Question Paper

Exam Date & Time: 29-Apr-2024 (02:00 PM - 05:00 PM)

MANIPAL ACADEMY OF HIGHER EDUCATION

Manipal School of Information Sciences (MSIS), Manipal
Second Semester Master of Engineering - ME (Artificial Intelligence and Machine Learning / Big Data Analytics) Degree Examination - April / May 2024

Advanced Applications of Probability and Statistics [AML 5201]

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

Monday, April 29, 2024

Answer all the questions.

1) [10 points] [L3, CO3] 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:

(a) mean-centered heart rates;

(b) standardized heart rates;

(c) mean-centered blood pressures;

(d) standardized blood pressures;

(e) covariance between heart rate and blood pressure;

(f) correlation between heart rate and pressure and interpret the result.

2) [10 points] [L5, CO2] Consider a dataset with 4 features with the following associated quantities: (10)

- 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) Which feature has the smallest mean?

(b) Justify whether the features are correlated or not.

(c) How would a scatter plot between the 1st and the 4th features of the data look like? Justify your plot briefly.

3) (10)

[10 points] [L5, CO2] Consider a classification model that separates passengers at an international airport checkpoint into two categories: "carrying dangerous items" or "not carrying dangerous items." Answer the following questions regarding precision and recall (a.k.a. sensitivity or true positive rate):

(a) Which is a more relevant performance metric in this case: recall or precision? Justify briefly why.

(b) Increasing the classification threshold generally $\underbrace{\text{increases/decreases}}_{\text{choose one}}$ FP.

(c) When the classification threshold increases, precision

$$\underbrace{\text{probably increases/probably decreases/definitely increases/definitely decreases}}_{\text{choose one}}.$$

(d) Keeping in mind that $TP + FP + TN + FN = n$, which is the number of samples, when the classification threshold is increased, what happens to the quantity TP?

(e) When the classification threshold is increased, the quantities TN and FN both

$$\underbrace{\text{uniformly/non-uniformly}}_{\text{choose one}}\ \underbrace{\text{increase/decrease}}_{\text{choose one}}.$$

(f) Decreasing the classification threshold generally $\underbrace{\text{increases/decreases}}_{\text{choose one}}$ FN.

(g) When the classification threshold is decreased, recall

$$\underbrace{\text{probably increases/probably decreases/definitely increases/definitely decreases}}_{\text{choose one}}.$$

(h) When the classification threshold is decreased, the quantities TP and FP both

$$\underbrace{\text{uniformly/non-uniformly}}_{\text{choose one}}\ \underbrace{\text{increase/decrease}}_{\text{choose one}}.$$

4) [10 points] [L3, CO2] 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] [L3, 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) 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] [L2, CO1] A multiple linear regression model for predicting house price (in dollars) as a function of living area (square feet) and type of fuel used for heating (a categorical variable) is built as follows: (10)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8411.608   5538.298   1.519  0.12899
livingArea    110.231      2.784  39.590  < 2e-16 ***
fuelgas     14630.007   4530.883   3.229  0.00127 **
fueloil      -252.581   6111.020  -0.041  0.96704
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 68830 on 1724 degrees of freedom
Multiple R-squared:  0.5119,    Adjusted R-squared:  0.5111
F-statistic: 602.8 on 3 and 1724 DF,  p-value: < 2.2e-16
```

(a) What is the name of the categorical variable before dummy encoding?

(b) How many levels does the categorical variable have?

(c) Identify the reference level for the categorical variable (pick one): solar, thermal, motor, electric, generator, wind, tidal.

(d) What are the non-reference levels of the categorical variable?

(e) What is the predicted house price of a gas-heated house?

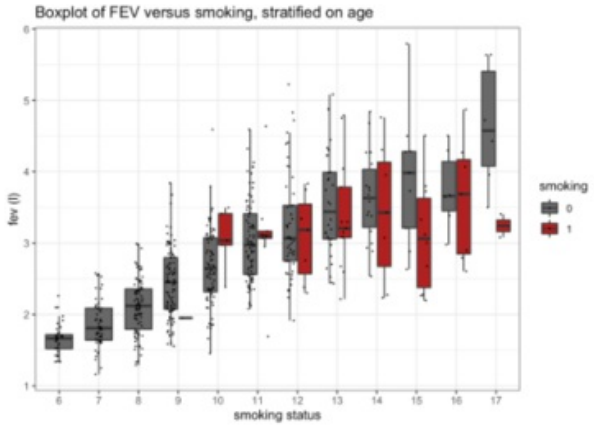
7) (10)

[10 points] [L5, C01] A simple linear regression model for how much air (in liters) a child can forcefully exhale from the lungs, referred to as the forced exhalation volume (FEV), as a function of smoking habit (no/yes) is built as follows:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.56614    0.03466  74.037  < 2e-16 ***
smokeyes     0.71072    0.10994   6.464 1.99e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8412 on 652 degrees of freedom
Multiple R-squared:  0.06023,   Adjusted R-squared:  0.05879
F-statistic: 41.79 on 1 and 652 DF,  p-value: 1.993e-10
```

According to this model, who has a greater FEV generally – smokers or non-smokers? Is the conclusion intuitively meaningful? Justify your answer briefly. You may use the following plot where the data is stratified based on age to answer this question:



8) [10 points] [L5, CO3] Consider the following data matrix where the feature **Gender** has 2 levels (female / male) and the feature **Education** has 4 levels (high school/ college/ post-graduate/ doctorate): (10)

Age	Gender	Education
26	male	college
32	female	college
28	female	post-graduate
27	male	doctorate
25	male	high school

26	female	high school
27	female	post-graduate

Numerically justify who the 1st sample (the 26 year old male) is most similar and most dissimilar to.

9) (10)

[10 points] [L5, CO1] Suppose we want to study the effect of **Smoking** on the 10-year risk of heart disease. The table below shows the summary of a logistic regression model for predicting the risk of contracting heart disease using **Smoking** as a predictor:

	Coefficient	Standard Error	p-value
Intercept	-1.93	0.13	<0.001
Smoking	0.38	0.17	0.03

Interpret the **Intercept** and the coefficient for **Smoking** in terms of odds and probabilities of contracting heart disease if:

(a) **Smoking** is a binary variable (no/yes);

(b) **Smoking** is a numerical variable (lifetime usage of tobacco in Kilograms);

(c) **Smoking** is an ordinal variable (0: non-smoker, 1: light smoker, 2: moderate smoker, 3: heavy smoker).

10) [10 points] [L3, CO4] Using a practical example, briefly explain what *autocorrelation* is and how it can be used to analyze time series data. (10)

-----End-----