~~By Sakthi Sir~~

DATE: ~~16/9/22~~
PAGE NO.:

Principal component analysis

19/9/22 Monday ~~Friday~~

PCA

$$X = \begin{matrix} \\ p_1 \\ p_2 \\ p_3 \end{matrix} \begin{matrix} f_1 = x & f_2 = y \end{matrix} \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ x_3 & y_3 \end{bmatrix}$$

1. Make mean-subtracted data

$$X \longrightarrow X', \quad X' = \begin{bmatrix} x_1 - \mu_x & y_1 - \mu_y \\ x_2 - \mu_x & y_2 - \mu_y \\ x_3 - \mu_x & y_3 - \mu_y \end{bmatrix}$$

2. Calculate the dispersion matrix $\Sigma = \frac{1}{N} X^T X$.

$$\Sigma = \begin{bmatrix} var(x) & cov(x,y) \\ cov(y,x) & var(y) \end{bmatrix}$$

$$var(x) = \frac{1}{N} \sum (x_i - \mu_x)^2$$

$$cov(x,y) = cov(y,x) = \frac{1}{N} \sum (x_i - \mu_x)(y_i - \mu_y)$$

$$\underset{D \times D}{\Sigma} = \frac{1}{N} \begin{bmatrix} x_1 - \mu_x & x_2 - \mu_x & x_3 - \mu_x \\ y_1 - \mu_y & y_2 - \mu_y & y_3 - \mu_y \end{bmatrix} \begin{bmatrix} x_1 - \mu_x & y_1 - \mu_y \\ x_2 - \mu_x & y_2 - \mu_y \\ x_3 - \mu_x & y_3 - \mu_y \end{bmatrix}$$

$$= \frac{1}{N} \begin{bmatrix} \sum_{i=1}^{3} (x_i - \mu_x)^2 & \sum_{i=1}^{3} (x_i - \mu_x)(y_i - \mu_y) \\ \sum_{i=1}^{3} (x_i - \mu_x)(y_i - \mu_y) & \sum_{i=1}^{3} (y_i - \mu_y)^2 \end{bmatrix}$$

3. Find D no. of eigen values of eigen vectors of Σ.

$\lambda_1 > \lambda_2 > \dots > \lambda_D$ ($D = 2$ in 2 attributes)

$\downarrow \quad \downarrow \qquad \downarrow$

$\vec{v}_1 \quad \vec{v}_2 \qquad \vec{v}_D$

$$\vec{v}_i = \begin{bmatrix} v_{i1} \\ v_{i2} \\ \vdots \\ v_{iD} \end{bmatrix}$$

$$W = \begin{bmatrix} v_1 & v_2 & \dots & v_D \\ v_{11} & v_{21} & & v_{D1} \\ v_{12} & v_{22} & & v_{D2} \\ \vdots & \vdots & & \vdots \\ v_{1D} & v_{2D} & & v_{DD} \end{bmatrix}_{D \times D}$$

$\lambda_1 > \lambda_2 > \dots > \lambda_D$

$Y = W^T X$

↑ data in another D-dimensional space

$$\text{Trace}(\Sigma) = \sum_{i=1}^{D} \lambda_i$$

for 'd' eigen values

$$\frac{\sum_{i=1}^{d} \lambda_i}{\text{Trace}(\Sigma)} \times 100 > (98)$$

$\sum_{i=1}^{D} \lambda_i$ or

DATE : / /
PAGE NO. :

let D=100, d=20.

∴ From adding 1st 20 eigen values, we are covering 98% of data acc to its significance

∴ other 80 eigen values are ~~[illegible]~~ not significant.

∴ discard D-d' eigen values.

∴ We consider $W = \begin{bmatrix} v_{11} & v_{d1} \\ v_{12} & v_{d2} \\ v_{1D} & v_{dD} \end{bmatrix}_{(D \times d)}$

$Y = W^T_{(d \times D)} \times X_{(D \times 1)}$

$\boxed{Y_{d \times 1}} =$

Eg:

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

Range of 1st dimension = 2 to 7

$\mu = (4.5) \quad (5)$

$$\therefore x' = \begin{bmatrix} -2.5 & -4 \\ -1.5 & 0 \\ -0.5 & -2 \\ 0.5 & 1 \\ 1.5 & 2 \\ 2.5 & 3 \end{bmatrix}$$

$$\Sigma = \frac{1}{6}\begin{bmatrix} -2.5 & -1.5 & -0.5 & 0.5 & 1.5 & 2.5 \\ -4 & 0 & -2 & 1 & 2 & 3 \end{bmatrix}\begin{bmatrix} -2.5 & -4 \\ -1.5 & 0 \\ -0.5 & -2 \\ 0.5 & 1 \\ 1.5 & 2 \\ 2.5 & 3 \end{bmatrix}$$

$$= \frac{1}{6}\begin{bmatrix} 17.5 & 22 \\ 22 & 34 \end{bmatrix} = \begin{bmatrix} 2.92 & 3.67 \\ 3.67 & 5.67 \end{bmatrix}$$

Now eigen values.

$$\begin{bmatrix} 2.92 & 3.67 \\ 3.67 & 5.67 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = 0$$

$\Rightarrow \lambda_1 = 8.22$ and $\lambda_2 = 0.38$

For eigen vectors:

$$\Sigma \bar{v} = \lambda_i \bar{v}$$

Trace $(\Sigma) = 2.92 + 5.67 = 8.59$

$= \lambda_1 + \lambda_2$

we see $\lambda_1 >> \lambda_2$

$$\frac{\lambda_1}{8.58} \times 100 = \boxed{95.6}$$

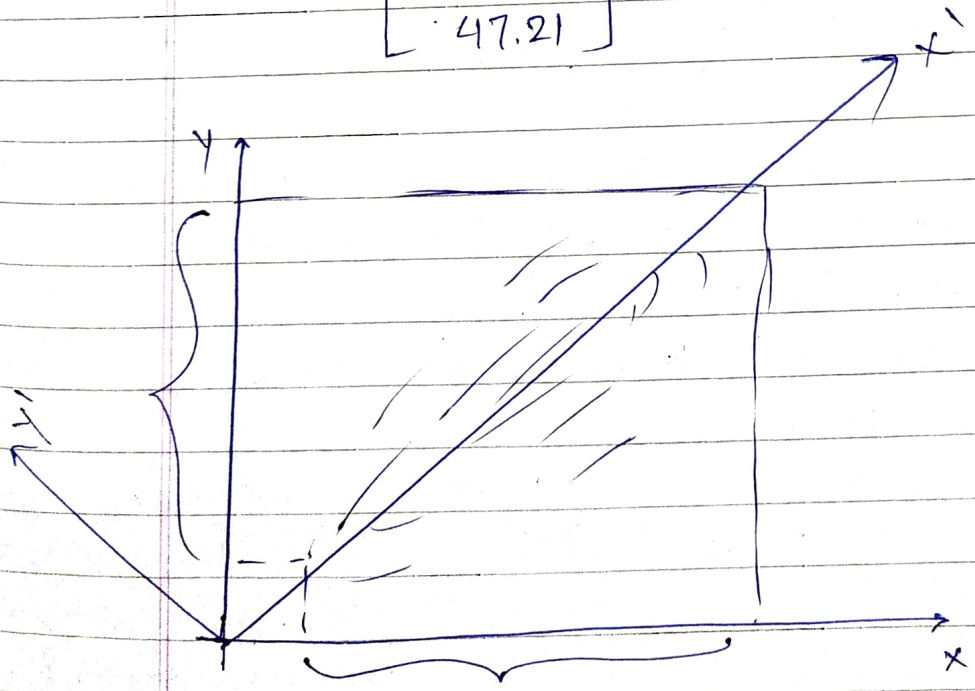
$$\Sigma \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 8.22 \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$

$v = \begin{bmatrix} 2.55 \\ 3.67 \end{bmatrix}$ ← eigen vector $= W$

$Y = W^T X$

$Y_1 = [2.55 \quad 3.67] \begin{bmatrix} 2 \\ 1 \end{bmatrix} = [8.77]$ ← (1st data entry)

$\therefore Y = \begin{bmatrix} 8.77 \\ 26 \\ 21.21 \\ 34.77 \\ 40.92 \\ 47.21 \end{bmatrix}$ → Range: 8.77 to 47.21



We are rotating the axes such that the variance of 1st dimension is increased.

DATE : / /
PAGE NO. :

Advantages

1. Computational complexity
   → By changing dimension space, we compute less differences.

2. In dataset, don't know the meaning of the attributes, we also don't know if any attribute is redundant or if atr 1 + atr 2 = atr 3.
   ∴ By this process, we can remove redundancy.

- For S.77, we have considered 2 & 1 both (both features),
- In transformed space, each value has considered every attribute's value.

Disadvantages

1. In OG dataset, all stored data is transformed in Y, although all values of OG set has contributions, the new values obtained, we don't know what this attribute denotes.

$$X = \begin{bmatrix} x & y \\ 2 & 1 \\ \vdots & \vdots \\ \vdots & \vdots \end{bmatrix} \longrightarrow Y = \begin{bmatrix} f' \\ S.77 \\ \vdots \end{bmatrix}$$

← don't know this attribute's meaning

→ Meaning of dataset is lost in transformed space.

DATE 24/9/22
PAGE NO.: Wed

2. If the weightage of attributes is given, no transformation required.

→ When machine improves performance, minimizes error, we call it machine learning.

Learning

Supervised

	f1	f2	Class label
P1	3	4	1
P2	4	8	2
P3	9	10	1

→ known: class label with data, we call it Training dataset

- we know data pattern belongs to what class.
- Then we can apply class labels to unknown data based on learning from training dataset.
- This is Supervised learning
- Lesser chance of error
- Error minimized.

Unsupervised

→ Data w/o class label info.
→ ∴ We 1st have to group that data, build an algo for it, based on similarity in patterns.
→ This is unsupervised learning
→ Done in clustering techniques.

DATE : / /
PAGE NO. :

- Supervised : Prior info given
- Unsupervised : No prior info given, clustering on basis of similarity.

# Classification	&	Clustering
→ Supervised learning → Given a new or unknown pattern and assign it into a predefined classes. → Generate and train a classification model with the help of training data (or other form of data.		→ Unsupervised learning → Based on similarity or dissimilarity measures among patterns, develop an algorithm to find the natural grouping of the data. (Look for intraclass (within) distance minimized, inter-class dist^n max) (between)

Classification Techniques

1. Probabilistic : Bayes classifier
2. Similarity-based : MDC, KNN, LDC, QDC
   (Min^n distance)
3. SVM

Probabilistic or Bayes' Rule of classification

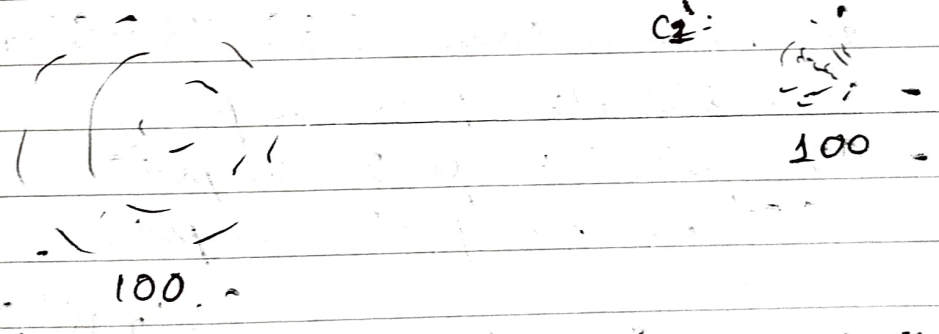
Prior info could be in form of level data or any other form.
conditional Prob. Density func^ given.

Dataset	Data Patterns
: C_1	100
: C_2	500
: C_3	1000

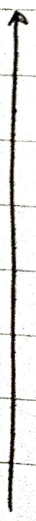
Now for a new pattern, P(class) to be in class $C_1, C_2, C_3 \Rightarrow \frac{1}{16}, \frac{5}{16}, \frac{10}{16}$ resp → Prior probability

~~Unique class~~

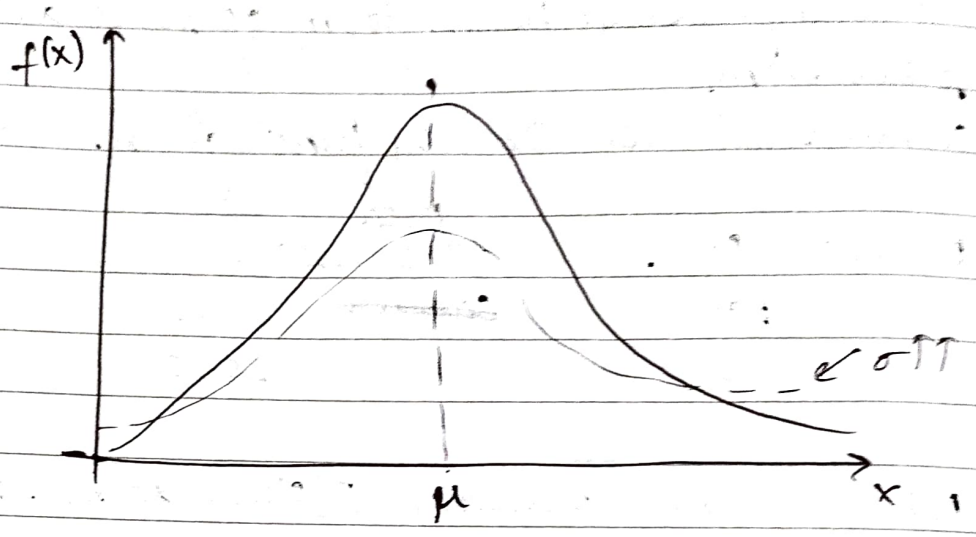
Eg: C_1: C_2:



We also need to know how data is distributed within a class.
Given by conditional Prob. Density function.



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



- If σ ↑↑, peak is lower.

Ex Empirical Rules (68-95-99.7% Rule)

- 68% lies b/w $\mu-\sigma$ to $\mu+\sigma$.
- 95% lies b/w $\mu-2\sigma$ to $\mu+2\sigma$
- 99.7% lies b/w $\mu-3\sigma$ to $\mu+3\sigma$.

$$f(x) = \frac{1}{(2\pi)^{d/2}\,|\Sigma|^{1/2}} e^{\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}}$$

d : No. of dimensions

Σ : Dispersion Matrix or variance

$\{\sqrt{\Sigma} = \sigma\}$ ← SD.

Bayes' Rule

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

→ Let $p_1(x), p_2(x), \ldots p_n(x)$ be the conditional probability density functions

→ Given a new or unknown pattern x_0,
x_0 be in class i if

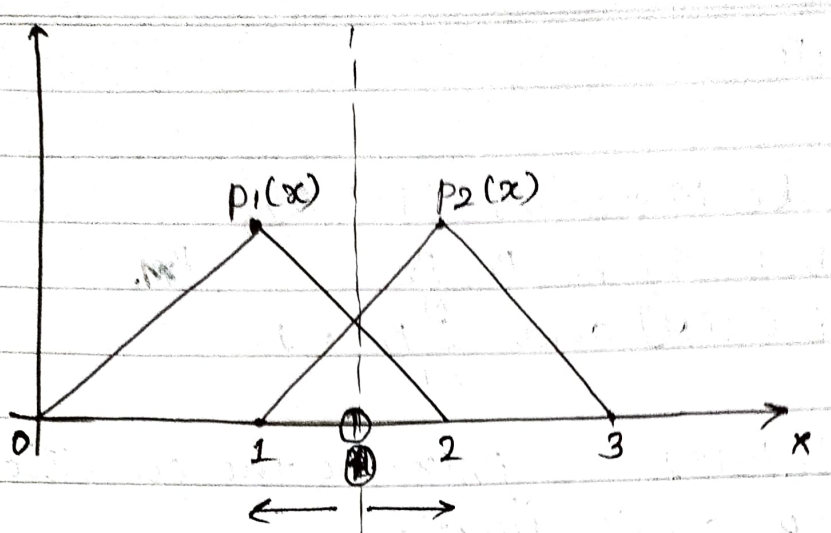
$$P_i p_i(x_0) \geq P_j p_j(x_0); \quad \forall\ i \neq j.$$

Eg: Let there be 2 classes ($M = 2$)
Prior Probability = $P, (1-P)$

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

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

Find the classification Model



$x = 1.5$
line of classification

∴ if $x < 1.5 \rightarrow$ class 1
$x > 1.5 \rightarrow$ class 2.

But for Bayes' Rule, we need to calculate & find the line of classification.

Domain of space : $x \in [0, 3]$

Break it : (0-1), (1- ——— ←——

0 to 1 : $P_1 = P$ $P_2 = 1-P$

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

$P_1 p_1(x) = Px$ $P_2 p_2(x) = 0$

$\therefore P_1 p_1(x) \geq P_2 p_2(x)$

∴ x should be in class 1
if $x \in [0, 1]$.

DATE: / /
PAGE NO:

case 2: 2 to 3:

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

$p_1(x) = 0$ $\quad$ $p_2(x) = 3-x$.

$P_1 p_1(x) = 0 \quad < \quad P_2 p_2(x) = (1-P)(3-x)$

$\therefore$ if $x \in [2,3]$, x should be in class 2.

case 3: 1 to 2:

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

$p_1(x) = 2-x$ $\quad$ $p_2(x) = x-1$.

If $P_1 p_1(x) > P_2 p_2(x) \longrightarrow x$ in class 1

$\Rightarrow P(2-x) \geq (1-P)(x-1)$

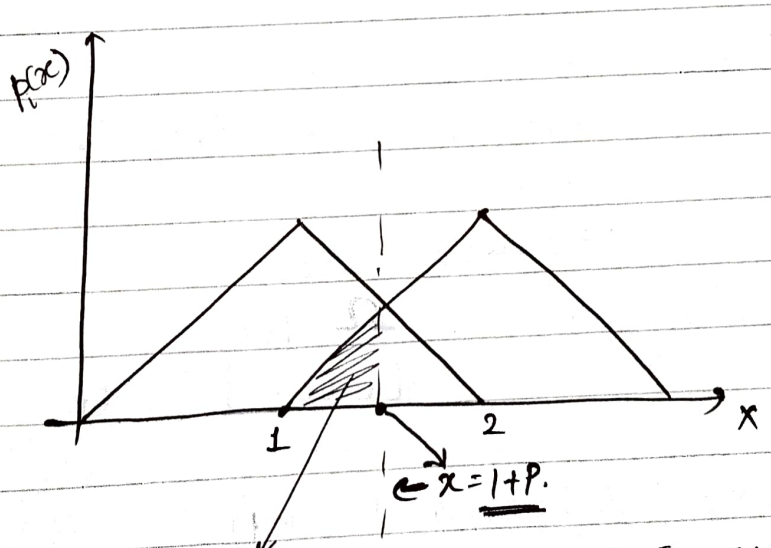
$2P - Px \geq x - 1 - Px + P$

$\boxed{x \leq P+1.}$

$\therefore x$ will be in class 1 if $x \leq 1+P$

and x will be in class 2 if $x > 1+P$

$\therefore$



- belong to class 2, but in our rule, they are counted in class 1.
- $\therefore$ Bayes Rule has some error.

Reason: Dataset ~~has data~~

- Data is not well-separated
- There is overlapping.

DATE : 23 / 9 / 22
PAGE NO. :
Friday

$P_i p_i(x) \geq P_j p_j(x)$, then x should be in class i.

- Overlapping of pattern from class 1 onto class 2 due to classification = ERROR.

ERROR: $P_1 \times \int_{1+P}^{3} p_1(x)\,dx$ { (1+P to 3) ← Outside class 1 }

1+P → No. of patterns that lie outside the class 1 and overlap in class 2

Let Ω ⇒ Domain ⇒ [0 to 3]

Ω_1 ⇒ Domain of class 1 in Ω.
⇒ [0 to 1.5]

Ω_2 ⇒ Domain of class 2 in Ω
⇒ [1.5 to 3]

$\therefore \Omega_1^c$ ⇒ Ω_1 complement
⇒ Range in Domain that is outside class 1
⇒ $\Omega - \Omega_1$
⇒ [1.5 to 3] = Ω_2

Similarly ⇒ $\Omega_2^c = \Omega - \Omega_2$
= [0 to 1.5] = Ω_1

error due to class 1 = e_1

$$e_i = P_i \int_{\Omega_i^c} p_i(x)\,dx$$

Total Error (E).

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

This is called Misclassification Probability or Error Rate

[In the given example

$$E = P_1 \int_{\Omega_1^c} p_1(x)\, dx + P_2 \int_{\Omega_2^c} p_2(x)\, dx$$

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

$$= P\left[\int_{1+P}^{2} (2-x)\, dx + \int_{2}^{3} 0\, dx\right] + (1-P)\left[\int_{0}^{1} 0 \cdot dx + \int_{1}^{1+P} (x-1)\, dx\right]$$

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

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

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

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

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

DATE : / /
PAGE NO. :

Classifiers

→ different types of classifiers divide in different categories

Eg: classifier 2.

class 1 → $x \leq 1.5$

class 2 → o/w

classifier 3

class 1 → $x \leq 1.4$

class 2 → o/w.

Now For classifier 2, E = ?.

$$E = P_1 \int_{1.5}^{2} p_1(x)\,dx + P_2 \int_{1}^{1.5} p_2(x)\,dx.$$

(1 → 1.5, 1+P →)

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

$$= P\left[1 - \frac{7}{8}\right] + (1-P)\left(\frac{5}{8} - \frac{1}{2}\right)$$

$$= P(1/8) + (1-P)(1/8)$$

$$= \boxed{1/8}$$

$$= \boxed{0.125}$$

- Bayes' classifier is the best classifier.
- No other classifier works better.
- $E_{Bayes'\ classifier} \Rightarrow$ is minimum.

i.e. $\frac{P(1-P)}{2} \leq$ (Any other error) (in the given example)

$\therefore \frac{P(1-P)}{2} \leq 0.125 \quad (\because E_{1.5} = 0.125)$

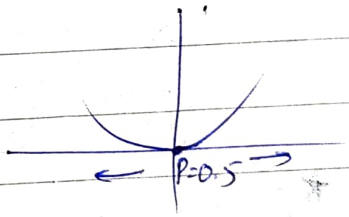
let $P = 0.4$

$\therefore E_{Bayes} = \frac{4}{10} \times \frac{6}{10} \times \frac{1}{2} \Longrightarrow 0.12$

let $P = 0.6$

$\therefore E_{Bayes} = 0.12$

$P = 0.5 \longrightarrow E_{Bayes} = 0.125 = E_{1.5}.$



If given data is in form of training data, like

Eg: f_1 f_2 f_3 → classes

C_1 C_2 C_3 } → can estimate prior probs.

No. of patterns: 100 200 300

$\Rightarrow P_1 = \frac{1}{6}, \quad P_2 = \frac{2}{6}, \quad P_3 = \frac{3}{6}$

P_1 ↓ for C_1

$p_1(x)$ = Assume ~~class~~ dataset forms Gaussian/Normal distribution.

DATE : / /
PAGE NO. :

$$p_i(x) = \frac{e^{\{-\frac{1}{2}(x-\mu_i)' \Sigma_i^{-1} (x-\mu_i)\}}}{(2\pi)^{d/2} |\Sigma_i|^{1/2}}$$

$\therefore$ $c_1 \rightarrow 100 \rightarrow \mu_1, \Sigma_1$
$c_2 \rightarrow 200 \rightarrow \mu_2, \Sigma_2$
$c_3 \rightarrow 300 \rightarrow \mu_3, \Sigma_3$
} Calculate.

- Then apply Bayes' Rule for classification

- Baye's' classifier used very less in Data science

Reason:

→ To solve real time prob, we get training dataset and we assume Gaussian Distribution but in real world probs, the dist^n is not normal.

$p_i(x) = ?$

* ① Estimation of conditional prob. density func^n is very difficult.

② It is computationally very complex.