Feature Engineering

DQ3 2:16 pm Tue, Apr 15

- Feature Transformation
  - missing value Imputation
  - Handling Categorical Features
  - Outlier Detection
  - Feature Scaling
- Feature Construction
  - Feature splitting
- Feature Selection
- Feature Extraction
  - PCA

DQ24 24

learn from Train data only | scale both train & test

Feature Scaling

- Standardization
  - Z-score Normalization
- Normalization
  - min max scaler
  - Robust scaler (outlier good)

when

- K-means
- KNN
- PCA
- ANN Artificial Neural Network
- Gradient descent → find minima

	x_i
x_1	15
x_2	27
x_3	10

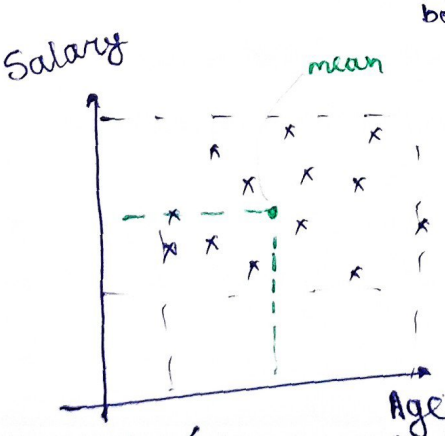
$$x_i' = \frac{x_i - \bar{x}}{\sigma}$$ ($\bar{x}$ → mean, σ → SD)

let $= 10\ \bar{x}$, $\mu = 30$

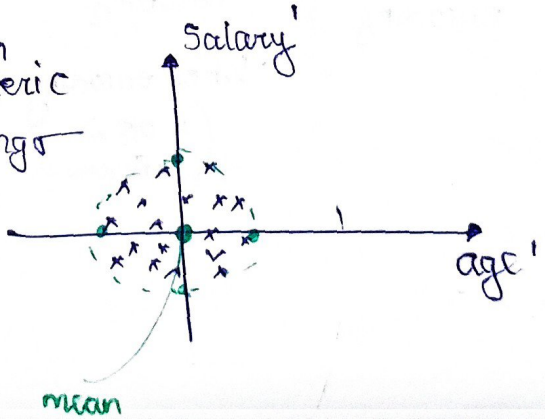
$$x_1' = \frac{15-30}{10}$$

$$x_2' = \frac{27-30}{10}$$

New Series: $\mu = 0$ (mean both) $\quad \sigma = 1$



- mean centeric
- Scaling σ



Normalization

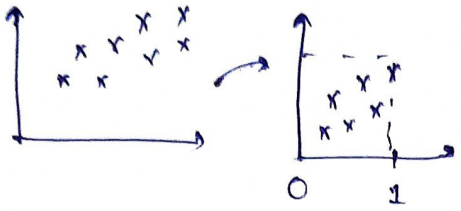
DO25 314

- min max Scaling
- Mean normalization
- max absolute
- Robust scaling ... more ex see

Elimate the unit.
makes into common scale.

$$X_i' = \frac{X_i - X_{min}}{X_{max} - X_{min}}$$

$x_i' \in [0, 1]$



Train data, but Transform both train & test

mean Norm.

$$x_i' = \frac{x_i - x_{mean}}{x_{max} - x_{min}}$$

$x_i' \in [-1 \text{ to } 1]$

no class
used in centeric data

Max Abs Scaling

$$X_i' = \frac{X_i}{|X_{max}|}$$

used → Spare data
when zero in more

Robust scaling

wt
200
300
100

$$X_i' = \frac{X_i - X_{median}}{IQR}$$

$\{75^{th} - 25^{th}\}$

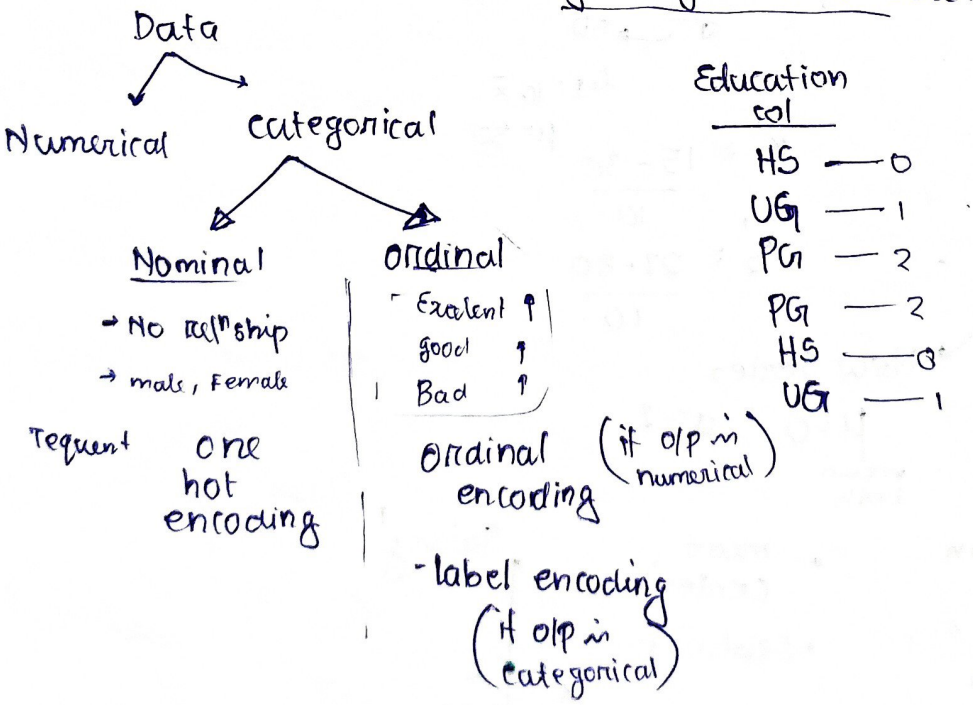
when outlier in more.

Norm V/s Stand.

1. Is feature scaling rqd ?
2. Stand is best
3. if you know min, max used MinMax
   Outlier Robust
   Spare MaxAbs

DO26 334

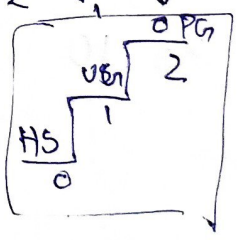
Encoding Categorical variables



Education col

HS — 0
UG — 1
PG — 2
PG — 2
HS — 0
UG — 1

PG > UG > HS
↓ original
2



age gender review education Purchased

Numerical ↑ nominal ↑ ordinal ↑ o/p ↑ label

one Pipeline Column Transformer

One Hot Encoding

D027

3:59 pm

Color	Target
Yellow	0
Y	1
Blue	1
Red	0
B	1
R	1

Y	B	R	Target
1	0	0	0
1	0	0	1
0	1	0	1
0	0	0	0
0	1	0	1
0	0	1	1

if we do like this

YBR	ML think
0 1 2	R > B > Y

But we convert into vector

1,0,0	→ Y
0 1 0	B
0 0 1	R

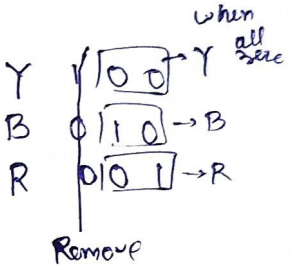
Dummy Variable Trap

I/p column must be independent for o/p

Multicollinearity

- no mathmetical Relationship

n categories
(n-1) categories



Column Transformer

D028 439

```
from sklearn.compose import ColumnTransformer

transformer = ColumnTransformer(transformers=[
        ('tnf1', SimpleImputer(), [fever]),
        ('tnf2', OrdinalEncoder(categories=[['mild','strong']]), [cough]),
        ('tnf3', OneHotEncoder(sparse=False, drop='first'), ['gender','city']),
], remainder='passthrough')

transformer.fit_transform(X_train)
transformer.transform(X_test)
```

name — object — tablecolumn

→ other col as it is

if you are not train the model then used fit_transformer.

if train the model used fit()

ML Pipeline


```
from sklearn import set_config
set_config(display = 'diagram')
```

Mathmatical Transformations

D030

Need Normal Distribution

↳ pd.skew()

sns.kdeplot()

QQ plot

Sklearn

Func Trans / [power] / Quantile

log Trans:	Box-Cox
Reciprocal	yeo-Johnson
power(sq / sqr)	

Log Transform

- Right skewed distribution

-don't use in -ve

log | log1p

x | x+1 (add each value w/ 1)

Reciprocal ($\frac{1}{x}$) | sq(x^2) | sqrt($\sqrt{x}$)

left skewed

Box-Cox Transform

D031

lambda

$$x_i^{(\lambda)} = \begin{cases} \frac{x_i^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln(x_i) & \lambda = 0 \end{cases} \qquad n>0$$

$\lambda \in [-5, 5]$

↳ maximum likehood

Basiyen stat

Yeo-Johnson Transform

$$x_i^{(\lambda)} = \begin{cases} [(x_i+1)^{\lambda}-1]/\lambda & \lambda \neq 0,\ x_i \geq 0 \\ \ln(x_i)+1 & \lambda = 0,\ x_i \geq 0 \\ -[(-x_i+1)^{2-\lambda}-1]/(2-\lambda) & \lambda \neq 2,\ x_i < 0 \\ -\ln(-x_i+1) & \lambda = 2,\ x_i < 0 \end{cases}$$

D032

Encoding Numerical Features

- Discretization (Binning)
- Binarization

Binning

- To handle Outliers
- To Improve the value spread

Binning
- unsupervised
  - Equal width (uniform)
  - Equal freq. (Quantile)
  - K-means
- sequensal
  - Decision Tree Binning
- Custom Binning

Equal width/uniform Binning

bins = 10

$\frac{max - min}{bins}$

23, 36, 84, 56 ...

(0-10) (10-20) ... (90-100)

bins 5 16 ... 5

Equal freq/Quantile Binning

uniform distribution

interval = 10

Each interval contains 10% of obs.

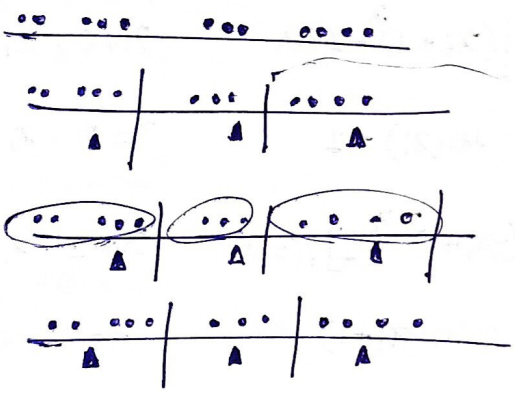
0-16, 16-20, 20-22 ...

no fixed width

10 4 2

K means Binning

centroid = 3 ~~apply~~ when numerical Discreate



1) Randomly take 5 centroids
2) find each point dist. w/ each centroid ~~of~~ then make into cluster.
3. or we can bisect the blw two ~~clust~~ centroid.
4. shif all centroid into mean
5. Repeat all until same

App

KBins Discreate()

- Bins = ?
- Strategy
  - uniform
  - quantile (def)
  - Kmeans
- encoding
  - Ordinal
  - One hot (def) encodig

Binarization

use during colour image → Black white

How EDA helps to disside how many bin in great pro accuracy?

Custom Domain → use Busines knowledge of pandas

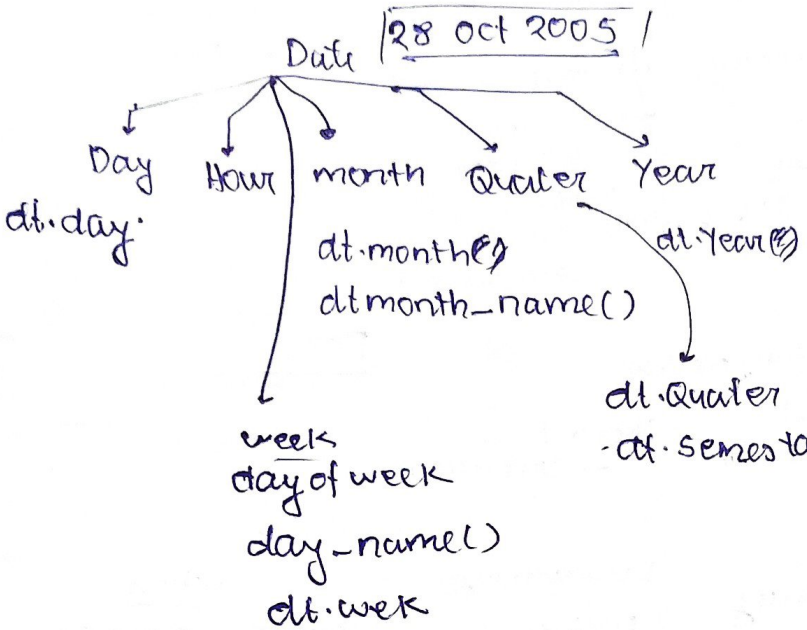
D033

10
Tue, Apr 17

Mixed Data (two types)

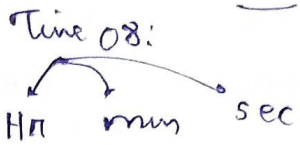
Cabin	→ two col: class	num
B5	B	5
C23	C	23
D41	D	41

	→	class	num
7		7	N/A
3		3	N/A
A		N/A	A
G1		N/A	G1



D034

Time 08:



11 09 am
FRI, Apr 18

1) convert into Datetime type

D035 8:51 pm

Missing Value

- Remove
  - Complete case Analysis (CCA)
- Impute
  - Univariate
    - numerical
      - mean/median
      - Random
      - Arbitrary
      - End of distribution
    - categorical
      - mode
      - missing Indicator
      - Automatic Selection (Both num cat)
  - Multivariate
    - KNN Impute
    - Iterative Impute MICE

CCA

Assumption complete at missing Randomly

MCAR

when

1. MCAR
2. 5% <

D036

1. mean / median

- Normal distribution
- Skewed distribution

Benfits

1. Simplicity

Disadvantages

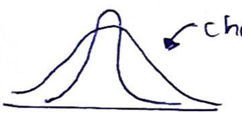
1. Shape of distribution
2. Outliers
3. co variance (change the Reln w/ other col)

when

1. MCAR
2. 5% <

Arbitrary

Create difference w/ other data in same column



when MCAR

not missing at Random

Normality

$\mu + 3\sigma$

$\mu - 3\sigma$

skewed

$Q_1 - 1.5\,IQR$ 25th

$Q_3 - 1.5\,IQR$ 75th

D037

* most frequenty value Imputation
  - replace w/ mode
  - because more efficient in

* Missing Category Imputation

  Replace NaN w/ "Missing" word

D038

11:43 AM
Bu

Random Imputation

- It take the train data on server
- If we use random Imputation then the for same i/p o/p an Imputation must be same

Ex: Fare = 50, Age missing 40 ← Randomly Created
again 50 → 44 ✗ → wrong model predict
50 → must be 40 same as before

Missing Indicator

Age	Fare	Indicator
27	31	F
41	35	F
NA	31	T
62	39	F

KDD
simple Indicator (add_indicator = True)
Rather than using missing Indicator

D039

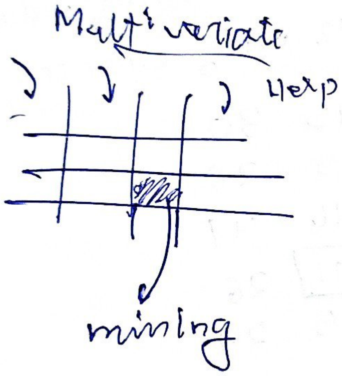
KNN

nan_euclidean

dist(x, y) = $\sqrt{wt * sq.\ dist\ from\ present\ co\text{-}ordinate}$

weight = $\frac{\text{Total \# of co-ordinate}}{\text{\# of present co-ordinates}}$

Multivariate



slno	F-1	F2	F3	F4
1	33	☐	67	21
2	☐	45	68	12
3	23	51	71	18
4	40	☐	81	☐

F2, F3, F4 for (2, F1)
But present only F3, F4 → $\sqrt{\frac{3}{2}\left[(68-67)^2 + (12-21)^2\right]}$

For $\sqrt{\frac{3}{3}\left[(51-45)^2 + (71-68)^2 + (18-12)^2\right]}$

calculate (4, F2)
present F1 F3 F4 → $\sqrt{\frac{3}{2}\left[(40-23)^2 + (81-71)^2\right]}$

KNN Imputer (weight = 'distance' or 'uniform')

1. more Accurate

distance (10) 1 → 2, (50) 2 → 3

$\frac{\left(\frac{1}{10} \times 33\right) + \left(\frac{1}{50} \times 23\right)}{2}$ ← $\frac{1}{...}$

(1,2) (2,3)
$\frac{33 + 23}{2}$

DO40

Iterative Imputer / MICE

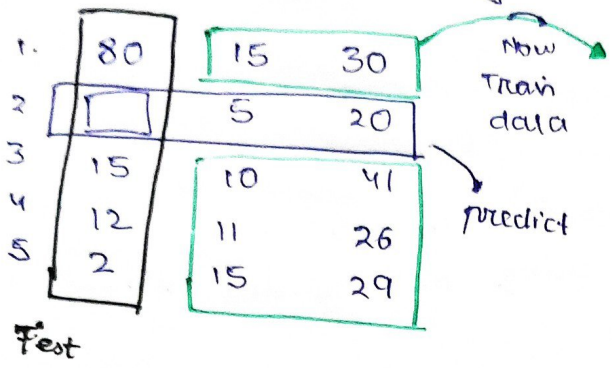
Multivariate Imputation by chained eqn

MCAR
MAR
MNAR

Assumption / accurate
MAR
memory: slow

S1. Fill NAN w/ mean of same col.

S2. Remove all col missing values

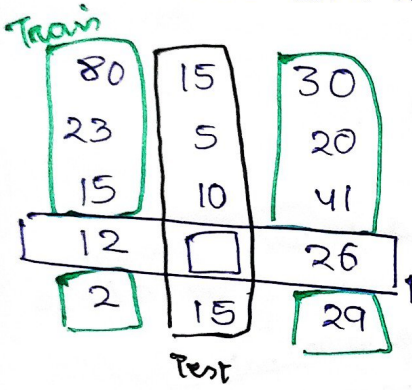


15	30
10	41
11	26
15	29

S3. Predict the missing values of col1 using other cols.

80	15	30
23	5	20
15	10	41
12	11	26
2	15	29

S4. Remove col2 missing values



Train data

80	30
23	20
15	41
2	29

S5. Predict the missing values of col2 using other col

80	15	30
23	5	20
15	10	41
12	11	26
2	15	29

⋮

so on . .

DOM

Outliers

what are Outliers?

when is outlier dangerous?

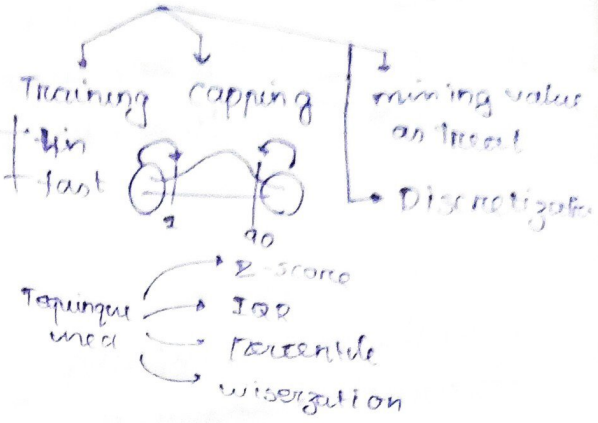
- Anamonally. outlier detection.

what to do with outlier?

Effect of outliers

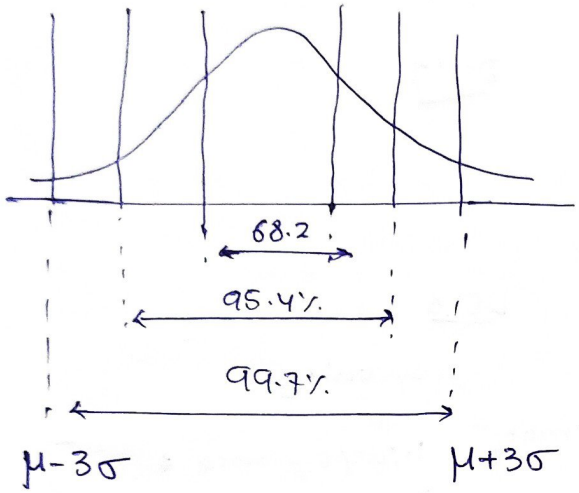
- Does we work w/ weight?

How to treat Outliers?

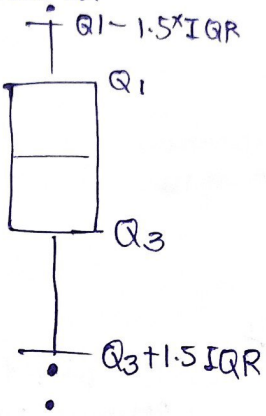


How to detect Outliers

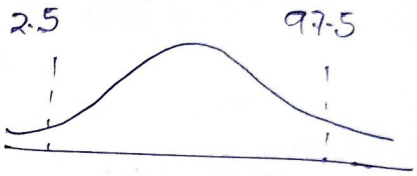
① If data in Normal dist



② Skewed



③ Other distributions
Percentile based



D042

Outlier removal using Z score

Z-Score

$$X_i' = \frac{X_i - \mu}{\sigma}$$

$\mu-\sigma$, $\mu+\sigma$ → 68.2%

$\mu-2\sigma$, $\mu+2\sigma$ → 95.4%

$\mu-3\sigma$, $\mu+3\sigma$ → 99.7%

Trimming → Remove all

Capping → 85 , 90 / 80 , 80 (covers)

D043/44

percentile method capping is called ~~mingation~~ winsorization

D045

Feature construction

Domain Knowledge v/s Experience on data

D046

Curse of Dimensionality

features

After a certain point the no of features in ~~use~~ not improve the model efficiency

How to find the optimal Dimensionality

sparsity

1. performance &
2. computation

Dimensionality Reduction

Feature selection → Feature Extraction

1. Forward selection
2. Backward elimination
3. Bidirection elimination

PCA
LDA
tsene

PCA

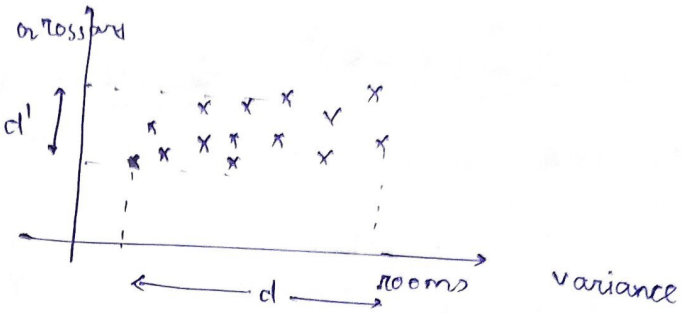
DO47 7:22pm Sun Apr 20

It is tecquine which can help to reduce higher dimensions to lower dimensio w/o lossing essence of data.

Benifits
1. faster execution
2. visualization

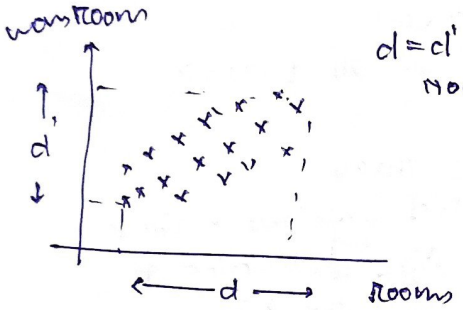
Geometric Intution

No of rooms	No of grossary shops	price ($)
3	2	60
4	0	130
5	6	170
2	10	90

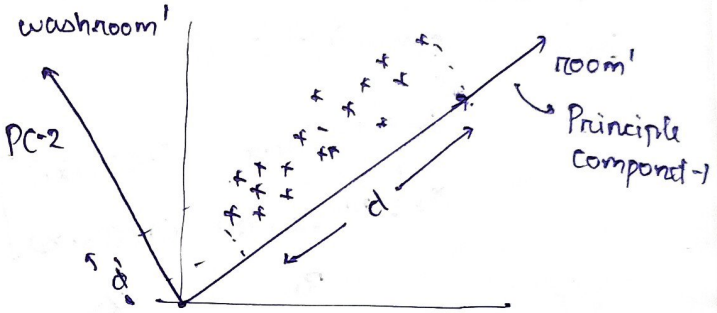


d > d' so I keep room colums

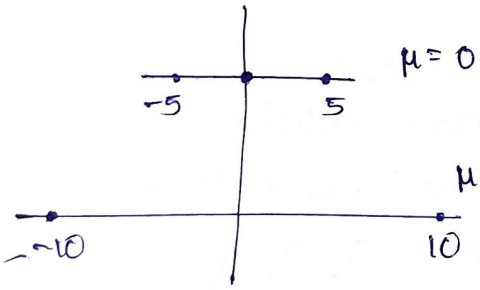
No of rooms	No washrooms	Price



d = d'
Now what ?

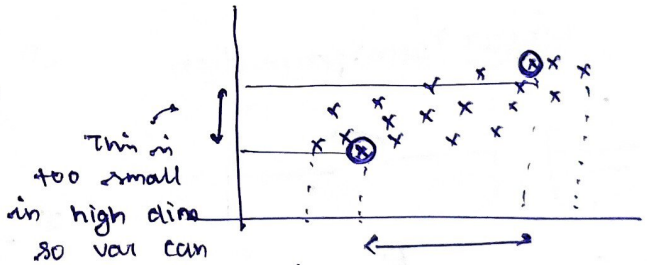


why variance is imp ?



$\mu = \frac{-10+0+10}{3} = 0$

Due to not mis match w/ other point distance we can't seen

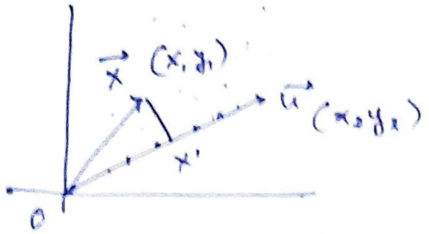


then how to compare

$$\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}$$ variance

$$var_1 = \frac{25^2 + 25^2}{3} = \frac{50}{3}$$

$$var_2 = \frac{100 + 0 + 100}{3} = \frac{200}{3}$$



$\vec{x}$ → point

$\vec{u}$ → unit vector

$\overline{ox'}$ projection : $\frac{\vec{x} \cdot \vec{u}}{|u|} = \vec{x} \cdot \vec{u} = u^T x$

$= \vec{x} \cdot \vec{u} = [x_1\ y_1]\ [x_2\ y_2]$

$[x_1\ y_1] \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = x_1 x_2 + y_1 y_2$

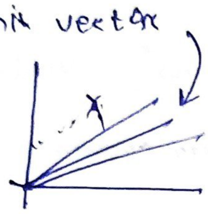
= Scalar

$[u^T x_1]\ [u^T x_2] \cdots [u^T x_n]$

$\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$ $\quad \frac{\sum_{i=1}^{n} (u^T x_i - u^T \bar{x})^2}{n}$ = variance

→ optimal ~~problem~~ func

Now PCA find the Max unit vector which line is best



sion → Relate cosent

$$\begin{array}{c|ccc} & x & y & z \\ \hline x & V_x & C_{xy} & C_{xz} \\ y & C_{yx} & V_y & C_{yz} \\ z & C_{zx} & C_{zy} & V_z \end{array}$$

↳ Spread orinlention of data

covariance matrix

$x_1 \mid x_2$ cov matrix [2×2]

$$\begin{array}{c} x_1 \\ x_2 \end{array} \begin{bmatrix} cov(x_1, x_1) & cov(x_2, x_1) \\ cov(x_1, x_2) & cov(x_2, x_2) \end{bmatrix}$$

$x_1 \quad x_2$

$$= \begin{bmatrix} var x_1 & cov(x_2 x_1) \\ cov(x_1, x_2) & var(x_2) \end{bmatrix}$$

↳ symatric

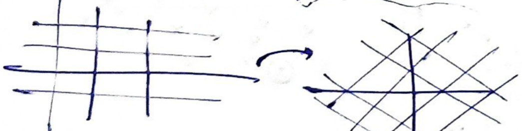
∵ $cov(x_1, x_1) = var(x_1)$

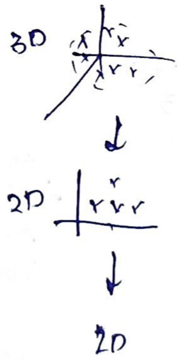
∵ $cov(x_1, x_2) = cov(x_2, x_1)$

Steps theory

1. find covariance matrix
2. Eige decomption & find the λ big value.
3. the eige vector is the unit vector

Linear Transformation





Steps

1. mean centrix
2. covariance matrix
3. find $\lambda_1\ \lambda_2\ \lambda_3$

$$\begin{bmatrix} 0.5 & 1 \\ 1 & 0.5 \end{bmatrix}$$

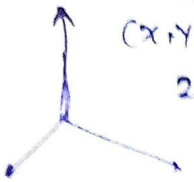
Eige vector

dir^n → same but mangitude changes

$A\vec{v} = \lambda \vec{v}$ ⟶ The largest Eige value that data have highest Eige vector & spread.

Eig vector ($\vec{v}$)

Eige value (λ) → mangitude changes.



if 100 vector (100,784) ↳ egie vector

vector

Finding optimum no. of Principle Com

$\lambda_1, \lambda_2, \lambda_3 \cdots \lambda_{784}$

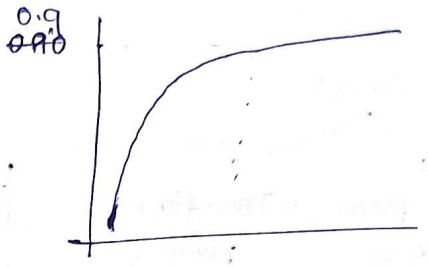
Cmin is for original egie value

but we found ^ → (90%)

so

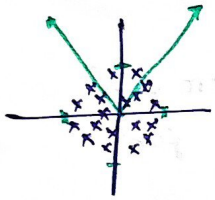
$\left(\frac{\lambda_1}{\pi\lambda}\right)$

plt.plot (np.cusum (pca.ex_var_ratio_))

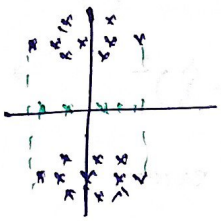


How much compone we need to explain 90% ?

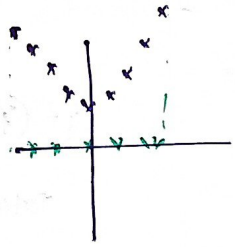
when PCA doesn't work



when data spread in same for all axis.



when projection in same



when higher dim. data lossing its originality