[231058002] MSIS - II Sessional Examination - Mar 2024 / AML 5201 - Advanced Applications of Probability and Stat... Total: 50

- Sec 1
- 1
- 2
- 3
- 4
- 5

Prev | 1 | Next

0 1 2 3 4 5 6 7 8 9 10

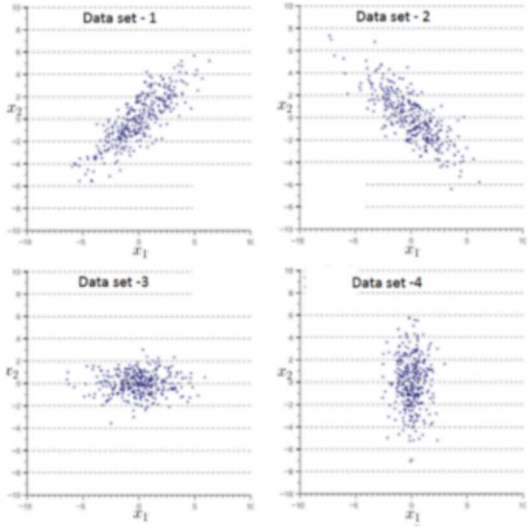
10

10.00

[10 points] [L5, CO3] Match the datasets in the scatter plots on the left with the covariance matrices on the right. Write your answers in the form

$$\text{Data set-}j = \text{Matrix-}\mathbf{M}_j,\ j = 1, 2, 3, 4$$

and justify them briefly:



$$\mathbf{M}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 5 \end{bmatrix},$$

$$\mathbf{M}_2 = \begin{bmatrix} 5 & 4 \\ 4 & 6 \end{bmatrix},$$

$$\mathbf{M}_3 = \begin{bmatrix} 5 & -4 \\ -4 & 6 \end{bmatrix},$$

$$\mathbf{M}_4 = \begin{bmatrix} 5 & 0 \\ 0 & 1 \end{bmatrix}.$$

Q: 1)
M: 10.00 / 10.00

Data set -1 = Matrix M_2

Because, from the scatter plot, we can see the positive co-variance between x_1 and x_2, and the matrix M2 denotes the same.

Dataset - 2 = Matrix M_3

Because, from the scatter plot, we can see the negative co-variance between x_1 and x_2, and the matrix M3 denotes the same.

Data set -3 = Matrix M4

Because, from the scatter plot, we see that there is more variance along x-axis (x_1) compared to the variance along y-axis (x_2). $var(x_1) > var(x_2)$.

Dataset - 4 = Matrix M1

Because, from the scatter plot, we see that there is more variance along y-axis (x_2) compared to the variance along x-axis (x_1). $var(x_2) > var(x_1)$.

[10 points] [L3, CO1] Consider the following frequency table:

regular drinker?	male	female	Total
yes	95	139	234
no	16	44	60
Total	111	183	294

(a) What are the odds that a woman is a regular drinker?

(b) What are the odds that a man is a regular drinker?

(c) What is the odds ratio? That is, compared to a man, what is the relative odds (odds ratio) that a woman is a regular drinker?

(d) Suppose we want to predict whether a person is a drinker or not based on the gender. Fill in the missing values in the table below:

hon	Coef.	Std. Err.	z	P>|z|
gendermale	?	.3414294	1.74	0.083
intercept	?	.2689555	-5.47	0.000

Q: 2)
M: 10.00 / 10.00

(a) odds that a woman is a regular drinker is given by

$$\frac{P(\text{woman is a regular drinker})}{1 - P(\text{woman is a regular drinker})}$$

$$= \frac{139/183}{1 - 139/183}$$

$$=$$

(b) Odds that a man is a regular drinker is given by:

$$\frac{P(\text{man is a regular drinker})}{1 - P(\text{man is a regular drinker})}$$

$$= \frac{95/111}{1 - 95/111}$$

$$=$$

(c) odds ratio : $\dfrac{\frac{139}{183}}{1-139/183}$

$= \dfrac{\frac{95/111}{}}{1-95/111}$

(d) $\log\left(\dfrac{\hat{p}}{1-\hat{p}}\right) = \hat{\beta}_0 + \hat{\beta}_1 * \text{gender male}$

to find $\hat{\beta}_0$, put gender male $=0$;

$\hat{\beta}_0 = \log\left(\dfrac{\hat{p}}{1-\hat{p}} \mid \text{gender female}\right)$

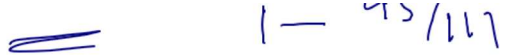
$$\text{intercept} = \hat{\beta}_0 = \log\left(\frac{139/183}{1 - 139/183}\right)$$

To find $\hat{\beta}_1$, put gendermale = 1

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{\beta}_0 + \hat{\beta}_1$$

$$\hat{\beta}_1 = \log\left(\frac{\frac{\hat{p}}{1-\hat{p}} \mid \text{genderfemale}}{\frac{\hat{p}}{1-\hat{p}} \mid \text{gendermale}}\right)$$

$$= \log\left(\frac{\frac{139}{183}}{\frac{1 - 139/183}{95/111}}\right)$$



[10 points] [L5, CO 1] Consider the performance shown below of two algorithms, A and B, for a binary classification task:

A		predicted Pos	predicted Neg
true	Pos	30	10
	Neg	10	30

B		predicted Pos	predicted Neg
true	Pos	38	2
	Neg	20	20

(a) For both algorithms, fill the entries of the table below:

	Accuracy	Recall	Precision	TNR	FPR
A	?	?	?	?	?
B	?	?	?	?	?

(b) In each one of the following scenarios, justify which algorithm you would use:

- fraud detection system for online transactions;
- airport security screening for prohibited items.

Q: 3)
M: 10.00 / 10.00

For table A;

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$= \frac{30 + 30}{30 + 30 + 10 + 10}$$

$$= 60/80 = 6/8 = 0.75$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$= \frac{30}{30 + 10} = \frac{3}{4} = 0.75$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$= \frac{30}{} = \frac{3}{} = 0.75$$

$30+10$ 4

$$TNR = \text{Specificity} = \frac{TN}{TN + FP}$$

$$= \frac{30}{30+10} = \frac{3}{4} = 0.75$$

$$FPR = 1 - TNR$$

$$= 1 - 0.75$$

$$= 0.25$$

For table B;

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$= \frac{38 + 20}{38 + 20 + 20 + 2}$$

$$= \frac{58}{80} = 0.725$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$= \frac{38}{38 + 2} = \frac{38}{40} = 0.95$$

$$\text{Precision} = \frac{TP}{}$$

$$TP + FP$$

$$= \frac{38}{38 + 20} = \frac{38}{58} = 0.655$$

$$TNR = \frac{TN}{TN + FP}$$

$$= \frac{20}{20 + 20} = \frac{20}{40} = 0.5$$

$$FPR = 1 - TNR$$

$$= 1 - 0.5$$

$$= 0.5$$

(a)

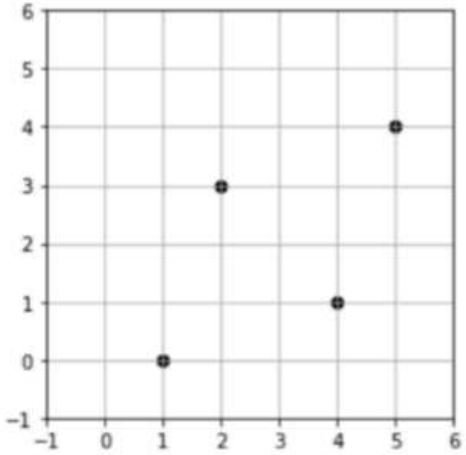
	Accuracy	Recall	Precision	TNR	FPR
A	0.75	0.75	0.75	0.75	0.25
B	0.725	0.95	0.655	0.5	0.5

(b) I use algorithm 'A' for fraud detection system because, it has good precision compared to algorithm 'B'. Given that the model predicts a transaction as fraud one, 75% of the times, it is true. i.e, it has low false positive rate. This reduces the chance of flagging a legit transaction as a fraud one and don't cause much inconvinience for the customers.

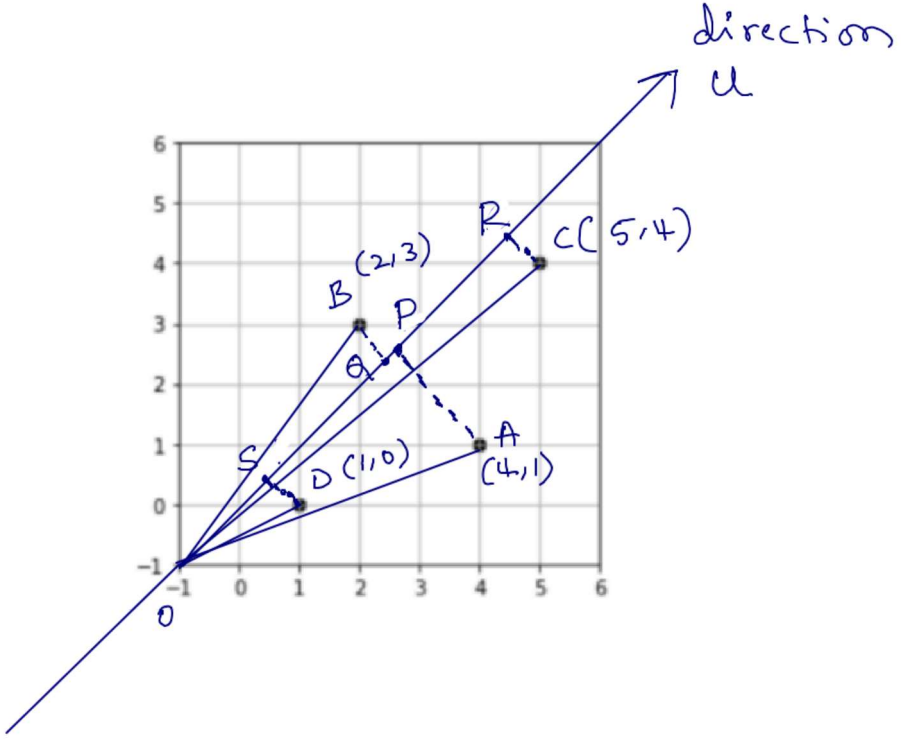
* for airport security screening, I use algorithm ~ . because it has highest Recall (in turn lowest false negative rate). I am ok with classifying someone who is not carrying any prohibited items as he is carrying some items, but not ok with the opposite.

[10 points] [L2, CO2] Consider the direction $u = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$. Using the image template below where the samples in $X = \begin{bmatrix} 4 & 1 \\ 2 & 3 \\ 5 & 4 \\ 1 & 0 \end{bmatrix}$ are shown,

- clearly draw the direction u;
- clearly show the projections of all samples in the data matrix X onto u;
- identify which two samples are nearest and farthest from each other after projection.



Q: 4)
M: 10.00 / 10.00



$$X = \begin{bmatrix} 4 & 1 \\ 2 & 3 \\ 5 & 4 \\ 1 & 0 \end{bmatrix} \begin{matrix} \rightarrow A \\ \rightarrow B \\ \rightarrow C \\ \rightarrow D \end{matrix}$$

$\overrightarrow{OP}$ is the projection of $\overrightarrow{OA}$ on u

$\overrightarrow{OQ}$ is the projection of $\overrightarrow{OB}$ on u

$\overrightarrow{OR}$ is the projection of $\overrightarrow{OC}$ on u

$\overrightarrow{OS}$ is the projection of $\overrightarrow{OD}$ on u

* After projection on 'u', points A (4,1) and B (2,3) are nearest points. [P and R are closer]

AND,

points C (5,4) and D (1,0) are the farthest points.

[R and S are farthest]

[10 points] [L6, CO2] At the beginning of the 20th century, one researcher obtained measurements on seven physical characteristics for each of 3000 convicted male criminals. The characteristics he measured are:

X_1: length of head from front to back (in cm.)

X_2: head breadth (in cm.)

X_3: face breadth (in cm.)

X_4: length of left forefinger (in cm.)

X_5: length of left forearm (in cm.)

X_6: length of left foot (in cm.)

X_7: height (in inches)

The sample correlation matrix, eigenvalues, and eigenvectors of the sample correlation matrix are shown below:

	X_1	X_2	X_3	X_4	X_5	X_6	X_7
X_1	1	0.402	0.395	0.301	0.305	0.399	0.340
X_2	0.402	1	0.618	0.150	0.135	0.206	0.183
X_3	0.395	0.618	1	0.321	0.289	0.363	0.345
X_4	0.301	0.150	0.321	1	0.846	0.759	0.661
X_5	0.305	0.135	0.289	0.846	1	0.797	0.800
X_6	0.399	0.206	0.363	0.759	0.797	1	0.736
X_7	0.340	0.183	0.345	0.661	0.800	0.736	1

	1	2	3	4	5	6	7
Eigenvectors	.285	-.351	877	-.088	-.076	.112	-.023
	.211	-.643	-.246	.686	-.098	-.010	.020
	.294	-.515	-.387	-.693	-.112	.029	-.074
	.435	.240	-.113	.126	-.604	.330	.500
	.453	.282	-.079	.127	-.024	.270	-.787
	.453	.167	.028	.023	-.065	-.873	.024
	.434	.182	-.027	-.090	.776	.208	.352
Eigenvalues	3.82	1.49	0.65	0.36	0.34	0.23	0.11

(a) Length of the left forearm has the highest correlation with which feature?

(b) What proportion of variance is explained by the first principal component?

(c) How many minimum principal components are needed to explain more than 90% of the variance in the data?

(d) Which two features are identically loaded for calculating the 1st principal component score?

(e) Which principal component assigns the greatest weight (in magnitude) to head breadth?

(f) The 2nd principal component assigns a maximum weight (in magnitude) to _____.

(g) Formulate a brief English interpretation of the second principal component.

Q: 5)
M: 10.00 / 10.00

(a) Length of the left forearm has the highest correlation with Length of left forefinger.

(b) $$\frac{3.82}{3.82 + 1.49 + 0.65 + 0.36 + 0.34 + 0.23 + 0.11}$$

$$= \frac{3.82}{7} = 0.5457 \text{ or } \boxed{54.57\%}$$

(c) Four principal components explain exactly 90% of the variance in the data.

To explain more than 90% of the variance, we need 5 PCs.

(d) Length of left forearm and length of left foot.

(e) Fourth principal component (PC-4)

(f) head breadth

(g) The second principal component gives a measure of 'Dissimilarity between the size of head and body'. i.e, the difference in size between the body (left finger, left forearm, left foot and height together) and head (head length, head breadth, face breadth together).

Save Changes Save and Exit

^ Top