Feature engineering :-

- continuous - scaling
  - standard scaler
  - min max scaler
- discrete - encoding
  - ordinal (ranked)
    - large, big, small,
    - Tier 1, Tier 3, Tier 2
    - high, medium, low
    - ↓
    - label encoder
  - nominal (unranked)
    - male, female
    - ⋮
    - ↓
    - one-hot-of dummies.

$$\bar{x} = \frac{x - \mu}{\sigma}$$

{mean = 0
variance = 1}

$$\bar{x} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

{max = 1, min = 0}

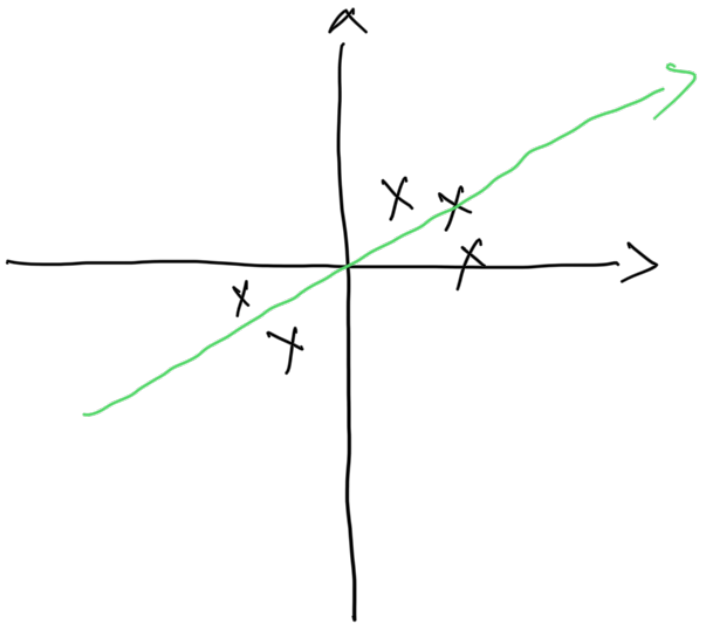
Dimensionality Reduction :-

Dimension → No. of features
input shape

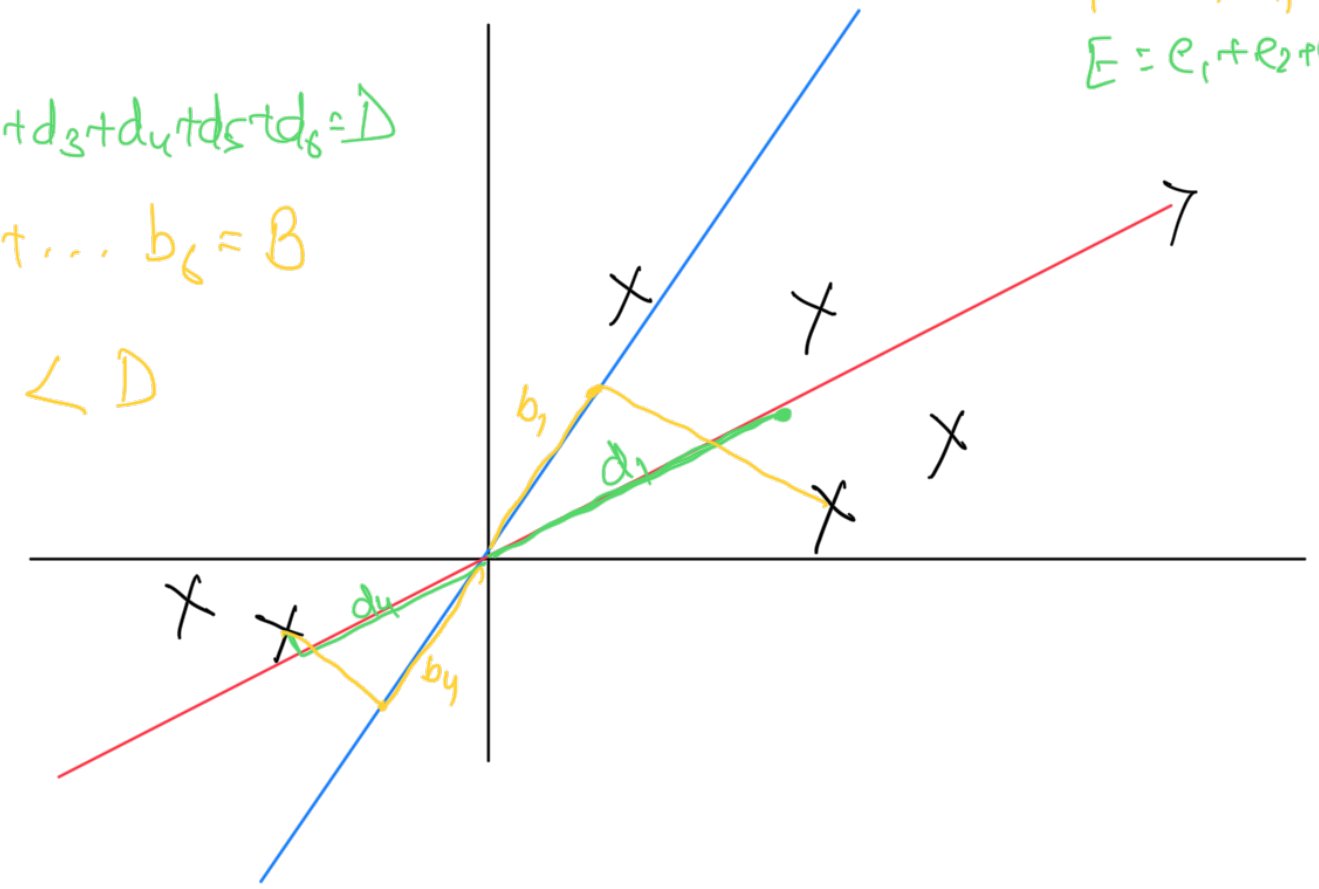
1) Feature selection.
2) PCA X (Feature summarization).

principal component analysis :-

2d → 1d



1. Scale to origin
2. Best fit line.



$d_1 + d_2 + d_3 + d_4 + d_5 + d_6 = D$

$b_1 + b_2 + \dots b_6 = B$

$B < D$

$F = f_1 + f_2 + f_3 + f_4 + f_5 + f_6$

$E = e_1 + e_2 + e_3 + e_4 + e_5 + e_6$

$E < F$

SSE(E) < SSE(F)

A best fit line is one

a. whose SSE is less.

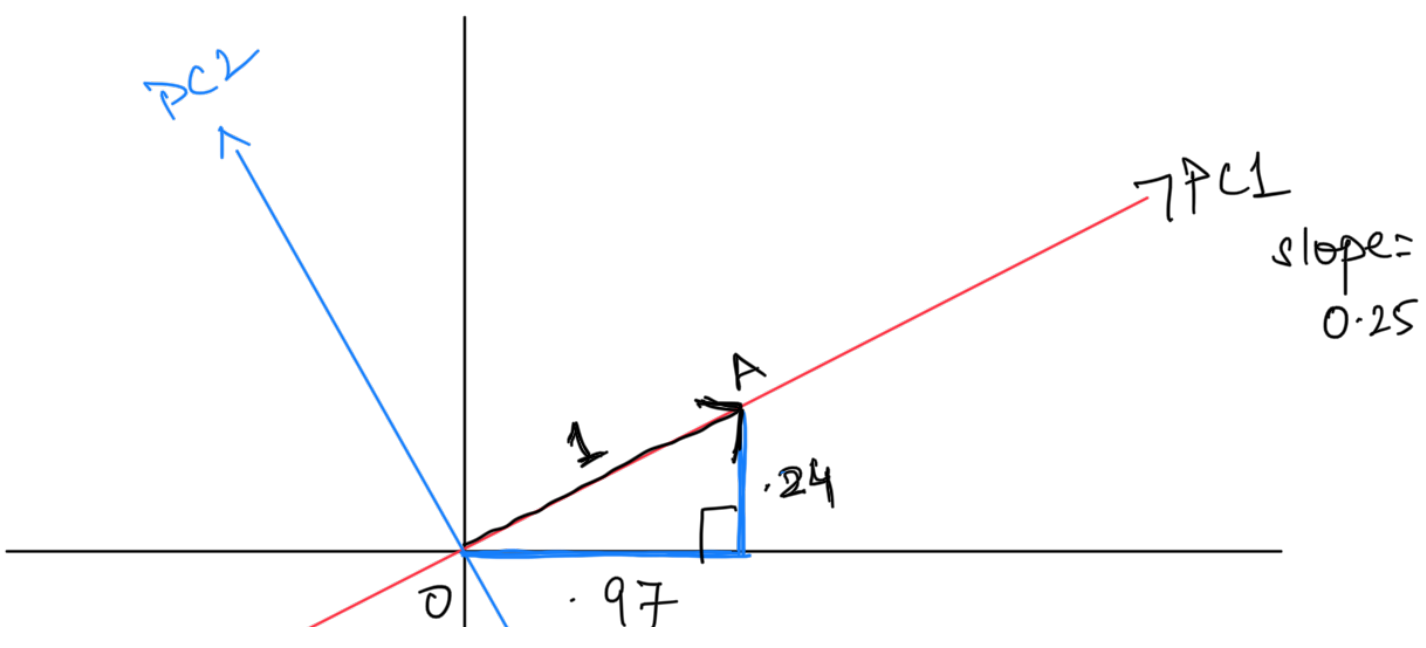
$E < F$ ∴ Red is best fit line.

b. whose (SSD) is more (variance)

$B < D$ ∴ Red is best fit line.

$d_1^2 + d_2^2 + d_3^2 + \dots d_n^2 = SSD$

This Best fit line is PC1

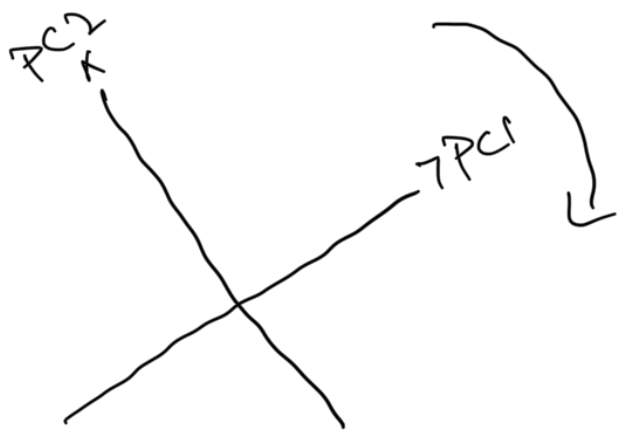

PC2
PC1
slope:
0.25
A
1
.24
.97
0

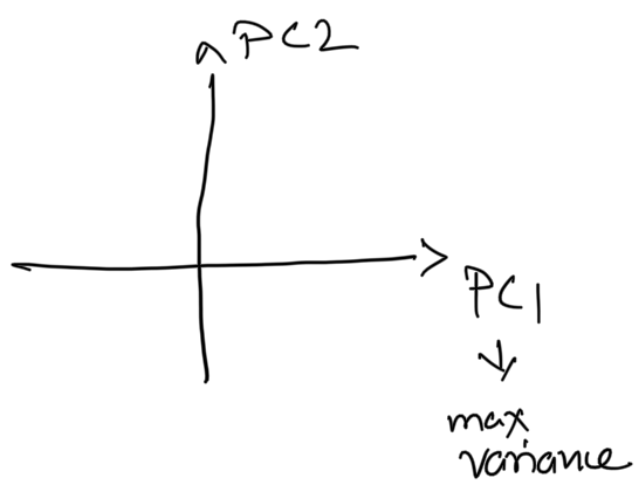
$4x + 1y \Rightarrow$ linear combination.

$\overrightarrow{OA}$ → unit vector
or
Eigenvector of PC1

and

$d_1^2 + d_2^2 + \dots d_n^2$ = eigenvalue of PC1





variation of PC1 = $\frac{d_1^2 + \dots d_n^2}{n-1} = 15$ → 15/18 %

" PC2 = $\frac{b_1^2 + \dots b_n^2}{n-1} = 3$ → 3/18 %

10 dim ⟶ 10 PC's. Plot variance ratio

[0.28 0.15 0.10 0.08 . . .] 1×10

> > > > >

(0.28 + 0.15 + 0.10 + 0.08)

n_components = 0.80

↓

10 PCs it will give x PC's < 10

s.t Σ [illegible]

(1) Scale origin

(2) find best fit line

min error or
max var

(3) PC1 is best fit line

(4) Normalize PC1
↳ ev of PC1
↳ evalue of PC1

(5) PC2 orthogonal to PC1

(6) Rotate the PC1 and PC2

(7) Calculate variance ratio

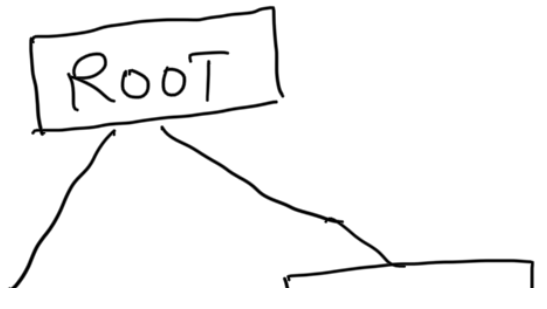
PC1 > PC2

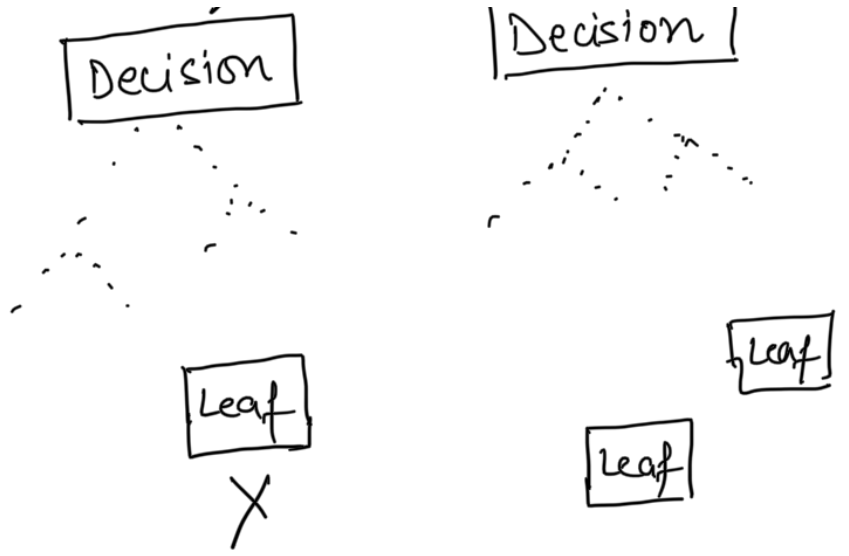
(8) Using a threshold select no. of PC's.

Total dim > no. of PC's

2 → 1
20 → 5/4/7
200 → 2/20/100..

Tree based methods.



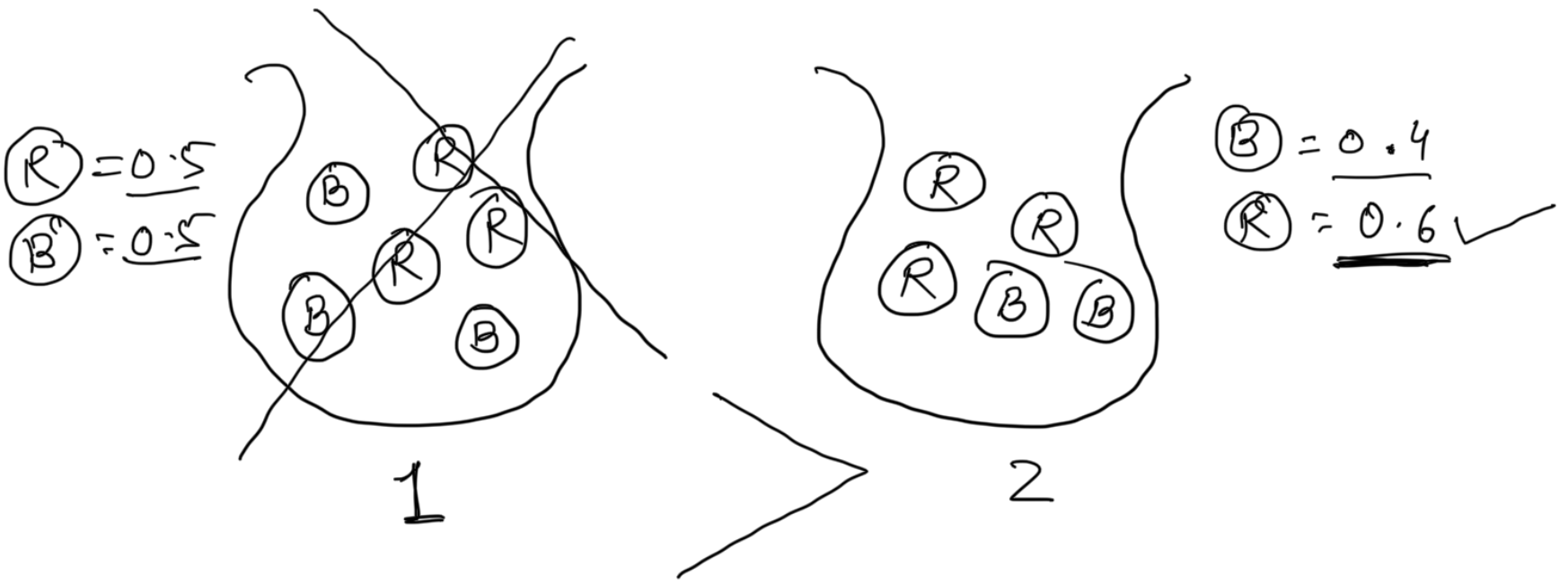


(1) What are these nodes and how do we split??

(2) The order of split...

Entropy

measure of impurity / uncertainity

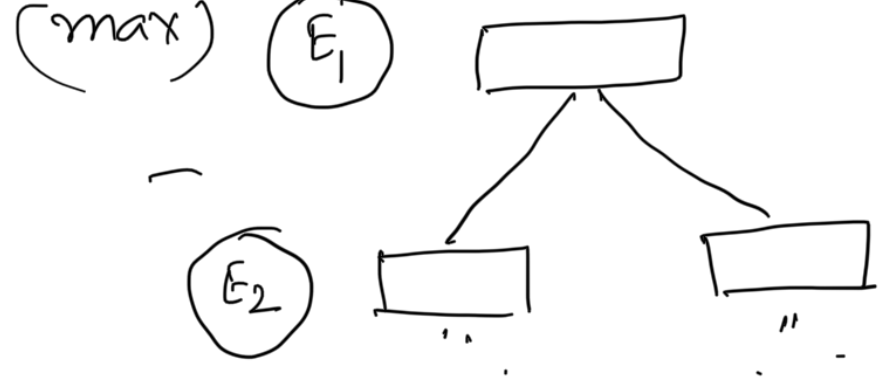


(1) Choose any bag, any color (Red | Blue)

(2) If you draw a ball of the chosen color from the chosen bag you win 10,000 $.
Which color, bag ??

Entropy

(max) E_1



$E_1 - E_2 = IG$

$IG \propto \frac{1}{E}$

$E_2 < E_1$

E_2

E_3

$\vdots$

E_n

$E_1 > E_2 > \dots E_n$

$$E = -\sum_{i=1}^{c} p_i \log_2 p_i$$

Initial Entropy $(9+, 5-)$

① ②

$$E = -\sum_{i=1}^{2} p_i \log_2 p_i$$

$$= -\left[p_1 \log_2 p_1 + p_2 \log_2 p_2\right]$$

$$= -\left[\frac{9}{9+5} \log_2 \frac{9}{9+5} + \frac{5}{9+5} \log_2 \frac{5}{9+5}\right]$$

$$= -\left[9/14 \log_2 9/14 + \frac{5}{14} \log_2 5/14\right]$$

$$= 0.940$$

$E_1 = 0.940$

Hum

Windy.