ROC (7 marks)

Spring : 2020

6) a) For threshold 0.75 :

if score ≥ 0.75 , prediction = 'P'
if score < 0.75 , prediction = 'N'

actual	score	prediction
p	0.9	p
p	0.8	p
n	0.7	n
p	0.6	n
p	0.55	n
p	0.54	n
n	0.53	n
n	0.52	n
p	0.51	n
n	0.505	n

Actual \ Prediction	1	0
1	2	4
0	0	4

$$\therefore TPR = \frac{TP}{TP+FN} = \frac{2}{2+4} = 0.33$$

$$\therefore FPR = \frac{FP}{FP+TN} = \frac{0}{0+4} = 0$$

$$\therefore TPR, FPR = (0.33, 0)$$

For threshold 0.57 :

if score ≥ 0.57, prediction = 'p'
if score < 0.57 , prediction = 'n'

actual	score	prediction
p	0.9	p
p	0.8	p
n	0.7	p
p	0.6	p
p	0.55	n
p	0.54	n
n	0.53	n
n	0.52	n
p	0.51	n
n	0.505	n

Actual \ Prediction	1	0
1	3	3
0	1	3

$$\therefore TPR = \frac{3}{3+3} = 0.5$$

$$\therefore FPR = \frac{1}{1+3} = 0.25$$

$$\therefore TPR, FPR = (0.5, 0.25)$$

For threshold 0.53 :

	1	0
1	5	1
0	2	2

p	p
p	p
n	p
p	p
p	p
p	p
n	p
n	n
p	n
n	n

$$\therefore TPR = \frac{5}{5+1} = 0.83$$

$$\therefore FPR = \frac{2}{2+2} = 0.5$$

$$\therefore TPR, FPR = (0.83, 0.5)$$

For threshold 0.51 :

	1	0
1	6	0
0	4	1

p	p
p	p
n	p
p	p
p	p
p	p
n	p
n	p
p	p
n	n

$\therefore TPR = \frac{6}{6+0} = 1$

$\therefore FPR = \frac{4}{4+1} = 0.8$

$\therefore TPR, FPR = (1, 0.8)$

For threshold 0.50 :

	1	0
1	6	0
0	4	0

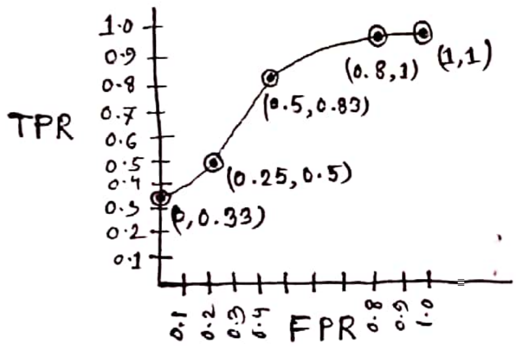
p	p
p	p
n	p
p	p
p	p
p	p
n	p
n	p
p	p
n	p

$\therefore TPR = \frac{6}{6+0} = 1$

$\therefore FPR = \frac{4}{4+0} = 1$

$\therefore TPR, FPR = (1, 1)$

Receiver Operating Characteristic (ROC)/Receiver Operating Curve :



Fall : 2018 :

4) c) Accuracy $= \frac{TP+TN}{TP+FP+FN+TN} = \frac{100+300}{100+60+40+300} = 0.8$

Precision $= \frac{TP}{TP+FP} = \frac{100}{100+60} = 0.625$

Recall $= \frac{TP}{TP+FN} = \frac{100}{100+40} \approx 0.7143$

F measure $= \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2}{\frac{1}{0.625} + \frac{1}{0.7143}} \approx 0.67$

$TP = 100$; $FN = 40$; $FP = 60$; $TN = 300$

(5-7 marks) PCA+LDA [can be a part of mandatory question 1]

Spring-2020:

1) b) step:01: Get some data :

Consider, X = Dhaka and Y = Rajshahi

Total X = 363 [## = 51]

$\therefore$ Average X, $\bar{X} = \frac{363}{9} \approx 40.33333333$

Total Y = 276

$\therefore$ Average Y, $\bar{Y} = \frac{276}{9} = 30.66666667$

step:02 : Subtract the mean :

$X_i - \bar{X}$	$Y_i - \bar{Y}$
10.66666667	9.33333333
19.66666667	-3.66666667
9.66666667	28.33333333
-10.33333333	-15.66666667
-38.33333333	-26.66666667
19.66666667	-9.66666667
-0.33333333	-2.66666667
21.66666667	40.33333333
-32.33333333	-19.66666667

step:03: calculate the co-variance matrix :

$$\text{cov}(x,x) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{4072}{8} = 509$$

$$\text{cov}(x,y) = \text{cov}(y,x) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{2806}{8} = 350.75$$

$$\text{cov}(y,y) = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{3974}{8} = 496.75$$

$$\therefore \text{cov} = \begin{bmatrix} \text{cov}(x,x) & \text{cov}(x,y) \\ \text{cov}(y,x) & \text{cov}(y,y) \end{bmatrix} = \begin{bmatrix} 509 & 350.75 \\ 350.75 & 496.75 \end{bmatrix}$$

step:04 : calculate eigenvector and eigenvalues:

$$(A-\lambda I)X=0$$

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

$$\Rightarrow \begin{bmatrix} 509 & 350.75 \\ 350.75 & 496.75 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = 0$$

$$\Rightarrow \begin{bmatrix} 509 & 350.75 \\ 350.75 & 496.75 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} = 0$$

$$\Rightarrow \begin{bmatrix} 509-\lambda & 350.75 \\ 350.75 & 496.75-\lambda \end{bmatrix} = 0$$

$$\Rightarrow (509-\lambda)(496.75-\lambda) - 350.75^2 = 0$$

$$\Rightarrow 252845.75 - 509\lambda - 496.75\lambda + \lambda^2 - 123025.5625 = 0$$

$$\Rightarrow \lambda^2 - 1005.75\lambda + 129820.1875 = 0$$

$$\therefore \lambda = \frac{-(-1005.75) \pm \sqrt{(-1005.75)^2 - 4\times1\times129820.1875}}{2\times1}$$

$$\Rightarrow \lambda = \frac{1005.75 \pm 701.6069502}{2}$$

$$\therefore \lambda_1 = 853.6784751$$

$$\therefore \lambda_2 = 152.0715249$$

$$\therefore \text{Eigenvalues} = \begin{bmatrix} 853.6784751 \\ 152.0715249 \end{bmatrix}$$

step:05: choosing components and forming a feature vector:

for $\lambda_1 = 853.6784751$:

$$\begin{bmatrix} 509-853.6784751 & 350.75 \\ 350.75 & 496.75-853.6784751 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = 0$$

$$\Rightarrow \begin{bmatrix} -344.6784751 & 350.75 \\ 350.75 & -356.9284751 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = 0$$

$$-344.6784751x_1 + 350.75y_1 = 0 \quad \dots \quad \dots \quad (\text{I})$$
$$350.75x_1 - 356.9284751y_1 = 0 \quad \dots \quad \dots \quad (\text{II})$$

From equation (I),

$$-344.6784751x_1 = -350.75y_1$$

$$\Rightarrow x_1 = \frac{-350.75}{-344.6784751} y_1$$

$$\therefore x_1 = 1.018\, y_1$$

From equation (II),

$$350.75x_1 = 356.9284751y_1$$

$$\Rightarrow x_1 = \frac{356.9284751}{350.75} y_1$$

$$\therefore x_1 = 1.018\, y_1$$

$$\therefore A = \sqrt{(1.018)^2 + (1)^2} \doteq 1.427$$

$$\therefore e_1 \sim \begin{bmatrix} \frac{1.018}{A} \\ \frac{1}{A} \end{bmatrix} \sim \begin{bmatrix} 0.7134 \\ 0.7008 \end{bmatrix}$$

For $\lambda_2 = 152.0715249$,

$$\begin{bmatrix} 509-152.0715249 & 350.75 \\ 350.75 & 496.75-152.0715249 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = 0$$

$$\Rightarrow \begin{bmatrix} 356.9284751 & 350.75 \\ 350.75 & 344.6784751 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = 0$$

$$356.9284751x_1 + 350.75y_1 = 0 \quad \dots \quad \dots \quad (\text{III})$$
$$350.75x_1 + 344.6784751y_1 = 0 \quad \dots \quad \dots \quad (\text{IV})$$

From equation (iii),

$$356.9284751 x_1 = -350.75 y_1$$

$$\Rightarrow x_1 = \frac{-350.75}{356.9284751} y_1$$

$$\therefore x_1 = -0.98269 y_1$$

From equation (iv),

$$350.75 x_1 = -344.6784751 y_1$$

$$\Rightarrow x_1 = \frac{-344.6784751}{350.75} y_1$$

$$\therefore x_1 = -0.98269 y_1$$

$$\therefore A = \sqrt{(-0.98269)^2 + 1^2} = 1.402$$

$$\therefore e_2 \sim \begin{bmatrix} \frac{-0.98269}{1.402} \\ \frac{1}{A} \end{bmatrix} \sim \begin{bmatrix} -0.7009 \\ 0.7133 \end{bmatrix}$$

$$\therefore \text{Eigenvector} = \begin{bmatrix} 0.7134 & -0.7009 \\ 0.7008 & 0.7133 \end{bmatrix}$$

$$\therefore \text{PCA (Principal Component Analysis)} = \begin{bmatrix} 0.7134 \\ 0.7008 \end{bmatrix} \text{ [maximum]}$$

$$\therefore \text{Eigen value} = 853.6784751$$

(ans:)

Fall-2020:

4) b) For class 1 ($\omega 1$):

X_1	X_2	$A = X_1 - \overline{X}_1$	$B = X_2 - \overline{X}_2$	A^2	B^2	AB
5	1	5/3	-4/3	25/9	16/9	-20/9
1	3	-7/3	2/3	49/9	4/9	-14/9
4	3	2/3	2/3	4/9	4/9	4/9

$$\overline{X}_1 = \frac{5+1+4}{3} = \frac{10}{3} = 3.333333333$$

$$\overline{X}_2 = \frac{1+3+3}{3} = \frac{7}{3} = 2.333333333$$

$$\therefore \text{cov}(x,x) = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1} = \frac{\Sigma A^2}{3-1} = \frac{\frac{26}{3}}{2} = 4.333333333$$

$$\therefore \text{cov}(x,y) = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{n-1} = \frac{\Sigma AB}{3-1} = \frac{-10/3}{2} = -1.666666667$$

$$\therefore \text{cov}(y,y) = \frac{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}{n-1} = \frac{\Sigma B^2}{3-1} = \frac{\frac{8}{3}}{2} = 1.333333333$$

$$\therefore S_1 = \begin{bmatrix} \text{cov}(x,x) & \text{cov}(x,y) \\ \text{cov}(y,x) & \text{cov}(y,y) \end{bmatrix} = \begin{bmatrix} 4.333333333 & -1.666666667 \\ -1.666666667 & 1.333333333 \end{bmatrix}$$

For class 2 ($\omega 2$):

X_1	X_2	$A = X_1 - \overline{X}_1$	$B = X_2 - \overline{X}_2$	A^2	B^2	AB
8	9	-2/3	7/3	4/9	49/9	-14/9
8	6	-2/3	-2/3	4/9	4/9	4/9
10	5	4/3	-5/3	16/9	25/9	-20/9

$$\overline{X}_1 = \frac{8+8+10}{3} = \frac{26}{3} = 8.666666667$$

$$\overline{X}_2 = \frac{9+6+5}{3} = \frac{20}{3} = 6.666666667$$

$$\therefore \text{cov}(x,x) = \frac{\Sigma A^2}{2} = \frac{\frac{8}{3}}{2} = 1.333333333$$

$$\therefore \text{cov}(x,y) = \frac{\Sigma AB}{2} = \frac{-10/3}{2} = -1.666666667$$

$$\therefore \text{cov}(y,y) = \frac{\Sigma B^2}{2} = \frac{26/3}{2} = 4.333333333$$

$$\therefore S_2 = \begin{bmatrix} 1.333333333 & -1.666666667 \\ -1.666666667 & 4.333333333 \end{bmatrix}$$

$$\therefore S_w = S_1 + S_2 = \begin{bmatrix} 5.666666666 & -3.333333334 \\ -3.333333334 & 5.666666666 \end{bmatrix}$$

$$\therefore S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T \quad \left[\mu_1 = \begin{bmatrix} \bar{X_1} \\ \bar{X_2} \end{bmatrix} \text{ for } \omega_1 ; \mu_2 = \begin{bmatrix} \bar{X_1} \\ \bar{X_2} \end{bmatrix} \text{ for } \omega_2 \right]$$

$$= \left(\begin{bmatrix} 3.333333333 \\ 2.333333333 \end{bmatrix} - \begin{bmatrix} 8.666666667 \\ 6.666666667 \end{bmatrix} \right) \left(\begin{bmatrix} 3.333333333 \\ 2.333333333 \end{bmatrix} - \begin{bmatrix} 8.666666667 \\ 6.666666667 \end{bmatrix} \right)^T$$

$$= \begin{bmatrix} -5.333333334 \\ -4.333333334 \end{bmatrix} \begin{bmatrix} -5.333333334 & -4.333333334 \end{bmatrix}$$

$$= \begin{bmatrix} 28.44444445 & 23.11111112 \\ 23.11111112 & 18.77777778 \end{bmatrix}$$

Now,

$$S_w^{-1} S_B \omega = \lambda \omega$$

$$\Rightarrow |S_w^{-1} S_B - \lambda I| = 0$$

$$\Rightarrow \begin{bmatrix} 5.666666666 & -3.333333334 \\ -3.333333334 & 5.666666666 \end{bmatrix}^{-1} \begin{bmatrix} 28.44444445 & 23.11111112 \\ 23.11111112 & 18.77777778 \end{bmatrix} - \lambda I = 0$$

$$\Rightarrow \begin{bmatrix} 0.2698412699 & 0.1587301588 \\ 0.1587301588 & 0.2698419699 \end{bmatrix} \begin{bmatrix} 28.44444445 & 23.11111112 \\ 23.11111112 & 18.77777778 \end{bmatrix} - \lambda I = 0$$

$$\Rightarrow \begin{bmatrix} 11.34391535 & 9.216931222 \\ 10.75133894 & 8.735462885 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = 0$$

$$\Rightarrow \begin{bmatrix} 11.34391535 - \lambda & 9.216931222 \\ 10.75133894 & 8.735462885 - \lambda \end{bmatrix} = 0$$

$$\Rightarrow (11.34391535 - \lambda)(8.735462885 - \lambda) - 9.216931222 \times 10.75133894 = 0$$

$$\Rightarrow 99.09435151 - 20.07937824\lambda + \lambda^2 - 99.09435155 = 0$$

$$\Rightarrow \lambda^2 - 20.07937824\lambda = 0$$

$$\Rightarrow \lambda(\lambda - 20.07937824) = 0$$

$$\therefore \lambda = 0, 20.07937824$$

for $\lambda = 0$,

$$\begin{bmatrix} 11.34391535-0 & 9.216931222 \\ 10.75133894 & 8.735462885-0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

$\therefore 11.34391535x_1 + 9.216931222x_2 = 0 \quad \dots \quad \dots \quad (\mathrm{I})$

$\therefore 10.75133894x_1 + 8.735462885x_2 = 0 \dots \quad \dots \quad (\mathrm{II})$

from equation (I),

$$x_1 = -\frac{9.216931222x_2}{11.34391535} = -0.8125x_2$$

From equation (II),

$$x_1 = \frac{-8.735462885x_2}{10.75133894} = -0.8125x_2$$

$$\therefore A = \sqrt{(-0.8125)^2 + (1)^2} = 1.288470508$$

$$\therefore e_1 \sim \begin{bmatrix} \frac{-0.8125}{A} \\ \frac{1}{A} \end{bmatrix} \sim \begin{bmatrix} -0.631 \\ 0.776 \end{bmatrix}$$

For $\lambda = 20.07937824$,

$$\begin{bmatrix} 11.34391535-20.07937824 & 9.216931222 \\ 10.75133894 & 8.735462885-20.07937824 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

$\therefore -8.73546289x_1 + 9.216931222x_2 = 0 \quad \dots \quad \dots \quad (\mathrm{III})$

$\therefore 10.75133894x_1 - 11.34391536x_2 = 0 \quad \dots \quad \dots \quad (\mathrm{IV})$

from equation (III),

$$x_1 = \frac{9.216931222}{8.73546289}x_2 = 1.055x_2$$

from equation (IV),

$$x_1 = \frac{11.34391536}{10.75133894}x_2 = 1.055x_2$$

$$\therefore A = \sqrt{(1.055)^2 + (1)^2} = 1.45362478$$

$$\therefore e_2 \sim \begin{bmatrix} \frac{1.055}{1.45362478} \\ \frac{1}{1.45362478} \end{bmatrix} \sim \begin{bmatrix} 0.726 \\ 0.688 \end{bmatrix}$$

$\therefore$ We get maximum value for $\lambda = 20.08$ (ans:)

Clustering [6-7 marks]

Fall-2020 :

5) b) step : 01 : $k=2$

$c1 = \{1.0, 1.0\}$ [ID = 1]

$c2 = \{5.0, 7.0\}$ [ID = 4]

step : 02 :

$d(c1, 1) = \sqrt{|1.0-1.0|^2 + |1.0-1.0|^2} = 0$

$d(c2, 1) = \sqrt{|5.0-1.0|^2 + |7.0-1.0|^2} \approx 7.21$

$d(c1, 2) = \sqrt{|1.0-1.5|^2 + |1.0-2.0|^2} \approx 1.12$

$d(c2, 2) = \sqrt{|5.0-1.5|^2 + |7.0-2.0|^2} \approx 6.10$

$d(c1, 3) = \sqrt{|1.0-3.0|^2 + |1.0-4.0|^2} \approx 3.61$

$d(c2, 3) = \sqrt{|5.0-3.0|^2 + |7.0-4.0|^2} \approx 3.61$

$d(c1, 4) = \sqrt{|1.0-5.0|^2 + |1.0-7.0|^2} \approx 7.21$

$d(c2, 4) = \sqrt{|5.0-5.0|^2 + |7.0-7.0|^2} \approx 0$

$d(c1, 5) = \sqrt{|1.0-3.5|^2 + |1.0-5.0|^2} \approx 4.72$

$d(c2, 5) = \sqrt{|5.0-3.5|^2 + |7.0-5.0|^2} = 2.5$

$d(c1, 6) = \sqrt{|1.0-4.5|^2 + |1.0-5.0|^2} \approx 5.32$

$d(c2, 6) = \sqrt{|5.0-4.5|^2 + |7.0-5.0|^2} \approx 2.06$

$\therefore$ ID 1, 2, 3 are under cluster 1

$\therefore$ ID 4, 5, 6 are under cluster 2

$\therefore$ New centroids are :

$c1 = \left(\frac{1}{3}(1+1.5+3), \frac{1}{3}(1+2+4)\right) = (1.83, 2.33)$

$c2 = \left(\frac{1}{3}(5+3.5+4.5), \frac{1}{3}(7+5+5)\right) = (4.33, 5.67)$

step:03:

$$d(c1,1) = \sqrt{|1.83-1.0|^2 + |2.33-1.0|^2} \approx 1.57$$

$$d(c2,1) = \sqrt{|4.33-1.0|^2 + |5.67-1.0|^2} \approx 5.74$$

$$d(c1,2) = \sqrt{|1.83-1.5|^2 + |2.33-2.0|^2} \approx 0.47$$

$$d(c2,2) = \sqrt{|4.33-1.5|^2 + |5.67-2.0|^2} \approx 4.63$$

$$d(c1,3) = \sqrt{|1.83-3.0|^2 + |2.33-4.0|^2} \approx 2.04$$

$$d(c2,3) = \sqrt{|4.33-3.0|^2 + |5.67-4.0|^2} = 2.14$$

$$d(c1,4) = \sqrt{|1.83-5.0|^2 + |2.33-7.0|^2} = 5.64$$

$$d(c2,4) = \sqrt{|4.33-5.0|^2 + |5.67-7.0|^2} = 1.49$$

$$d(c1,5) = \sqrt{|1.83-3.5|^2 + |2.33-5.0|^2} = 3.15$$

$$d(c2,5) = \sqrt{|4.33-3.5|^2 + |5.67-5.0|^2} = 1.07$$

$$d(c1,6) = \sqrt{|1.83-4.5|^2 + |2.33-5.0|^2} = 3.78$$

$$d(c2,6) = \sqrt{|4.33-4.5|^2 + |5.67-5.0|^2} = 0.69$$

$\therefore$ ID 1,2,3 will be in cluster 1

$\therefore$ ID 4,5,6 will be in cluster 2

Therefore, there is no change in the clusters.

$\therefore$ Two final clusters are : $\{1,2,3\}$, $\{4,5,6\}$

Spring-2020:

3) b) i) Manhattan distance $= |22-20| + |1-0| + |42-36| + |10-8|$

$= 2+1+6+2$

$= 11$ (ans:)

ii) Minkowski distance $= \sqrt[3]{|22-20|^3 + |1-0|^3 + |42-36|^3 + |10-8|^3}$

$= \sqrt[3]{2^3 + 1^3 + 6^3 + 2^3}$

$= \sqrt[3]{233}$

≈ 6.15 (ans:)

iii) Supremum distance $= \max(|22-20|, |1-0|, |42-36|, |10-8|)$

$= \max(2, 1, 6, 2)$

$= 6$ (ans:)

Spring-2020:

7) a) step: 01:

	P1	P2	P3	P4	P5
P1	0	2	6	10	9
P2	$\boxed{2}$	0	5	9	8
P3	6	5	0	4	5
P4	10	9	4	0	3
P5	9	8	5	3	0

min

step: 02: P1, P2 are closest

	P1,P2	P3	P4	P5
P1,P2	0	5.5	9.5	8.5
P3	5.5	0	4	5
P4	9.5	4	0	3
P5	8.5	5	$\boxed{3}$	0

min

Scanned with CamScanner

step :03 :

	P1, P2	P3	P4, P5
P1, P2	0	5.5	9
P3	5.5	0	4.5
P4, P5	9	$\boxed{4.5}$	0

min

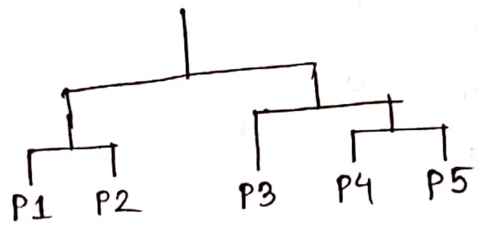
step :04:

	P1, P2	P3, (P4, P5)
P1, P2	0	7.25
P3, (P4, P5)	7.25	0

∴ Average distance = 7.25

∴ Clusters : ((P1, P2), (P3, (P4, P5)))

Dendogram:



Scanned with CamScanner

Spring 2019:

4) a) Manhattan distance : $|x_1 - x_2| + |y_1 - y_2| + |z_1 - z_2|$

Distance calculation :

$1,2 = |0-1| + |0-1| + |0-0| = 1+1+0 = 2$

$1,3 = |0-1| + |0-1| + |0-1| = 1+1+1 = 3$

$1,4 = |0-1| + |0-1| + |0-0| = 1+1+0 = 2$

$2,3 = |1-1| + |1-1| + |0-1| = 0+0+1 = 1$

$2,4 = |1-1| + |1-1| + |0-0| = 0+0+0 = 0$

$3,4 = |1-1| + |1-1| + |1-0| = 0+0+1 = 1$

step:01 :

	1	2	3	4
1	0	2	3	2
2	2	0	1	0
3	3	1	0	1
4	2	$\boxed{0}$ → min	1	0

step:02 :

	1	2,4	3
1	0	2	3
2,4	2	0	1
3	3	$\boxed{1}$ → min	0

step:03 :

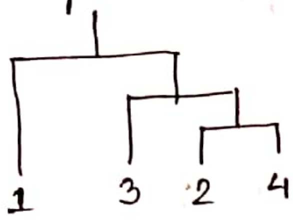
	1	(2,4),3
1	0	2
(2,4),3	2	0

∴ Minimum distance = 2

∴ Clusters : (1 , ((2,4),3))

Scanned with CamScanner

Dendogram :



Linear Discriminant Function [7-14 marks]

Fall-2020:

1) b) Given, class 1, $\omega 1 = (-2)$

class 2, $\omega 2 = (51, 2)$

phi function, $y = [x^2 \; x \; 1]$

learning rate, $\eta = 0.45$

initial weight, $w(0) = [1 \; 1]^T$

For $\omega 1$: $y = [4 \; -2 \; 1]$ $[x_1 = -2]$

For $\omega 2$: $y = [2601 \; 51 \; 1]$ $[x_2 = 51]$

For $\omega 2$: $y = [4 \; 2 \; 1]$ $[x_3 = 2]$

Using normalization,

For $\omega 1$: $y = [4 \; -2 \; 1]$

For $\omega 2$: $y = [-2601 \; -51 \; -1]$

$y = [-4 \; -2 \; -1]$

We know: $\omega(t+1) = \omega(t) + \eta \sum_{\forall y \text{ misclassified}} y$

$\begin{cases} \omega^T y > 0 \rightarrow cc \\ \omega^T y \leq 0 \rightarrow mc \end{cases}$

y			ω^T			$\omega^T y$
4	-2	1	1	1	1	3 → cc
-2601	-51	-1	1	1	1	-2653 → mc
-4	-2	-1	-1169.45	-21.95	0.55	4699.2 → cc
4	-2	1	-1169.45	-21.95	0.55	-4633.35 → mc
-2601	-51	-1	-1167.65	-22.85	1	3038222 → cc
-4	-2	-1	-1167.65	-22.85	1	4715.3 → cc
4	-2	1	-1167.65	-22.85	1	-4623.9 → mc
-2601	-51	-1	-1165.85	-23.75	1.45	3033585.65 → cc
-4	-2	-1	-1165.85	-23.75	1.45	4709.45 → cc

$\therefore \omega^T = [-1165.85 \quad -23.75 \quad 1.45]$

$\therefore g(y) = \omega_1 y_1 + \omega_2 y_2 + \omega_0 = 0$

$\Rightarrow -1165.85 y_1 - 23.75 y_2 + 1.45 = 0$

$\therefore$ Decision Boundary equation: $-1165.85 y_1 - 23.75 y_2 + 1.45 = 0$

(ans:)

Spring-2020:

2) a)

$$
\begin{array}{cc|cc}
\multicolumn{2}{c|}{\omega_1} & \multicolumn{2}{c}{\omega_2} \\
\hline
\begin{bmatrix}2\\3\end{bmatrix} & \begin{bmatrix}1\\51\end{bmatrix} & \begin{bmatrix}3\\1\end{bmatrix} & \begin{bmatrix}-2\\-4\end{bmatrix} \\
\hline
\begin{bmatrix}4\\9\\-1\\6\end{bmatrix} & \begin{bmatrix}1\\2601\\-50\\51\end{bmatrix} & \begin{bmatrix}9\\1\\2\\3\end{bmatrix} & \begin{bmatrix}4\\16\\2\\8\end{bmatrix} \\
\hline
\begin{bmatrix}4\\9\\-1\\6\end{bmatrix} & \begin{bmatrix}1\\2601\\-50\\51\end{bmatrix} & \begin{bmatrix}-9\\-1\\-2\\-3\end{bmatrix} & \begin{bmatrix}-4\\-16\\-2\\-8\end{bmatrix}
\end{array}
$$

[y]

[normalization]

$$
\begin{cases} \omega^T y > 0 \rightarrow cc \\ \omega^T y \leq 0 \rightarrow mc \end{cases}
$$

We know: $\omega(t+1) = \omega(t) + \eta \sum_{\forall y} y$ misclassified

Here, $\eta = 2$, $\omega = \begin{bmatrix}0.5\\0.5\\0.5\\0.5\end{bmatrix}$

y				ω^T				$\omega^T y$
4	9	-1	6	0.5	0.5	0.5	0.5	9 $\longrightarrow$ cc
1	2601	-50	51	0.5	0.5	0.5	0.5	1301.5 $\rightarrow$ cc
-9	-1	-2	-3	0.5	0.5	0.5	0.5	-7.5 $\longrightarrow$ mc
-4	-16	-2	-8	-17.5	-1.5	-3.5	-5.5	145 $\longrightarrow$ cc
4	9	-1	6	-17.5	-1.5	-3.5	-5.5	-113 $\longrightarrow$ mc
1	2601	-50	51	-9.5	16.5	-5.5	6.5	43513.5 $\rightarrow$ cc
-9	-1	-2	-3	-9.5	16.5	-5.5	6.5	60.5 $\longrightarrow$ cc
-4	-16	-2	-8	-9.5	16.5	-5.5	6.5	-267 $\longrightarrow$ mc

$\therefore \omega^T = [-9.5 \quad 16.5 \quad -5.5 \quad 6.5]$

$\therefore g(y) = -9.5 y_1 + 16.5 y_2 - 5.5 y_3 + 6.5 = 0$

$\therefore$ Decision boundary equation: $-9.5 y_1 + 16.5 y_2 - 5.5 y_3 + 6.5 = 0$

(Ans:)

Chapter 02+03 [18 - 32 marks]

Fall-2019:

1) a) Here,

$$\mu_1 = 0$$

$$\mu_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\Sigma_1 = \begin{bmatrix} 3 & 0.5 \\ 0.5 & 3 \end{bmatrix}$$

$$\Sigma_2 = \begin{bmatrix} 6 & 5 \\ 5 & 6 \end{bmatrix}$$

decision boundary :

$$g_i(x) = g_j(x)$$

$$\Rightarrow -\frac{1}{2}(x-\mu_1)^t \Sigma_1^{-1}(x-\mu_1) = -\frac{1}{2}(x-\mu_2)^t \Sigma_2^{-1}(x-\mu_2) \quad [P(\omega_1) = P(\omega_2)]$$

$$\Rightarrow \frac{-\|x-\mu_1\|^2}{2\Sigma_1} = \frac{-\|x-\mu_2\|^2}{2\Sigma_2}$$

$$\Rightarrow \frac{\|x\|^2 - 2\cdot x\cdot \mu_1^t + \|\mu_1\|^2}{2\Sigma_1} = \frac{\|x\|^2 - 2\cdot x\cdot \mu_2^t + \|\mu_2\|^2}{2\Sigma_2}$$

$$\Rightarrow \frac{x^t\cdot x}{2\Sigma_1} - \frac{2x\mu_1^t}{2\Sigma_1} + \frac{\mu_1^t\cdot\mu_1}{2\Sigma_1} = \frac{x^t\cdot x}{2\Sigma_2} - \frac{2x\mu_2^t}{2\Sigma_2} + \frac{\mu_2^t\cdot\mu_2}{2\Sigma_2}$$

$$\Rightarrow -\frac{2x\mu_1^t}{2\Sigma_1} + \frac{\mu_1^t\cdot\mu_1}{2\Sigma_1} = -\frac{2x\mu_2^t}{2\Sigma_2} + \frac{\mu_2^t\cdot\mu_2}{2\Sigma_2} \quad [x^t\cdot x = \text{constant}]$$

$$\Rightarrow \frac{2x\mu_1^t}{2\Sigma_1} - \frac{2x\mu_2^t}{2\Sigma_2} - \frac{\|\mu_1\|^2}{2\Sigma_1} + \frac{\|\mu_2\|^2}{2\Sigma_2} = 0$$

$$\Rightarrow \frac{2x\cdot 0}{\Sigma_1} - \frac{2x\mu_2^t}{2\Sigma_2} - \frac{\sqrt{0^2}}{2\Sigma_1} + \frac{\mu_2^t\cdot\mu_2}{2\Sigma_2} = 0$$

$$\Rightarrow \Sigma_2^{-1} x\mu_2^t - \frac{\Sigma_2^{-1}\mu_2^t\cdot\mu_2}{2} = 0$$

$$\Rightarrow \Sigma_2^{-1}\mu_2^t\left(x - \frac{\mu_2}{2}\right) = 0 \quad \ldots\ldots\ldots\ldots (1)$$

if $\Sigma_2^{-1} \mu_2^t = w^t$; $\frac{\mu_2}{2} = x_0$ then from equation ①,

$$w^t (x - x_0) = 0$$

which is the equation of decision boundary.

The hyperplane won't be perpendicular with w^t because there is $\Sigma_2^{-1} \cdot \Sigma_2^{-1}$ explains that the direction isn't towards hyperplane.

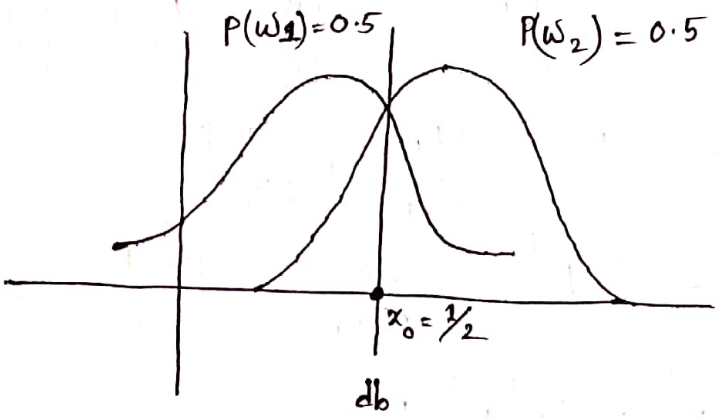
putting the values in equation ①,

$$\begin{bmatrix} 6 & 5 \\ 5 & 6 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 1 \end{bmatrix} \left(x - \frac{\begin{bmatrix} 1 \\ 1 \end{bmatrix}}{2} \right) = 0$$

$$\Rightarrow \frac{1}{36-25} \begin{bmatrix} 6 & -5 \\ -5 & 6 \end{bmatrix} \begin{bmatrix} 1 & 1 \end{bmatrix} \left(x - \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix} \right) = 0$$

$$\Rightarrow \frac{1}{11} \begin{bmatrix} 6 & -5 \\ -5 & 6 \end{bmatrix} \begin{bmatrix} 1 & 1 \end{bmatrix} \left(x - \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix} \right) = 0$$

$$\Rightarrow \underbrace{\begin{bmatrix} 6/11 & -5/11 \\ -5/11 & 6/11 \end{bmatrix} \begin{bmatrix} 1 & 1 \end{bmatrix}}_{w^T} \left(x - \underbrace{\begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}}_{x_0} \right) = 0$$



Fall-2019:

1) b)

Instance	F1	F2	Class
x1	0.1	0.4	+
x2	0.3	0.6	+
x3	0.2	0.2	+
	$\mu=0.2$	$\mu=0.4$	
	$\sigma^2=0.01$	$\sigma^2=0.04$	

$$\therefore P(F1=0.9|+) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(0.2-0.9)^2}{2\sigma^2}}$$

$$= \frac{1}{\sqrt{2\pi \times 0.01}} e^{-\frac{0.49}{2\times 0.01}}$$

$$= 9.133\times 10^{-11}$$

$$\therefore P(F2=0.4|+) = \frac{1}{\sqrt{2\pi\times 0.04}} e^{-\frac{(0.4-0.4)^2}{2\times 0.04}}$$

$$= 1.994310088$$

$$\therefore P(F1=0.9, F2=0.4|+) = 9.133\times 10^{-11} \times 1.994310088$$

$$= 1.82\times 10^{-10}$$

$$\therefore P(x6|+)\,P(+) = 1.82\times 10^{-10} \times \frac{3}{5} = 1.092\times 10^{-10}$$

Instance	F1	F2	Class
x4	0.7	0.1	−
x5	0.8	0.3	−
	$\mu=0.75$	$\mu=0.2$	
	$\sigma^2=0.005$	$\sigma^2=0.02$	

$$\therefore P(F1=0.9|-) = \frac{1}{\sqrt{2\pi\times 0.005}} e^{-\frac{(0.75-0.9)^2}{2\times 0.005}} = 0.5945318088$$

$$\therefore P(F2=0.4|-) = \frac{1}{\sqrt{2\pi\times 0.02}} e^{-\frac{(0.2-0.9)^2}{2\times 0.02}} = 1.35\times 10^{-5}$$

$$\therefore P(F1=0.9, F2=0.4|-) = 0.5945318088 \times 1.35\times 10^{-5}$$

$$= 8.024\times 10^{-6} \qquad P(x6|-)\,P(-) = 3.21\times 10^{-6}$$

Though $P(x6|+)P(+) < P(x6|-)P(-)$, class will be = "−"

Scanned with CamScanner

Fall-2020:

4) a) i) $P(B) = P(B,A) + P(B,\sim A)$

$= P(B|A)\,P(A) + P(B|\sim A)\;P(\sim A)$

$= 0.7 \times 0.3 + 0.5 \times (1-0.3)$

$= 0.21 + 0.35$

$= 0.56$

$\therefore P(A|B) = \dfrac{P(B|A)\,P(A)}{P(B)} = \dfrac{0.7 \times 0.3}{0.56} = 0.375$

ii) $P(C|B) = P(C|A,B)\,P(A|B) + P(C|\sim A,B)\;P(\sim A|B)$

$= P(C|A)\,P(A|B) + P(C|\sim A)\,P(\sim A|B)$

$= 0.2 \times 0.375 + 0.6 \times 0.625$

$= 0.45$

$P(\sim A|B) = \dfrac{P(B|\sim A)\;P(\sim A)}{P(B)} = \dfrac{0.5 \times (1-0.3)}{0.56} = 0.625$

iii)

Spring-2020:

5) b) $P(HD=Yes) = P(HD \mid E=Yes, D=Unhealthy) \times P(E=Yes) \times P(D=Unhealthy)$

$= 0.45 \times 0.7 \times (1-0.25)$

$= 0.23625$

$P(HD=No) = P(HD \mid E=Yes, D=Unhealthy) \times P(E=Yes) \times P(D=Unhealthy)$

$= (1-0.45) \times 0.7 \times (1-0.25)$

$= 0.28875$

$P(Hb=Yes) = P(Hb \mid D=Unhealthy) \times P(D=Unhealthy)$ [Hb independent to E]

$= 0.85 \times (1-0.25)$

$= 0.6375$

$P(Hb=No) = P(Hb \mid D=Unhealthy) \times P(D=Unhealthy)$

$= (1-0.85) \times (1-0.25)$

$= 0.1125$

Spring-2020:

7) b) i)

A	B	C	Class
0	0	0	+
0	0	1	+
1	0	1	+
1	1	1	+
1	0	1	+

A	B	C	Class
0	0	1	−
0	1	1	−
0	1	1	−
1	0	1	−
1	0	1	−

$\therefore P(A \mid +) = P(A=0 \mid +) \times P(A=1 \mid +)$

$= \frac{2}{5} \times \frac{3}{5} = 0.24$

$\therefore P(B \mid +) = P(B=0 \mid +) \times P(B=1 \mid +)$

$= \frac{4}{5} \times \frac{1}{5} = 0.16$

$\therefore P(C \mid +) = P(C=0 \mid +) \times P(C=1 \mid +)$

$= \frac{1}{5} \times \frac{4}{5} = 0.16$

$\therefore P(A \mid -) = P(A=0 \mid -) \times P(A=1 \mid -)$

$= \frac{3}{5} \times \frac{2}{5} = 0.24$

$\therefore P(B \mid -) = P(B=0 \mid -) \times P(B=1 \mid -)$

$= \frac{3}{5} \times \frac{2}{5} = 0.24$

$\therefore P(C \mid -) = P(C=0 \mid -) \times P(C=1 \mid -)$

$= \frac{0}{5} \times \frac{5}{5} = 0$

ii) For class = '+'

$P(A=0, B=1, C=0) = \frac{2}{5} \times \frac{1}{5} \times \frac{1}{5}$

$= 0.016$

For class = '−'

$P(A=0, B=1, C=0) = \frac{3}{5} \times \frac{2}{5} \times \frac{0}{5}$

$= 0$

$\therefore P(0,1,0 \mid +) > P(0,1,0 \mid -)$; class will be = '+'

Fall-2020:

Size	Color	Target
XS	green	Yes
L	green	Yes
XL	green	Yes
XS	white	Yes
M	green	Yes

Size	Color	Target
XS	white	No
M	black	No
L	black	No

$P(L, white | Yes) = P(L | Yes) \times P(white | Yes)$

$= \frac{1}{5} \times \frac{1}{5}$

$= 0.04$

$P(L, white | No) = P(L | No) \times P(white | No)$

$= \frac{1}{3} \times \frac{1}{3}$

≈ 0.111111

$P(L, white | Yes) \times P(Yes) = 0.04 \times \frac{5}{8} = 0.025$

$P(L, white | No) \times P(No) = 0.111 \times \frac{3}{8} = 0.042$

$\therefore$ Target will be = 'No'

Adaboost (4-7 marks)

Algorithm :

1) For $i=1$ to N initialize the data weight $w_1^{(i)} = \frac{1}{N}$

2) For $t=1$ to T.

a) Find a classifier $h_t(x)$ by minimizing the weighted error function :

$$J_t = \sum_{i=1}^{N} w_t^{(i)} \times I\left(y^{(i)} \neq h_t(x^{(i)})\right)$$

b) Find the weighted error of $h_t(x)$:

$$\epsilon_t = \frac{\sum_{i=1}^{N} \omega_t^{(i)} \times I\left(y^{(i)} \neq h_t(x^{(i)})\right)}{\sum_{i=1}^{N} w_t^{(i)}}$$

and the new component is assigned votes based on it's error :

$$\alpha_t = \ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$$

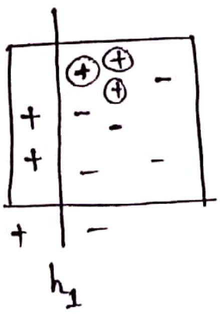
c) The normalized weights are updated :

$$w_{t+1}^{(i)} = \omega_t^{(i)} e^{\alpha_t I\left(y^{(i)} \neq h_t(x^{(i)})\right)}$$

3) Combined classifier $\hat{y} = \text{sign}(H_T(x))$ where $H_T(x) = \sum_{t=1}^{M} \alpha_t h_t(x)$

Example :

step : 01 :



$$\epsilon_1 = \frac{3}{10}$$

$$\alpha_1 = \frac{1}{2} \ln \frac{1-\frac{3}{10}}{\frac{3}{10}} = \ln\sqrt{\frac{7}{3}} \approx 0.42$$

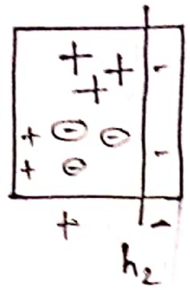
upweighting misclassified circled points :

$$W_1^{(i)} = \frac{1}{10} e^{\ln\sqrt{7/3}} \approx 0.1528$$

downweighting correctly classified points:

$$W_1^{(i)} = \frac{1}{10} e^{-\ln\sqrt{7/3}} \approx 0.0655$$

step:02:



$$\epsilon_2 = 3 \times 0.0655 = 0.1965$$

$$\alpha_2 = \frac{1}{2} \ln \frac{1 - 0.1965}{0.1965} = 0.7042$$

upweighting misclassified circled points:

$$W_2^{(i)} = 0.0655 e^{0.7042} \approx 0.1325$$

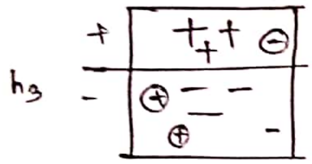
downweighting correctly classified (small +/-):

$$W_2^{(i)} = 0.0655 e^{-0.7042} \approx 0.0324$$

downweighting correctly classified (large +):

$$W_2^{(i)} = 0.1325 e^{-0.7042} \approx 0.0655$$

step:03:



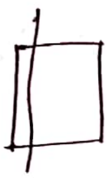
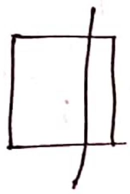
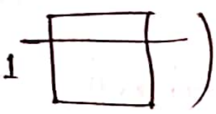
$$\epsilon_3 = 3 \times 0.0324 = 0.0972$$

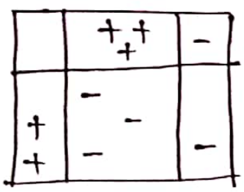
$$\alpha_3 = \frac{1}{2} \ln \frac{1 - 0.0972}{0.0972} \approx 1.1144$$

upweighting misclassified circled points:

$$W_3^{(i)} = 0.0324 e^{1.1144} \approx 0.0988$$

Final:

$H_{final} = \text{sign}\Big(0.42$  $+ 0.7$  $+ 1.1$  $\Big)$

$=$ 

SVM [7 marks]

Fall-2020

3) b) $L(x,y,\lambda) = f(x,y) - \lambda_1 g_1(x,y) - \lambda_2 g_2(x,y)$

$= (x^2+y^2) - \lambda_1(x+1) - \lambda_2(y+1)$

$= x^2+y^2 - \lambda_1 x - \lambda_1 - \lambda_2 y - \lambda_2$

We solve from the gradient of the lagrangian,

$\nabla L(x,y,\lambda) = \nabla f(x,y) - \lambda_1 \nabla g_1(x,y) - \lambda \nabla g_2(x,y) = 0$

$\frac{\partial}{\partial x}(L) = 2x - \lambda_1 = 0 \quad \dots \dots \dots \dots \text{(I)}$

$\frac{\partial}{\partial y}(L) = 2y - \lambda_2 = 0 \quad \dots \dots \dots \dots \text{(II)}$

$\frac{\partial}{\partial \lambda_1}(L) = -x-1 = 0 \quad \dots \dots \dots \dots \text{(III)}$

$\frac{\partial}{\partial \lambda_2}(L) = -y-1 = 0 \quad \dots \dots \dots \dots \text{(IV)}$

From equation (III),

$-x = 1$

$\therefore x = -1$

From equation (IV),

$-y = 1$

$\therefore y = -1$

From equation (I),

$2 \times (-1) = \lambda_1$

$\therefore \lambda_1 = -2$

From equation (II),

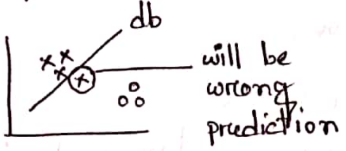
$2 \times (-1) = \lambda_2$

$\therefore \lambda_2 = -2$

$\therefore f(x,y) = x^2 + y^2 = (-1)^2 + (-1)^2 = 1+1 = 2$

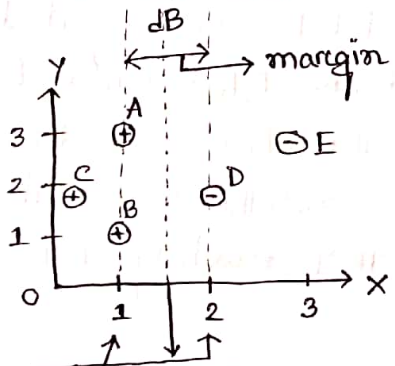
(ans:)

SVM (Support Vector Machine) : এটা একটা classifier type algorithm.

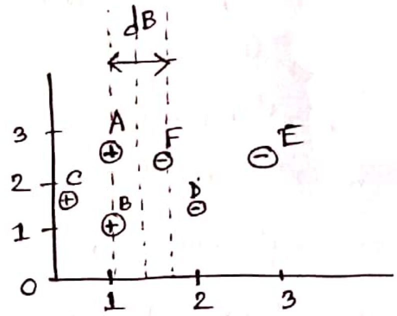


তাই decision boundary (db) এমনভাবে set করতে হবে যেন সব class থেকে same distance maintain করে। db এর দুই পাশে gap consider করা হয় SVM এ যা perceptron classifier এ করা হয় না।

margin বাড়ানোর try করাটাই SVM এর উদ্দেশ্য। মানে classify করার পাশাপাশি both class এর margin increase করে।



যদি new point F আসে



dB = 1.5

1 এর উপর আঁকা dotted line = ~~margin~~ line/gutter

2 " " " " = "

margin = 1 থেকে 2 এর মধ্যের distance

A, B, C, D, E points are vector

যে pointগুলো margin এর উপর দিয়ে গিয়েছে আর যে pointগুলো db কে decide করতে help করে ঐ point গুলোই support vector।

So, A, B, D are support vectors whereas C, E are non-support vectors

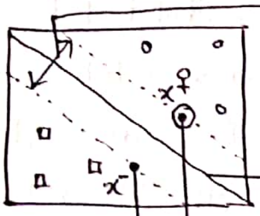
gutter : • a line through closest + and − points near dB
• parallel to db
• not a region, it's a line

SV : • points that decide db

margin : • distance between the two gutter

* if we change the points and it won't effect on db then those points are not SV.

কোন db better? ans: যে hyperplane margin কে maximize করবে এটা।



margin

linear function

$\underbrace{\text{db এর } eq^n}$

$$\vec{w}\cdot\vec{x}+b=0$$

$$\vec{w}\cdot\vec{x^+}+b=+1 \quad \dots\dots (i)$$

$$\vec{w}\cdot\vec{x^-}+b=-1 \quad \dots\dots (ii)$$

imagine, $\circ = (+ve)$; $\square = (-ve)$

$$\therefore \text{margin} = \frac{2}{||\vec{w}||^2}$$

from eq^n (i) and (ii),

$$\vec{w}\cdot\vec{x^+}+b-1=0 \quad \dots\dots (iii)$$

$$\vec{w}\cdot\vec{x^-}+b+1=0 \quad \dots\dots (iv)$$

from eq^n (iii) and (iv),

$$\vec{w}\cdot\vec{x^+}+b-1=\vec{w}\cdot\vec{x^-}+b+1$$

$$\Rightarrow \vec{w}\cdot\vec{x^+}-1=\vec{w}\cdot\vec{x^-}+1$$

$$\Rightarrow \vec{w}\cdot\vec{x^+}-\vec{w}\cdot\vec{x^-}=1+1$$

$$\Rightarrow \vec{w}(\vec{x^+}-\vec{x^-})=2 \quad \dots\dots (v)$$

from figure,

$$\text{margin} = (x^+-x^-)\cdot\frac{w}{||w||}$$

$$= \frac{2}{||w||} \quad [\text{from } (v)]$$

We know,

$$w^Tx+w_0=0$$

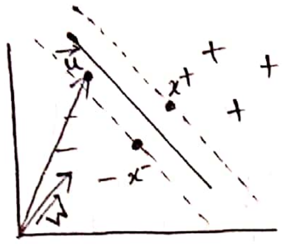
$$\Rightarrow w\cdot x+\underbrace{w_0}_{\text{bias}}=0 \quad [w^Tx=w.x]$$

কোন point dB এর উপর থাকলে সেটার eq^n এ 0 হয়। মানে dB এর eq^n এ 0 হয়।

কোন ভেক্টরকে তার magnitude দিয়ে divide করলে unit ভেক্টর পাওয়া যায়। কোন line এর direction পাওয়ার জন্যে unit vector দিয়ে multiply করা লাগে।

distance এর জন্য: (x^+-x^-)

direction এর জন্য: $\frac{w}{||w||}$



যদি $\vec{u}$ এর class বের করতে চাই, তখন

$\vec{w}\cdot\vec{u}+b \geq 0$ হলে (+ve) [decision rule]

put additional constraits to calculate w and b

$\vec{w}\cdot\vec{x^+}+b \geq 1 \longrightarrow$ (+ve) ...... (vi)

$\vec{w}\cdot\vec{x^-}+b \leq -1 \longrightarrow$ (-ve) ...... (vii)

2 eq^n are painful to carry. So, introducing new variable to make the math convenient.

y_i such that: $y_i = +$ for +ve samples

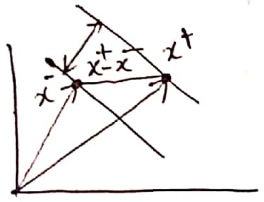
$y_i = -$ for -ve samples

$\therefore y_i(\vec{w}\cdot\vec{x_i}+b) \geq 1$

$y_i(\vec{w}\cdot\vec{x_i}+b)-1 \geq 0$

$\therefore y_i(\vec{w}\cdot\vec{x_i}+b)-1 = 0$

[for x_i in gutter/margin]



$$\therefore \text{margin/width} = (x^+-x^-)\cdot\frac{w}{\|w\|}$$

$$= \frac{x^+w - x^-w}{\|w\|}$$

$$= \frac{(1-b)-(-1-b)}{\|w\|} \quad [\text{from (vi), (vii)}]$$

$$= \frac{1-b+1+b}{\|w\|}$$

$$= \frac{2}{\|w\|}$$

max $\frac{2}{\|w\|}$ দরকার কারণ margin max করাই SVM এর target।

math কে easy করার জন্য:

$$\max \frac{2}{\|w\|}$$

$$\approx \max \frac{1}{\|w\|}$$

$$\approx \min \|w\|$$

$$\approx \min \frac{\|w\|^2}{2}$$

[constraints: $y_i(\vec{w}\cdot\vec{x_i}+b) \geq 1$]

optimization problem + শর্ত তাই differentiation দিয়ে solve করা যাবেনা। সেক্ষেত্রে lagrange multiplier লাগে।

[max/min এর সাথে constraint থাকলে lagrange multiplier use হয়]

Lagrange multiplier:

Given, $f(x,y) = 2 - x^2 - 2y^2$ (function) — Find extreme value.

$g(x,y) = x + y - 1 = 0$ (constraint)

$=$: তাই equal constraint
$>/<$: inequal constraint

Solution:

$$L(x,y,\lambda) = f(x,y) - \lambda\, g(x,y)$$

$[\lambda = \text{multiplier}]$

lagrange format এ transferred

$$= 2 - x^2 - 2y^2 - \lambda(x + y - 1)$$

We solve from the gradient of the lagrangian,

$$\nabla L(x,y,\lambda) = \nabla f(x,y) - \lambda \nabla g(x,y) = 0$$

$$\frac{\partial}{\partial x}(L) = -2x - \lambda = 0 \quad \dots \dots \dots \text{(I)}$$

$$\frac{\partial}{\partial y}(L) = -4y - \lambda = 0 \quad \dots \dots \dots \text{(II)}$$

$$\frac{\partial}{\partial \lambda}(L) = x + y - 1 = 0 \quad \dots \dots \dots \text{(III)}$$

From (I) and (II),

$\lambda = -2x$

$\lambda = -4y$

$\therefore -2x = -4y$

$\Rightarrow x = \frac{-4y}{-2}$

$\therefore x = 2y$

Putting $x = 2y$ in (III),

$2y + y - 1 = 0$

$\Rightarrow 3y = 1$

$\therefore y = \frac{1}{3}$

$\therefore x = \frac{2}{3}$

$\therefore \lambda = \frac{-4}{3}$

$$\therefore f(x,y) = 2 - \left(\frac{2}{3}\right)^2 - 2\left(\frac{1}{3}\right)^2 = 2 - \frac{4}{9} - \frac{2}{9} = \frac{18-4-2}{9} = \frac{12}{9} = \frac{4}{3}$$

For our topic:

$$L = \frac{1}{2}\|\vec{w}\|^2 - \Sigma\, \alpha_i [y_i(\vec{w}\cdot\vec{x_i} + b) - 1]$$

$\therefore \frac{\partial L}{\partial \vec{w}} = \vec{w} - \Sigma\alpha_i y_i x_i = 0 \qquad \therefore \vec{w} = \Sigma\alpha_i y_i \vec{x_i}$

$\therefore \frac{\partial L}{\partial b} = -\Sigma\alpha_i y_i = 0 \qquad \therefore \Sigma\alpha_i y_i = 0$

$\alpha = \lambda$

Σ কারন i বলতে কোনটা সেটা জানি না

$y_i = +ve/-ve$

putting the value of $\vec{w}$ in L,

$$L = \frac{1}{2}\left(\Sigma \alpha_i y_i \vec{x}_i\right)\cdot\left(\Sigma \alpha_j y_j \vec{x}_j\right) - \Sigma \alpha_i y_i x_i \left(\Sigma \alpha_j y_j x_j\right) - \Sigma\, \underbrace{\alpha_i y_i}\, b + \Sigma \alpha_i$$

→ dual optimization এর আমরা solve করেছি যেখানে dot product থাকে।

$$\Rightarrow\ = \Sigma \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j \underbrace{\vec{x}_i \cdot \vec{x}_j}_{K(x_i, x_j)} \quad [\because \Sigma \alpha_i y_i = 0] \quad [\text{quadratic eq}^n \downarrow \text{example } ax^2+bx+c=0]$$

unknown $= \alpha$; feature $= x$; label $= y$ $\{x_i^T x = K(x_i, x)\}$

$\therefore f(x) = \Sigma \alpha_i y_i x_i^T x + b$

Here, parameters are multiplied to each other in L.
So, it's a quadratic optimization problem. QP solver is used.

KKT condition

qp solver : quadratic problem solver [for finding α]

all the math depends on dot product.

decision rule : $\vec{w}\cdot\vec{u} + b \geq 0$

$$\Rightarrow \Sigma \alpha_i y_i \vec{x}_i \cdot \vec{u} + b = 0 \quad [\because \vec{w} = \Sigma \alpha_i y_i \vec{x}_i]$$

যে point এর জন্য $\alpha > 0$: support vector

" " " $\alpha = 0$: non-support vector

Problem

A_1	A_2	y	λ_i
0.38	0.47	+	65.52
0.49	0.61	-	65.52
0.92	0.41	-	0

Sol^n: $\lambda_i = 0$ not considerable because $\lambda_i > 0$ are SV only.

$\therefore w_1 = \Sigma \lambda_i y_i x_{i1}$

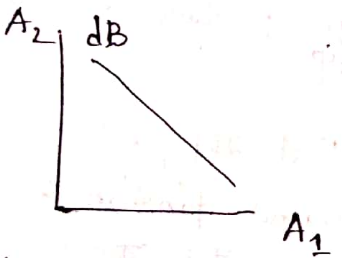
$= 65.52 \times 1 \times 0.38 + 65.52 \times (-1) \times 0.49$

$= -6.64$

$\therefore w_2 = \Sigma \lambda_i y_i x_2$

$= 65.52 \times 1 \times 0.47 + 65.52 \times (-1) \times 0.61$

$= -9.32$



$\therefore$ dB: $W_1 x_1 + W_2 x_2 + b = 0$

$\therefore \underbrace{\boxed{-6.64x_1 - 9.32x_2 + 8.9212 = 0}}_{\text{MMH = maximum margin hyperplane}}$

$b_1 = 1 - Wx_1$

$= 1 - \{(-6.64) \times 0.38 + (-9.32) \times 0.47\}$

$= 7.9036$

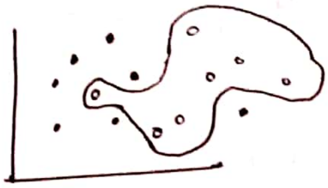
$b_2 = 1 - \{(-6.64) \times 0.49 + (-9.32) \times 0.61\} = 9.9388$

$\therefore$ Average, $b = \frac{b_1 + b_2}{2} = 8.9212$

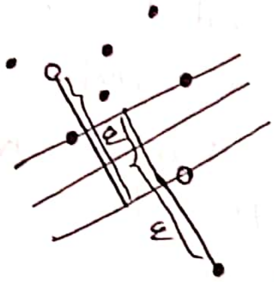
$X = (0.5, 0.5)$, তাহলে কোন class?

$$g(x) = w \cdot x + b$$
$$= -6.64 \times 0.5 - 9.32 \times 0.5 + 8.9212$$
$$= 0.9412$$
$$= +ve$$

Hard margin : ~~overfiffi~~ overfitting →
SV



Soft margin :
SV



$\forall_i \; (w^T x_i + b) \ge 1 - \epsilon_i \; [\epsilon_i \ge 0]$

$w^2 = w^T w$

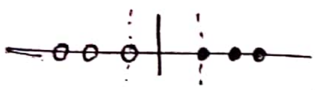
For hard : $\phi(w) = \frac{1}{2} w^T w$

$\forall_i \; (w^T x_i + b) \ge 1$

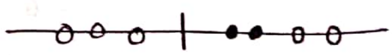
For soft : $\phi(w) = \frac{1}{2} w^T w + C \sum \epsilon_i$ → misclassify হলে cost বাড়বে। C = control overfitting

$\forall_i (w^T x_i + b) \ge 1 - \epsilon_i$

linearly separable ↓



non-linearly separable ↓



তাই → higher dimentional space এ নিব (ফাই function এর মাধ্যমে) ϕ

each point কে high dimention এ নিতে অনেক computation লাগে। φ function use করে 2টা point কে আগে high dimention এ নেয়া হয়। এখন kernel function এ ঐ point দিয়ে দিলে ফাই function এর কাজটাই explicitely করবে। ফলে computation low হয়ে যাবে।

শর্ত হচ্ছে kernel function এ parameter পাঠাতে চাইলে ঐ দুইজনকে dot product এর relation এ থাকতে হবে else kernel function কাজ করবেনা। $K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$

Bayes Theorem : Model দুই ধরনের হয় : generative, discriminative/linear।

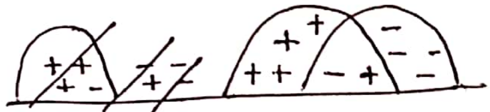
perceptron, SVM → discriminative → { এমনভাবে db draw করে যেন error or misclassify কম হয়

আগে data distribution বের করে, then সেটাকে classify করে : generative

Gaussian/Bernouli/Poison
↓
normal

mean, variance/standard deviation এর মাধ্যমে বের করে

step-1 : mean + sd দিয়ে ২টা class বের করে :



step-2 : threshold value set করে :

Bayes theorem, Naive bayes → generative

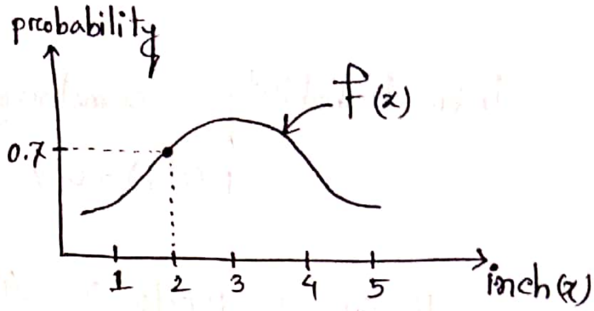
Random process : যা accurately বলা যায় না, guess করা লাগে।
Example : Flip a coin, rolling a dice, measuring the rain
Random process যে output দিবে তাকে quantify করে random variable।
Capital letter দিয়ে random variable represent করা হয়। যেমন: X, Y
For a coin toss : $X = \{h,t\}$ means coin toss is a random process যার output h,t এর মধ্যে সীমাবদ্ধ যাকে quantify করছি 'x' দিয়ে।
একটা random process হওয়ার probability কত সেটাই হচ্ছে event।
$P(x)$ $P(X=x)$ মানে X এর value x আসার probability কত।
$P(X=h)$ মানে X এর value h আসার probability কত

Discrete random variable : finite : single value এর উপর কাজ করে : countable
Continuous random variable : infinite : অসীম value এর উপর কাজ করে : uncountable

PDF (Probability Density Function) : একে area under the curve ও বলে।
continuos variable small letter দিয়ে প্রকাশ করে : $p(x|w_j)$
probability capital letter দিয়ে প্রকাশ করে : $P(w_j|x)$

Scanned with CamScanner

$\underbrace{\rightarrow Y}_{\text{continuous random variable}}$ = exact amount of rain tomorrow

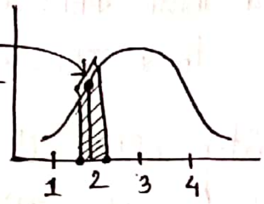


কতটুকু rain হবে assume করলাম

∴ $P(Y=2) = 0.7$ বলা যায়না কারন এটা continuous random variable

তাই ques এ 'Y=2' এটাই বলা যায় না।

ques can be → $\underbrace{P(1.9 < Y < 2.1)}_{\text{area under curve}}$



area of line = 0
↓ এজন্য
area of range বের করা লাগে

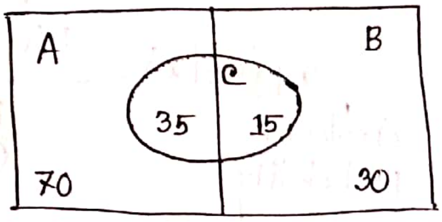
$$\therefore P(1.9<Y<2.1) = \int_{1.9}^{2.1} f(x) = \text{PDF}$$

joint probability
↓

$P(A) = \frac{70}{100} = 0.7$

$P(B) = \frac{30}{100} = 0.3$

$$\left\{ \begin{aligned} P(A,C) &= \frac{35}{100} = 0.35 \\ P(B,C) &= \frac{15}{100} = 0.15 \end{aligned} \right\}$$



Conditional Probability :

A এর respect এ C এর probability $P(C|A) = \frac{35}{70} = 0.5$

B এর respect এ C এর probability $P(C|B) = \frac{15}{30} = 0.5$

$$\left.\begin{aligned} \therefore P(A,C) &= P(C|A)\,P(A) \\ \Rightarrow 0.35 &= 0.5 \times 0.7 \\ \Rightarrow 0.35 &= 0.35 \quad [\text{proved}] \end{aligned}\right\} \rightarrow \text{Product Rule}$$

Likelihood Probability : $P(x|\omega_i)$; [x = feature ; ω = class]

ω_i class এ x feature যাওয়ার probability

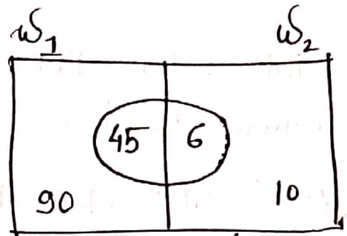
{if probability of x_1 event ω_1 greater than probability of x_1 event ω_2 হয় তাহলে x_1 belong করে ω_1 class এ

→ if $P(x_1|\omega_1) > P(x_1|\omega_2)$ then $x_1 \in \omega_1$

এটা kind of decision rule এর মতো কাজ করে

Prior Probability : something given in advance যেমন:

$P(\omega_1) = 0.7$; $P(\omega_2) = 0.3$



likelihood : $P(x|\omega_1) = \frac{45}{90} = 0.5$

$P(x|\omega_2) = \frac{6}{10} = 0.6$

যদি শুধু Likelihood think করি: ω_2 class এ যাবে $(0.6 > 0.5)$

যদি শুধু prior think করি: ω_1 class এ যাবে $(0.7 > 0.3)$

এটা ambiguity তৈরি হচ্ছে। মানে শুধু likelihood নিলে একটা class আবার শুধু prior নিলে আরেক class। তাই bayes theorem এ posterior probability আনা হয়।

$$P(\omega_i|x) = \frac{\overbrace{P(x|\omega_i)}^{\text{likelihood}}\ \overbrace{P(\omega_i)}^{\text{prior}}}{\underbrace{P(x)}_{\text{(evidence/normalization factor)}}} \longrightarrow \text{এটাই bayes theorem/rule}$$

(posterior probability)

<u>decision rule:</u> if $P(\omega_1|x) > P(\omega_2|x)$ then $x \in \omega_1$

<u>Bayesian Decision Theory</u> : এটা একটা classifier। decision make করে classifier design করার জন্য যা risk minimize করে। risk = classification error। ω = class/state of nature

conditional probability density = likelihood = class conditional probability

conditional probability = posterior

যদি ω_1 class এ যাওয়ার কথা হলেও ω_2 class এ যায় / ω_2 class এ যাওয়ার কথা হলেও ω_1 class এ যায়

$$P(\text{error}) = \begin{cases} P(\omega_1) & \text{if we decide } \omega_2 \\ P(\omega_2) & \text{if we decide } \omega_1 \end{cases}$$

or

২টা probability এর মধ্যে যে minimum সেটাই আমার error ← $P(\text{error}) = \min[P(\omega_1), P(\omega_2)]$

ω_1 if $P(\omega_1/x) > P(\omega_2|x)$; else ω_2

ω_1 if $P(x|\omega_1)\,P(\omega_1) > P(x|\omega_2)\,P(\omega_2)$; else ω_2

ω_1 if $\underbrace{P(x|\omega_1)/P(x|\omega_2)}_{\text{likelihood ratio}} > \underbrace{P(\omega_2)/P(\omega_1)}_{\text{threshold}}$; else ω_2 ⟶ Bayes Rules

$$P(\text{error}|x) = \begin{cases} P(\omega_1|x) & \text{if we decide } \omega_2 \\ P(\omega_2|x) & \text{if we decide } \omega_1 \end{cases}$$

$\therefore P(\text{error}|x) = \min\,[P(\omega_1|x), P(\omega_2|x)]$

} probability of error

$$P(\text{error}) = \int_{-\alpha}^{\alpha} P(\text{error}, x)\,dx = \int_{-\alpha}^{\alpha} P(\text{error}|x)\,p(x)\,dx$$

} average probability of error

Minimum error classifier বলে কারণ error টাকে যতটা সম্ভব minimum করে } Bayes Theorem

যদি এক এর বেশি feature/class থাকে তখন একটা general error function বানানো যাবে যার নাম হবে expected risk যা cost এর সাথে associate থাকবে আর cost তৈরি হবে loss function থেকে। Loss function হচ্ছে classify correct হলে '1' দিবে, misclassify হলে '0' দিবে এমন।

Feature vector $= x =$ Real number

$\omega_1, \omega_2, \omega_3 \ldots \omega_c = c$ সংখ্যক class

action = কোন class এ যাবে/যাবে না $= l$ সংখ্যক class $= \alpha_1, \alpha_2 \ldots \alpha_l$

loss function $= \lambda(\alpha_i|\omega_j)$ [α_i = action; ω_j = class]

মানে decide করেছি α_i class এ যাবে
কিন্তু আসলে যাবে j class এ

x কে পাঠালাম i class এ কিন্তু আসলে হবে j class

loss function

expected loss/ conditional risk

$$R(\alpha_i|x) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) P(\omega_j|x)$$

actually j class এ যাবে তার probability

পাঠালাম i class এ

actually যাবে j class এ

যেহেতু 'c' সংখ্যক class তাই c times add হবে

overall risk : $R = \int R(\alpha(x)|x)\, p(x)\, dx$

এখানে $\alpha(x)$ = general decision rule = determine করে কোন action নিবে x এর জন্য মানে α_1? নাকি α_2? নাকি α_k?

Bayes rule overall risk কে কমায়। loss কে কমায়।

step : 1 - সব $R(\alpha_i|x)$ calculate করব for given x

step : 2 - minimum $R(\alpha_i|x)$ choose করব

যে class এ error কম feature টা ঐ class এ যাবে +

যে class এ probabity বেশি feature টা ঐ class এ যাবে

$\lambda_{ij} = \lambda(\alpha_i|\omega_j)$ } যাওয়ার জন্য যে loss টা সেটাই λ_{ij}

মানে λ_{ij} তে $i=j$ হলে error নাই মানে predict করেছি i but real এ j এবং $i=j$। means সত্যিটাই predict করেছি।

conditional risks

$$R(\alpha_i|x) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) P(\omega_j|x)$$

দুইটা class চিন্তা করলে

$$R(\alpha_1|x) = \lambda_{11} P(\omega_1|x) + \lambda_{12} P(\omega_2|x)$$

$$R(\alpha_2|x) = \lambda_{21} P(\omega_1|x) + \lambda_{22} P(\omega_2|x)$$

Two category classification

minimum error

$\lambda_{11}, \lambda_{22}$ = no error ; $\lambda_{12}, \lambda_{21}$ = error

ω_1 if $R(\alpha_1|x) < R(\alpha_2|x)$; else ω_2
মানে যে class এ loss কম ঐ class এ যাচ্ছে

} minimum risk decision rule

ω_1 if $(\lambda_{21} - \lambda_{11}) P(\omega_1|x) > (\lambda_{12} - \lambda_{22}) P(\omega_2|x)$; else ω_2

$$\omega_1 \text{ if } \underbrace{\frac{p(x|\omega_1)}{p(x|\omega_2)}}_{\text{likelihood ratio}} > \underbrace{\frac{(\lambda_{12} - \lambda_{22})}{(\lambda_{21} - \lambda_{11})} \frac{P(\omega_2)}{P(\omega_1)}}_{\text{threshold}} \text{ ; else } \omega_2$$

Zero-One loss function

loss function : $i = j$ হলে 0, $i \neq j$ হলে 1 $\xrightarrow{\text{তখন}}$ $\frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} = \frac{1-0}{1-0} = 1$

Minimum risk classification rule

$$\lambda(\alpha_i|\omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$$

$$\therefore R(\alpha_i|x) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) P(\omega_j|x) = \sum_{i \neq j} P(\omega_j|x) = 1 - P(\omega_i|x)$$

i, j সমান না হলে 1 হবে

So R কমাতে চাইলে $P(\omega_i|x)$ বাড়াতে হবে, R কমলে Probability বাড়ে।
মানে একই সাথে loss কমাচ্ছি + max probability নিচ্ছি।
So, Minimum risk classification এর objective fulfil হচ্ছে।

Discriminant function for Bayesian theory:
general loss function : $g_i(x) = -R(\alpha_i|x)$
zero-one loss function : $g_i(x) = P(\omega_i|x)$

$$g_i(x) = \frac{p(x|\omega_i) P(\omega_i)}{p(x)}$$

$= p(x|\omega_i) P(\omega_i)$ (যে class গুলো থাকে তাদের evidence same হয় তাই $P(x)$ বাদ দেয়া যায়)

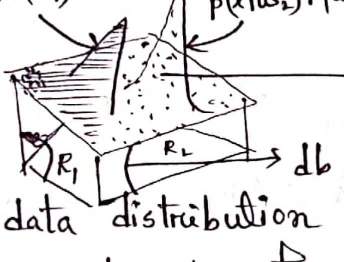
new discriminant function } $\rightarrow = \ln p(x|\omega_i) + \ln P(\omega_i)$

$\swarrow$ dichotomizer

ω_1 if $g(x) > 0$: $g(x) = g_1(x) - g_2(x)$ (db এর equation)
else ω_2

$$= P(\omega_1|x) - P(\omega_2|x)$$

$$= \ln \underbrace{\frac{p(x|\omega_1)}{p(x|\omega_2)}}_{\text{likelihood ratio}} + \ln \underbrace{\frac{P(\omega_1)}{P(\omega_2)}}_{\text{priore ratio}}$$ ← ln applied যেন math করতে easy হয়



উঁচু বোঝানো হয়েছে: ঘনত্ব / density এর জন্য মানে উচ্চতা দিয়ে বুঝিয়েছে বেশি data।

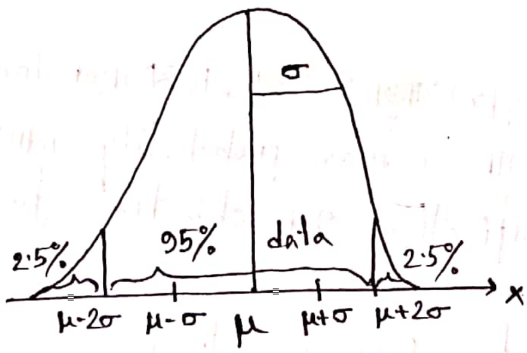
data distribution describe করে figure

একে Gaussian/normal distribution বলে

১টা feature এর উপর base করে data distribution করব সেটাকে বলে univariate normal distribution।
১টার বেশি feature হলে তাকে multivariate বলে।

p(x) —— pdf

normal distribution →



মিউ/mean

bell shape

σ^2 = variation/variance
σ = standard deviation

mean এ data বেশি থাকে তাই mean এর জায়গাটা narrow থাকে।

variance = mean থেকে data গুলো কতখানি scattered।
mean এর কাছে data density maximum

only 'x' feature এর উপর base করে draw করা তাই এটা univariate normal distribution এর equation :

univariate normal density equation

$$P(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad / \quad P(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

continuous density

μ আর x এর দূরত্ব $= x-\mu$

σ দিয়ে ভাগ = normalize এর জন্য

multivariate normal density equation

probability density/distribution function — normal distribution function

$\hookrightarrow P(x) \sim N(\mu, \Sigma)$

convert করছি with the help of মিউ (μ) + Capital sigma (Σ)

co-variance matrix

এক এর বেশি feature থাকলে

$$P(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^t \cdot \Sigma^{-1}(x-\mu)\right]$$

transpose

$(x-\mu)^t \cdot (x-\mu) = (x-\mu)^2$

$[a^T a = a^2 \text{ formula}]$

$|\Sigma|$ = determinant

$\Sigma = d \times d$ covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_{1d} \\ \vdots & & \vdots \\ \sigma_{d1} & \cdots & \sigma_d^2 \end{bmatrix}$$

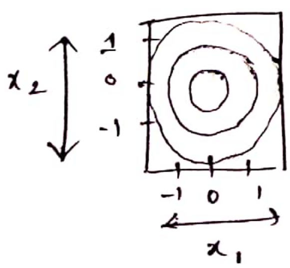
variance

co-variance/co-relation

normal distribution

$$\hookrightarrow \mathcal{N}(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right]$$

correlated মানে ১টা বাড়লে আরেকটা বাড়বে means x_1 আর x_2 correlated হবে যদি x_1 বাড়লে x_2 বাড়ে, x_1 কমলে x_2 কমে।



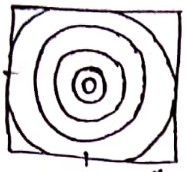
x_1, x_2 aren't correlated.

circle shape এ correlated হয়না, oval shape এ হয়।

top view দেখলে যেটাকে contour বলে।

2d multivariate normal density

x_2 0



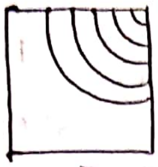
0 x_1

$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

মানে diagonal { 1, 1 → তাই variance = সমান

0, 0 = covariance means independent (correlation নাই)

$\mu = [0, 0]$ means mean 0,0 তে



$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

Σ like before

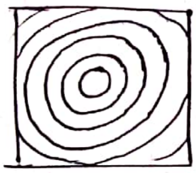
$\mu = [1, 1]$ মানে just mean changed

$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}$

x_2 এর জন্য variance changed। এজন্য shape টা x_2 বরাবর লম্বা হয়েছে। means x_2 এর দিকে density বেশি

$\Sigma = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$

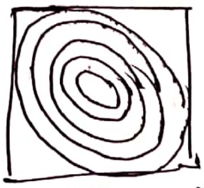
x_1 এর জন্য সব হবে



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

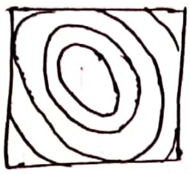
$\Sigma = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$

co-variance বাড়ানোতে আরও squeezed হয়েছে



$\Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$

negative এর জন্য figure ঘুরে গিয়েছে



$\Sigma = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 4 \end{bmatrix}$

co-variance দেখে visualize করা যায় যে কোন feature রাখা দরকার আর কোনটা রাখা দরকার নাই।

circular {
- circular মানে variance সমান + co-variance নাই
- circular টাই যেকোন ১ দিকে squeeze হওয়া মানে variance এ difference আছে

non circular {
- angle করে plot হলে variance + co-variance both আছে।
- diagonal বরাবর। যত বেশি squeezed তত বেশি co-variance

discriminant function: $g_i(x) = \ln p(x|\omega_i) + \ln P(\omega_i)$

if $p(x|\omega_i) \sim N(\mu_i, \Sigma_i)$ then $g_i(x) = -\frac{1}{2}(x-\mu_i)^t \Sigma^{-1}(x-\mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$

multivariate gaussian density

case: I : $\Sigma_i = \sigma^2 I$ (diagonal matrix)

(variance আছে but co-variance নাই)

যেহেতু co-variance নাই, তাই statistically independent + প্রত্যেক feature এর variance সমান

$\therefore g_i(x) = -\frac{\|x-\mu_i\|^2}{2\sigma^2} + \ln P(\omega_i)$ [disregarding $\frac{d}{2}\ln 2\pi$ and $\frac{1}{2}\ln|\Sigma_i|$]

$\|x-\mu_i\|^2 = (x-\mu_i)^T (x-\mu_i)$

$\therefore g_i(x) = -\frac{1}{2\sigma^2}[x^t x - 2\mu_i^t x + \mu_i^t \mu_i] + \ln P(\omega_i)$ [$\|x\|^2 = x^t \cdot t$]

disregarding $x^t x$ (constant),

$$g_i(x) = -\frac{1}{2\sigma^2}(-2\mu_i^t x + \mu_i^t \mu_i) + \ln P(\omega_i) \quad \dots \dots (1)$$

$$= \frac{-2\mu_i^t x}{-2\sigma^2} - \frac{\mu_i^t \mu_i}{2\sigma^2} + \ln P(\omega_i)$$

$$= \underbrace{\frac{\mu_i^t}{\sigma^2}}_{w_i^t} x + \underbrace{\left[-\frac{\mu_i^t \mu_i}{2\sigma^2} + \ln P(\omega_i)\right]}_{w_{i0}}$$

$$= w_i^t x + w_{i0}$$ [which is a linear discriminant]

for db: $g_i(x) = g_j(x)$

$w^t(x - x_0) = 0$

hyperplane টা w এর উপর perpendicular।

$w = \mu_i - \mu_j$ ← reason of orthogonal

এবং x_0 দিয়ে যাবে।

{μ_1 থেকে μ_2 তে line draw করলে তার mid point টাই x_0

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{\|\mu_i - \mu_j\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (\mu_i - \mu_j)$$

From equation ①,

$$g_j(x) = -\frac{1}{2\sigma^2}\left(-2\mu_j^t x + \mu_j^t \mu_j\right) + \ln P(\omega_j)$$

$$\therefore g_i(x) = g_j(x) \quad [\text{for db}]$$

$$\Rightarrow -\frac{1}{2\sigma^2}\left(-2\mu_i^t x + \mu_i^t \mu_i\right) + \ln P(\omega_i) = -\frac{1}{2\sigma^2}\left(-2\mu_j^t x + \mu_j^t \mu_j\right) + \ln P(\omega_j)$$

$$\Rightarrow \frac{-2\mu_i^t x}{-2\sigma^2} - \frac{\mu_i^t \mu_i}{2\sigma^2} + \ln P(\omega_i) = \frac{-2\mu_j^t x}{2\sigma^2} - \frac{\mu_j^t \mu_j}{2\sigma^2} + \ln P(\omega_j)$$

$$\Rightarrow \frac{\mu_i^t x}{\sigma^2} - \frac{\mu_i^t \mu_i}{2\sigma^2} + \ln P(\omega_i) = \frac{\mu_j^t x}{\sigma^2} - \frac{\mu_j^t \mu_j}{2\sigma^2} + \ln P(\omega_j)$$

$$\Rightarrow \frac{\mu_i^t x}{\sigma^2} - \frac{\mu_j^t x}{\sigma^2} - \frac{\mu_i^t \mu_i}{2\sigma^2} + \frac{\mu_j^t \mu_j}{2\sigma^2} + \ln P(\omega_i) - \ln P(\omega_j) = 0$$

$$\Rightarrow \frac{x}{\sigma^2}\left(\mu_i^t - \mu_j^t\right) - \frac{1}{2\sigma^2}\left[\|\mu_i\|^2 - \|\mu_j\|^2\right] + \ln \frac{P(\omega_i)}{P(\omega_j)} = 0$$

$$\Rightarrow x\left(\mu_i^t - \mu_j^t\right) - \frac{(\mu_i + \mu_j)(\mu_i - \mu_j)}{2} + \sigma^2 \ln \frac{P(\omega_i)}{P(\omega_j)} = 0$$

$$\Rightarrow \left(\mu_i^t - \mu_j^t\right)\left[x - \frac{(\mu_i + \mu_j)}{2} + \frac{\sigma^2}{(\mu_i - \mu_j)} \ln \frac{P(\omega_i)}{P(\omega_j)}\right] = 0$$

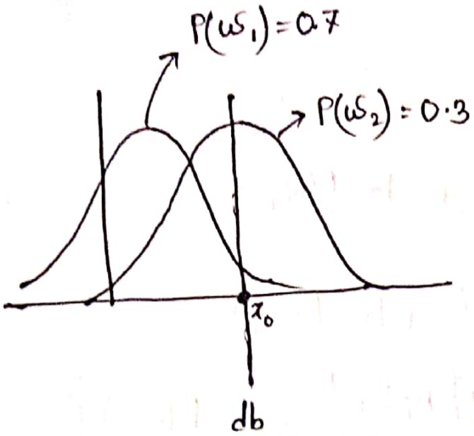
$$\Rightarrow \underbrace{\left(\mu_i^t - \mu_j^t\right)}_{w^t}\left[x - \underbrace{\left\{\frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{\|\mu_i - \mu_j\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (\mu_i - \mu_j)\right\}}_{x_0}\right] = 0$$

$$\therefore w^t (x - x_0) = 0$$

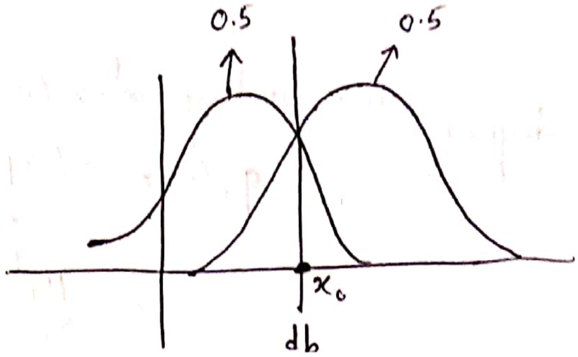
যদি ২টা class এর prior probability equal হয়, তখন $P(\omega_i) = P(\omega_2)$ মানে $\ln(1)$ which is 0, তাই $\frac{\sigma^2}{\|\mu_i - \mu_j\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (\mu_i - \mu_j) = 0$ হবে।

তখন $x_0 = \frac{1}{2}(\mu_i + \mu_j)$ হবে।

$P(\omega_i) \neq P(\omega_j)$ হলে, যার priority বেশি x_0 সেদিকের মানের থেকে দূরে সরে যাবে।



যেহেতু $P(\omega_1) > P(\omega_2)$ তাই x_0 টা $P(\omega_1)$ থেকে দূরে সরে গিয়েছে



যখন ২টা class এর priority সমান তখন middle দিয়ে intersect করবে।

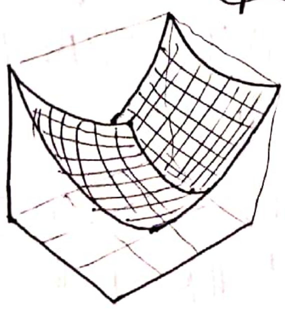
prior সমান হলে তখন minimum distance classifier এর কাছে যায়।

discriminant becomes : $$g_i(x) = -\frac{\|x-\mu_i\|^2}{2\sigma^2} + \ln P(\omega_i)$$

$$= -\|x-\mu_i\|^2 \quad \text{(euclidean distance)}$$

Case : II : ২টা class এর co-variance matrix সমান : $\Sigma_i = \Sigma$

clusters have hyperellipsoidal shape and same size (centered at μ)



disregarding $\frac{d}{2}\ln 2\pi$ and $\frac{1}{2}\ln|\Sigma_i|$ constants :

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t \Sigma^{-1}(x-\mu_i) + \ln P(\omega_i)$$

$$= \frac{(x-u_i)^t(x-\mu_i)}{-2\Sigma} + \ln P(\omega_i)$$

$$= \frac{\|x-u_i\|^2}{-2\Sigma} + \ln P(\omega_i)$$

$$= \frac{\|x\|^2 - 2\cdot x\cdot\mu_i^t + \|\mu_i\|^2}{-2\Sigma} + \ln P(\omega_i)$$

$$= \frac{x^t\cdot x - 2x\mu_i^t + \mu_i^t\cdot\mu_i}{-2\Sigma} + \ln P(\omega_i)$$

disregarding $x^t \cdot x$ (consant) :

$$g_i(x) = \frac{-2x\mu_i^t + \mu_i^t \mu_i}{-2\Sigma_i} + \ln P(\omega_i)$$

$$= \frac{\mu_i^t}{\Sigma_i} x - \frac{\mu_i^t \mu_i}{2\Sigma_i} + \ln P(\omega_i)$$

$$= \underbrace{\Sigma_i^{-1} \mu_i^t}_{w_i^t} x + \underbrace{\left\{ -\frac{\mu_i^t \mu_i \Sigma_i^{-1}}{2} + \ln P(\omega_i) \right\}}_{w_{i0}}$$

$$= w_i^t x + w_{i0}$$ [which is a linear discriminent]

for db : $g_i(x) = g_j(x)$

$$\Rightarrow \frac{\mu_i^t}{\Sigma} x - \frac{\mu_i^t \mu_i}{2\Sigma} + \ln P(\omega_i) = \frac{\mu_j^t}{\Sigma} x - \frac{\mu_j^t \mu_j}{2\Sigma_i} + \ln P(\omega_j)$$

$$\Rightarrow \frac{\mu_i^t}{\Sigma_i} x - \frac{\mu_j^t}{\Sigma_i} x - \frac{\mu_i^t \mu_i}{2\Sigma_i} + \frac{\mu_j^t \mu_j}{2\Sigma_i} + \ln P(\omega_i) - \ln P(\omega_j) = 0$$

$$\Rightarrow x \left\{ (\mu_i^t - \mu_j^t) \Sigma_i^{-1} \right\} - \left(\frac{\mu_i^t \mu_i - \mu_j^t \mu_j}{2} \right) \Sigma_i^{-1} + \ln \frac{P(\omega_i)}{P(\omega_j)} = 0$$

$$\Rightarrow x \left\{ \Sigma_i^{-1} (\mu_i^t - \mu_j^t) \right\} - \Sigma_i^{-1} \left\{ \frac{1}{2} (\|\mu_i\|^2 - \|\mu_j\|^2) \right\} + \frac{\|\mu_i - \mu_j\|^2}{\|\mu_i - \mu_j\|^2} \cdot \ln \frac{P(\omega_i)}{P(\omega_j)} = 0$$

$$\Rightarrow \Sigma^{-1} (\mu_i^t - \mu_j^t) \left[x - \frac{1}{2}(\mu_i + \mu_j) + \frac{\ln [P(\omega_i)/P(\omega_j)] (\mu_i - \mu_j)}{\Sigma^{-1} \|\mu_i - \mu_j\|^2} \right] = 0$$

$$\Rightarrow \underbrace{\Sigma^{-1} (\mu_i^t - \mu_j^t)}_{w^t} \left[x - \underbrace{\left\{ \frac{1}{2}(\mu_i + \mu_j) - \frac{\ln [P(\omega_i)/P(\omega_j)] (\mu_i - \mu_j)}{\Sigma^{-1} \|\mu_i - \mu_j\|^2} \right.}_{x_0} \right] = 0$$

$\therefore w^t (x - x_0) = 0$

hyperplane টা perpendicular হবে না w^t এর সাথে। বরং আড়াআড়ি ভাবে যাবে কেননা w^t এর মধ্যে Σ^{-1} আছে। x_0 point দিয়েই যাবে।

$\downarrow$ reason of not orthogonal

Σ^{-1} বলছে যে direction টা hyperplane এর দিকে না।

prior equal হলে, $P(\omega_i) = P(\omega_j)$

discriminant : $g_i(x) = -\frac{1}{2}(x-\mu_i)^t \Sigma^{-1}(x-\mu_i) + \ln P(\omega_i)$

$= -\frac{1}{2}(x-\mu_i)^t \Sigma^{-1}(x-\mu_i)$ [mahalanobis distance classifier]

higher dimension এ ২টা point এর distance বের করতে চাইলে mahalanobis distance classifier use হয়।

Case : III : Most General Case : Σ_i = arbitrary

Clusters have different shapes and sizes

disregarding $\frac{d}{2}\ln 2\pi$ (constant) :

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t \Sigma_i^{-1}(x-\mu_i) - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

$$= -\frac{1}{2}\Sigma_i^{-1}\|x-\mu_i\|^2 - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

$$= -\frac{1}{2}\Sigma_i^{-1}\{\|x\|^2 - 2\cdot x\cdot\mu_i + \|\mu_i\|^2\} - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

$$= -\frac{1}{2}\Sigma_i^{-1}\|x\|^2 + x\mu_i\Sigma_i^{-1} - \frac{1}{2}\Sigma_i^{-1}\|\mu_i\|^2 - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

$$= x^t\cdot\underbrace{\left(-\frac{1}{2}\Sigma_i^{-1}\right)}_{W_i}\cdot x + \underbrace{(\Sigma_i^{-1}\mu_i)}_{w_i}\cdot x + \underbrace{\left(-\frac{1}{2}\mu_i^t\Sigma^{-1}\cdot\mu - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)\right)}_{w_{i0}}$$

$= x^t W_i x + w_i + w_{i0}$ [quadratic discriminant] [not linear]

db is determined by hyperquadrics : $g_i(x) = g_j(x)$

e.g. hyperplanes, pairs of hyperplanes, hyperspheres, hyperellipsoids, hyperparaboloids etc.

Classifier based on :

MLE = maximum likelihood event

MAP = maximum a-posterior probability

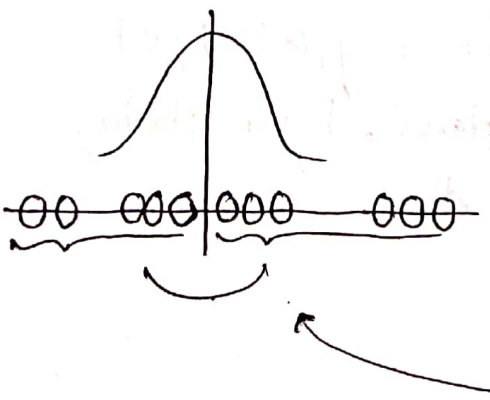
MAP : $$P(\omega_i|x) = \frac{P(x|\omega_i) \cdot P(\omega_i)}{P(x)}$$ (bayes theorem)

MLE : $$P(x|\omega_i) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$ (Gaussian density function)

Parameter Estimation : (MLE)

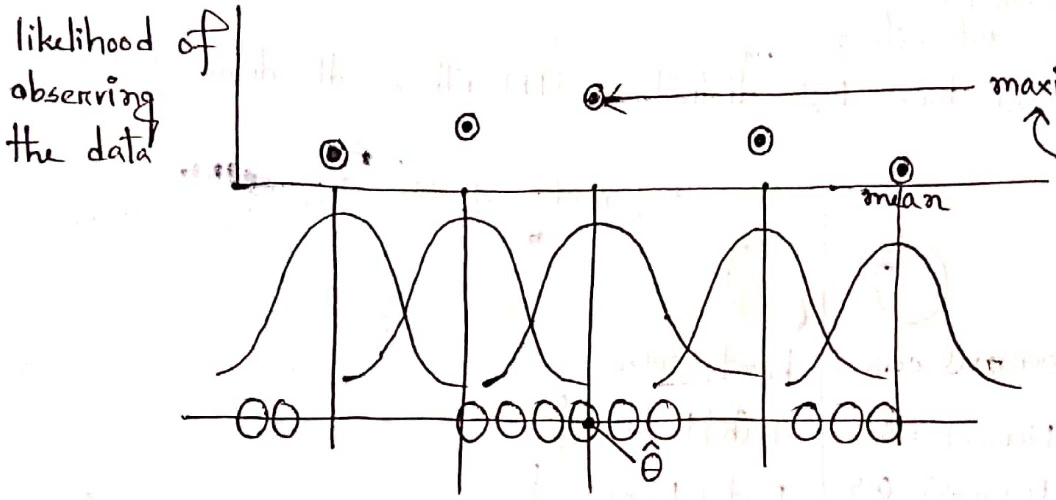
class conditional densities = likelihood = $p(x|\omega_i)$

minimum error classifier/optimal classifier বানানো সম্ভব যদি likelihood আর prior জানা থাকে। কিন্তু real life এ prior পাওয়া গেলেও likelihood পাওয়া difficult. So, complete knowledge না থাকা সত্ত্বেও কিভাবে probability বের করব সেটাই দেখব। সবসময় Likelihood জানা থাকেনা, তাই bayesian distribution best হওয়া সত্ত্বেও আমাদের gaussian/bernoulli/poison/exponential distribution use করা লাগে। prior এর knowledge ও জানা না থাকতে পারে। high dimension এ গেলে এমন problem আসে।



gaussian distribution করলে mean বের করা হয়। দেখা যাচ্ছে max data mean এর কাছাকাছি। এখানে একটা symmetric structure maintain করে।

distribution কে left এ সরালে, mean এর কাছাকাছি max value থাকবেনা



best fit মানে যেখানে mean ধরলে max data কাছাকাছি থাকে

gaussian distribution এর parameter ২টা : mean, variance

Feature vector x বেশি বড় হলে কোন distribution হবে জানা কঠিন।

MLE : Parameter জানব কিন্তু parameter এর value জানি না।

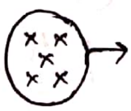
MAP/Bayesian method : শুধু হয়তো prior দেয়া থাকবে।

যেমন: mean (μ_i), covariance matrix (Σ_i) জানা থাকলেও μ_i, Σ_i এর value জানা থাকবে না।

Linear Discriminant Analysis

Generative Method

Bayesian Decision Rule



Model: $N(\mu, \Sigma)$

↓

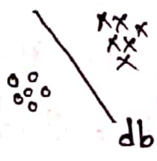
decision

$g(x) = \omega^T x + \omega_0$

আগে distribution বের করি then সেটার উপর base করে decision make করি যা linear/quadratic হয়

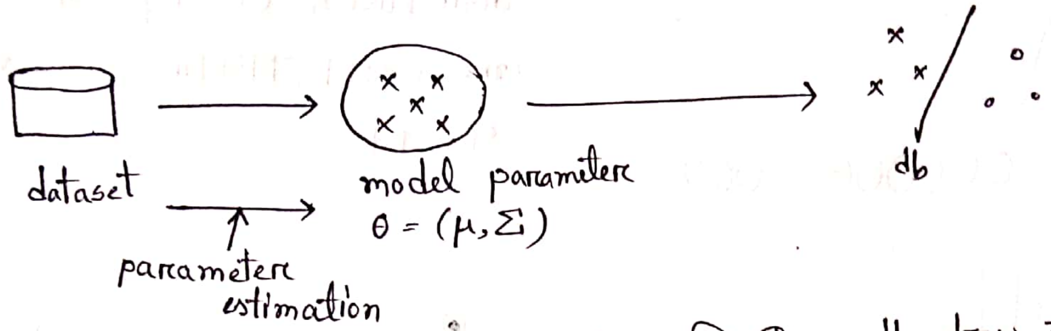
Discriminative method

- logistic regression
- SVM
- Deep networks

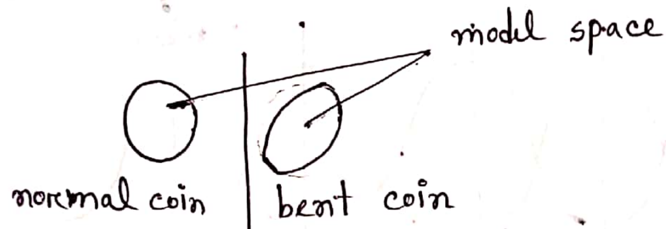


$g(x) = \omega^T x + \omega_0$

(মার্জিন max করে db draw করি)



θ এর উপর base করে distribute করি। Then db draw করি।



$P(h \mid Normal) = 0.5$	$P(h \mid Bent) = \frac{4}{5}$
$P(t \mid Normal) = 0.5$	$P(t \mid Bent) = \frac{1}{5}$

event

এখন যদি output এমন হয়: HTH HHT THH HTH
তাহলে normal coin বেশি flip করা হয়েছে নাকি bent coin?
ans: bent coin কারন output এ H বেশি এসেছে আর bent coin এই H আসার probability বেশি।

MLE : $$\underset{coin \in (bent, normal)}{\arg\max} \; P(HTHHHTTHHHTH \mid Coin)$$

$$\underset{model \in model\ space}{\arg\max} \; P(data \mid model)$$

মানে given model এর respect এ যত data দেয়া আছে তাদের মধ্যে max বের করলেই ans চলে আসে।

আগে feature, পরে class, তাই এটা likelihood। So, the whole thing is maximum likelihood কারন max বের করছি

$$\therefore P(data \mid bent) = \frac{4}{5}\cdot\frac{1}{5}\cdot\frac{4}{5}\cdot\frac{4}{5}\cdot\frac{4}{5}\cdot\frac{1}{5}\cdot\frac{1}{5}\cdot\frac{4}{5}\cdot\frac{4}{5}\cdot\frac{4}{5}\cdot\frac{4}{5}\cdot\frac{1}{5}\cdot\frac{4}{5} \approx 0.000268$$

$$\therefore P(data \mid normal) = \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2} \approx 0.000244$$

∴ max = bent (ans:) (এ কাজটাই করছে likelihood) i.i.d = independent $\underbrace{\text{identically}}$ distributed

২ জনের parameter value same

} samples should be i.i.d in MLE

$p(x|\omega_j)$ কে লিখা যাবে $p(x|\omega_j, \theta_j)$

likelihood এর value ঠিক হবে বের করব but সাথে দিয়ে দিয়েছি θ_j

$p(x|\omega_j)$ এই likelihood এর equivalence $N(\mu_j, \Sigma_j)$ normal distribution

parameter এর নাম θ_j

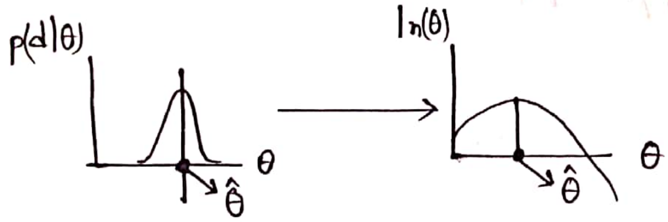
θ_j এর ভিতর μ_j, Σ_j

θ এর respect এ x sample গুলো বের করে গুন করছি

likelihood এর equation $\rightarrow$ $\{ F(\theta) = p(D|\theta) = \prod_{k=1}^{n} p(x_k|\theta) \longrightarrow$ (if D = fixed ; $p(D|\theta)$ isn't density)

$p(D|\theta)$ = likelihood of θ with respect to dataset

θ এর যে value এর জন্য $P(D|\theta) = max$ এটাই $\hat{\theta}$ (mean value for gaussian distribution)



{ ln নিলে curve টা spread করে + math এও সুবিধা। spread করলেও max mean same place এই।

$p(x|\theta) \rightarrow \theta$ এর জন্য x বের করছি
তাই it's not conditional density

MLE : $\hat{\theta} = \arg_{\theta} \max \left(p(D|\theta) \right)$

easier

$\ln \left(p(D|\theta) \right)$

$= \arg_{\theta} \max \left(\ln \prod_{k=1}^{n} p(x_k|\theta) \right)$

$= \arg_{\theta} \max \left(\sum_{k=1}^{n} \ln p(x_k|\theta) \right)$

ln করায় product (Π) converted into sum (Σ)

derivative বের করব : $\nabla_\theta = \left[\frac{\partial}{\partial \theta_1}, \frac{\partial}{\theta_2}, \dots \frac{\partial}{\partial \theta_p}\right]^t$

θ এর gradient

যেই derivative এর value 0 হবে সেটাই max value।

$\nabla_\theta I = 0 \qquad [I(\theta) = \ln(p(D|\theta))]$.

Suppose, μ = unknown

σ^2 = variance = known

$\therefore$ unknown needs to be estimated. So, $\theta = \mu$

$$\hat{\mu} = \arg\max_{\mu} I(\mu) = \arg\max_{\mu}\left(\sum_{k=1}^{n} \ln p(x_k|\mu)\right)$$

$$= \arg\max_{\mu}\left(\sum_{k=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_k-\mu)^2}{2\sigma^2}}\right)\right) \quad [\because p(x|\mu) = N(\mu, \sigma^2)]$$

$$= \arg\max_{\mu} \sum_{k=1}^{n}\left(-\frac{(x_k-\mu)^2}{2\sigma^2} - \ln\sqrt{2\pi}\sigma\right)$$

$$\therefore \frac{d}{d\mu}(I(\mu)) = 0$$

$$\Rightarrow \sum_{k=1}^{n} \frac{1}{\sigma^2}(x_k - \mu) = 0$$

$$\Rightarrow \sum_{k=1}^{n} x_k - n\mu = 0$$

$$\Rightarrow \hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k \quad \text{(mean এর equation)}$$

MLE দিয়ে prove করলাম যে, সব data sum করে data সংখ্যা দিয়ে ভাগ করলে mean আসে যা normally আমরা জানি।

if unknown $= \sigma^2$ then

$$\frac{d}{dt}(I(t)) = 0 \qquad (t = \sigma^2 \text{ suppose})$$

$$\Rightarrow \frac{d}{dt}\sum_{k=1}^{n}\left(-\frac{(x_k-\mu)^2}{2t} - \ln\sqrt{2\pi t}\right) = 0$$

$$\Rightarrow \sum_{k=1}^{n} \left\{ -\frac{(x_k - \mu)^2}{2} \cdot (-1) \cdot t^{-2} - \frac{1}{\sqrt{2\pi t}} \cdot \sqrt{2\pi} \cdot \frac{1}{2} \cdot t^{-\frac{1}{2}} \right\} = 0$$

$$\Rightarrow \sum_{k=1}^{n} \left(\frac{(x_k - \mu)^2}{2t^2} - \frac{1}{2\sqrt{t} \cdot \sqrt{t}} \right) = 0$$

$$\Rightarrow \sum_{k=1}^{n} \left(\frac{(x_k - \mu)^2}{2t^2} - \frac{1}{2t} \right) = 0$$

$$\Rightarrow \sum_{k=1}^{n} \left(\frac{(x_k - \mu)^2 - t}{2t^2} \right) = 0$$

$$\Rightarrow \sum_{k=1}^{n} \left\{ (x_k - \mu)^2 - t \right\} = 0$$

$$\Rightarrow \sum_{k=1}^{n} (x_k - \mu)^2 - nt = 0$$

$$\Rightarrow nt = \sum_{k=1}^{n} (x_k - \mu)^2$$

$$\therefore \sigma^2 = \frac{1}{n} \sum_{k=1}^{n} (x_k - \hat{\mu})^2 \quad [t = \sigma^2]$$

Expected value আসলেই mean।

$$E(X) = \mu = \begin{cases} \sum_x x f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x f(x) dx & \text{if } X \text{ is continuous} \end{cases}$$

$\{ f(x) =$ probability distribution

1	2	3	4	5	6	$\rightarrow x$
0.1	0.1	0.1	0.1	0.1	0.5	$\rightarrow p$

$\therefore \sum xp = 1\times0.1 + 2\times0.1 + 3\times0.1 + 4\times0.1 + 5\times0.1 + 6\times0.5$

$= 4.5$

$\therefore \mu = 4.5$ (expected value)

probability distribution $f(x)$

The variance of a random variable X is defined as the expect squared deviation from the expected value.

$$\sigma^2 = Var(X)$$
$$= E[(X-\mu)^2]$$
$$= E[(X - E(X))^2]$$
$$= \begin{cases} \sum_x (x-\mu)^2 f(x) & \text{if } X \text{ is discrete} \\ \int_{-\alpha}^{\alpha} (x-\mu)^2 f(x)\,dx & \text{if } X \text{ is continuous.} \end{cases}$$

$$E[(X-\mu)^2] = E(X^2 - 2\mu X + \mu^2)$$
$$= E(X^2) - 2\mu E(X) + \mu^2$$
$$= E(X^2) - 2\mu\cdot\mu + \mu^2$$
$$= E(X^2) - 2\mu^2 + \mu^2$$
$$= E(X^2) - \mu^2$$
$$\therefore \sigma^2 = Var(X) = E(X^2) - [E(X)]^2$$

$$Var(X) = E[(X-\mu)^2]$$
$$= \sum_x (x-\mu)^2 p(x)$$
$$= \sum_x (x^2 - 2x\mu + \mu^2) p(x)$$
$$= \sum_x x^2 p(x) - 2\mu \sum x\, p(x) + \mu^2 \sum_x p(x)$$
$$= E[X^2] - 2\mu^2 + \mu^2$$
$$= E[X^2] - \mu^2$$

$$\therefore Var(X) = \sum x^2 p - \mu^2$$
$$= 23.5 - (4.56)^2$$
$$= 3.25$$

$$\sum x^2 p = 0.1 + 0.4 + 0.9 + 1.6 + 2.5 + 18$$
$$= 23.5$$

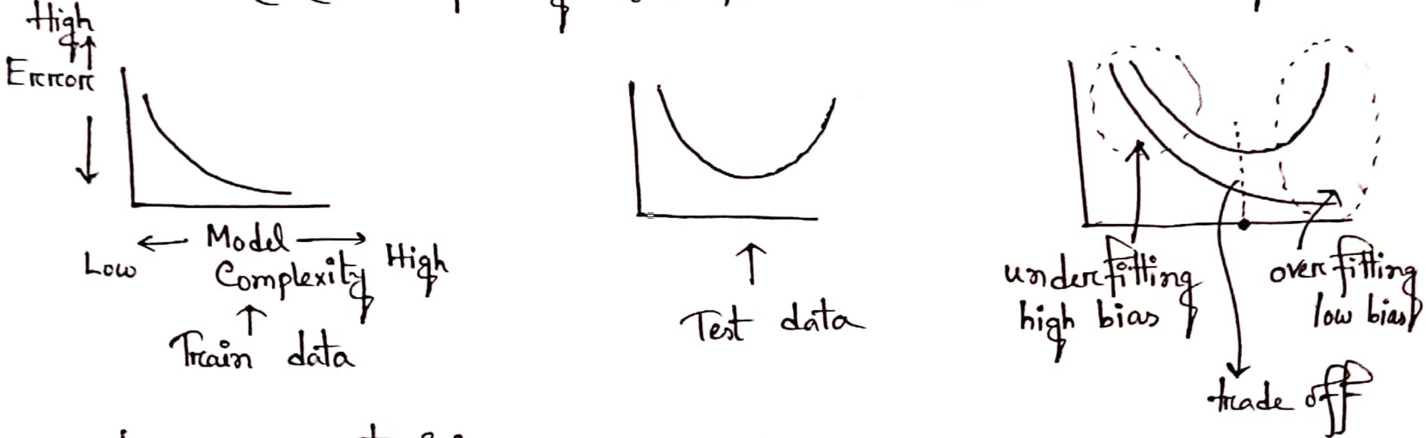
an estimator of a parameter is biased if the expected value of the estimate is different from the true value of the parameters

$$E[\hat{\sigma}^2] = \frac{n-1}{n}\sigma^2 \; ; \; E[\hat{\Sigma}] = \frac{n-1}{n}\Sigma$$

an estimator of a parameter is unbiased if the expected value of the estimate is same as the true value of the parameters.

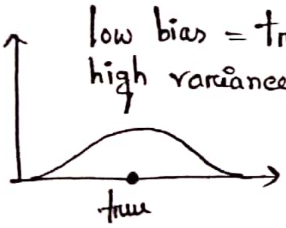
$$\Sigma[\hat{\theta}] = \theta \; ; \; E[\hat{\mu}] = \mu$$

estimation কতটুকু ভালো হয়েছে সেটা বের করি bias + variance দিয়ে। true value থেকে আমার estimated value কতটুকু দূরে সেটাই bias। bias high মানে distance বেশি। আমার estimated value এর মধ্যে data গুলো বেশি একসাথে থাকে তাহলে low variance, দূরে ছড়িয়ে থাকলে high variance। Trade off মানে balance। যে point এ bias+ variance দুই টাই low পাব এটাই trade off. Trade off এর right side এ গেলে overfitting হয়, left side এ গেলে underfitting হয়।



loss curve = training error curve
loss function = error function

low bias = true value এর কাছাকাছি
high variance = value গুলো scattered



high bias = true value থেকে দূরে
low variance = value গুলো কাছাকাছি