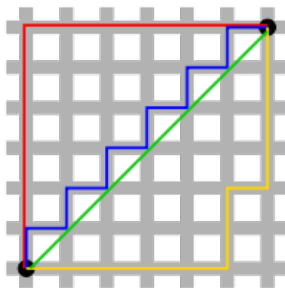


COMS20011 – Data-Driven Computer Science



January 2023

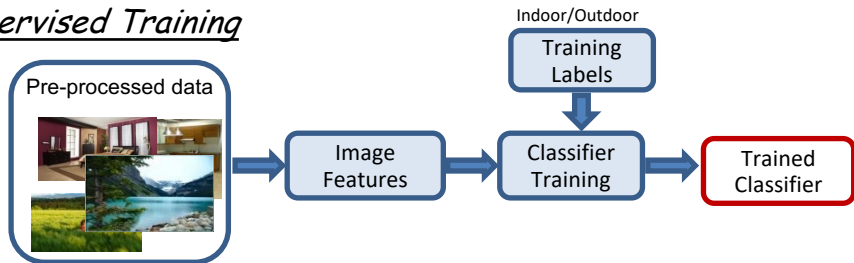
Majid Mirmehdi

with slides from Rui Ponte Costa & Dima Damen

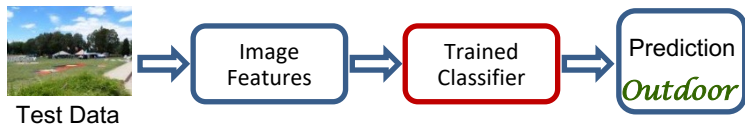
Last lecture: Typical Data Analysis Problem

1. Pre-processing
2. Feature Selection
3. Modelling & Classification

Supervised Training



Testing



This lecture



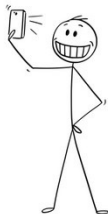
Analog Signal



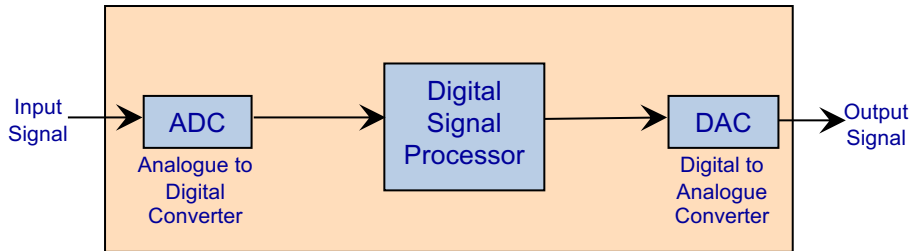
Digital Signal

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

Data Acquisition – Example Data Journey



<https://www.vectortock.com/>



Data Acquisition - Analogue to Digital Conversion

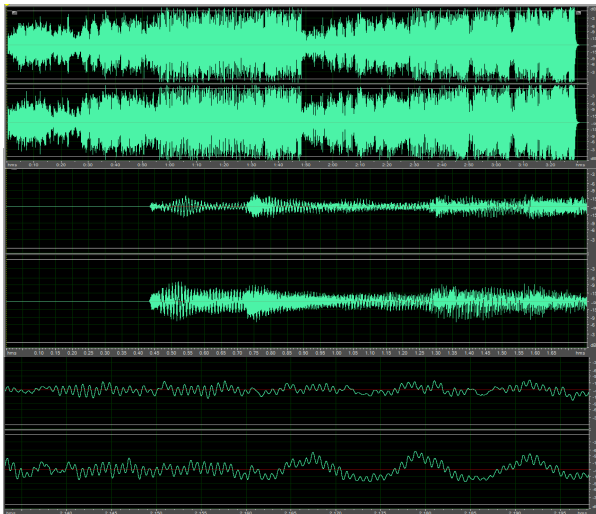
Analogue to Digital conversion involves *Sampling & Quantisation*

e.g. a 1D Audio Signal

low zoom

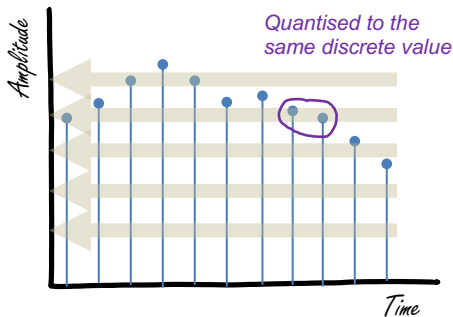
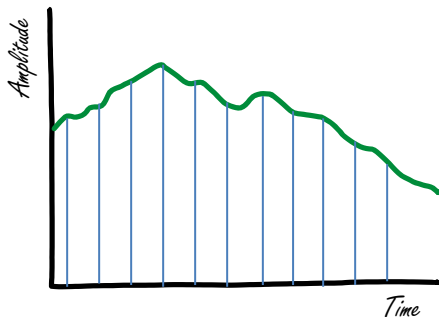
medium zoom

large zoom



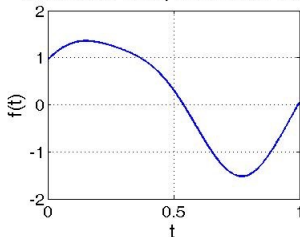
Data Acquisition - Analogue to Digital Conversion

Analogue to Digital conversion involves *Sampling* & *Quantisation*

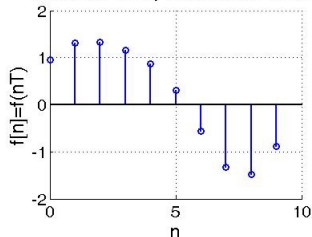


Sample and Quantise

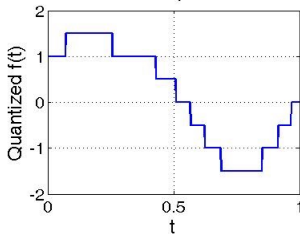
Continuous Time, Continuous Value



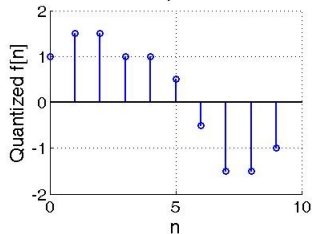
Discrete Time, Continuous Value



Continuous Time, Discrete Value

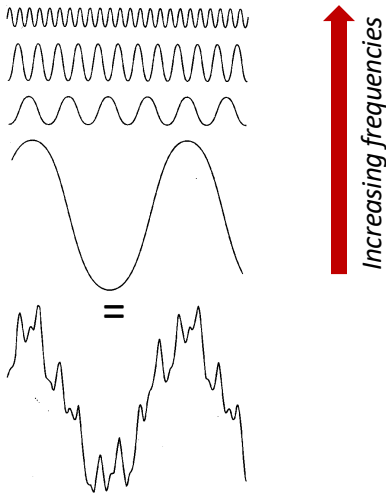


Discrete Time, Discrete Value



Nyquist-Shannon Sampling Theory

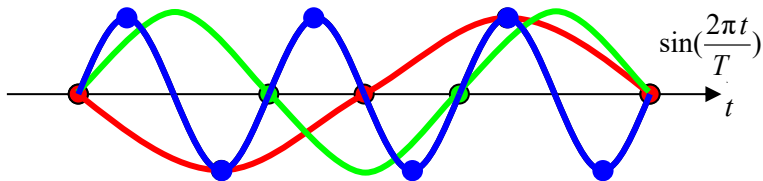
"An analogue signal containing components up to some maximum frequency u (Hz) may be completely reconstructed by regularly spread samples, provided the sampling rate is at least $2u$ samples per second"



Nyquist-Shannon Sampling Theory

"An analogue signal containing components up to some maximum frequency u (Hz) may be completely reconstructed by regularly spread samples, provided the sampling rate is at least $2u$ samples per second"

Also referred to as the Nyquist-Shannon criterion: sampling rate s should be at least twice the highest spatial frequency u .



$$\text{sampling period } T \leq \frac{1}{2u}$$

$$\text{equivalent to sampling rate } s \geq 2u$$

Data Acquisition - Analogue to Digital Conversion

Examples of sampling and quantisation of standard audio formats:

- Speech (e.g. phone call)
 - Sampling: 8 KHz samples
 - Quantisation: 8 bits / sample
- Audio CD
 - Sampling: 44 KHz samples
 - Quantisation: 16 bits / sample
 - Stereo (2 channels)

Higher sampling and quantisation levels achieves better signal quality, but at the expense of larger memory and storage.

Data Acquisition - Analogue to Digital Conversion

Examples of sampling and quantisation of Images - Multi-Dimensional:

- Sampling: Resolution in digital photography
- Quantisation: Representation of each pixel in the image
 - 8 Mega Pixel Camera: 3264 x 2448 pixels
 - Colour images: 3 channels: Red, Green, Blue (8 bits per colour)
 - Greyscale images: 1 channel: intensity = $aR+bG+cB$ where $a+b+c=1.0$
 - Binary images: Black/White 1 bit per pixel

Sampling – visual example

The effect of sparser sampling...is **ALIASING**



256x256



64x64



32x32

Anti-aliasing is achieved by filtering to remove frequencies above the Nyquist limit.

Quantisation – visual example

This results from representing a continuously varying function $f(x)$ with a discrete one using quantisation levels



16 levels



6 levels



2 levels

Next...



Analog Signal

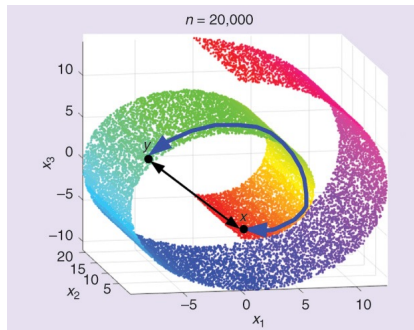


Digital Signal

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

Data Characteristics: Distance Measures

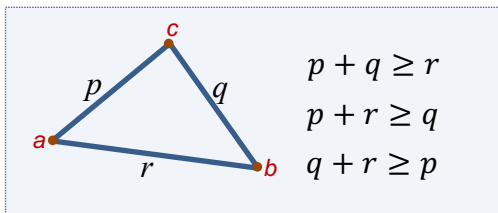
- Distance is measure of separation between data.
- Distance is important as it:
 - enables data to be ordered
 - allows numeric calculations
 - enables measuring similarity and dissimilarity
- Without defining a distance measure, almost all statistical and machine learning algorithms will not function!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



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: $D(a,b) \leq D(a,c) + D(c,b)$



Distance (Numerical)

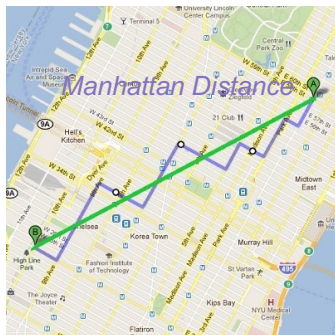
To find the distance between numerical data points $\mathbf{x} = (x_1, x_2, \dots, x_n)$ and $\mathbf{y} = (y_1, y_2, \dots, y_n)$ in Euclidean space \mathbb{R}^n , the **Minkowski Distance** of order p (p -norm distance) is defined as:

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

- $p = 1$
- $1 - \text{norm distance}$ (L_1)

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

- Also known as the *Manhattan Distance*



Distance (Numerical)

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

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

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

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

Can be expressed in vector form:

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



Distance (Numerical)

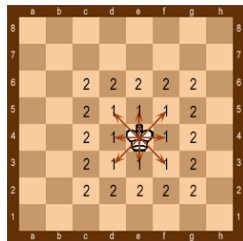
To find the distance between numerical data points $\mathbf{x} = (x_1, x_2, \dots, x_n)$ and $\mathbf{y} = (y_1, y_2, \dots, y_n)$ in Euclidean space \mathbb{R}^n , the **Minkowski Distance** of order p (p -norm distance) is defined as:

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

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

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

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

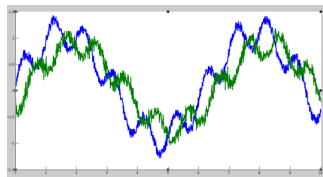


Distance (Numerical Series)

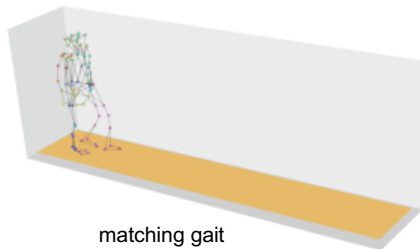
- Time Series: successive measurements made over a time interval

p-norm distances can only

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



matching audio signal of two people saying the same word

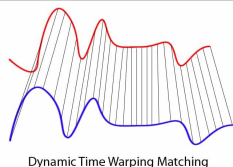
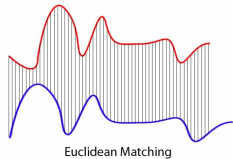


matching gait

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, length, and scaling
- Dynamic Time Warping (DTW) can be defined **recursively** to tell us how two signals align with each other:



For two time series $\mathbf{X} = (x_1, \dots, x_n)$ and $\mathbf{Y} = (y_1, \dots, y_m)$

$$DTW(\mathbf{X}, \mathbf{Y}) = D(x_1, y_1) + \min\{DTW(\mathbf{X}, \text{REST}(\mathbf{Y})), DTW(\text{REST}(\mathbf{X}), \mathbf{Y}), DTW(\text{REST}(\mathbf{X}), \text{REST}(\mathbf{Y}))\}$$

$$\text{where } \text{REST}(\mathbf{X}) = (x_2, \dots, x_n)$$

DTW builds a distance matrix between two time series and then finds the minimum path (alignment cost) for an optimal match.

OPTIONAL: for more details, watch: <https://www.youtube.com/watch?v=ERKDHZyZdWA> (2 parts)

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

cawh|
cash

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 $D(\text{'Bristol'}, \text{'Burton'}) = 4$

➤ *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. $D(\text{'fish'}, \text{'first'}) = 2$

'fish' $\xrightarrow{\text{insertion}}$ 'firsh' $\xrightarrow{\text{substitution}}$ 'first'

- used in spelling correction, DNA string comparisons, etc.

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 contrast)
e.g. wet ↔ dry
- online: <http://ws4jdemo.appspot.com/>