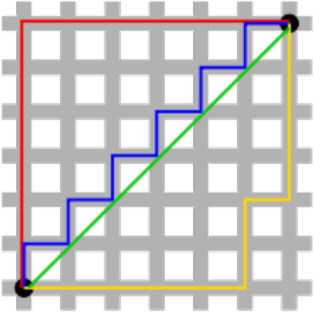
COMS20011 – Data-Driven Computer Science



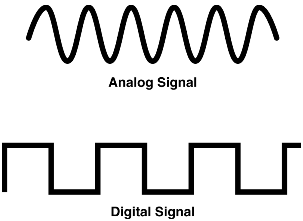
February 2021

Majid Mirmehdi, Rui Ponte Costa & Dima Damen

Image from Wikipedia

Lecture Video #4

This lecture



- Data acquisition
- **Data characteristics: distance measures**
- Data characteristics: summary statistics [*reminder*]
- Data normalisation and outliers

Data Characteristics: Distance Measures

- Distance is measure of separation between data.
- Distance is important as it:
  - enables data to be ordered
  - allows numeric calculations
  - enables measuring similarity and dissimilarity
- Without defining a distance measure, almost all statistical and machine learning algorithms will not function!
- Can be defined between single-dimensional data, multi-dimensional data or data sequences.

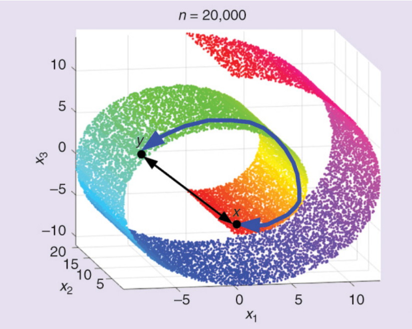
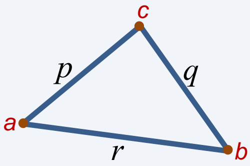


Image from https://www.embs.org/pulse/articles/what-is-the-distance-between-objects-in-a-data-set/

Distance

A valid distance measure $D(a,b)$ between two components a and b has the following properties

- non-negative: $D(a,b) \geq 0$
- reflexive: $D(a,b) = 0 \iff a = b$
- symmetric: $D(a,b) = D(b,a)$
- satisfies triangular inequality: $D(a,b) \leq D(a,c) + D(c,b)$



$$p + q \geq r$$
$$p + r \geq q$$
$$q + r \geq p$$

Distance (Numerical)

Distances between numerical data points in Euclidean space $\mathbb{R}^n$, for a point $x = (x_1, x_2, .., x_n)$ and a point $y = (y_1, y_2, .., y_n)$, the Minkowski distance of order p (p-*norm* distance) is defined as:

$$D(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}}$$

- $p = 1$
- $1 - norm$ distance (L_1)
- Also known as the *Manhattan Distance*

$$D(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

Distance (Numerical)

Distances between numerical data points in Euclidean space $\mathbb{R}^n$, for a point $x = (x_1, x_2, .., x_n)$ and a point $y = (y_1, y_2, .., y_n)$, the Minkowski distance of order p (p-*norm* distance) is defined as:

$$D(x, y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}}$$

- $p = 2$
- $2 - norm$ distance (L_2)
- Also known as the *Euclidean Distance*

$$D(x, y) = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2}$$

Can be expressed in vector form:

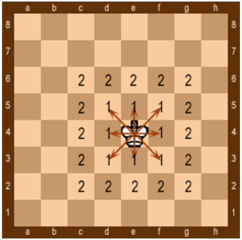
$$D(x, y) = \| \mathbf{x} - \mathbf{y} \|$$
$$= \sqrt{(\mathbf{x} - \mathbf{y})^T (\mathbf{x} - \mathbf{y})}$$

Distance (Numerical)

Distances between numerical data points in Euclidean space $\mathbb{R}^n$, for a point $x = (x_1, x_2, .., x_n)$ and a point $y = (y_1, y_2, .., y_n)$, the Minkowski distance of order p (p-*norm* distance) is defined as:

$$D(x,y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}}$$

- $p = \infty$
- $\infty - norm$ distance (L_∞)
- Also known as the *Chebyshev Distance*



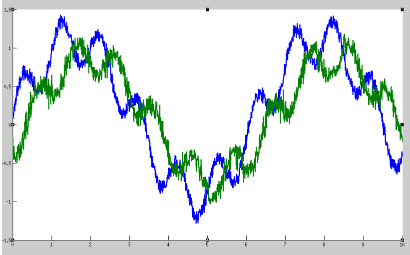
$$D(x,y) = \lim_{p \to \infty} (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}}$$

$$= max\ (|x_1 - y_1|, |x_2 - y_2|, \ldots, |x_n - y_n|)$$

Image from https://www.translatorscafe.com/

Distance (Numerical Series)

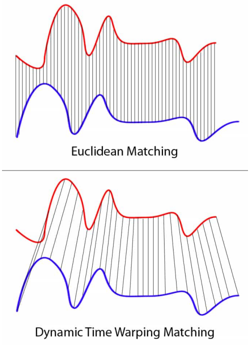
- Time Series: successive measurements made over a time interval
- Consider an audio signal of two people saying the same word



p-norm distances can only

- compare time series of the same length
- very sensitive to signal transformations:
  - shifting
  - uniform amplitude scaling
  - non-uniform amplitude scaling
  - uniform time scaling

Distance (Numerical Time Series)



Dynamic Time Warping (Berndt and Clifford, 1994)

- Replaces Euclidean one-to-one comparison with many-to-one
- Recognises similar shapes even in the presence of shifting and/or scaling
- Dynamic Time Warping (DTW) can be defined ***recursively***:

For two time series $\boldsymbol{X} = (x_0, .., x_n)$ and $\boldsymbol{Y} = (y_0, .., y_m)$

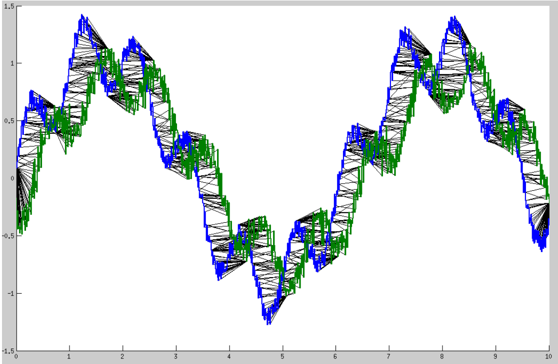
$$DTW(\boldsymbol{X}, \boldsymbol{Y}) = D(x_0, y_0) + \min\{DTW(\boldsymbol{X}, REST(\boldsymbol{Y})), DTW(REST(\boldsymbol{X}), \boldsymbol{Y}), DTW(REST(\boldsymbol{X}), REST(\boldsymbol{Y}))\}$$

where $REST(\boldsymbol{X}) = (x_1, .., x_n)$

Distance (Numerical Time Series)

Dynamic Time Warping (Berndt and Clifford, 1994)

- Can be used for aligning sequences



Distance (Symbolic)

- Distance is not always between numerical data
- Distance between symbolic data is less well-defined (e.g. text data)
- Distance in text could be:
  - syntactic
  - semantic

Distance (Symbolic)

Syntactic - e.g. Hamming Distance

- Defined over symbolic data of *the same* length
- Measures the number of substitutions required to change one string/number into another
  - *B r i s t o l*
    B u r t t o n
    D('Bristol', 'Burtton') = 4
  - 5 2 4 3
    6 2 1 3
    D(5243, 6213) = 2
  - 1011101
    1001001
    D(1011101, 1001001) = 2
- For binary strings, Hamming Distance equals L_1

Distance (Symbolic)

Syntactic - e.g. Edit Distance

- Defined on text data of *any* length
- Measures the *minimum* number of ‘operations’ required to transform one sequence of characters into another
- ‘Operations’ can be: insertion, substitution, deletion

- e.g. D(‘fish’, ‘first’) = 2

$$\text{‘fish’} \xrightarrow{\textit{insertion}} \text{‘firsh’} \xrightarrow{\textit{substitution}} \text{‘first’}$$

- used in spelling correction, DNA string comparisons

Distance (Symbolic)

Semantic - e.g. WUP Relatedness Measure

- Built on top of a hierarchy of word semantics
- Most commonly used is WordNet (Princeton)
  - http://wordnet.princeton.edu/
- WordNet use directed relationships (parent-child hierarchies)
  - hyponymy (is-a relationship)

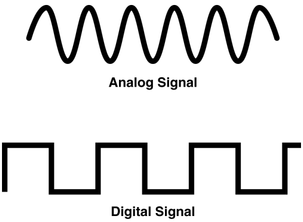
    e.g. furniture → bed
  - meronymy (part-of relationship)

    e.g. chair → seat
  - troponymy [for verb hierarchies] (specific manner)

    e.g. communicate → talk → whisper
  - antonymy (strong contract)

    e.g. wet ↔ dry
- online: http://ws4jdemo.appspot.com/

Next lecture video



- Data acquisition
- Data characteristics: distance measures
- **Data characteristics: summary statistics [*reminder*]**
- **Data normalisation and outliers**