STEVENS INSTITUTE of TECHNOLOGY
School of Business

FE 582

Lecture 4: Similarity and Distances

Dragos Bozdog

Similarity and Distances

Formal statement of similarity or distance quantification problem:

Given two objects O_1 and O_2, determine a value of the similarity Similarity(O_1,O_2) (or distance Distance(O_1,O_2)) between the two objects.

Observations:

- In similarity functions, larger values imply greater similarity
- In distance functions, smaller values imply greater similarity

Multidimensional Quantitative Data

The L_p-norm between two data points $\bar{X} = (x_1 \cdots x_d)$and $\bar{Y} = (y_1 \cdots y_d)$ is defined as follows:

$$Dist(\bar{X}, \bar{Y}) = \left(\sum_{i=1}^{d} |x_i - y_i|^p \right)^{1/p}$$

Special cases of the L_p-norm:

- *Euclidean* (*p* = 2)
- *Manhattan* (*p* = 1)

Multidimensional Quantitative Data

Impact of Domain-Specific Relevance

Assign weights for feature components a_i in L_p-norm:

$$Dist(\bar{X}, \bar{Y}) = \left(\sum_{i=1}^{d} a_i \cdot |x_i - y_i|^p \right)^{1/p}$$

Known as *generalized Minkowski distance.*

Multidimensional Quantitative Data

Impact of High Dimensionality

Effect of increase of data dimensionality:

Decrease of efficiency of distance-based clustering, classification, and outlier detection

- Example: Unit cube of dimensionality *d,* fully located in the nonnegative quadrant, with one corner at the origin *O.*

Manhattan distance:

$$Dist(\bar{O}, \bar{X}) = \sum_{i=1}^{d} (Y_i - 0)$$

$\mu = d/2$ and $\sigma = \sqrt{d/12}$

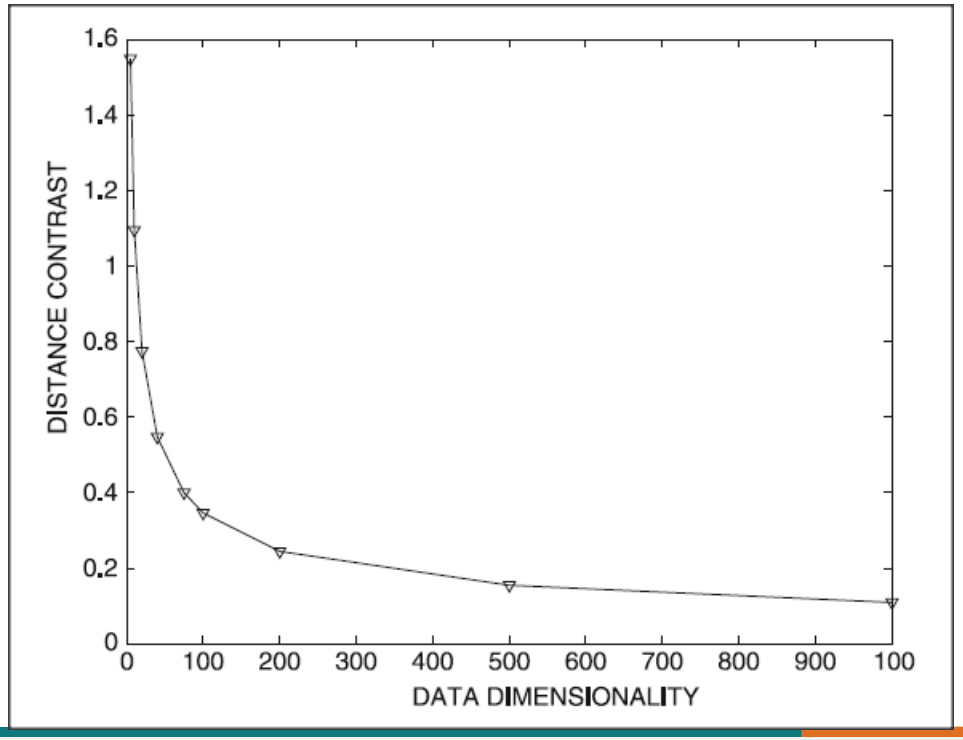
- By Law of Large Numbers, majority of points $[D_{min}, D_{max}] = [\mu - 3\sigma, \mu + 3\sigma]$. Range $D_{max} - D_{min} = 6\sigma = \sqrt{3d}$

Multidimensional Quantitative Data

Impact of High Dimensionality

$E[Dist(\bar{O}, \bar{X})]$ grows with dimensionality at a rate that is linearly proportional to *d*.

$$Contrast(d) = \frac{D_{max} - D_{min}}{\mu} = \sqrt{12/d}$$



Multidimensional Quantitative Data

Impact of Locally Irrelevant Features

Observation: Many features are likely to be irrelevant in a typical high-dimensional data set.

- Example: Euclidean metric may unnecessarily contribute a high value from the more noisy components because of its square-sum approach.

Problem cannot be solved by global feature subset selection:

- Relevance of features is *locally* determined by the pair of objects that are being considered.
- *Globally* all features may be relevant.

Multidimensional Quantitative Data

Impact of Different L_p-Norms

Different L_p-norms behave differently in terms of the impact of irrelevant features or the distance contrast.

- Example: extreme case when $p = \infty$. This translates to using only the dimension where the two objects are the most *dissimilar*.

A single irrelevant attribute on which the two objects are very different will throw off the distance value in the case of the L_∞metric.

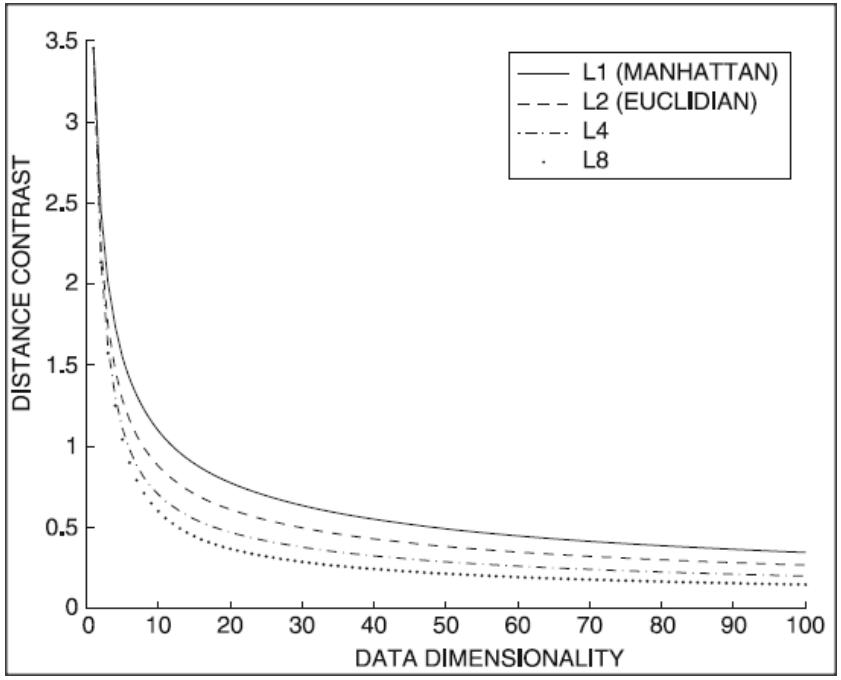
- Local similarity properties of the data are de-emphasized by L_∞.

Multidimensional Quantitative Data

Impact of Different L_p-Norms

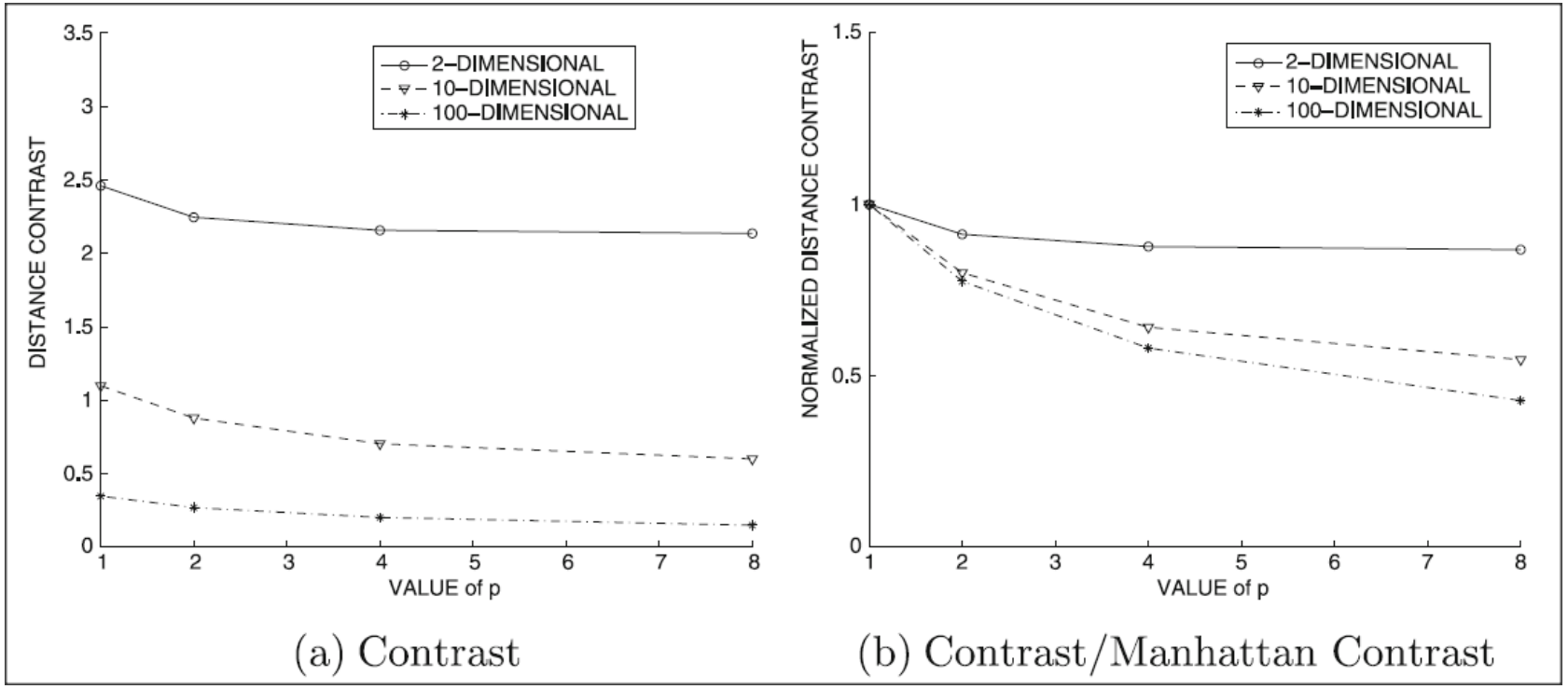
Distance contrasts illustrated for different values of *p* for the L_p-norm over different dimensionalities.

- L_p-norms degrade with increasing dimensionality
- Faster degradation is much for larger values of *p*.



Multidimensional Quantitative Data

Impact of Different L_p-Norms



(a) Contrast (b) Contrast/Manhattan Contrast

Multidimensional Quantitative Data

Match-Based Similarity Computation

With increasing dimensionality, a record is likely to contain both relevant and irrelevant features.

Observations:

- A pair of semantically similar objects may contain feature values that are dissimilar because of the noisy variations in irrelevant features.

- A pair of objects are unlikely to have similar values across many attributes, just by chance, unless these attributes were relevant.

Solution:

- De-emphasize individual attributes
- Evaluate the cumulative match across many dimensions

Multidimensional Quantitative Data

Match-Based Similarity Computation

Possible solution: *proximity thresholding* in a dimensionality-sensitive way.

Proximity Thresholding:

- Discretize data into equi-depth buckets.
- Divide each dimension into k_d equi-depth buckets, containing a fraction $1/k_d$ of the records.

Multidimensional Quantitative Data

Match-Based Similarity Computation

Let $\bar{X} = (x_1 \cdots x_d)$ and $\bar{Y} = (y_1 \cdots y_d)$ be two *d*-dimensional records.

- For dimension *i*, if both x_i and y_i belong to the same bucket, the two records are said to be in *proximity* on dimension *i*.

- The subset of dimensions on which *X* and *Y* map to the same bucket is referred to as the *proximity set*, and it is denoted by $S(\bar{X}, \bar{Y}, k_d)$.

- For each dimension $i \in S(\bar{X}, \bar{Y}, k_d)$, let m_i and n_i be the upper and lower bounds of the bucket in dimension *i*.

Multidimensional Quantitative Data

Match-Based Similarity Computation

The similarity $PSelect(\bar{X}, \bar{Y}, k_d)$ is defined as follows:

$$PSelect(\bar{X}, \bar{Y}, k_d) = \left[\sum_{i \in S(\bar{X}, \bar{Y}, k_d)} \left(1 - \frac{|x_i - y_i|}{m_i - n_i} \right)^p \right]^{1/p}$$

Observations:

- $PSelect(\bar{X}, \bar{Y}, k_d)$ will take values between 0 and $|S(\bar{X}, \bar{Y}, k_d)|$
- Each individual expression in the summation lies between 0 and 1.

This is a *similarity* function because larger values imply greater similarity.

Multidimensional Quantitative Data

Impact of Data Distribution

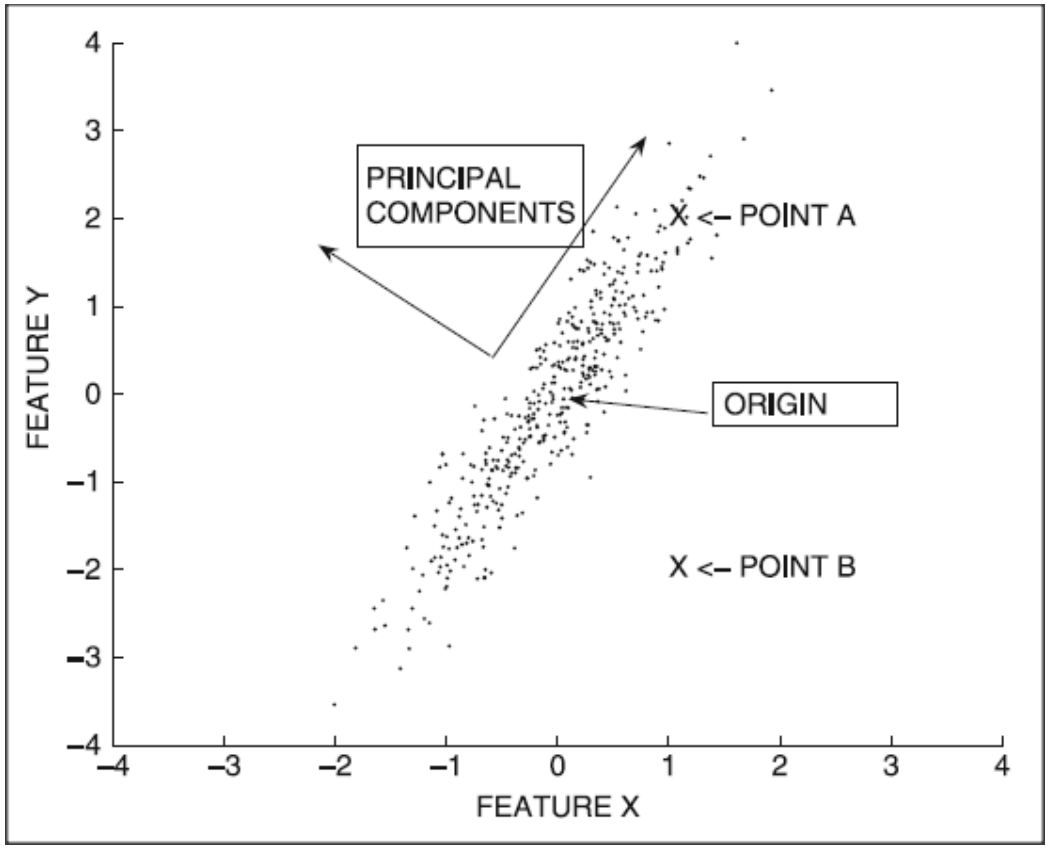
L_p-norm:

- depends only on the two data points in its argument
- It is invariant to the global statistics of the remaining data points.

Should distances depend on the underlying data distribution of the remaining points in the data set?

Multidimensional Quantitative Data

Impact of Data Distribution



Multidimensional Quantitative Data

Impact of Data Distribution

Given points A= (1, 2) and B= (1,−2):

- A and B are equidistant from the origin according to any L_p-norm.
- Statistically, it is much less likely for *B* to be so far away from *O* along this direction. Therefore, the distance from *O* to *A should* to be less than that of O to B.

Multidimensional Quantitative Data

Impact of Data Distribution

Let Σ be its $d \times d$ covariance matrix of the data set.

The Mahalanobis distance between two *d*-dimensional data points $\bar{X}$ and $\bar{Y}$ is as follows:

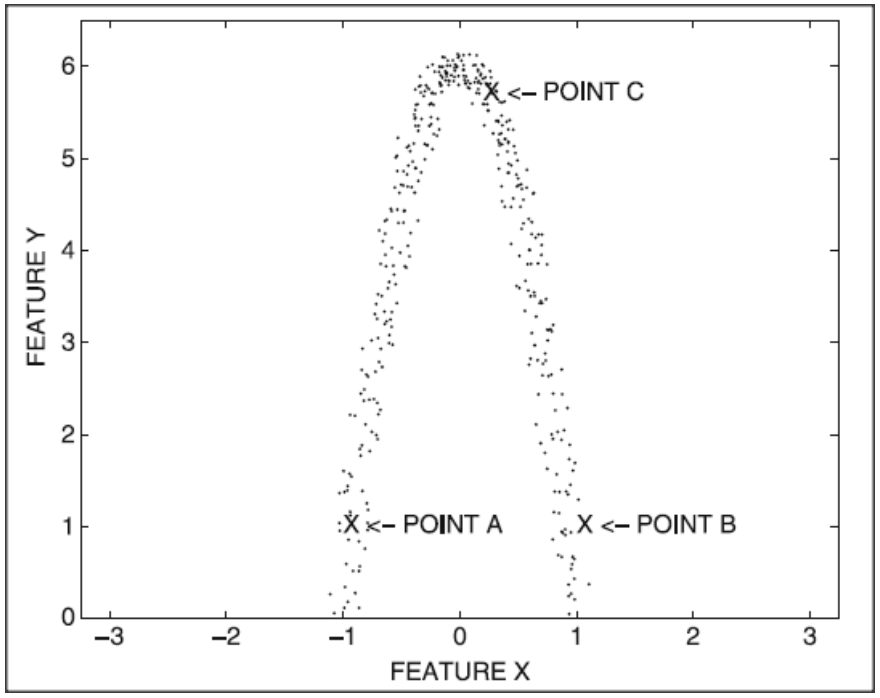
$$Maha(\bar{X}, \bar{Y}) = \sqrt{(\bar{X} - \bar{Y})\Sigma^{-1}(\bar{X} - \bar{Y})^T}$$

- Similar to the Euclidean distance, except that it normalizes the data on the basis of the inter-attribute correlations.

Multidimensional Quantitative Data

Nonlinear Distributions: ISOMAP

Consider the global distribution:



- Among the three data points *A*, *B*, and *C*, which pair are the closest to one another?

Multidimensional Quantitative Data

Nonlinear Distributions: ISOMAP

Distances as the shortest length of the path from one data point to another:

- using only point-to-point jumps from data points to one of their *k*-nearest neighbors
- based on a standard metric such as the Euclidean measure.

The overall sum of the point-to-point jumps reflects the aggregate change (distance) from one point to another (distant) point more accurately than a straight-line distance between the points.

- Such distances are referred to as *geodesic distances*.

Multidimensional Quantitative Data

Nonlinear Distributions: ISOMAP

ISOMAP consists of two steps:

- Compute the *k*-nearest neighbors of each point. Construct a weighted graph *G* with nodes representing data points, and edge weights (costs) representing distances of these *k*-nearest neighbors.

- For any pair of points $\bar{X}$ and $\bar{Y}$, report $Dist(\bar{X}, \bar{Y})$ as the shortest path between the corresponding nodes in *G*.

Multidimensional Quantitative Data

Nonlinear Distributions: ISOMAP

For improved computational performance:

- Additional step of embedding the data into a multidimensional space makes *repeated* distance computations between many pairs of points much faster, while losing some accuracy.

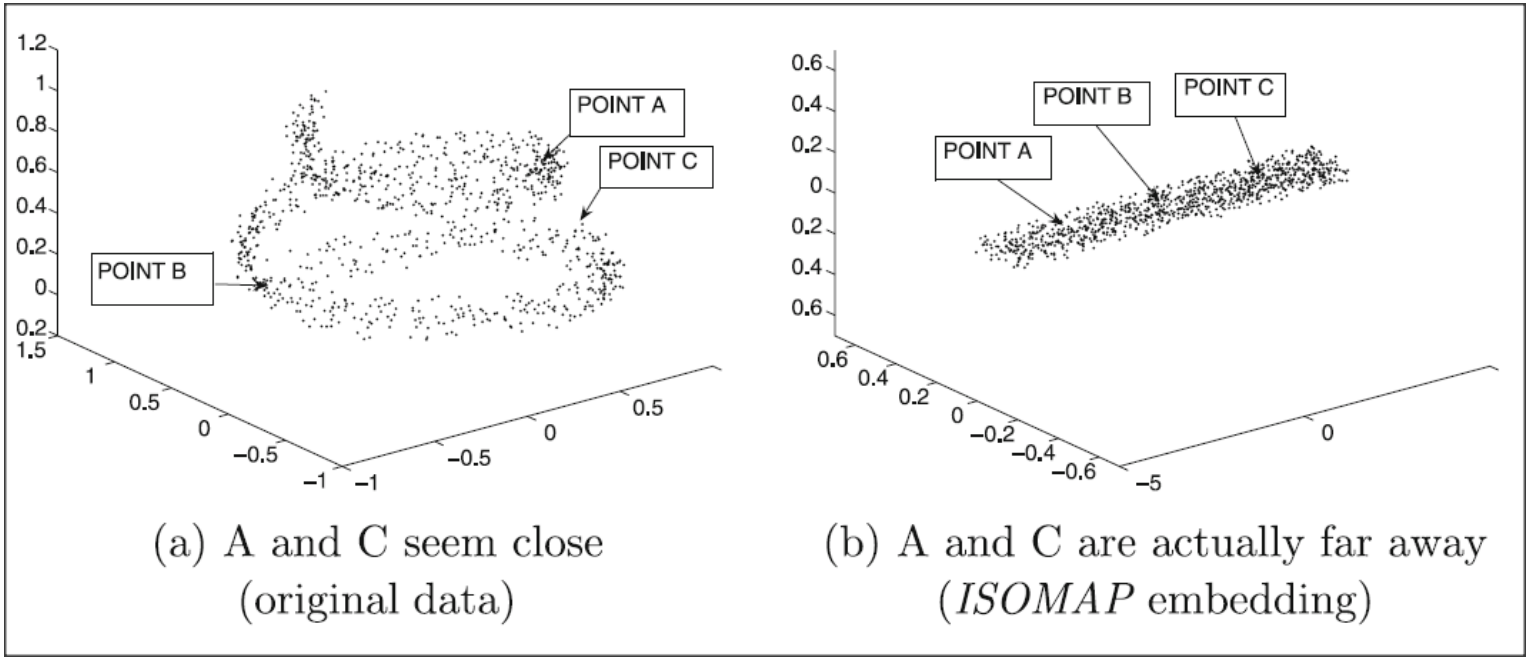
Method:

- Use the all-pairs shortest-path problem to construct the full set of distances between any pair of nodes in *G*.
- Apply *multidimensional scaling (MDS)* to embed the data into a lower dimensional space.
    - Effect: “straighten out” the nonlinear shape into a flat stip.
    - After *MDS*, Euclidian metric can be used.

Multidimensional Quantitative Data

Nonlinear Distributions: ISOMAP

- Example: data is distributed along a spiral. Data points A and C seem much closer to each other than data point B.
- In the *ISOMAP* embedding of Figure (right), the data point B is much closer to each of A and C.



(a) A and C seem close (original data)

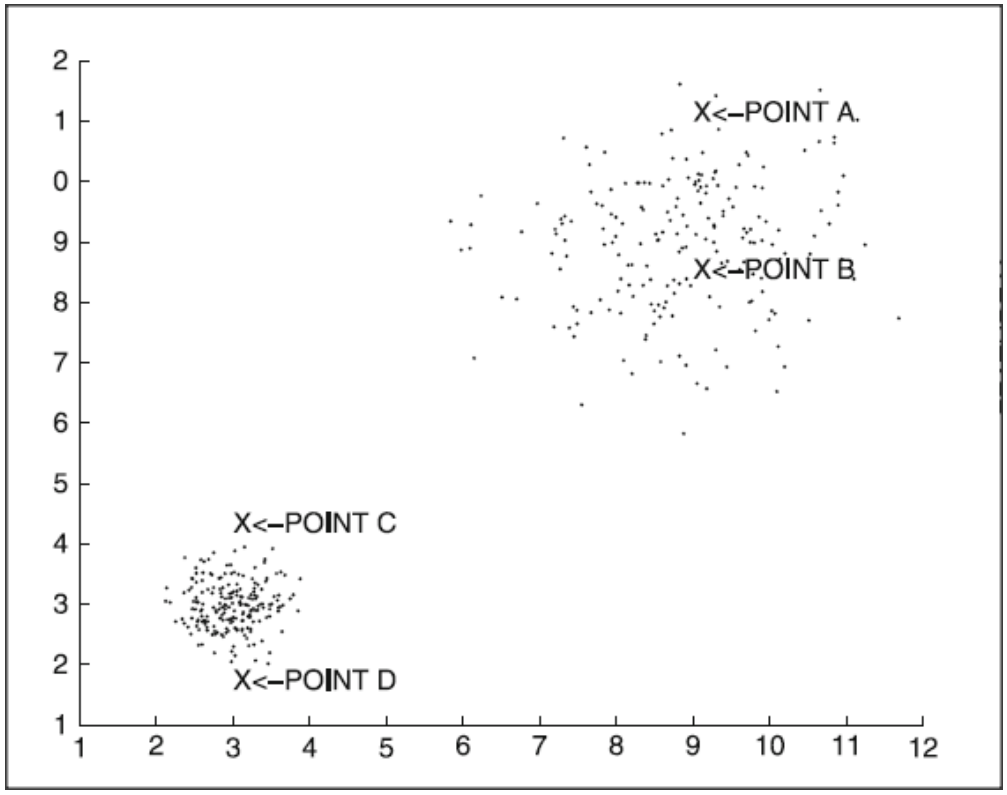
(b) A and C are actually far away (*ISOMAP* embedding)

Multidimensional Quantitative Data

Impact of Local Data Distribution

Two types:

Local density variation



Impact of Local Data Distribution

Shared Nearest-Neighbor Similarity

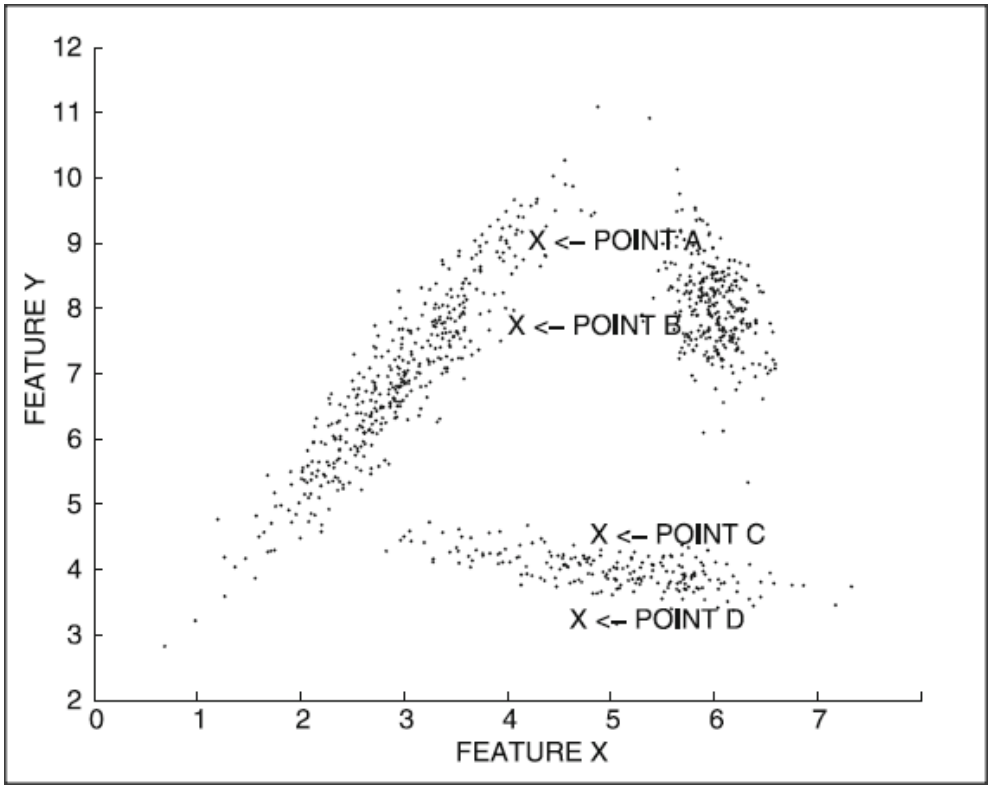
Steps:

- *Calculate k*-nearest-neighbors of each data point.
- The shared nearest-neighbor *similarity* is equal to the number of common neighbors between the two data points.
- This metric is locally sensitive because it depends on the number of common *neighbors*, and not on the absolute values of the distances.

Multidimensional Quantitative Data

Impact of Local Data Distribution

Local orientation variation



Multidimensional Quantitative Data

Impact of Local Data Distribution

Generic Methods:

Steps:

- Partition the data into a set of local regions.
- For any pair of objects:
  - Determine the most relevant region for the pair
  - Compute the pairwise distances using the local statistics of that region.
  - (Example: local Mahalanobis distance may be used in each local region.)

Multidimensional Categorical Data

Distance functions are naturally computed as functions of value differences along dimensions in numeric data, which is *ordered*.

However, no ordering exists among the discrete values of categorical data. How can distances be computed?

For the case of categorical data, it is more common to work with similarity functions rather than distance functions because discrete values can be matched more naturally.

Multidimensional Categorical Data

Consider two records $\bar{X} = (x_1 \cdots x_d)$ and $\bar{Y} = (y_1 \cdots y_d)$.

- The simplest possible similarity between the records $\bar{X}$ and $\bar{Y}$ is the sum of the similarities on the individual attribute values.
- If $S(x_i, y_i)$ is the similarity between the attributes values x_i and y_i, then the overall similarity is defined as follows:

$$Sim(\bar{X}, \bar{Y}) = \sum_{i=1}^{d} S(x_i, y_i)$$

Multidimensional Categorical Data

The simplest possible choice:

$$S(x_i, y_i) = \begin{cases} 1 & \text{if} \quad x_i = y_i \\ 0 & \text{otherwise} \end{cases}$$

- This is also referred to as the *overlap* measure.

Important drawback:

- Does not account for relative frequencies among the different attributes

Alternative:

- Compute similarity by using aggregate statistical properties of the data

Multidimensional Categorical Data

The *inverse occurrence frequency* is a generalization of the simple matching measure.

Let $p_k(x)$ be the fraction of records in which the k^{th} attribute takes on the value of x in the data set.

$$S(x_i, y_i) = \begin{cases} 1/p_k(x_i)^2 & \text{if } x_i = y_i \\ 0 & \text{otherwise} \end{cases}$$

Multidimensional Categorical Data

A related measure is the *Goodall* measure.

In a simple variant of this measure, the similarity on the k^{th} attribute is defined as $1 - p_k(x_i)^2$, when $x_i = y_i$ and 0 otherwise.

$$S(x_i, y_i) = \begin{cases} 1 - p_k(x_i)^2 & \text{if } x_i = y_i \\ 0 & \text{otherwise} \end{cases}$$

Multidimensional Mixed Data

Mixed Quantitative and Categorical Data

Generalize the approach to mixed data by adding the weights of the numeric and quantitative components.

Consider two records $\bar{X} = (\overline{X_n}, \overline{X_c})$ and $\bar{Y} = (\overline{Y_n}, \overline{Y_c})$ where $\overline{X_n}$, $\overline{Y_n}$ are the subsets of numerical attributes and $\overline{X_c}$, $\overline{Y_c}$ are the subsets of categorical attributes.

- The overall similarity between *X* and *Y* is defined as follows:

$$Sim(\bar{X}, \bar{Y}) = \lambda \cdot NumSim(\overline{X_n}, \overline{Y_n}) + (1 - \lambda) \cdot CatSim(\overline{X_c}, \overline{Y_c})$$

Multidimensional Mixed Data

Mixed Quantitative and Categorical Data

Normalization may be required to meaningfully compare the similarity value components on the numerical and categorical attributes that may be on completely different scales.

$$Sim(\bar{X}, \bar{Y}) = \lambda \cdot NumSim(\overline{X_n}, \overline{Y_n})/\sigma_n + (1 - \lambda) \cdot CatSim(\overline{X_c}, \overline{Y_c})/\sigma_c$$

- By performing this normalization, the value of λ becomes a true *relative weight* between the two components.

Text Similarity Measures

Text can be considered quantitative multidimensional data when it is treated as a list of words.

Frequency of each word:
- quantitative attribute

Base lexicon:
- full set of attributes.

Problem:
- structure of text is *sparse* in which most attributes take 0 values.
- L_p-norms do not perform well with varying lengths of documents

Text Similarity Measures

How can one normalize for such irregularities?

Cosine measure.

- Computes the *angle* between the two documents, which is insensitive to the absolute length of the document.

Let $\bar{X} = (x_1 \cdots x_d)$ and $\bar{Y} = (y_1 \cdots y_d)$ be two documents on a lexicon of size d.

$$\cos(\bar{X}, \bar{Y}) = \frac{\sum_{i=1}^{d} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{d} x_i^2} \cdot \sqrt{\sum_{i=1}^{d} y_i^2}}$$

- The measure simply uses the raw frequencies between attributes.

Text Similarity Measures

Possible improvement:

- Use global statistical measures to improve the similarity computation.

The *inverse document frequency* id_i is a decreasing function of the number of documents n_i in which the *i*th word occurs:

$$id_i = \log(n/n_i)$$

- Commonly used for normalization

Text Similarity Measures

Normalized frequency $h(x_i)$ for the *i*th word may be defined as follows:

$$h(x_i) = f(x_i) \cdot id_i$$

where $f(x)$ is a damping function

- Examples:

$$f(x_i) = \sqrt{x_i}$$

$$f(x_i) = \log(x_i)$$

Text Similarity Measures

The cosine measure with normalized frequencies:

$$\cos(\bar{X}, \bar{Y}) = \frac{\sum_{i=1}^{d} h(x_i) \cdot h(y_i)}{\sqrt{\sum_{i=1}^{d} h(x_i)^2} \cdot \sqrt{\sum_{i=1}^{d} h(y_i)^2}}$$

Another measureless commonly used for text is the *Jaccard coefficient* $J(\bar{X}, \bar{Y})$:

$$\mathrm{J}(\bar{X}, \bar{Y}) = \frac{\sum_{i=1}^{d} h(x_i) \cdot h(y_i)}{\sqrt{\sum_{i=1}^{d} h(x_i)^2} \cdot \sqrt{\sum_{i=1}^{d} h(y_i)^2} - \sum_{i=1}^{d} h(x_i) \cdot h(y_i)}$$

- Rarely used for the text domain, but it is used commonly for sparse binary data sets.

Binary and Set Data

Binary multidimensional data are a representation of set-based data, where value of 1 indicates the presence of an element in a set.

If S_X and S_Y are two sets with binary representations *X* and *Y*, then it can be shown that applying Jaccard coefficient to the raw binary representation of the two sets is equivalent to:

$$\mathrm{J}(\bar{X},\bar{Y}) = \frac{\sum_{i=1}^{d} h(x_i)\cdot h(y_i)}{\sqrt{\sum_{i=1}^{d} h(x_i)^2}\cdot\sqrt{\sum_{i=1}^{d} h(y_i)^2} - \sum_{i=1}^{d} h(x_i)\cdot h(y_i)} = \frac{|S_X \cap S_Y|}{|S_X \cup S_Y|}$$

Temporal Similarity Measures

Temporal data contains

- a single contextual attribute representing time
- one or more behavioral attributes that measure the properties varying along a particular time period.

Temporal data can be represented:

- continuous time series
- discrete sequences

Temporal Similarity Measures

Time-Series Similarity Measures

Design of time-series similarity measures is application specific.

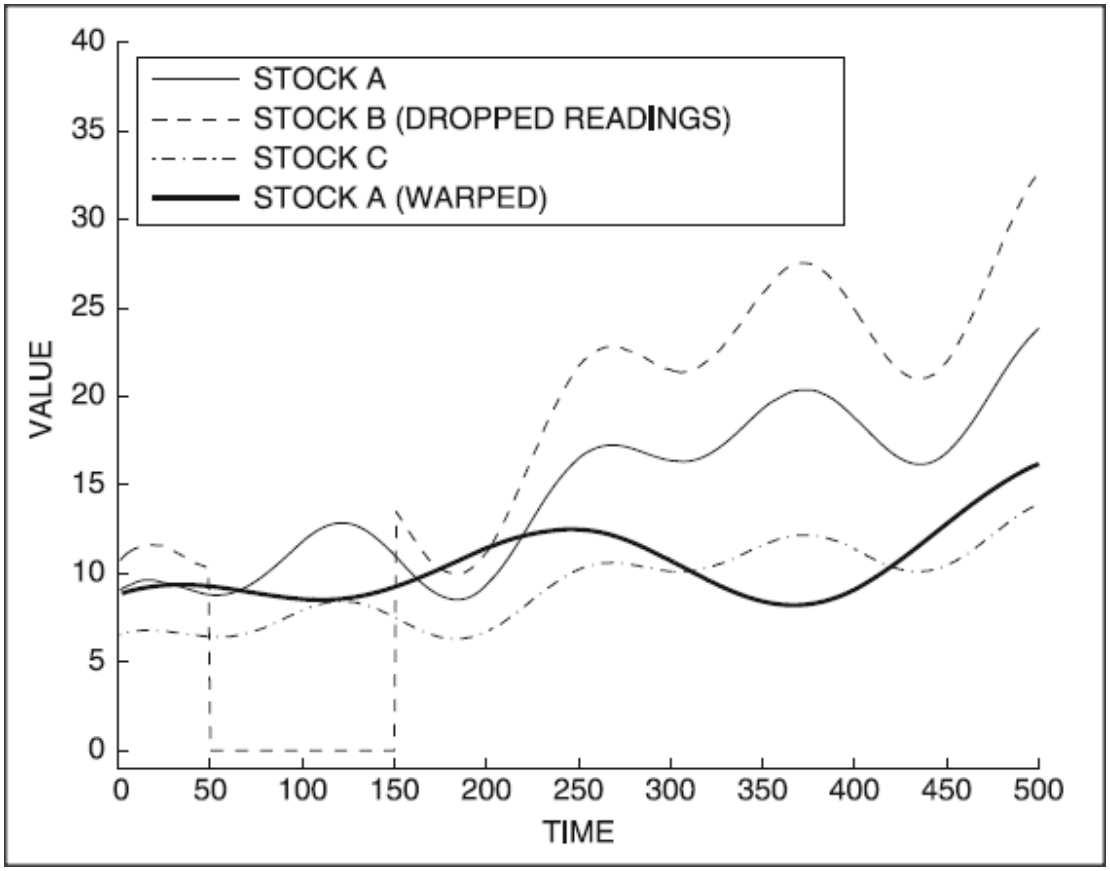
- The Euclidean metric may work well in many scenarios
- Issue: several distortion factors common to many applications.

Temporal Similarity Measures

Time-Series Similarity Measures

Distortion factors:

Behavioral attribute scaling and translation:



Temporal Similarity Measures

Time-Series Similarity Measures

Distortion factors (cont.):

Temporal (contextual) attribute translation:

- In some applications, such as real-time analysis of financial markets, the different time series may represent the same periods in time.

Temporal (contextual) attribute scaling:

- *Time Warping*.
- *Dynamic Time Warping* (DTW)

Noncontiguity in matching:

- Presence of noisy segments. The distance function may need to be robust to such noise.

Temporal Similarity Measures

Time-Series Similarity Measures

Solution for distortion factors: *attribute normalization*

Behavioral attribute translation:

- The behavioral attribute is mean centered during preprocessing.

Behavioral attribute scaling:

- The standard deviation of the behavioral attribute is scaled to 1 unit.

Temporal Similarity Measures

L_p-Norm

Already discussed. The L_p-norm between two data points $\bar{X} = (x_1 \cdots x_d)$and $\bar{Y} = (y_1 \cdots y_d)$ is defined as follows:

$$Dist(\bar{X}, \bar{Y}) = \left(\sum_{i=1}^{d} |x_i - y_i|^p \right)^{1/p}$$

- L_p-norm can be applied to wavelet transform

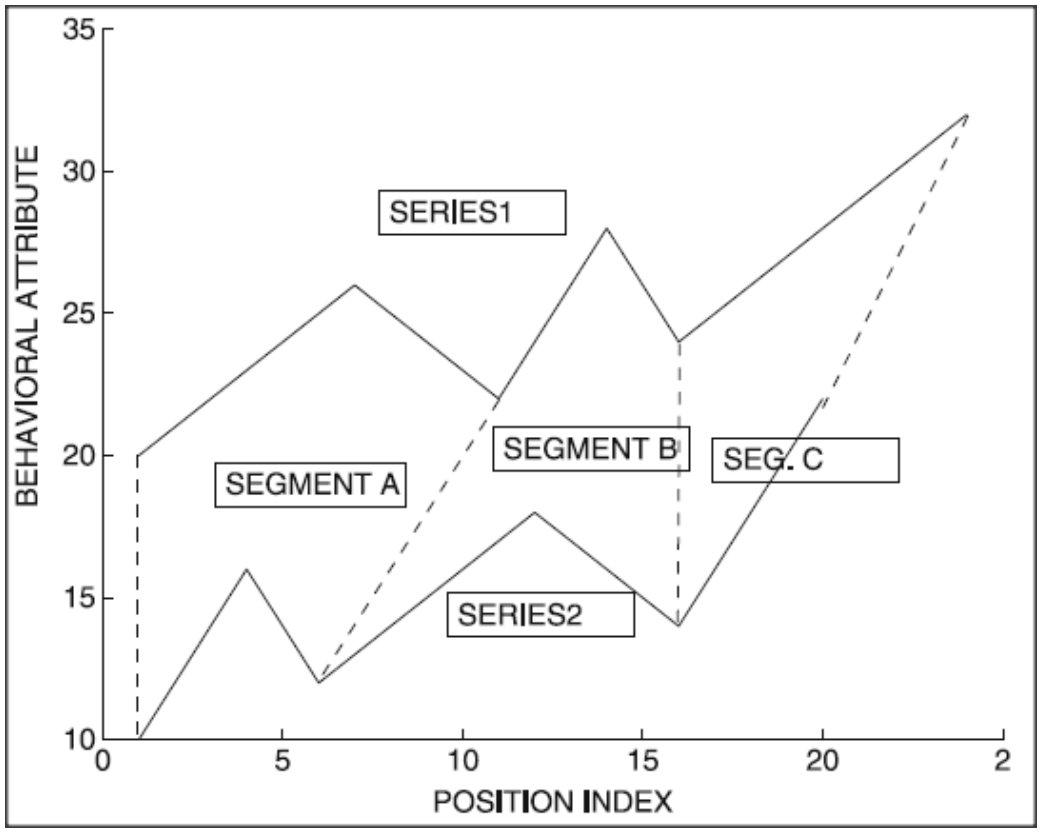
The major problem:

- L_p-norms are designed for time series of equal length
- Cannot address distortions on the temporal (contextual) attributes.

Temporal Similarity Measures

Dynamic Time Warping (DTW)

- stretches the series along the time axis in a varying (or *dynamic*) way over different portions to enable more effective matching.

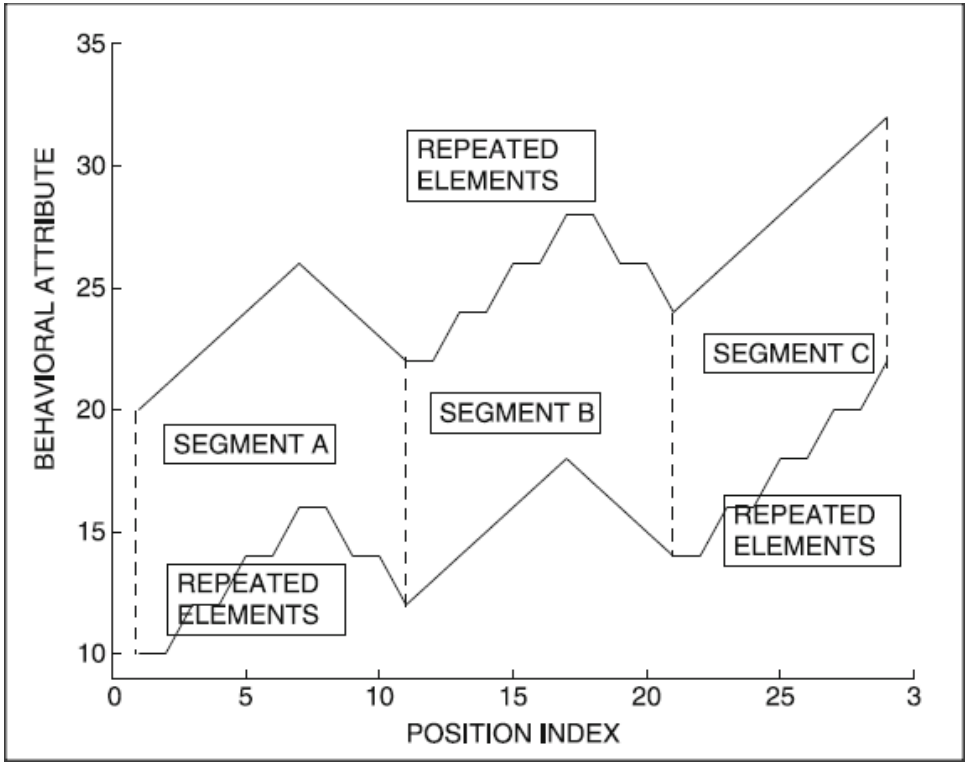


Temporal Similarity Measures

Dynamic Time Warping (DTW)

Many-to-one mapping by

- Repeating some elements in segments of either time series
- Artificially create two time series of same lengths
- Use distance measures (ex: L_p-norm) on warped series



Temporal Similarity Measures

Dynamic Time Warping (DTW)

Example: DTW generalization to one-to-one distance metric such as the L_p-norm

Consider the L_1 (Manhattan) metric $M(\bar{X}_i, \bar{Y}_i)$, computed on the first *i* elements of two time series $\bar{X} = (x_1 \cdots x_d)$and $\bar{Y} = (y_1 \cdots y_d)$ of equal length.

Value of $M(\bar{X}_i, \bar{Y}_i)$ can be written *recursively* as follows:

$$M(\bar{X}_i, \bar{Y}_i) = |x_i - y_i| + M(\overline{X_{i-1}}, \overline{Y_{i-1}})$$

- Note: In DTW many-to-one mapping allowed. In right hand side:
  - Any one or both indices may reduce by 1
  - Based on the *best match* between the two time series

Temporal Similarity Measures

Dynamic Time Warping (DTW)

Choice of index reduction result of an *optimization*.

Let $DTW(i,j)$ be the optimal distance between the first i and first j elements of two time series $\bar{X} = (x_1 \cdots x_m)$ and $\bar{Y} = (y_1 \cdots y_n)$.

- The value of $DTW(i,j)$ is defined recursively:

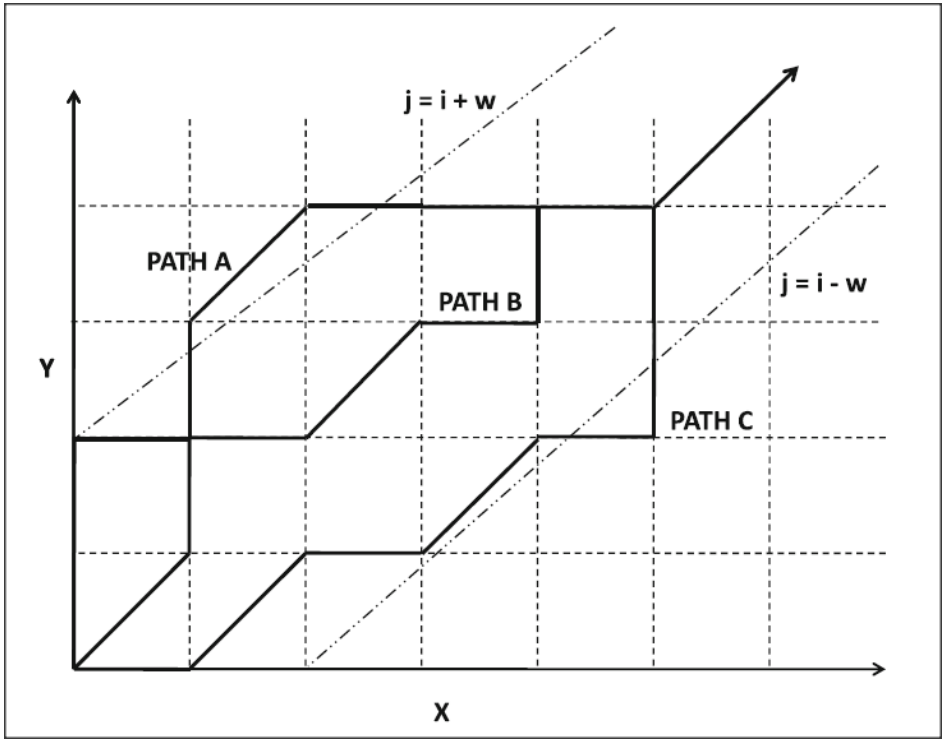
$$DTW(i,j)$$

$$= distance(x_i, y_j) + \min \begin{cases} DTW(i, j-1) & repeat\ x_i \\ DTW(i-1, j) & repeat\ y_j \\ DTW(i-1, j-1) & repeat\ neither \end{cases}$$

Temporal Similarity Measures

Dynamic Time Warping (DTW)

- Algorithm computes $DTW(i,j)$ with increasing index values of i and j.
- Optimal warping is the optimal path through different values of i and j.



Temporal Similarity Measures

Window-Based Methods

Window-based schemes attempt to decompose the two series into windows and then "stitch" together the similarity measure.

Consider two time series $\bar{X}$ and $\bar{Y}$, and let $\overline{X_1} \cdots \overline{X_r}$ and $\overline{Y_1} \cdots \overline{Y_r}$ be

- Temporally ordered
- Non-overlapping windows extracted from the respective series.
- Noise segments dropped

The overall similarity between $\bar{X}$ and $\bar{Y}$ can be computed as follows:

$$Sim(\bar{X}, \bar{Y}) = \sum_{i=1}^{r} Match(\overline{X_i}, \overline{Y_i})$$

- $Match(\overline{X_i}, \overline{Y_i})$ difficult to chose

Discrete Sequence Similarity Measures

Discrete sequence similarity measures are based on the same general principles as time series similarity measures.

Discrete sequence data may or may not have a one-to-one mapping between the positions.

When a one-to-one mapping does exist:

- many of the multidimensional categorical distance measures can be adapted to this domain, (similar to L_p-norm adapted to continuous time series).

STEVENS
INSTITUTE of TECHNOLOGY
School of Business

Thank You

Dragos Bozdog

For academic use only.