Machine learning - E1 SLOT

1) Given,

confusion matrix

	Predicted +ve	Predicted -ve	
Actual +ve	TP	FN	→ sensitivity
Actual -ve	FP	TN	→ specificity

people who really have diabetes = actual positive = 6%

$$TP + FN = 6\% = 0.06 \quad \text{--- (1)}$$

$$\text{sensitivity} = \frac{TP}{TP+FN} = 90\% = 0.9 \quad \text{--- (2)}$$

$$\text{specificity} = \frac{TN}{TN+FP} = 85\% = 0.85 \quad \text{--- (3)}$$

probability that a person who test positive = ? = TP
activity has the diabetes

From (2), $TP = 0.9TP + 0.9FN$

$$0.1TP = 0.9FN$$

$$TP = 9FN \quad \text{--- (4)}$$

From (3), $TN = 0.85TN + 0.85FP$

$$0.15TN = 0.85FP$$

$$15TN = 85FP \Rightarrow 3TN = 18FP$$

$$TN = 6FP \quad \text{--- (5)}$$

we know that $TP + FN + FP + TN = 100\% = 1$ and

$$TP + FN = 0.06$$

From (4), $TP + \frac{TP}{9} = 0.06$

$$\frac{10TP}{9} = 0.06 \Rightarrow \boxed{TP = 0.054}$$

∴ probability that a person who Tests +ve have diabetes is 0.054

4) Given, data

x_1	x_2	Y
200	30	50
150	25	40
250	35	60
300	50	70
100	20	30

a)

We take Regression line as $Y = a_1 x_1 + a_2 x_2 + c$

substitute data in points, we get

$200a_1 + 30a_2 + c = 50$

$150a_1 + 25a_2 + c = 40$

$250a_1 + 35a_2 + c = 60$

$300a_1 + 50a_2 + c = 70$

$100a_1 + 20a_2 + c = 30$

Into the matrix form as

$$\begin{bmatrix} 200 & 30 & 1 \\ 150 & 25 & 1 \\ 250 & 35 & 1 \\ 300 & 50 & 1 \\ 100 & 20 & 1 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ c \end{bmatrix} = \begin{bmatrix} 50 \\ 40 \\ 60 \\ 70 \\ 30 \end{bmatrix}$$

It is in the form of $AX = B$

where $X = \begin{bmatrix} a_0 \\ a_1 \\ c \end{bmatrix}$

$$X = (A^T A)^{-1} A^T B$$

$$A^T A = \begin{bmatrix} 200 & 150 & 250 & 300 & 100 \\ 30 & 25 & 35 & 50 & 20 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 200 & 30 & 1 \\ 150 & 25 & 1 \\ 250 & 35 & 1 \\ 300 & 50 & 1 \\ 100 & 20 & 1 \end{bmatrix}$$

$$A^TA = \begin{bmatrix} 225000 & 35500 & 1000 \\ 35500 & 5650 & 160 \\ 1000 & 160 & 1 \end{bmatrix}$$

$$(A^TA)^{-1} = \begin{bmatrix} 5.1\times10^{-4} & -3\times10^{-3} & -7\times10^{-4} \\ -3\times10^{-3} & 0.0198 & 0.0128 \\ -7\times10^{-4} & 0.0128 & -0.282 \end{bmatrix}$$

$$A^TB = \begin{bmatrix} 200 & 150 & 250 & 300 & 100 \\ 30 & 25 & 35 & 50 & 20 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 50 \\ 40 \\ 60 \\ 70 \\ 30 \end{bmatrix}$$

$$= \begin{bmatrix} 55000 \\ 8700 \\ 250 \end{bmatrix}$$

$$x = (A^TA)^{-1} \cdot A^TB = \begin{bmatrix} 0.1692 \\ 0.5128 \\ -1.282 \end{bmatrix}$$

Regression line, $y = 0.1692\, x_1 + 0.5128\, x_2 - 1.282$

b)

$y = 0.1692 \times 255 + 0.5128 \times 40 - 1.282$

$y = 62.376$

c)

x_1	x_2	y_{pred}	y_{true}	$(y_{true}-y_{pred})$	$(y_{true}-y_{pred})^2$
200	30	47.942	50	2.058	4.235
150	25	36.918	40	3.082	9.498
250	35	58.966	60	1.034	1.069
300	50	75.118	70	-5.118	26.193
100	20	25.894	30	4.106	16.859

$$MSE = \frac{1}{n}\sum_{i=1}^{5}(y_{true} - y_{pred})^2 = \frac{58.304}{5}$$

$$= 11.6608$$

2) Given

x_1	x_2	x_1^2	x_2^2	x_1x_2	$x_1-\bar{x}_1$	$x_2-\bar{x}_1$
5	7	25	49	35	-3.5	-2.75
4	6	16	36	24	-4.5	-3.75
11	17	121	289	187	2.5	7.25
14	9	196	81	126	5.5	-0.75

a) $\bar{x}_1 = 8.5$

$\bar{x}_2 = 9.75$

$\overline{x_1x_2} = 93$

b) covariance matrix $= \begin{bmatrix} cov(x_1,x_1) & cov(x_1,x_2) \\ cov(x_2,x_1) & cov(x_2,x_2) \end{bmatrix}$

$cov(x_1,x_1) = \sigma^2(x_1) = 17.25$

$cov(x_2,x_2) = \sigma^2(x_2) = 18.6875$

$cov(x_1,x_2) = \overline{x_1 \cdot x_2} - \bar{x}_1 \cdot \bar{x}_2$

$= 93 - 9.75 \times 8.5 = 10.125$

$cov(x_1,x_2) = cov(x_2,x_1) = 10.125$

$$S = \begin{bmatrix} 17.25 & 10.125 \\ 10.125 & 18.6875 \end{bmatrix}$$

c) Eigen values and eigen vectors of S are

$$\lambda_1 = -\frac{-575 + \sqrt{105505}}{32} \qquad \lambda_2 = \frac{\sqrt{105505} + 575}{32}$$

$$x_1 = \begin{bmatrix} -\left(\frac{23 + \sqrt{105505}}{324}\right) \\ 1 \end{bmatrix} \qquad x_2 = \begin{bmatrix} \frac{-23 + \sqrt{105505}}{324} \\ 1 \end{bmatrix}$$

$$\Downarrow$$

d)

$$\lambda_1 = \cancel{0.777}\ 7.818 \qquad \lambda_2 = 28.119$$

$$x_1 = \begin{bmatrix} -1.073 \\ 1 \end{bmatrix} \qquad x_2 = \begin{bmatrix} 0.931 \\ 1 \end{bmatrix}$$

Since λ_2 has the highest value

x_2 is the principle component.

$$PC1 = \begin{bmatrix} 0.931 \\ 1 \end{bmatrix}$$

e)

$$X_{scaled} = \begin{bmatrix} -3.5 & -2.75 \\ -4.5 & -3.75 \\ 2.5 & 7.25 \\ 5.5 & -0.75 \end{bmatrix}$$

dimensionally reduced data = $X_{scaled} \times PC1$

$$= \begin{bmatrix} -6.0085 \\ -7.939 \\ 9.5775 \\ 4.3705 \end{bmatrix}$$