Curse of dimensionality

- When dimensionality increases data become sparse (Information content is less)
- Density and distance between points, which is critical to clustering, outlier analysis becomes less meaningful
- The possible combinations of subspaces will grow exponentially.

Dimensionality Reduction

- Avoid the curse of dimensionality.
- Help eliminate irrelevant features and reduce noise
- Reduce time and space required for data mining
- Allow easier visualization

Techniques used

PCA - Principal component Analysis
FA - Factor Analysis
DFT - Discrete Fourier Transform
DCT - Discrete cosine Transform
DWT - Discrete Wavelet Transform.

Time Domain - Analysis of signal or data w.r.t Time

Frequency Domain - Analysis of signal or data w.r.t Frequency

Example

ECG - Electro cardiogram

If a doctor maps the heartbeat with time. Say, the recording is done for 2 min. we call it time domain signal.

In ECG a number of peaks are there. Say, in on heart beat 4 types of peaks or variation in amplitude occurs.

In frequency domain representation → How many times each peak comes is recorded over the entire time period of recording.

A given function or signal can converted between the time and frequency domain with a pair of mathematical operations is called a transform.

Discrete Fourier Transform (DFT)

Using DFT we can determine the frequency content of a signal.

$$\xrightarrow{DFT} \quad X(K) = \sum_{n=0}^{N-1} x(n)\, e^{\frac{-j2\pi Kn}{N}}, \quad 0 \le K \le N-1$$

$$\xrightarrow{IDFT} \quad x(n) = \frac{1}{N} \sum_{K=0}^{N-1} X(K)\, e^{j2\pi Kn/N}, \quad 0 \le n \le N-1$$

Q Find the DFT of a sequence $x(n) = \{1, 1, 0, 0\}$

Sol^n

$N = 4$ = length of $x(n)$

$K = 0, 1, \ldots N-1 = 0, 1, 2, 3$

$K=0$

$$X(0) = \sum_{n=0}^{3} x(n) e^{0} = x(0)+x(1)+x(2)+x(3) \quad (2)$$
$$= 1+1+0+0 = 2$$

$K=1$

$$X(1) = \sum_{n=0}^{3} x(n) e^{\frac{-j2\pi n}{4}} = x(0)e^{0} + x(1)e^{-j\pi/2}$$
$$+ x(2)e^{-j\pi} + x(3)e^{\frac{-j3\pi}{2}}$$
$$= 1+1(\cos\tfrac{\pi}{2} - j\sin\pi/2) = 1-j$$

$K=2$

$$X(2) = \sum_{x=0}^{3} x(n) e^{\frac{-j2\pi\cdot 2\cdot n}{4}} = \sum_{x=0}^{3} x(n) e^{-j\pi n}$$
$$= x(0)+x(1)e^{-j\pi} = 1+\cos\pi - j\sin\pi$$
$$= 1-1 = 0$$

$K=3$

$$X(3) = \sum_{n=0}^{3} x(n) e^{\frac{-j2\pi\times 3\times n}{4}} = \sum_{n=0}^{3} x(n) e^{\frac{-j3\pi n}{2}}$$
$$= x(0) + x(1)e^{\frac{-j3\pi}{2}} + x(2)e^{-j3\pi/2} + x(3)e^{-j9\pi/2}$$
$$= 1+\cos\tfrac{3\pi}{2} - j\sin\tfrac{3\pi}{2} = 1+j$$

$$X(K) = \{2,\ 1-j,\ 0,\ 1+j\}.$$

IDFT $(X(K)) \rightarrow x(n)$

$X(K) = \{2, 1-j, 0, 1+j\}$ IDFT

$e^{j\phi} = \cos\phi + j\sin\phi$

$e^{-j\phi} = \cos\phi - j\sin\phi$

$$y(n) = \frac{1}{N}\sum_{K=0}^{N-1} Y(K)\, e^{j2\pi nK/N}, \quad n = 0, 1, \cdots N-1$$

n=0

$$y(0) = \frac{1}{4}\sum_{K=0}^{3} Y(K)\, e^{0} = \frac{1}{4}[Y(0) + Y(1) + Y(2) + Y(3)]$$
$$= \frac{1}{4}[2 + 1 - j + 0 + 1 + j] = \frac{4}{4} = 1$$

n=1

$$y(1) = \frac{1}{4}\sum_{K=0}^{3} Y(K)\, e^{\frac{j2\pi K}{4}\times 2} = \frac{1}{4}\sum_{K=0}^{3} Y(K)\, e^{\frac{j2\pi K}{2}}$$
$$= \frac{1}{4}\left[Y(0)\cdot e^{0} + Y(1)\, e^{j\pi/2} + Y(2)\, e^{j\pi} + Y(3)\, e^{j3\pi/2}\right]$$
$$= \frac{1}{4}\left[2\cdot 1 + (1-j)\left(\cos\tfrac{\pi}{2} + j\sin\tfrac{\pi}{2}\right) + 0 + (1+j)\left(\cos\tfrac{3\pi}{2} + j\sin\tfrac{3\pi}{2}\right)\right]$$
$$= \frac{1}{4}[2 + (1-j)(0+j) + (1+j)(0-j)]$$
$$= \frac{1}{4}[2 + j - j^2 + 1 - j - j^2]$$
$$= \frac{1}{4}[2 - 2(-1)] = \frac{4}{4} = 1.$$

n=2

$$y(2) = \frac{1}{4}\sum_{K=0}^{3} Y(K)\, e^{\frac{j2\pi\times 2K}{4}} = \frac{1}{4}\sum_{K=0}^{3} Y(K)\, e^{jK\pi}$$
$$= \frac{1}{4}\left[Y(0)e^{0} + Y(1)e^{j\pi} + Y(2)e^{j2\pi} + Y(3)e^{j3\pi}\right]$$
$$= \frac{1}{4}[2\times 1 + (1-j)(\cos\pi + j\sin\pi) + 0 + (1+j)(\cos 3\pi - j\sin 3\pi)]$$
$$= \frac{1}{4}[2 + (1-j)(-1+0) + (1+j)(-1+0)]$$
$$= \frac{1}{4}[2 - 1 + j - 1 - j] = \frac{0}{4} = 0.$$

n=3

$$y(3) = \frac{1}{4}\sum_{K=0}^{3} Y(K)\, e^{\frac{j6\pi K}{4}\times 2} = \frac{1}{4}\sum_{K=0}^{3} Y(K)\, e^{j3\pi K/2}$$
$$= \frac{1}{4}\left[Y(0)e^{0} + Y(1)e^{j3\pi/2} + Y(2)e^{j3\pi} + Y(3)e^{j9\pi/2}\right]$$
$$= \frac{1}{4}\left[2 + (1-j)\left(\cos\tfrac{3\pi}{2} + j\sin\tfrac{3\pi}{2}\right) + 0 + (1+j)\left(\cos\tfrac{9\pi}{2} + j\sin\tfrac{9\pi}{2}\right)\right]$$
$$= \frac{1}{4}[2 + (1-j)(0-j) + (1+j)(0+j)]$$
$$= \frac{1}{4}[2 - j + (-1) + j + (-1)] = \frac{1}{4}[2-2] = 0$$

Discrete cosine Transform (DCT)

1-D DCT

2-D DCT → used in images.

Image compression

Let the size of an image is 256 X 256. we divide it into 32 by 32 square blocks of 8×8 pixels each and treat each one independently. DCT is applied to 8×8 pixel blocks of the image. The 64 pixel values of each block are transformed by the DCT into a new set of 64 values known as DCT coefficients.

DC coefficient

Low frequency coefficients

4.32	3.12	3.01	2.41
2.74	2.11	1.92	1.55
2.11	1.32	0.32	0.11
1.62	0.44	0.03	0.02

→ High frequency (AC coefficients)

DCT values for 4×4 image block.

JPEG standard for Lossy compression

1) First the lowest weights are trimmed by setting to zero.

2) The remaining weights are quantized (i.e Round off to the nearest integer value), set the most of the high frequency values to zero.

3) Then use Zig-Zag coding.

4 3 3 2 2 2 1 2 2 0 0 0 0 0 0 0

4→3→3→2→2→2→1→2→2→0→0→0→0→0→0→0

Discrete Wavelet Transform (DWT)

① $\vec{X} \rightarrow DWT \rightarrow \vec{X'}$ (wavelet coefficient)

The size of $(\vec{X})$ = size of $(\vec{X'})$

For Dimensionality reduction, we will keep the wavelet coefficients larger than some threshold value and truncate the small values or set them to zero value.

Given the wavelet coefficients, an approximation of the original data can be constructed by applying IDWT.

Method

1. The length L, of the input data vector must be an integer power of 2. This condition can be met by padding the data vector with zeros as necessary

2) DWT(X) — L

This results in two sets of data of length L/2. Generally it contains a low frequency version of the input data and high frequency content of it respectively.

3) The low frequency content again divided into two sets and the process continues until the resulting data set obtained are of length 2.

4) selected values from the data sets obtained in above are Wavelet coefficients.

Advantages of DWT over DFT

1) DWT achieves better lossy compression.

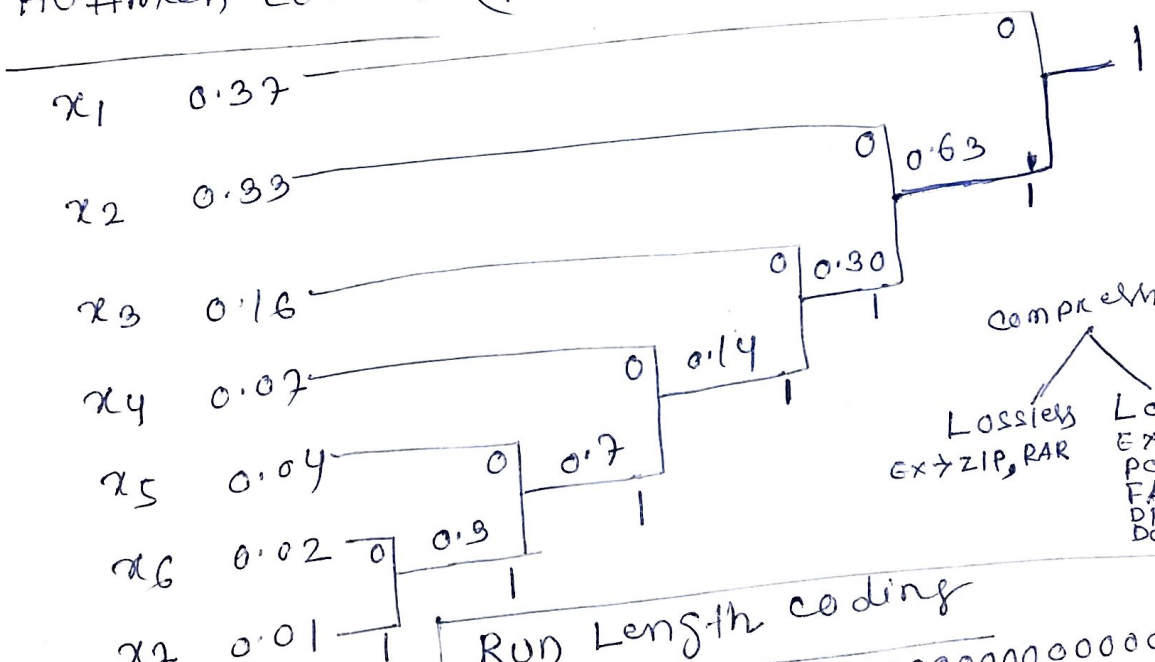
2) Requires less space than DFT

3) There is only one DFT, but there are several families of DWT.

4) popular wavelet transforms include Haar-2, Daubechies-2, Daubechies-6.

5) DWT uses a hierarchical pyramid algorithm that halves the data at each iterations resulting in fast computation.

Huffman coding (for Data compression)



Compression

Lossless
Ex → ZIP, RAR

Lossy
Ex →
PCA
FA
DFT
DCT, DWT

codeword

$x_1 = 0$
$x_2 = 10$
$x_3 = 110$
$x_4 = 1110$
$x_5 = 11110$
$x_6 = 111110$
$x_7 = 1111111$

Run Length coding

s = 111111111111111000000000000000000
00 1111

(15→1), (19→0), (4→1)

Max. no. of repition = 19, which can represented with 5-bits.

(01111, 1), (10011, 0), (00100, 1)

Compression ratio = $\frac{18}{38} = 1 : 2.11$