time Series Data

- Found Time Series : time series data we put together ourselves from data sources in the wild.
- Finding time series data in structured data not explicitly stored as a time series can be easy in the sense that timestamping is ubiquitous.
- Here are a few examples of where you'll see timestamps in your database:
  - Timestamped recordings of events
  - "Timeless" measurements where another measurement substitutes for time
    - For example, if you have a dataset of customer transactions, you can use the order number to represent the time at which each transaction took place.

Cleaning Your Data

- Handling missing data
- Changing the frequency of a time series (that is, upsampling and downsampling)
- Smoothing data
- Addressing seasonality in data
- Preventing unintentional lookaheads

Handling Missing Data

- Missing data is surprisingly common. For example, in healthcare, missing data in medical time series can have a number of causes:
  - The patient didn't comply with a desired measurement.
  - The patient's health stats were in good shape, so there was no need to take a particular measurement.
  - The patient was forgotten or undertreated.
  - A medical device had a random technical malfunction.
  - There was a data entry error

Handling Missing Data

The most common methods to address missing data in time series are:

- Imputation
  - When we fill in missing data based on observations about the entire data set.
  - Forward fill, backward fill, moving average
- Interpolation
  - When we use neighboring data points to estimate the missing value.

    Interpolation can also be a form of imputation.
- Deletion of affected time periods
  - When we choose not to use time periods that have missing data at all.

Upsampling and Downsampling

- Often, related time series data from different sources will not have the same sampling frequency.
- This is one reason, among many, that you might wish to change the sampling frequency of your data.
- Of course you cannot change the actual rate at which information was measured, but you can change the frequency of the timestamps in your data collection.
- This is called upsampling and downsampling, for increasing or decreasing the timestamp frequency, respectively.

When to downsample

The original resolution of the data isn't sensible.

- There can be many reasons that the original granularity of the data isn't sensible.
- For example, you may be measuring something too often.
- Suppose you have a data set where someone had measured the outside air temperature every second.
- Common experience dictates that this measurement is unduly frequent and likely offers very little new information relative to the additional data storage and processing burden.
- In fact, it's likely the measurement error could be as large as the second-to-second air temperature variation.
- So, you likely don't want to store such excessive and uninformative data.
- In this case—that is for regularly sampled data—downsampling is as simple as selecting out every nth element.

When to downsample

Focus on a particular portion of a seasonal cycle.

- Instead of worrying about seasonal data in a time series, you might choose to create a subseries focusing on only one season.
- For example, we can apply downsampling to create a subseries, as in this case, where we generate a time series of January measurements out of what was originally a monthly time series.
- In the process, we have downsampled the data to a yearly frequency.

When to downsample

Match against data at a lower frequency.

- You may want to downsample data so that you can match it with other low-frequency data.
- In such cases you likely want to aggregate the data or downsample rather than simply dropping points.
- This can be something simple like a mean or a sum, or something more complicated like a weighted mean, with later values given more weight.
- For example in the donation data the idea of summing all donations over a single week is more appropriate, since it was the total amount donated that was likely to be most interesting.
- In contrast, for our economic data, what is most likely to be interesting is the yearly average.

When to upsample

Irregular time series.

- A very common reason to upsample is that you have an irregularly sampled time series and you want to convert it to a regularly timed one.
- This is a form of upsampling because you are converting all the data to a frequency that is likely higher than indicated by the lags between your data.

When to upsample

Inputs sampled at different frequencies.

- Sometimes you need to upsample low frequency information simply to carry it forward with your higher-frequency information in a model that requires your inputs to be aligned and sampled simultaneously.
- You must be vigilant with respect to lookaheads (predictions), but if we assume that known states are true until a new known state comes into the picture, we can safely upsample and carry our data forward.
- For example, suppose we know it's (relatively) true that most new jobs start on the first of the month.
- We might decide we feel comfortable using the unemployment rate for a given month indicated by the jobs report for the entire month (not considering it a lookahead because we make the assumption that the unemployment rate stays steady for the month).

When to upsample

Knowledge of time series dynamics.

- If you have underlying knowledge of the usual temporal behavior of a variable, you may also be able to treat an upsampling problem as a missing data problem.
- In that case, all the techniques we've discussed already still apply.
- An interpolation is the most likely way to produce new data points, but you would need to be sure the dynamics of your system could justify your interpolation decision.

Smoothing Data

Purposes of smoothing

- You can remove outliers with a moving average to eliminate measurement spikes, errors of measurement, or both.
  - Even if the spikes indeed exist in the data, they may not reflect the underlying process and may be more a matter of instrumentation problems; this is why it's quite common to smooth data.
- Smoothing data is strongly related to imputing missing data, and so some of those techniques are relevant here as well.
  - For example, you can smooth data by applying a rolling mean, with or without a lookahead, as that is simply a matter of the point's position relative to the window used to calculate its smoothed value.

Purpose of smoothing (cont'd)

Feature generation

- It is the practice of taking a sample of data, like characteristics about a person, image, or anything else, and summarizing it with a few metrics.
- In his way a fuller sample is collapsed along a few dimensions or down to a few traits.
- Feature generation is especially important for machine learning.
- Smoothing helps in accurate feature generation.

Purpose of smoothing (cont'd)

Prediction

- The simplest form of prediction for some kinds of processes is mean reversion, which you get by making predictions from a smoothed feature.

Visualization

- Do you want to add some signal to what seems like a noisy scatter plot? If so, what is your intention in doing so? How will your outcomes be affected by smoothing or not smoothing?
- Does your model assume noisy and uncorrelated data, whereby your smoothing could compromise this assumption?
- Will you need to smooth in a live production model?
- If so, you need to choose a smoothing method that does not employ a lookahead.
- Do you have a principled way to smooth, or will you simply do a hyperparameter grid search?If the latter, how will you make sure that you use a time-aware form of cross-validation such that future data does not leak backward in time?

Exponential smoothing

- You often won't want to treat all time points equally when smoothing.
- In particular, you may want to treat more recent data as more informative data, in which case exponential smoothing is a good option.
- Exponential smoothing is temporally aware, weighting more recent points higher than less recent points.
- So, for a given window, the nearest point in time is weighted most heavily and each point earlier in time is weighted exponentially less (hence the name).

Questions?