13.1 what is dimensionality Reduction

we can Visualize data in 2D, 3D using scatterplot

for 4D, 5D, 6D we use pair plot

what about 10-D, 100-D, 1000-D, How do we visualize & understand data?

we will try to reduce the data from nD → to 2-D & 3-D

we use PCA & t-SNE

13.2 Row Vector and column Vector

for any given flow we are given 4 variables

$$\text{flower}[\underbrace{SL, PL, SW, PW}_{\text{real values}}]$$

We write the i^{th} datapoint as x_i we will represent as $x_i \in R^d$

belongs

Column Vector

Real space

dimension

$$x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{id} \end{bmatrix} \rightarrow \text{d dimensional Vector} \rightarrow \text{Colum Vector}$$

$d \times n = 1$

$$f_1 = \begin{bmatrix} 2.1 \\ 3.3 \\ 1.6 \\ 4.3 \end{bmatrix} \begin{matrix} \leftarrow PL \\ \leftarrow PW \\ \leftarrow SL \\ \leftarrow SW \end{matrix}$$

Default vector notation is a Column Vector.

$x_i = [2.1, 3.3, 4.6, 1.2]_{1\times 4}$ ← row Vector.

13.3 How to represent a dataset?

Dataset

$\mathcal{D} = \{x_i, y_i\}_{i=1}^{n}$ datapoints, class labels

$x_i \in R^d$, $x_i \in R^4$ $\begin{bmatrix} SL \\ SW \\ PL \\ PW \end{bmatrix}$ $y_i \in \{\text{Setsa, veri, verg.}\}$

$y_i \in \{\text{Setosa, versicolo, Virginica}\}$

13.4 How to represent a dataset as a matrix

$\mathcal{D} : \{x_i, y_i\}_{i=1}^{n} \quad x_i \in \mathbb{R}^d \quad y_i = \{s, vi, ve\}$

$$X = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ n \end{array} \begin{array}{c} f_1, f_2, f_3 \cdots f_d \\ \left[\begin{array}{c} \\ \\ \\ \longleftarrow x_i^T \longrightarrow \\ \\ \end{array} \right]_{n \times d} \end{array} \longrightarrow \begin{array}{c} Y \\ \left[\begin{array}{c} \\ \\ \\ \rightarrow y_i \rightarrow \\ \\ \end{array} \right]_{n \times 1} \end{array} \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ n \end{array}$$

Each data point : row , Each column represent a feature

$$X = \begin{array}{c} \\ f_1 \\ f_2 \\ f_3 \\ \vdots \\ f_d \end{array} \begin{array}{c} 1 \quad 2 \quad 3 \quad 4 \cdots i \cdots n \\ \left[\begin{array}{c} \uparrow \\ x_i \\ \downarrow \end{array} \right]_{d \times n} \end{array}$$

13.5 Data Preprocessing : Feature normalization

obtain data $\rightarrow$ Pre-processing $\rightarrow$ data modeling
↳ column normalization

what is column normalization

$\rightarrow$ Take all the value of feature j which f_j (PL)

Now we have $a_1, a_2, \ldots a_n \rightarrow$ n Values of feature j

$max(a_i)$ = maximum value of $a_i = a_{max}$

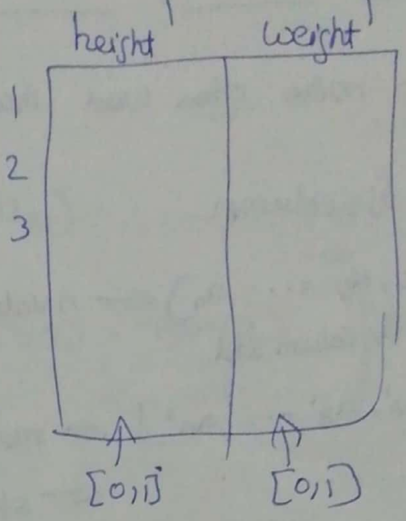
$min(a_i)$ = minimum " $a_i = a_{min}$

$$a_i' = \frac{a_i - a_{min}}{a_{max} - a_{min}}$$

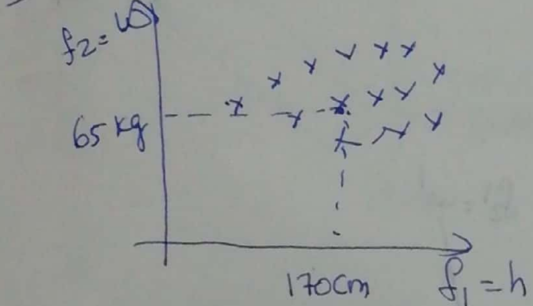
By doing this a_i will lie b/w 0 and 1 why Column normalization?

height (cm)	weight (kg)	
162	56	1
172	72	2
⋮		3
182	84	⋮
⋮		
160	58	⋮
⋮	⋮	
		n

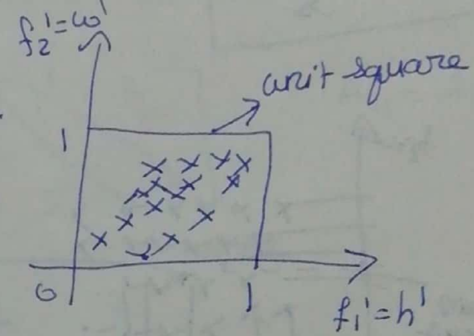
→ col normalizat (getting rid of scale) (same scale)



Geometry



→ col normalizat



any where ndim space $\xrightarrow[\text{normalization}]{\text{Column}}$ unit hype cube in the same n-dimensional space

13.6 Mean of a data matrix

Let us Consider we have two vectors $x_1; x_2$

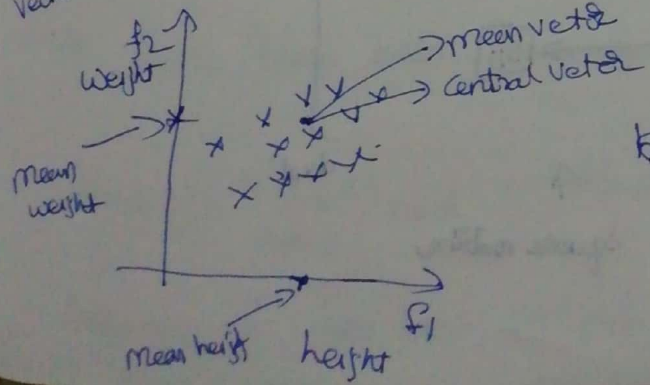
$$x_1 = \begin{array}{cc} f_1 & f_2 \\ [2.2, & 4.2] \end{array} \in \mathbb{R}^2$$

$$x_2 \; [1.2, \; 3.2] \in \mathbb{R}^2 \qquad x_1 + x_2 = [3.4, 7.4]$$

mean vector

$$\bar{x} \in \mathbb{R}^d = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{n} (x_1 + x_2 + \cdots + x_n)$$

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



$$\bar{x} = [h_{\bar{x}}, w_{\bar{x}}]$$

$$h_{\bar{x}} = \text{mean}(h_i)_{i=1}^{n}$$

$$w_{\bar{x}} = \text{mean}(w_i)_{i=1}^{n}$$

13.7 Data Preprocessing : Column Standardization.

Col. std is more often used than the normalization.

Consider f_j column

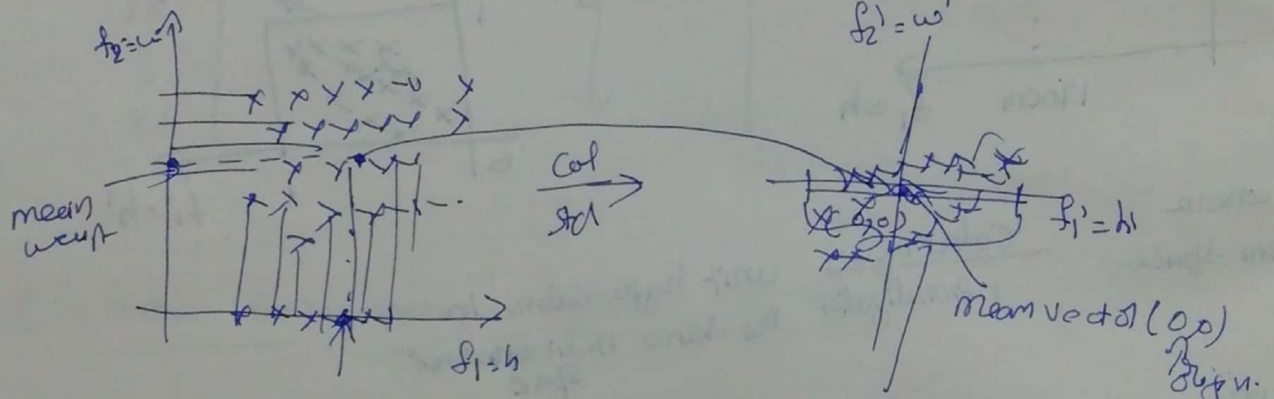
$[a_1, a_2, a_3 \cdots a_n] \leftarrow$ n values of f_j

$\downarrow$ Column std

$[a_1', a_2', a_3' \cdots a_n'] \leftarrow$ mean of $\{a_i'\}_{i=1}^{n} = 0 = \bar{a} \leftarrow$ sample mean

$\leftarrow$ std dev of $\{a_i'\}_{i=1}^{n} = 1 = s \leftarrow$ sample mean.

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



Std on any axis = 1

(1) moving the Mean vector to origin

(2) squashing / ~~Expand~~ Expanding the point.

13.8 Co-variance of a Matrix

$$X = \begin{array}{c} \\ 1 \\ 2 \\ \vdots \\ x_i \\ n \end{array} \overset{f_1, f_2 \cdots f_j \cdots f_d}{\begin{bmatrix} \\ \\ \\ \leftarrow x_i^T \rightarrow \\ \\ \end{bmatrix}}_{n \times d} \quad ; \quad S = \text{cov-matrix of } X \quad \begin{array}{c} \\ i \end{array} \overset{j}{\begin{bmatrix} & & \\ \rightarrow & \boxed{S_{ij}} & \\ & & \end{bmatrix}}_{d \times d}$$

$\uparrow$ Square matrix

S_{ij} = i^{th} row & j^{th} column Element in S

$S_{ij} = \text{cov}(f_i, f_j)$

$i \to 1 \to d$

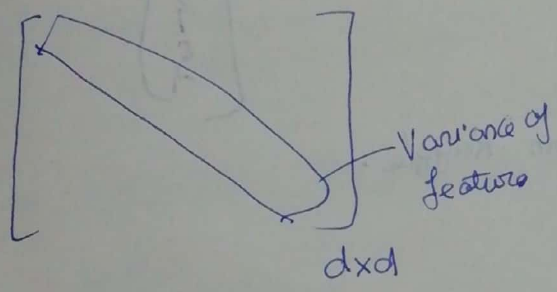
$j : 1 \to d$

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

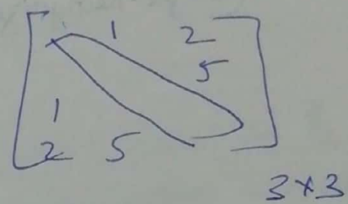
→ $\text{cov}(f_i, f_j) = \text{Var}(f_i)$ or $\text{cov}(x, x) = \text{Var}(x)$

→ $\text{cov}(f_i, f_j) = \text{cov}(f_j, f_i)$

Properties of cov



→ Symmetric matrix



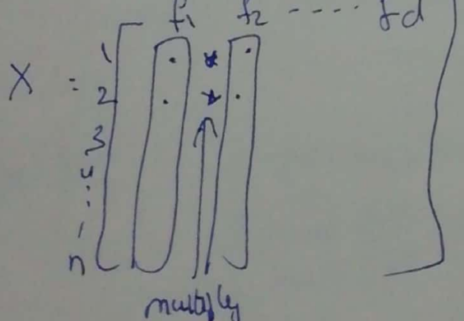
$A_{21} = A_{12}$

Let X is col standardized ⇒ mean $\{f_i\} = 0$
std-d $\{f_i\} = 1$

$$\text{what is cov of } (f_1, f_2) = \frac{1}{n} \sum_{i=1}^{n} (x_{i1} - \mu_1)(x_{i2} - \mu_2)$$

μ_1 → mean(f_1) → 0; μ_2 → mean (f_2) → 0

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



$\text{cov}(f_1, f_2) = (f_1^T f_2)(1/n)$

If f_1 & f_2 have been standardz $\text{cov}(f_1, f_2) = (f_1^T \cdot f_2)(1/n)$

$$S_{d \times d} = \frac{1}{n} \underbrace{(X^T)}_{d \times n} (X)_{n \times d} = d \times d$$

Assuming X has been col-std.