Dimensionality Reduction

2D, 3D :→ Scatter plot

4D, 5D, 6D :→ Pair plot

10D, 100D, 1000D ⟶ dimension reduction techniques

- PCA
- t-SNE

* Dataset :→

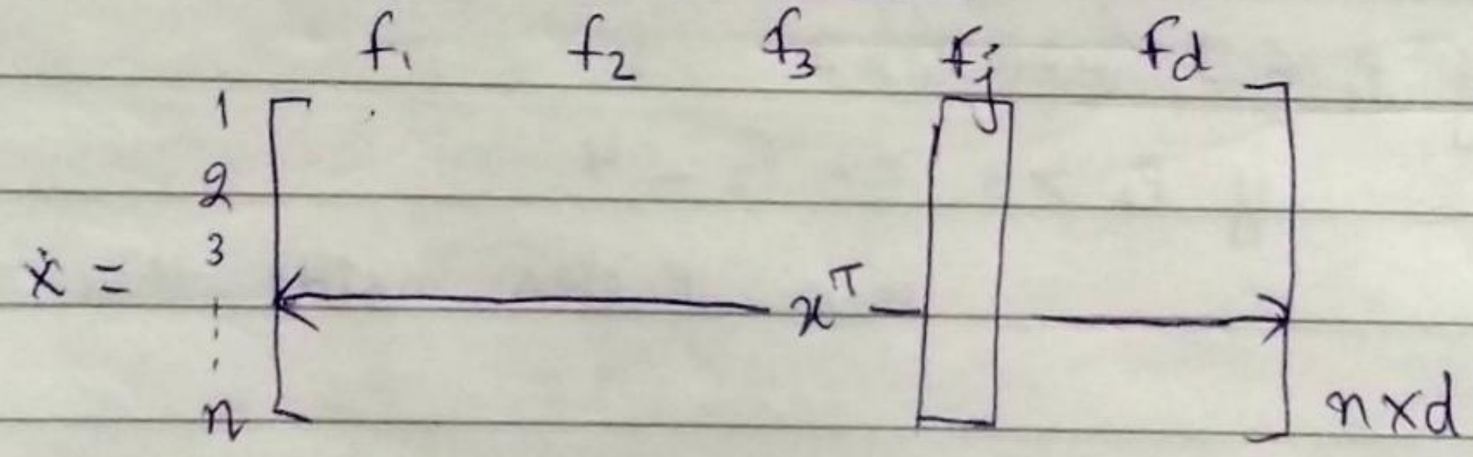
Iris :→

$D = \{x_i, y_i\}^n$ → data points (x_i → data points, y_i → class labels)

$x_i \in \mathbb{R}^d$; $x_i \in \mathbb{R}^4$

$y_i \in$ {setosa, versicolor, viginica}

* Dataset as a data-matrix :→



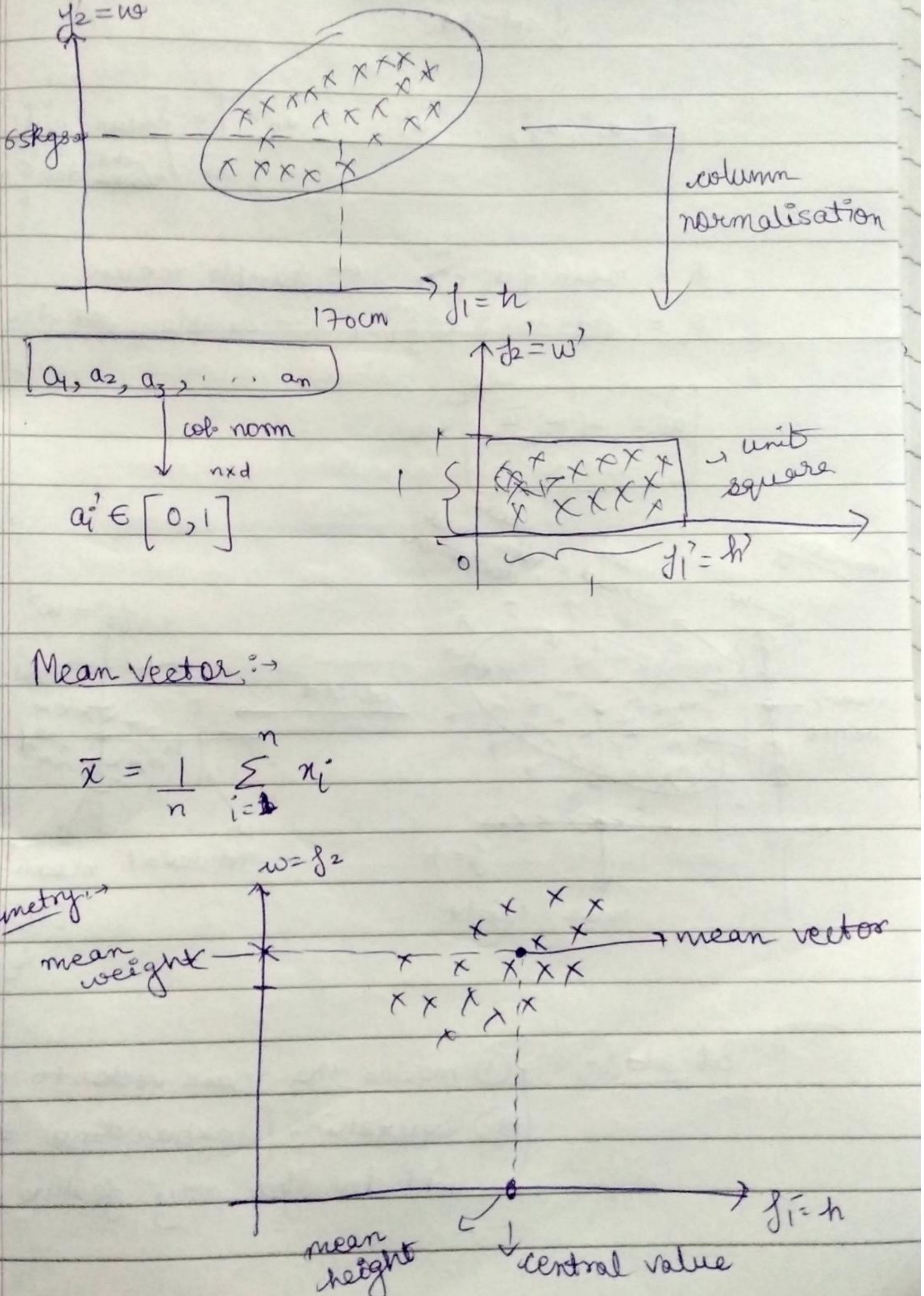
each datapoint = row

each column = feature.

Or we can also represent it into d x n representation.

* Data Pre-processing : column normalization :→

Squashing the data into unit cube or unit cuboid.



* Mean Vector :→

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Geometry :→

Date : / /
Page No.

* Data preprocessing - Column standardization

↓

more often used

$[a_1, a_2, a_3, \ldots, a_n]$ → n values of f_j

↓ col-std

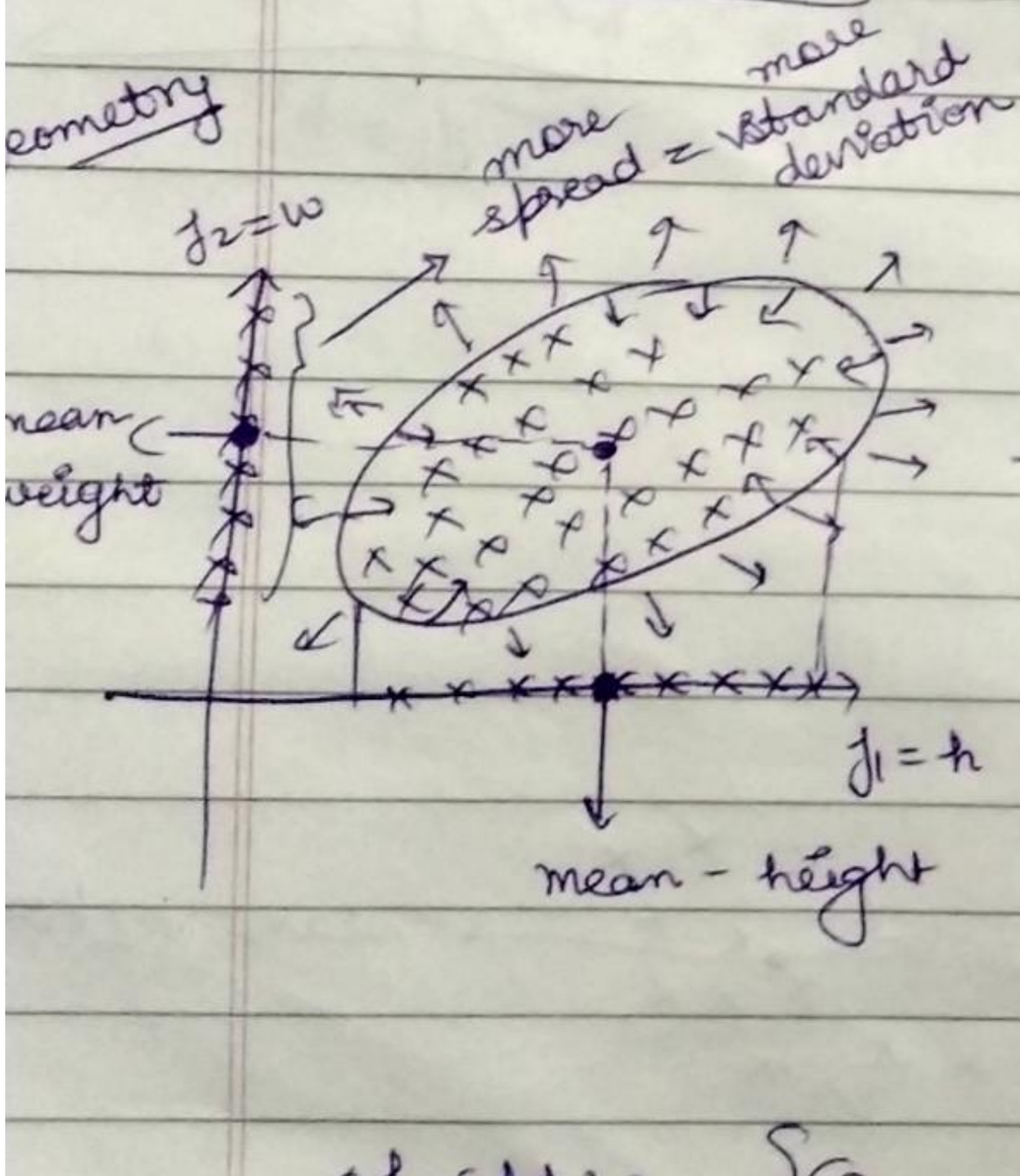
$[a_1', a_2', a_3', \ldots a_i', \ldots a_n']$ ← $\{\text{mean}\{a_i'\}_{i=1}^{n} = 0 \;\&\; \text{std-dev}\{a_i'\}_{i=1}^{n} = 1\}$

$\bar{a} = \text{mean}\{a_i\}_{i=1}^{n}$ ← Sample mean

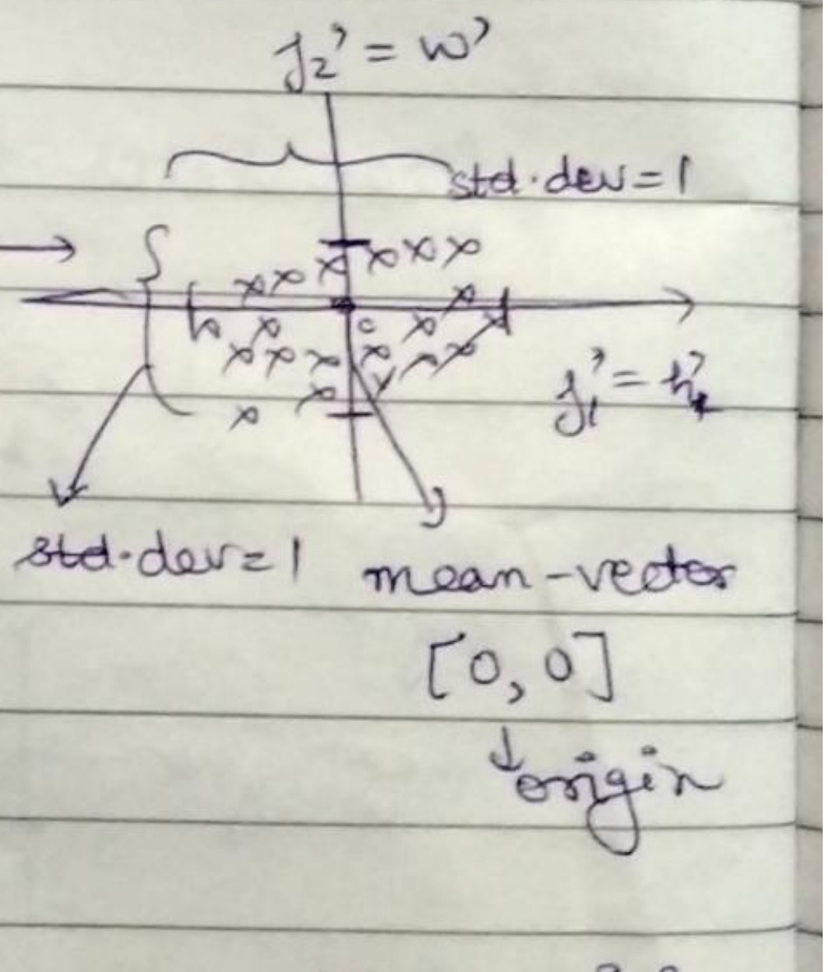
$s = \text{std-dev}\{a_i\}_{i=1}^{n}$ ← sample std-dev.

$$a_i' = \frac{a_i - \bar{a}}{s}$$

Geometry



col-std



col-std :→ { ① moving the mean vector to origin
② Squashing / expanding so that std-dev for any feature is 1.

Date : / /
Page No.

col-standardization → mean-centering → origin
+ scaling
↓
ensuring that std-dev for all the features is 1.

* Co-variance Matrix :→

$$cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$

S_{ij} → cov matrix = $cov(f_i, f_j)$

$cov(f_i, f_j) = cov(f_j, f_i)$

→ It is a symmetric matrix as $S_{ij} = S_{ji}$ and diagonal elements are variance of features.

→ It is a square matrix, as it has d rows and d columns.

$$cov(f_1, f_2) = \frac{1}{n}\sum_{i=1}^{n} x_{i1} * x_{i2}$$

$$cov(f_1, f_2) = (f_1^T f_2) * \frac{1}{n}$$

$f_1^T f_2$ → dot product.

So, if f_1 & f_2 have been column standardised,

$$cov(f_1, f_2) = \frac{f_1^T f_2}{n}$$

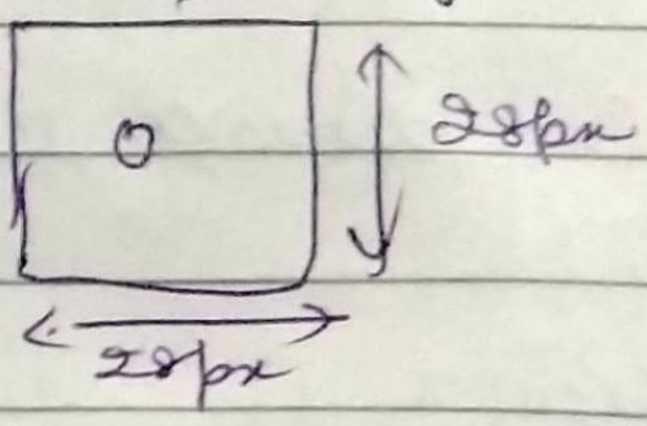
Date : / /
Page No.

$$S_{d\times d} = \underset{d\times n}{X^T} \star \underset{n\times d}{X}$$ → if x has been col. std.

* MNIST ~~dataset~~ digits :→ Dataset for visualising

→ image

$D = \{x_i, y_i\}_1^{60K}$, x_i :



$y_i = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$