Data Stream Mining- Lecture 10

Streaming Time Series Mining

Chandresh Kumar Maurya, Research Assistant Professor

Eötvös Loránd University, Budapest, Hungary

November 25, 2019

Time Series

Definition

A time series is defined as sequence of data points which have some notion of time or put another way, a non-random sequence which are equally spaced in time.

Note: In time series, time is an independent variable.

Time Series

Definition

A time series is defined as sequence of data points which have some notion of time or put another way, a non-random sequence which are equally spaced in time.

Note: In time series, time is an independent variable. Examples:

- Electricity demand over a period of time
- customer buying behavior over time
- browsing over www
- Stock market fluctuations over the day.
- Any other data which has a notion of time.

An example of time series data

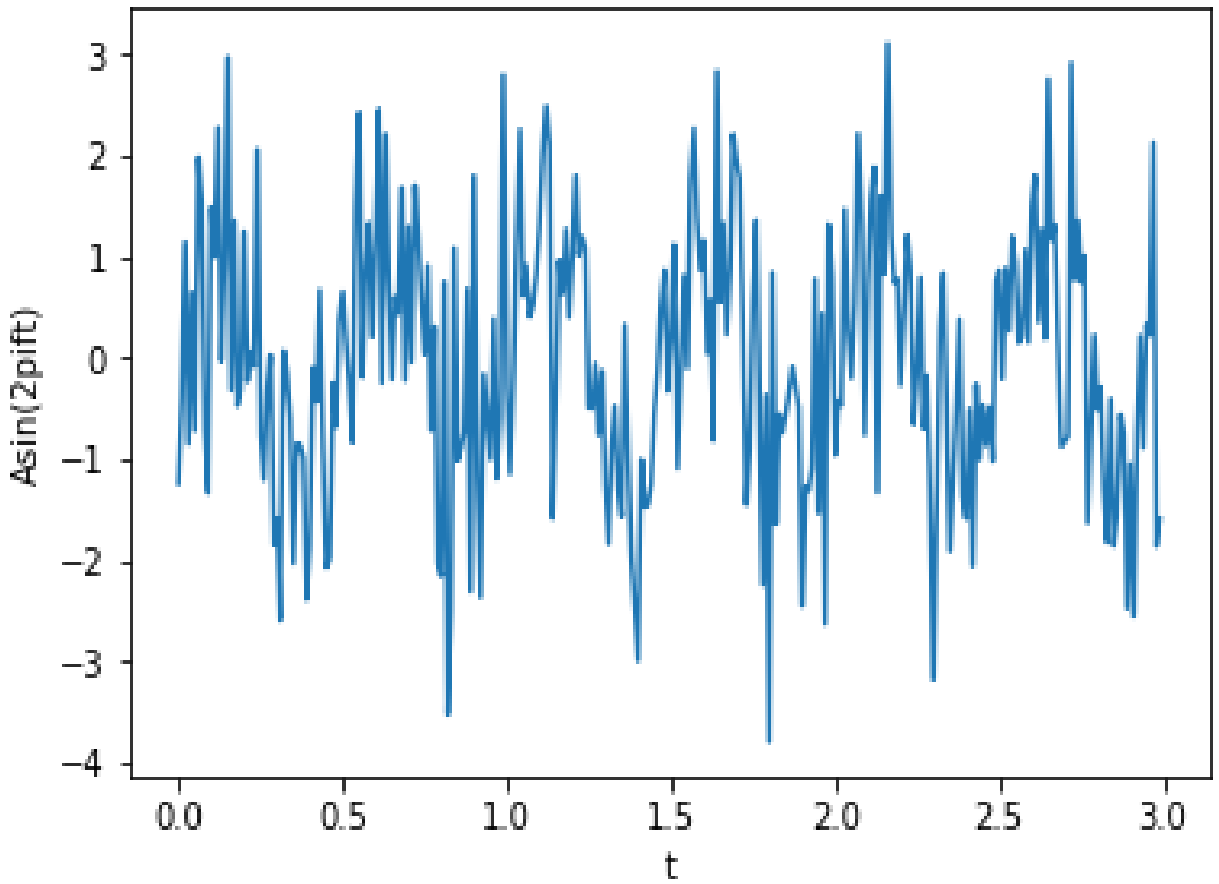

Figure: An example of time series (sinusoidal signal $2sin(2\pi * 2 * t) + np.random.rand(t)$

Time series Analysis and Forecasting

Definition

Time series analysis comprises methods for analyzing time series data for extracting meaningful insights.

Time series Analysis and Forecasting

Definition

Time series analysis comprises methods for analyzing time series data for extracting meaningful insights.

Definition

Time series forecasting is prediction at future time points using a model learned from previously seen data.

Time series Analysis and Forecasting

Definition

Time series analysis comprises methods for analyzing time series data for extracting meaningful insights.

Definition

Time series forecasting is prediction at future time points using a model learned from previously seen data.

Time series tasks

Time series tasks include the following:-

- Time series forecasting
- Time series similarity
- Time series classification
- Time series clustering and so on

Time series similarity

Definition

Time series similarity problem is as follows: Given a query time series Q and a similarity measure $d(\cdot, \cdot)$, find the most similar time series in a given database.

This is also known as Indexing or Query by Content.

Why study time series similarity?

Time series similarity

Definition

Time series similarity problem is as follows: Given a query time series Q and a similarity measure $d(\cdot, \cdot)$, find the most similar time series in a given database.

This is also known as Indexing or Query by Content.

Why study time series similarity?

Because it is a pre-step in other time series mining tasks such as clustering, classification etc.

Two distance measures

- Euclidean distance
- Dynamic Time Warping (DTW)

Euclidean distance between two time-series

Given two time series $P = (p_1, p_2, \ldots, p_n)$ and $Q = (q_1, q_2, \ldots, q_n)$. The Euclidean distance between them is:

$$d(P, Q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)}$$

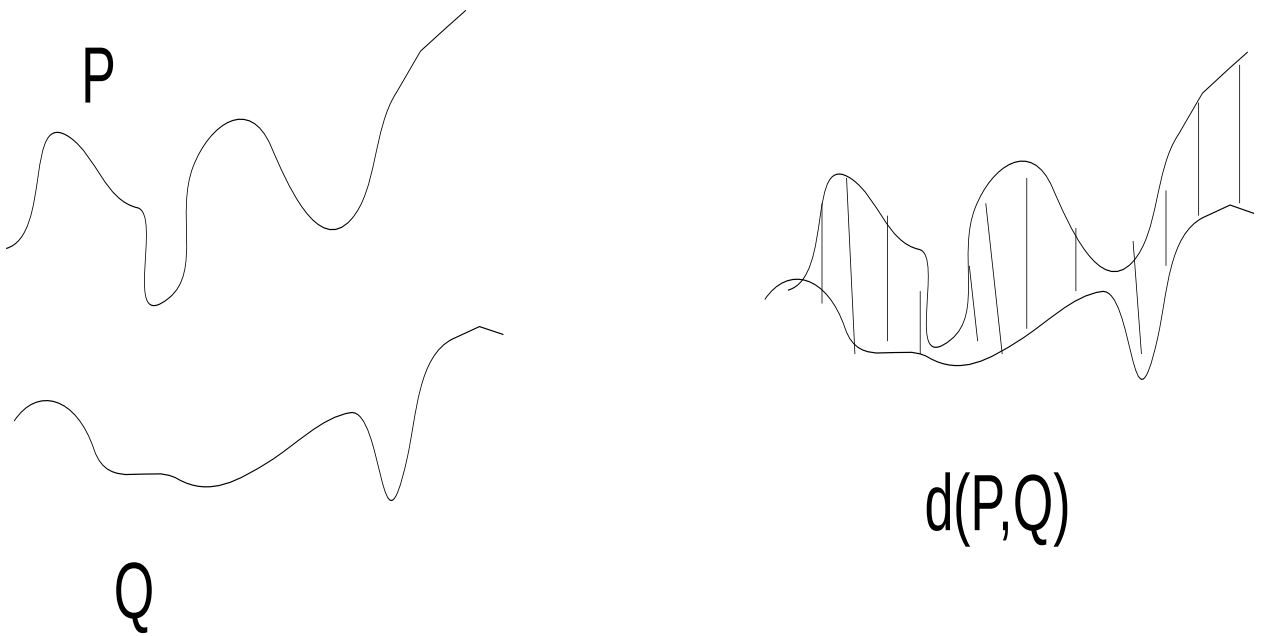

Figure: Euclidean distance between two time series

Contd...

What is the problem of Euclidan distance?

Contd...

What is the problem of Euclidan distance?
Not suitable for time series shifted in time domain.

Contd...

What is the problem of Euclidan distance?
Not suitable for time series shifted in time domain.
Solution: Use DTW

Dynamic Time Warping (DTW)

1. An algorithm to find optimal alignment between two time series.
2. Time series may vary in length in time or speed.
3. It is optimal in the sense that it optimizes the Euclidean distance.
4. It optimizes the Euclidean distance using dynamic programming.

Working of DTW Algorithm [Salvador and Chan, 2007]

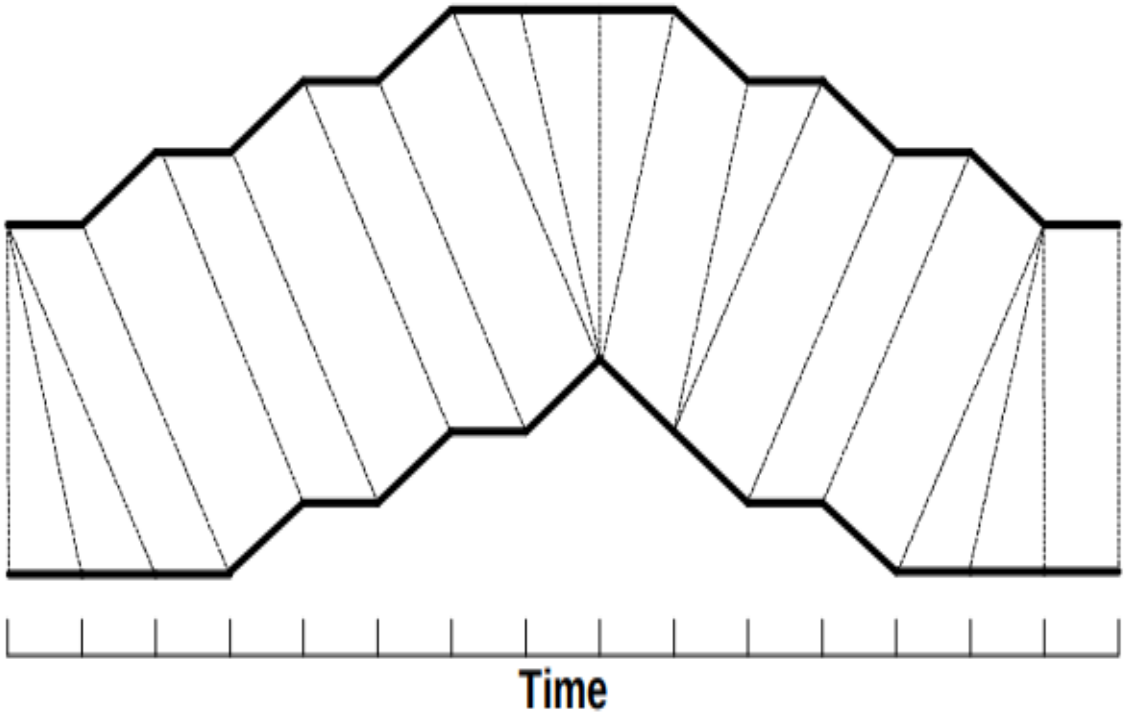

Figure: Warping between two time series

Contd...

Intuitively, the algorithm proceeds as follows:

1. Take first time series as reference time series.
2. Pick the 1^{st} point in the reference times series and calculate its distance wrt to all the points in the second time series. Store the min. distance
3. Now, pick the 2^{nd} point in the reference time series and compute its distance wrt to all the points in the second time series. Store the min. distance
4. Keep repeating the above steps until all the points in the reference time series are exhausted.
5. Now, take the second time series as reference time series and repeat the above steps.
6. Finally, add the minimum distances together to find the warp distance between the two time series.

Contd...

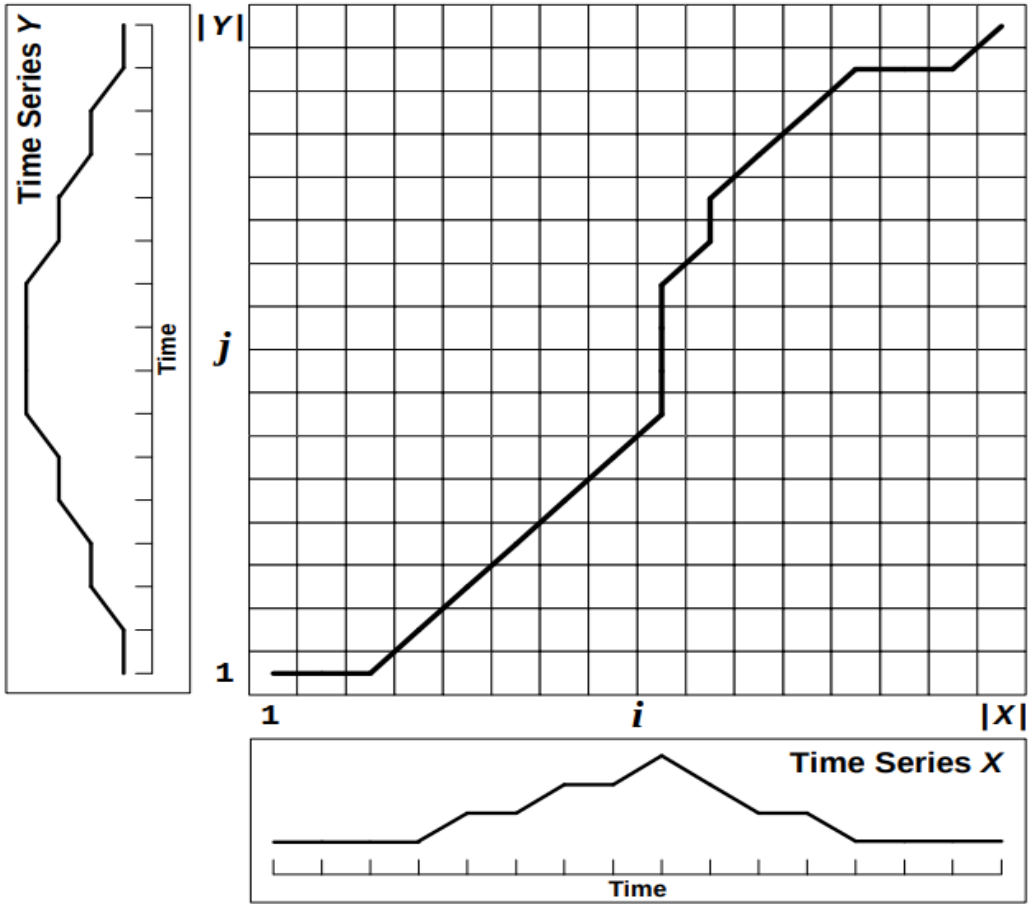

Figure: A cost matrix with the minimum-distance warp path traced through it.

Formal Definition of DTW

DTW is defined as follows. Given two time series $P = (p_1, p_2, \ldots, p_n)$ and $Q = (q_1, q_2, \ldots, q_m)$ of lengths $|P|$ and $|Q|$. Construct a warp path $W = (w_1, w_2, \ldots, w_K)$ such that

$$\max(|P|, |Q|) \leq K < |P| + |Q| \tag{1}$$

Where K the warp path length and each element of the warp path W is found as:

$$w_k = (i, j) \tag{2}$$

where i and j are indices in the time series P and Q respectively. Further, $w_1 = (1, 1)$ and $w_K = (|p|, |Q|)$. Constraints on i and j are:

$$w_k = (i, j), w_{k+1} = (i', j'), \quad i \leq i' \leq i' + 1, j \leq j' \leq j' + 1. \tag{3}$$

The DTW optimization problem

The optimal warp path is the warp path with minimum distance.

$$d(W) = \sum_{k=1}^{K} d(w_{ki}, w_{kj}) \tag{4}$$

Where in the above eq., on the left is the warp path distance and on the right is the warp path distance between two data point indexes-(k, i) and (k, j).

The DTW optimization problem

The optimal warp path is the warp path with minimum distance.

$$d(W) = \sum_{k=1}^{K} d(w_{ki}, w_{kj}) \tag{4}$$

Where in the above eq., on the left is the warp path distance and on the right is the warp path distance between two data point indexes-(k, i) and (k, j). The solution by DP is to fill the cost matrix D using the formula:

$$D(i,j) = d(i,j) + \min(D(i-1,j), D(i,j-1), D(i-1,j-1)) \tag{5}$$

The DTW optimization problem

The optimal warp path is the warp path with minimum distance.

$$d(W) = \sum_{k=1}^{K} d(w_{ki}, w_{kj}) \tag{4}$$

Where in the above eq., on the left is the warp path distance and on the right is the warp path distance between two data point indexes-(k, i) and (k, j). The solution by DP is to fill the cost matrix D using the formula:

$$D(i,j) = d(i,j) + \min(D(i-1,j), D(i,j-1), D(i-1,j-1)) \tag{5}$$

Ex. The time complexity of the solution by DP is $O(n^2)$.

Optimizing DTW: FastDTW

How to make DTW fast?

Optimizing DTW: FastDTW

How to make DTW fast?

- **Constraints** – Limit the number of cells that are evaluated in the cost matrix.
- **Data Abstraction** – Perform DTW on a reduced representation of the data.
- **Indexing** – Use lower bounding functions to reduce the number of times DTW must be run during time series classification or clustering.

FastDTW: The main idea

Combines the best of two worlds: constraints and data abstraction. FastDTW has 3 main steps:

- **Coarsening** – Shrink a time series into a smaller time series that represents the same curve as accurately as possible with fewer data points.
- **Projection** – Find a minimum-distance warp path at a lower resolution, and use that warp path as an initial guess for a higher resolution's minimum-distance warp path.
- **Refinement** Refine the warp path projected from a lower resolution through local adjustments of the warp path.

FastDTW has time complexity of $O(n)$.

Contd...

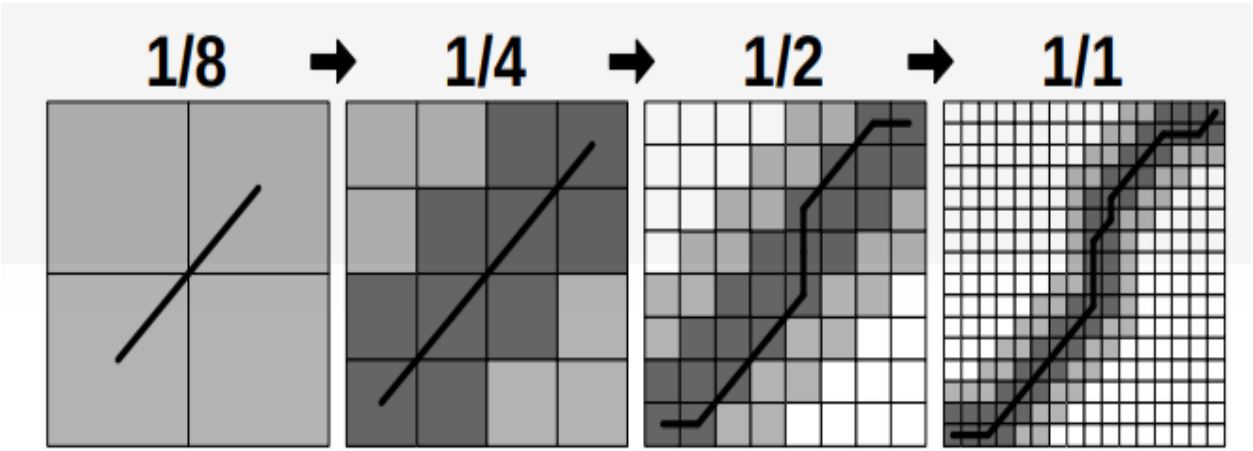

Figure: The four different resolutions evaluated during a complete run of the FastDTW algorithm.

Bibliography I

Salvador, S. and Chan, P. (2007).
Toward accurate dynamic time warping in linear time and space.
Intelligent Data Analysis, 11(5):561–580.