Question Paper

Exam Date & Time: 27-Jun-2022 (02:00 PM - 05:00 PM)

MANIPAL ACADEMY OF HIGHER EDUCATION

Manipal School of Information Sciences (MSIS), Manipal
Second Semester Master of Engineering - ME (Big Data Analytics / Artificial Intelligence and Machine Learning / Big Data and Data Analytics) Degree Examination - June 2022

Advanced Applications of Probability and Statistics [AML 5201]

Marks: 100 **Duration: 180 mins.**

Monday, June 27, 2022

Answer all the questions.

1) [10 points] [TLO 1.1, CO 3] Consider the following data matrix X: (10)

	HR	BP	Temp
Patient-1	76	126	38.0
Patient-2	74	120	38.0
Patient-3	72	118	37.5
Patient-4	78	136	37.0

Calculate the following quantities, and explain in plain English what they signify:

(a) $X^{\mathrm{T}}e_2$

(b) Xe_3

(c) $e_3^{\mathrm{T}}(Xe_1)$

(d) $\|x^{(1)} - x^{(3)}\|$

(e) $\frac{1}{4}X^{\mathrm{T}}\mathbf{1}$

2) (10)

[10 points] [TLO 1.2, CO 2] Consider a dataset with 4 features with the following associated quantities:

- the mean sample $\mu = \begin{bmatrix} 8 \\ 6 \\ 4 \\ 12 \end{bmatrix}$;
- the sample covariance matrix $S = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1/4 & 0 & 0 \\ 0 & 0 & 1/8 & 0 \\ 0 & 0 & 0 & 1/16 \end{bmatrix}$.

Answer the following questions:

(a) *True/false*: feature-1 has the largest mean.

(b) *True/false*: the features are correlated.

(c) Express the Mahalanobis distance between sample $x^{(i)}$ and the mean sample in terms of the features of that sample.

(d) In two words, state an application of Mahalanobis distance.

3) [10 points] [TLO 2.2, CO 1] Consider the performance shown below of two algorithms, A and B, for a binary classification task: (10)

A		predicted Pos	predicted Neg
true	Pos	3	1
	Neg	1	3

B		predicted Pos	predicted Neg
true	Pos	4	0
	Neg	2	2

For both algorithms, calculate (a) accuracy (b) precision (c) recall (d) true positive rate (e) false positive rate.

4) [10 points] [TLO 3.1, CO 2] Consider the data matrix (10)

$$X = \begin{bmatrix} 5 & 4 \\ 2 & 3 \\ 1 & 0 \\ 4 & 1 \end{bmatrix}.$$

(a) Calculate X_m, the mean-centered version of X.

(b) Calculate $\frac{1}{4}X_m^{\mathrm{T}}X_m$. What does this matrix represent?

(c) Project the samples onto the direction $u = [-1, 1]^{\mathrm{T}}$. Show the projections graphically.

5) (10)

[10 points] [TLO 3.1, CO 2] 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) Head breadth has the highest correlation with which feature?

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

(c) How many minimum principal components are needed to explain more than 95% of the variance in the data?

(d) Which features are negatively loaded for calculating the 2nd principal component score?

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

(f) The 5th principal component assigns a maximum weight (in magnitude) to _____.

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

6) [10 points] [TLO 2.1, CO 1] Consider a simple linear regression model between sales (in 1000s of units) and radio advertisement budget (in 1000s of Rupees). We have the following output: (10)

	Coefficient	Std. error	t-statistic	p-value
Intercept	9.312	0.563	16.54	< 0.0001
radio	0.203	0.020	9.92	< 0.0001

(a) Fill in the question marks: sales $\approx$? $\times$ radio + ?

(b) According to this model, how many units are expected to be sold even when no money is invested in radio advertisement?

(c) According to this model, how many units will be sold when the radio advertisement budget is 5000 Rupees?

(d) In the absence of any radio advertising, sales is expected to be between and units.

(e) What is the increase in sales associated with a 1000 Rupees increase in spending on radio advertising?

7) (10)

[10 points] [TLO 2.1, CO 1] Suppose we are interested in a linear model of *instructor evaluation score* as a function of *age* and *gender*. Assume there are two genders: female and male. The output of fitting such a model is shown below:

term	estimate	std_error	statistic	p_value
intercept	4.484	0.125	35.79	0.000
age	-0.009	0.003	-3.28	0.001
gendermale	0.191	0.052	3.63	0.000

Write down the predicted instructor evaluation scores for a male and female instructor; simplify as much as possible. Quantify the effect of age on instructor evaluation score for both genders.

8) [10 points] [TLO 2.1, CO 1] Continuing from the previous question, now consider an interaction model whose output is shown below: (10)

term	estimate	std_error	statistic	p_value
intercept	4.883	0.205	23.80	0.000
age	-0.018	0.004	-3.92	0.000
gendermale	-0.446	0.265	-1.68	0.094
age:gendermale	0.014	0.006	2.45	0.015

Write down the predicted instructor evaluation scores for a male and female instructor; simplify as much as possible. Quantify the effect of age on instructor evaluation score for both genders.

9) [10 points] [TLO 2.2, CO 1] Consider the following frequency table: (10)

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

10) (10)

[10 points] [TLO 4.1, CO 4] Using a practical example, briefly explain what *autocorrelation* is.

-----End-----