08/08/22

Data Science

The Great Hack

In God we trust, all others bring data

— W. E. Deming.

→ Data acts as evidence, which helps us trust others

* 3V's of Data → Velocity, Volume & Variety.

* Data Collection → Data Storage → Processing.

* OLTP vs OLAP
  - Storing data according to use

10/08/22

Case Study of Netflix →

organised competation in 2006, challenging to make recommendation system.

- In 2010, deanonimization problem; two UG students find which person watch which movie.
  - → fined.
- Throttling Issue → ISP caps the network to slow the speed of network.
  - took place in US in 2010
    - found using Data Analysis
  - { In case of higher usage, it done of Netflix. }

Data Science in the Business Context {Business Analytics}

- change is constant in business Environment.
- So, companies have to agile & take quick strategic, tactical & operational decision.

22/08/22

Mode → . not affected by Outlayers.

• Only talk abt value occuring max^m no. of times.

* Using in Election or in situation where you want to cater max^m no. of people.

Median → • Value that divides the sorted data in equal half. (50^th percentile)

• Not affected by outlayers.

↳ To get Median from data (Even → avg does not belong to data (of 2))

↳ The 50^th percentile of a list of no. is the smallest no. that is at least as large as 50% of list.

* Less Informative as not effected by Extremes; so if after median = 3 a value = 100 present ↳ Not able to give info abt this

Eg → list → 0, 2, 4, 7, 7, 12

find 50^th & 25^th percentile.

50% of 6 = 3 ⇒ Atleast 3 values should be less than the median

Median = 7

25^th percentile = 4

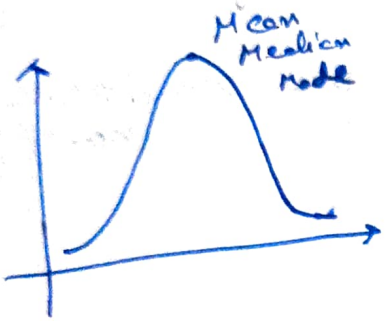
Eg → To decide the salary of a Role → Median of salary for the particular role in different competitive companies.

Mean → • Affected by the outlayer

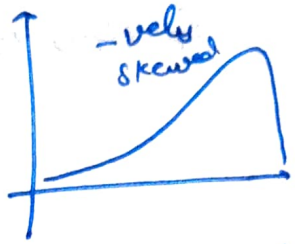
• Holistic picture of set of values.

Some questions to be Answered

① Skewness → A long tail at a particular end of graph/distribution



Normal Curve
Perfect Distribution
Not Skewed



(Many smaller values → Outlayers)

② Cross Sectional vs. Longitudinal.

↓

Can't compare.

Eg → Diff set of people of diff age grp

Measures of Location → If I know the avg. Can I answer questions related to distribution of Data.

Eg → If avg of class is 30%. What portion of students scored more than 80%

Ans → 1/3 at max

Can be answered by Markov's Inequality

→ If the list consists of only non negative numbers then the proportion of entries that are at least as large as k times the average is at most 1/k.

Q) How much of the portion of people are of greater than half of average?

Ans → 1/1/2 = 200% of people

It is correct But is quite obvious ⇐ gives loose bound Sometimes.

* So, to get the tighter bound only the mean is not enough.

Measure of Spread → How much data is deviated from mean

Standard Deviation

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

i/a rms of deviation from average.

root mean square.

** Why Square Instead Of Modulus??

Chebychev's Inequality

The proportion of entries that are k or more SDs away from the average is at most $1/k^2$.

9) Suppose the class avg is 30% & SD = 10

What portion of students scored more than 90%

$\left\{\frac{1}{36}\right\}$

24/08/22

Given that a value is greater than median. Can we predict about mean & mode

Correlation Coefficient

$$r = \frac{1}{n} \sum_{i=1}^{n} \frac{(x_i - \mu_x)}{\sigma_x} \frac{(y_i - \mu_y)}{\sigma_y}$$

Steps → ① Given X & Y & $|X| = |Y| = n$

x_i y_i

② Convert X & Y to standard unit.

- find mean (μ) & SD (σ)
- To convert to std. unit $\left\{ \frac{x_i - \mu_x}{\sigma_x} \right\}$ → How far it is from mean in terms of std. deviation

Why Std. Unit??

- also known as Z-Score (need this to compare values of diff units)
- have no unit.
- So call it Std. Unit.
- $r \in [-1, 1]$
  - 1 → Strong +ve relation
  - 0 → No Relation

③ Mult. Corresponding pairs of std. units.

④ r is avg of above products.

↓

Correlation coefficient.

* Association is not causation → Just Happening, Not like {Not Causal Relationship}! {Eg → Rain & Exam} {Due to inc of x, y is inc}
* r measures linear association
* Even one outlier might affect r.

Given mean = m & std. deviation σ

if n added to all values in list → mean = $m+n$ & Std. Dev. = σ

if all values mult by n → mean = mn & std Dev = σn

Random Variables →

{68-95-99 rule} : Normal Distribution
↓ σ, ↓ 2σ, ↓ 3σ

Through list of degree of nodes → Power large Distribution

26/08/22 Types of Prob. Distributions

Eg → Use least significant digit to sort the delivery Code
143 → 3
196 → 6
Delivery Code, Counters No.

- Uniform → $f(x) = \frac{1}{b-a} \;\forall\; x \in [a,b]$
  $F(x) = (x-a)/b-a$; $\mu = \frac{b+a}{2}$; $\sigma^2 = \frac{1}{12}(b-a)^2$
- Bernoulli → toss coin once $f(0) = 1-p$ & $f(1) = p$.
- Binomial → Repeat several times independently. $B(n,p) \to f(k) = {}^nC_k\, p^k (1-p)^k$; $\mu = np$ & $\sigma^2 = np(1-p)$
- Poisson → for very large n in B(n, p) $f(k) = \frac{\lambda^k}{k!} e^{-\lambda}$; $\mu = \lambda$ & $\sigma^2 = \lambda$

  * Here, no. of success/people can be any no.
- Geometric → no. of attempts till we get success. $f(k) = (1-p)^{k-1} p$; $F(k) = 1-(1-p)^k$; $\mu = 1/p$; $\sigma^2 = \lambda$
- Exponential → $f(x) = \lambda e^{-\lambda x}$; $F(x) = 1 - e^{-\lambda x}$; $\mu = 1/\lambda$; $\sigma^2 = 1/\lambda^2$
- Normal Distribution → $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

  bell Shaped curve

  $= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ * z-table given

31/08/22

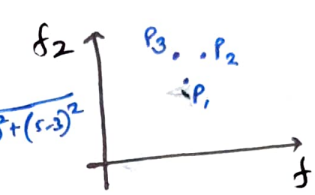
* A pattern is a pt. on D - dimensional space
where D → no. of features present in Dataset

→ To find the closeness/similarity among the two patterns
↳ measure distance b/w two Patterns; $d(P_1, P_2)$
↳ Larger distance ⇒ Much dissimilar.

* Many types of distances can be measured → i) Euclidean Distance

Eg →

$d(P_1, P_2) = \sqrt{(170-160)^2 + (5-3)^2}$
$= 10.$



	f_1	f_2
P_1	160	3
		5
P_2	170	
		5
P_3	160	

Here first attribute, f_1, is dominating on the dist. value
⇒ f_2 is insignificant over distance
So, f_2 has larger effect on the outcome

So, To avoid the dominancy Effect, we need to normalize all the data → so that they come on same scale.

Normalization → $\{x_1, x_2, \ldots x_n\} \in R^D$

$Max_j = Max\{x_{1j}, x_{2j} \ldots x_{nj}\}$ where j → attribute
So x_{ij} → Value of P_i at attribute j

$Min_j = Min\{x_{1j}, x_{2j} \ldots x_{nj}\}$

$y_{ij} = \frac{x_{ij} - Min_j}{Max_j - Min_j}$; $i = 1$ to n. $j = 1$ to D here $y_{ij} \in [0,1]$

* Remove/Reduce the effect of dominancy of a attribute due to diff. of scale in values.

if $y_{ij} \in [-1, 1]$ → $y_{ij} = 2\left[\frac{x_{ij} - Min_j}{Max_j - Min_j}\right] - 1$

Metric → Let $X \in R^D$ be a set

(gives measurement)

$d: X \cdot X \longrightarrow [0, \infty)$

d be a metric if

1> $d(x,y) = d(y,z)$; $\forall (x,y) \in X$

2> $d(x,y) = 0$; if $x = y$

3> $d(x,y) + d(y,z) \geq d(x,z)$; $\forall x, y, z \in X$

Minkowski Metric →

$$d_p(x,y) = \left(\sum_{i=1}^{D} |x_i - y_i|^p \right)^{1/p} \quad \text{where } p \geq 1$$

① $p=1$; $d_1(x,y) = \sum_{i=1}^{D} |x_i - y_i|$ → Called as City Block Distance OR Manhattan Distance

② $p=2$; $d_2(x,y) = \left(\sum_{i=1}^{D} |x_i - y_i|^2 \right)^{1/2}$ → K/a Euclidean Distance.

③ $p=\infty$; $d_\infty(x,y) = \text{Max}\{|x_i - y_i|\}$

* $d_1 > d_2 > d_3 \cdots > d_\infty$

Eg → $d_1(P_1, P_2) = |170-160| + |5-5| = 12$

$d_2(P_1, P_2) = \sqrt{10^2 + 2^2} = \sqrt{104}$

$d_\infty(P_1, P_2) = \text{Max}(10, 2) = 10$

Weighted distance

If we know that a attribute (given) is very discriminant for a particular task

Then we can give the higher weight to the higher priority attribute.

$$d_w(P_1, P_2) = \sum_{i=1}^{D} w_i |x_i - y_i|$$

$$= w_1(10) + w_2(2)$$

Let $w_1 = 1$ & $w_2 = 10$

$$= 1 \times 10 + 10 \times 2 = 30$$

Equation of Distance in Matrix form

$$x = \begin{bmatrix} 160 \\ 3 \end{bmatrix}_{D \times 1} \qquad y = \begin{bmatrix} 170 \\ 5 \end{bmatrix}_{D \times 1}$$

for considered eg. $D = 2$

* $x' = [160 \ 3]_{1 \times 2}$

* $d_2^2(x, y) = (x-y)'(x-y)$

* $x - y = \begin{bmatrix} -10 \\ -2 \end{bmatrix}$

* $(x-y)'(x-y) = \underset{1\times2}{[-10 \ -2]} \underset{2\times1}{\begin{bmatrix} -10 \\ -2 \end{bmatrix}} = [100 + 4] = \underset{1\times1}{[104]} = d_2^2(x, y)$

* $(x-y)' I_{D\times D} (x-y) = \underset{1\times2}{[-10 \ -2]} \underset{2\times2}{\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}} \underset{2\times1}{\begin{bmatrix} -10 \\ -2 \end{bmatrix}} = [-10 \ -2] \begin{bmatrix} -10 \\ -2 \end{bmatrix}$

$= [104]$

In case of any other matrix (not I) → weighted euclidean distance

02/09/22

$\sigma_{Sample} = \frac{\sigma}{\sqrt{n}}$

↳ Std. Error

9) Avg = 67 in SD = 3 in

63 in. to 67 in.

$\frac{63-67}{3} = -\frac{4}{3}$

$= -1.33$ to 0

$0.5000 - 0.0918 \Rightarrow 50 - 9.18 = 40.82\%$

~~Central Limit Theorem~~

Central Limit Theorem → ~~Single Digit~~

Eg → Dice → Avg. of 3 dice

More the no. of dice, more close to Normal Distribution

Eg → Student

Let there be any sample / distribution.

Mean of the sample Mean will be Mean of Population

Distribution of sample Mean will be ~~Mean of~~ the Normal Distribution

05/08/22

Inferential Statistics

* Not the sample size that matters but the quality of sample is that matters.

Simple Random Sample → all need to be equally likely.
Eg → Tossing a coin to select.

Random Sample → all need not be equally likely. {But same proportion}
Eg → Population divided in equal classes

Estimation & Std. Error

Q) Given avg age = 37 } for population.
std dev = 15 }

find avg age of sample of size 200 from population

Sol:

Expected age for sample = 37 {as avg age of population}

Std. error = $\frac{15}{\sqrt{200}}$ = 1.06

07/09/22

Testing of Hypothesis

* one sample z-test → Size of sample ≥ 30

* Significance Level / Power

09/09/22 **t-test** (for sample of size < 30 ⟹ Use t statistics)

Eg ⟹ population of weights → app. normal.

μ & σ of population → Unknown

Data → 31.8, 30.9, 34.2, 32.1, 28.8 (gm) → So small sample

Sample mean = 31.56 gm

* Guessed population mean = 30

Sol^n → $H_0: \mu = 30$ & $H_A: \mu > 30$

If H_0 true ⟹ sample mean = 30

SE of sample mean = $\sigma/\sqrt{5}$

$SE = \frac{\sigma}{\sqrt{5}} \Rightarrow \frac{1.96}{\sqrt{5}} = 0.87$

$t\text{score} = \frac{(31.56-30)}{0.87} = 1.79 \Rightarrow P = 7.2\%$

* Degree of freedom = 1 (sample size)

$\sigma = 1.96$

cal from data (is 37) but divide by (n-1) as sample is too small

Two Sample Test → Eg — Apple & Egg
df Test H0 A

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

→ for Z test

Unpaired Test — diff sets of 100 eggs taken before & after ~~cold~~ temp

Paired Test → same set of 100 eggs throughout

* for t-test

If var(σ^2) of both same same

$$t = \left(\frac{(\bar{x}_1 - \bar{x}_2) - d_0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right)$$

$$S_p^2 = \frac{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}{n_1 + n_2 - 2}$$

$$df = n_1 + n_2 - 2.$$

If var are diff

What if More than Two sample? {Anova}

Method-Between

$$\text{Sum of Sq. B/w (SSB)} = n \sum_{i=1}^{a} (\bar{y}_i - \bar{\bar{y}})^2$$

(a → no. of populations)

$$\text{DoF} = a - 1$$

$$\text{Mean Sq. Between (MSB)} = \text{SSB} / \text{DoF}$$

Method - Within

$$\text{SSE (Error)} = \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_i)^2$$

(y_{ij} → every individual value)

$$\text{DoF} = N - a$$

(N → total no. of items)

$$\text{MSE} = \text{SSE} / \text{DoF}$$

$$\text{fScore} = \text{MSB} / \text{MSE}$$

$\bar{\bar{y}}$ (Global Mean)

#					
A	x_{11}	x_{12} ...		x_{1n}	mean $\bar{y}_1$
B	x_{21}	x_{22} ...		x_{2n}	$\bar{y}_2$
C	x_{31}	x_{32} ~		x_{3n}	$\bar{y}_3$
D	x_{41}	x_{42} ~		x_{4n}	$\bar{y}_4$

Method-Total

$$\text{SST} = \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{\bar{y}})^2$$

DoF : N-1

$$\text{MST} = \text{SST} / \text{DoF}$$

If MSB, MSE & MST are same (statistically) Null.

Else if not Null hypo

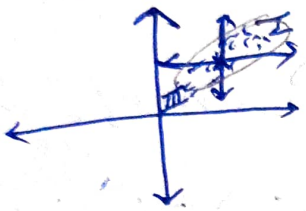
MSB > MST > MSE

12/09/22

Mean → $\mu_x = \frac{1}{n}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} p_i x_i$

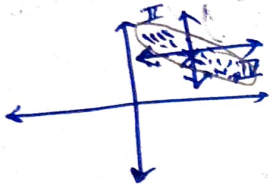
Variance → $\sigma^2(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(x_i - \mu_x)$

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



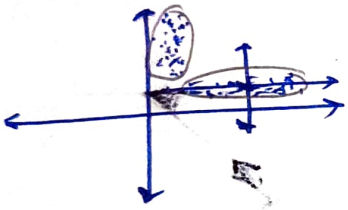
$\Rightarrow$ $\boxed{Cov(x,y) > 0}$ → if most of Data pts lies in I & III quad of Shifted origin

$x\uparrow \Leftrightarrow y\uparrow$



$\Rightarrow$ $\boxed{Cov(x,y) < 0}$ → if most of Data pts lies in II & IV quad of shifted origin

$x\uparrow \Leftrightarrow y\downarrow$



$\Rightarrow$ $\boxed{Cov(x,y) = 0}$ → x&y are unrelated

Correlation → Correlation of two variables (x, y)

$$\boxed{\rho(x,y) = \frac{Cov(x,y)}{\sqrt{Var(x)\,Var(y)}}}$$

* $-1 \le \rho(x,y) \le 1$

Measuring dist. b/w a datapoint & a cluster/class of pts.

i) Measuring dist. b/w datapt. & the mean of class.

$d_E(x, \mu_i)$ → if dist more → not in class else Belong to class

Problem → Distribution/Spreadness of classes is ignored

ii) So, we need another measure to consider spread

$$\Sigma = \begin{array}{c} \\ f_1 \\ f_2 \\ \vdots \\ f_D \end{array} \begin{array}{c} \begin{array}{cccc} f_1 & f_2 & \cdots & f_D \end{array} \\ \begin{bmatrix} \text{cov}(f_1,f_1) & \text{cov}(f_1,f_2) & \cdots & \text{cov}(f_1,f_D) \\ \text{cov}(f_2,f_1) & \text{cov}(f_2,f_2) & \cdots & \text{cov}(f_2,f_D) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(f_D,f_1) & \text{cov}(f_D,f_2) & \cdots & \text{cov}(f_D,f_D) \end{bmatrix} \end{array}$$

Each cell i,j contains covariance of f_i, f_j

$D \times D$ → no. of features

Dispersion Matrix

Variance-Covariance Matrix

Covariance Matrix

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

* Σ is real, symmetric, positive definite Matrix.

- all cells contains real values only
- Symm wrt diagonal as $\text{cov}(f_i,f_j) = \text{cov}(f_j,f_i)$
- A matrix is positive definite iff $x'Ax > 0$ for any vector x

→ $|\Sigma| > 0$

→ Σ^{-1} also be positive definite

→ All the eigenvalues are positive

$Ax = \lambda x$ (eigen matrix)

$|A - \lambda I| = 0$

Let $A = \begin{bmatrix} 2 & -1 \\ -1 & 3 \end{bmatrix}$, $x = \begin{bmatrix} a \\ b \end{bmatrix}$

$x'Ax = [a\ b] \begin{bmatrix} 2 & -1 \\ -1 & 3 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}$

$= [a\ b] \begin{bmatrix} 2a - b \\ -a + 3b \end{bmatrix}$

$= 2a^2 - 2ab + 3b^2$

$= 2\left[a - \frac{b}{2}\right]^2 + \frac{5}{2}b^2$ (>0, >0)

* Mahalanobis Dist b/w class & pt.

$$d_M^2(C_1, x) = (x - \mu_1)' \Sigma_1^{-1} (x - \mu_1)$$

* Mahalanobis dist b/w two classes

$$d_M^2(C_1, C_2) = (\mu_2 - \mu_1)' \Sigma^{-1} (\mu_2 - \mu_1)$$

$$\Sigma = \frac{\Sigma_1 + \Sigma_2}{2}$$

14/09/22

Find $\cos\theta$ b/w two vectors.

$$X_1 = \begin{bmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1D} \end{bmatrix} \quad X_2 = \begin{bmatrix} x_{21} \\ x_{22} \\ \vdots \\ x_{2D} \end{bmatrix}$$

$$X_1 \cdot X_2 = |X_1| \cdot |X_2| \cos\theta = X_1^T X_2$$

$$\Rightarrow \cos\theta = \frac{X_1^T X_2}{\sqrt{\sum_{i=1}^{D} x_{1i}^2} \sqrt{\sum_{i=1}^{D} x_{2i}^2}}$$

~~A~~ $A = ai + bj = \begin{bmatrix} a \\ b \end{bmatrix} \rightarrow x_1$

$B = ci + dj = \begin{bmatrix} c \\ d \end{bmatrix} \rightarrow x_2$

$A \cdot B = ac + bd$.

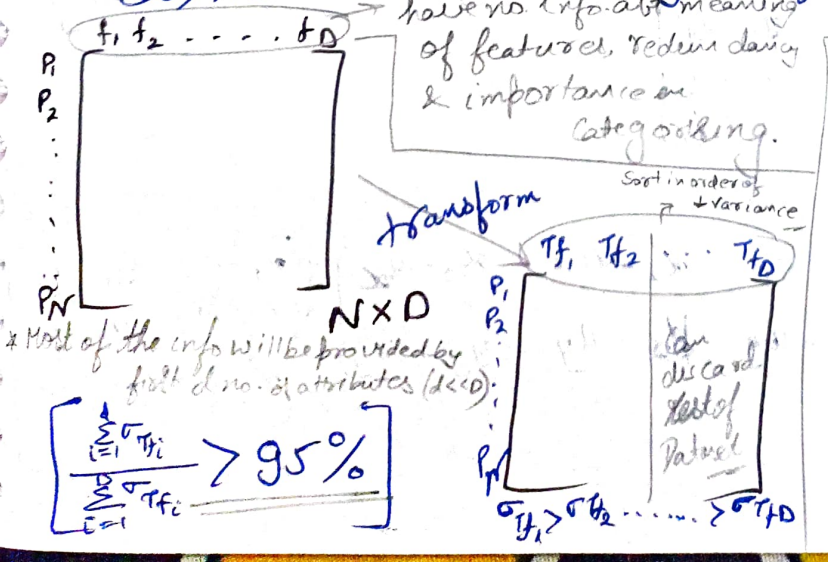
$X_1^T X_2 = [a\ b] \begin{bmatrix} c \\ d \end{bmatrix}$

$= [ac + bd]$

* Attribute with high variance is better to be considered for categorisation.

* ~~Eg~~ Choosing all the features ~~eg~~ for categorization will just increase the complexity.

So, need to select some very imp. features. But HOW??



M-1 → Transform the dataset (of which we don't know the feature meaning) into the dataset whose features are ordered in order of decreasing variance of values of that features.

$[x_i - \mu_x ; y_i - \mu_y]$

1. Let X be the mean subtracted data i.e. matrix of N×D.

   N → no of pattern.

   D → no of attributes.

2. Calculate Dispersion matrix

$$\Sigma = \frac{1}{N}(X^T X)$$

Eg →

	$x = f_1$	$y = f_2$
P_1	x_1	y_1
P_2	x_2	y_2
P_3	x_3	y_3

N×D [3×2]

$$\mu_x = \frac{x_1 + x_2 + x_3}{3} \quad \& \quad \mu_y = \frac{y_1 + y_2 + y_3}{3}$$

~~Cov(x,y)~~ $\frac{1}{N}\left[(x_1 - \mu_x)(y_1 - \mu_y) + (x_2 - \mu_x)(y_2 - \mu_y) + (x_3 - \mu_x)(y_3 - \mu_y)\right]$

$N = 3$

* $\Sigma = \begin{bmatrix} Var(x) & Cov(x,y) \\ Cov(x,y) & Var(y) \end{bmatrix}$

$$Var(x) = \frac{1}{N}\left[(x_1 - \mu_x)^2 + (x_2 - \mu_x)^2 + (x_3 - \mu_x)^2\right]$$

$N = 3$

* $X = \begin{bmatrix} x_1 - \mu_x & y_1 - \mu_y \\ x_2 - \mu_x & y_2 - \mu_y \\ x_3 - \mu_x & y_3 - \mu_y \end{bmatrix}$

Product = N × Var(x)

$$X^T X = \begin{bmatrix} x_1 - \mu_x & x_2 - \mu_x & x_3 - \mu_x \\ y_1 - \mu_y & y_2 - \mu_y & y_3 - \mu_y \end{bmatrix} \begin{bmatrix} x_1 - \mu_x & y_1 - \mu_y \\ x_2 - \mu_x & y_2 - \mu_y \\ x_3 - \mu_x & y_3 - \mu_y \end{bmatrix}$$

* Σ is a $D \times D$ matrix.

we need to find eigenvalues & eigen vectors.

"D" eigen Values will be there

$(\lambda_1, \lambda_2, \ldots \lambda_D) > 0$

$\rightarrow \lambda_1 > \lambda_2 > \ldots > \lambda_D$ be the ordered eigen values.

$v_1 \quad v_2 \quad v_D \rightarrow$ eigen vectors

$\forall i, \lambda_i > 0$

& $v_i = \begin{bmatrix} v_{i1} \\ v_{i2} \\ \vdots \\ v_{iD} \end{bmatrix}_{D \times 1}$

$\rightarrow$ Trace(Σ) $\rightarrow$ Sum of Variances w.r.t. all features

$$= \boxed{\sum_{i=1}^{D} var(f_i) = \sum_{i=1}^{D} \lambda_i}$$

W = Transformation Matrix

$$= \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1D} \\ v_{21} & v_{22} & \cdots & v_{2D} \\ \vdots & & & \vdots \\ v_{d1} & v_{d2} & \cdots & v_{dD} \end{bmatrix}_{d \times D}$$

* Can arrange in form of rows also.

Select largest d Eigen values

$(\lambda_1, \ldots \lambda_d)$

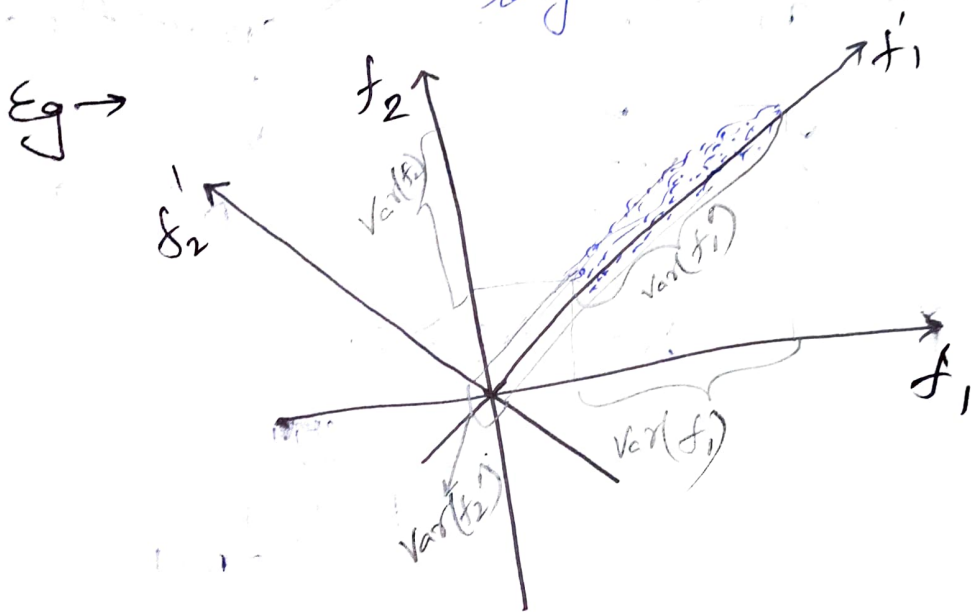
$(v_1, \ldots v_d)$

$\hookrightarrow$ first d eigen vectors from W.

19/09/22

* $\frac{\sum_{i=1}^{d} \lambda_i}{\sum_{i=1}^{D} \lambda_i} > 98\%$ → Choose d & remove all others from f_d to f_D

→ After transformation → variance of a feature is much larger than the others.



→ $Y_{d \times N} = W_{d \times D} \times_{D \times N}$

- $d \times D$ → Transformation matrix
- $D \times N$ → initial
- $Y_{d \times N}$ → transformed Dataset

This Concept is Principal Component Analysis

Eg →

$$X = \begin{bmatrix} 2 & 1 \\ 3 & 5 \\ 4 & 3 \\ 5 & 6 \\ 6 & 7 \\ 7 & 8 \end{bmatrix}$$

$\mu_1 = 4.5 \quad \mu_2 = 5$

$$\rightarrow X' = \begin{bmatrix} -2.5 & -4 \\ -1.5 & 0 \\ -0.5 & -2 \\ 0.5 & 1 \\ 1.5 & 2 \\ 2.5 & 3 \end{bmatrix}_{6\times 2}$$

$$\Sigma = \frac{1}{N}\left(X'^{T}_{2\times 6} \; X'_{6\times 2}\right)$$

here $N = 6$

$$= \frac{1}{N}\left[10 + 0 + 1 + 0.5 + 3 + 7.5 = 24\right]$$

$$\Sigma = \frac{1}{N}\begin{bmatrix} 6.25 + 2.25 + 0.25 + 0.25 + 2.25 + 6.25 & 22 \\ 22 & 16 + 0 + 4 + 1 + 4 + 9 \end{bmatrix}$$

$$\Sigma = \frac{1}{N}\begin{bmatrix} 17.5 & 22 \\ 22 & 34 \end{bmatrix} = \frac{1}{6}\begin{bmatrix} 17.5 & 22 \\ 22 & 34 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 2.92 & 3.67 \\ 3.67 & 5.67 \end{bmatrix}$$

$$|A - \lambda I| = 0$$

Is Y the final DS

How can we be assure of no redundancy

$$\Rightarrow \begin{vmatrix} 2.92-\lambda & 3.67 \\ 3.67 & 5.67-\lambda \end{vmatrix} = 0$$

$$\Rightarrow (2.92-\lambda)(5.67-\lambda) - (3.67)^2 = 0$$

$$\Rightarrow 2.92 \times 5.67 - 5.67\lambda - 2.92\lambda + \lambda^2 - (3.67)^2 = 0$$

$$\Rightarrow$$

$$\Rightarrow \boxed{\lambda_1 = 8.22} \quad \& \quad \boxed{\lambda_2 = 0.37}$$

$$v_1 = \begin{bmatrix} 2.55 \\ 3.67 \end{bmatrix}$$

* $\frac{\lambda_1}{\lambda_1 + \lambda_2}$

As λ_1 is higher $\Rightarrow W = \overset{v_1}{\begin{bmatrix} 2.55 \\ 3.67 \end{bmatrix}}_{2\times 1}$

"final DS" but lost the characteristics/meaning of all the attributes of Initial DS

$$Y = \begin{bmatrix} 2.55 \\ 3.67 \end{bmatrix} \begin{bmatrix} 2 & 3 & 4 & 5 & 6 & 7 \\ 1 & 5 & 3 & 6 & 7 & 8 \end{bmatrix}$$

$$Y = \begin{bmatrix} 2.55 & 3.67 \end{bmatrix} \begin{bmatrix} 2 & 3 & 4 & 5 & 6 & 7 \\ 1 & 5 & 3 & 6 & 7 & 8 \end{bmatrix}$$

$1\times 2 \qquad 2\times 6$

$$= \begin{bmatrix} 8.77 & 26 & 21.21 & 34.77 & 40.92 & 41.21 \end{bmatrix}$$

$$Y = \begin{bmatrix} 8.77 & 26 & 21.21 & 34.77 & 40.92 & 41.21 \end{bmatrix}$$

* If we know the meaning of the features ⇒ we know their importance, so we do not need to perform PCA.

So, need to perform PCA, to know the importance which we do based on variance

Use only when no info abt meaning/characteristics of features.

Measurement of Compactness of Classes.

→ Within classes Distance

Intra class Distance

* Should be small

↑ better as within class Distance (dist of every pt. from mean) is less

Measurement of Better Dataset

→ Between class Distance

Inter class Distance

* Should be larger
  → properly distributed / categorised
  → Less Error

Dataset-1

Dataset-2

21/09/22 **Learning** → Why? → to get success & make less mistakes

↳ getting some knowledge & use it to increase your performance

- ~~Supervised~~ Supervised → prior info. given → Parents teach their child
  ↳ training info/Data given.
- Unsupervised → learn from own experience → no prior info
  ↓
  need to learn directly.

* **Training Data** → Along with patterns (of diff features), class labels are also given.

* **Classification Task** → Supervised learning using training data

* **clustering Task** → Unsupervised learning based on similarities & Dissimilarities of patterns of features.

  → need to find Natural grouping

  look on within & B/w class Dist

→ Generate and train a classification model by using Training Data (Pattern + Classes)

  ↓ finding the decision Rule

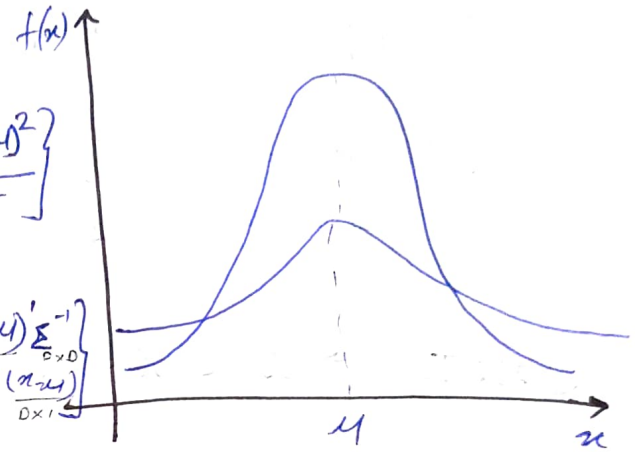
→ for an unknown pattern, assign it to a predefined classes based on the classifier.

Various types of Classification Models

① Bayes' Decision Rule / Bayes' classifier.

prior prob., P_i / $P(\omega_i)$ → Prob. of any new data to be a part of C_i

* Gaussian Dist. / Normal Dist



● Exponential Rule
↓
68% - 95% - 99.7%

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{\frac{-1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\}$$

$$f(x) = \frac{1}{(2\pi)^{d/2}\,|\Sigma|^{1/2}} \exp\left\{\frac{-1}{2}\underset{1\times D}{(x-\mu)'}\underset{D\times D}{\Sigma^{-1}}\underset{D\times 1}{(x-\mu)}\right\}$$

→ Let there be M classes ($M \geq 2$). The prior probabilities be $P_1, P_2, \ldots P_M$, where $P_i \in [0,1]$ & $\sum_{i=1}^{M} P_i = 1$.

(can be found from Dataset)

Let $p_1(x), p_2(x), \ldots p_M(x)$ be the Cond^{nal} prob. density func. of M classes.

(Need to know the type of dist. of Data)

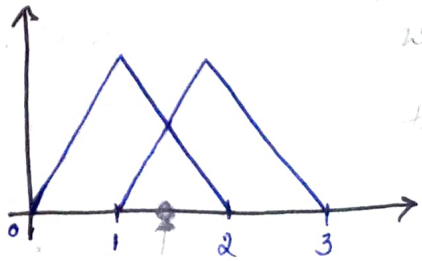
Now, any unknown pattern x_0 be in class i if

$$P_i\, p_i(x_0) \geq P_j\, p_j(x_0) \;;\; \forall i, j \;\&\; i \neq j.$$

Example → $M = 2$; Prior Prob: $P, 1-P$.

$$p_1(x) = \begin{cases} x & ; 0 \leq x \leq 1 \\ 2-x & ; 1 \leq x \leq 2 \\ 0 & ; \text{otherwise} \end{cases}$$

$$p_2(x) = \begin{cases} x-1 & ; 1 \leq x \leq 2 \\ 3-x & ; 2 \leq x \leq 3 \\ 0 & ; \text{otherwise} \end{cases}$$



need to find x which divides the upcoming data into two diff. Classes

Case-I → $0 \le x \le 1$

$P_1 = P \qquad P_2 = 1-P$

$p_1(x) = x \qquad p_2(x) = 0$

→ $P_1 p_1(x) \ge P_2 p_2(x)$ ($p_2(x) \to 0$) ⇒ x be in class 1

⇒ $\boxed{Px \ge 0}$ → as P is const. (> 0) ⇒ $\boxed{x \ge 0}$

Case-II → $2 \le x \le 3$

$P_1 = P \qquad P_2 = 1-P$

$p_1(x) \neq 0 \qquad p_2(x) = 3-x$

→ $P_2 p_2(x) \ge P_1 p_1(x)$ ($p_1(x) \to 0$) ⇒ x be in class 2

⇒ $(1-P)(3-x) \ge 0$ ⇒ $\boxed{Px - 3P - x + 3 \ge 0}$ → as $P \in (0,1)$ is const. ⇒ $\boxed{x \le 3}$

Case-III → $1 \le x \le 2$

$P_1 = P \qquad P_2 = 1-P$

$p_1(x) = 2-x \qquad p_2(x) = x-1$

for x be in class 1 → $P_1 p_1(x) \ge P_2 p_2(x)$

⇒ $P(2-x) \ge (1-P)(x-1)$

⇒ $2P - Px \ge x - 1 - Px + P$

⇒ $\boxed{x \le P+1}$

So → if $\boxed{x \le P+1}$ → in class 1

& if $\boxed{x \ge P+1}$ → x in class 2

23/09/22

Error, $E = P_1 \underbrace{\int_{1+P}^{2} p_1(x)\,dx}_{\text{class 2 element in the region of class I}} + P_2 + \int_{1}^{1+P} p_2(x)\,dx$ for Case - III

In general,

$$E = \sum_{i=1}^{c} P_i \int_{\Omega_i^c} p_i(x)\,dx$$

c → no. of classes

Ω_i^c → Complement of class-i region

$\Omega = [0 \text{ to } 3]$

$\Omega_1 = [0 \text{ to } 1+P]$

$\Omega_1^c = [1+P \text{ to } 3]$

$\Omega_2 = [1+P \text{ to } 3]$

$\Omega_2^c = [0 \text{ to } 1+P]$

So, overall,

$$E = \sum_{i=1}^{2} P_i \int_{\Omega_i^c} p_i(x)\,dx$$

$$E = P_1 \int_{1+P}^{3} p_1(x)\,dx + P_2 \int_{0}^{1+P} p_2(x)\,dx$$

$$E = P_1 \int_{1+P}^{2} p_1(x)\,dx + P_1 \cancel{\int_{2}^{3} p_1(x)\,dx} + P_2 \cancel{\int_{0}^{1} p_2(x)\,dx} + P_2 \int_{1}^{1+P} p_2(x)\,dx$$

$$E = P_1 \int_{1+P}^{2} p_1(x)\,dx + P_2 \int_{1}^{1+P} p_2(x)\,dx$$

$$E = P \int_{1+P}^{2} (2-x)\,dx + (1-P) \int_{1}^{1+P} (x-1)\,dx$$

$$E = P\left[2x - \frac{x^2}{2}\right]_{1+P}^{2} + (1-P)\left[\frac{x^2}{2} - x\right]_{1}^{1+P} = P\left[4 - 2 - 2(1+P) + \frac{(1+P)^2}{2}\right] + \left[\frac{(1+P)^2}{2} - 1 - P\right]$$

$$E = P(1-P) + \frac{(P^2+2P+1)}{2} - P - 1 = \frac{-2P^2+2P+P^2+2P+1-2P-2}{2}$$

$$E = \frac{-P^2+2P-1}{2} = -\left(\frac{P-1}{\sqrt{2}}\right)^2$$

So, for bayes' Decision Rule

$$E_B = \frac{P(1-P)}{2}$$

$\Omega_1 = [0 \text{ to } 1+p]$

$\Omega_2 = [1+p \text{ to } 3]$

Let there be another classifier

$\Omega_1 = [0 \text{ to } 1.5]$ Here: ~~$E = 0.125$~~

$\Omega_2 = [1.5 \text{ to } 3]$

$$E = P\left[2x - \frac{x^2}{2}\right]_{1.5}^{2} + (1-P)\left[\frac{x^2}{2} - x\right]_{1}^{1.5}$$

$$E = P\left[4 - 2 - 3 + \frac{(1.5)^2}{2}\right] + (1-P)\left[\frac{(1.5)^2}{2} - 1.5 - \frac{1}{2} + 1\right]$$

$$E = P\left[-1 + \frac{(1.5)^2}{2} - \frac{(1.5)^2}{2} + 1\right] + \left[\frac{(1.5)^2}{2} - 1\right]$$

$$E = \frac{2.25 - 2}{2} = \frac{25}{200} = \frac{1}{8}.$$

$$E_{1.5} = 0.125$$

→ Is $E_B \leq E_{1.5}$? → Yes

E_B is less than $E_{1.5}$ ∀ values of $P \neq 0.5$

E_B is equal to $E_{1.5}$ for $P = 0.5$

$$E_B \leq E_{cl}$$

E_{cl} of any classifier

NOTE → Minimum Error classifier / Classifier with minimum error prob. is Bayes' Classifier

So, Bayes' Classifier is the best classifier.

★ What is the need of any other classifiers then

↳ Real Time Data's distribution cannot be found / Real Time Data does not follow any fixed Distribution. {In Most of the Cases}

Estimating $p_i(x)$, ~~Conditional~~ Probability Density function, is very difficult & computationally very complex for the Real Time Datasets.