BT307 Biological Data Analysis

Total Marks: 30 | **End-semester Examination** Jan-May 2024 | **Duration: 3 hr**

Instructions: In all questions, you **MUST** show all the calculation steps. **Mark** the final answer clearly.

Section A: Ten questions with one mark each

Q1. **X** and **Y** are two data vectors. Prove that when the $\text{var}(\mathbf{X}) = \text{var}(\mathbf{Y}) = 1$, and **X** is independent of **Y**, the Euclidean distance between **X** and **Y** is equal to the Mahalanobis distance between **X** and **Y**.

Q2. For the given data, calculate the Spearman correlation between Height and Weight.

Individual	Height	Weight
A	68	150
B	72	175
C	65	160
D	71	170
E	67	165

Q3. For the same data `D`, we created two models using R:

```
reg1 <- lm(y ~ x + 0, data = D)
reg2 <- lm(y ~ x, data = D)
```

What is the difference between these two models?

Q4. We project data on new coordinates in principal component analysis (PCA). Like standard coordinate systems, these new coordinates should be orthogonal. How does PCA ensure that the new coordinates or principal components would be orthogonal?

Q5. *Y* is a dependent variable, and *X* is an independent variable. We have (*X*, *Y*) pairwise data: (2, 8), (3, 10), (4, 12). Fit these three data points to a linear model $Y = a + bX$ by regression. What is the slope of the regression line?

Q6. For the data and linear regression in Q5, what is the degree of freedom of RSS?

Q7. We want to fit *u* vs. *v* data to the equation $v = \frac{mu}{1+u}$. Here, *u* and *v* are the predictor and dependent variable, respectively. Can we use linear regression for this? If yes, explain how you would perform the linear regression.

Q8. We use the following code to perform k-means clustering in R. Explain the utility or meaning of each argument used for the `kmeans()` function in this code.

```
c <- kmeans(data, 4, nstart = 10, iter.max = 10000)
```

Q9. We are using logistic regression for a classifier. The outcome or dependent variable depends upon three predictors – X, Y, and Z. Write the sigmoid function that will be used for logistic regression in this classifier.

Q10. The Pearson correlation between two variables, X and Y, is 0.8. The p-value of the t-test for this correlation is 0.1. Based on this information, what is your conclusion on the association between X and Y? Justify your answer.

Section B: Five questions with four marks each

Q11. Fit the equation $y = bx^2$ to the given data and report the value of *b*. You should show details of your calculation.

x	y
1	4
2	10
3	28
4	40
5	75

Q12. We performed PCA for the data shown here. The loading matrix is given. Project the data on Principal Components 1 and 2 (PC1 and PC2). Calculate the Manhattan distance between samples 1 and 3 in the PC1-PC2 space.

Data:

	Drug 1	Drug 2	Drug 3
Sample 1	2	1	1
Sample 2	1	3	2
Sample 3	1	2	4
Sample 4	6	2	1

Loading matrix:

$$\begin{bmatrix} -4 & -1 & 2 \\ 1 & 1 & 7 \\ 1 & -4 & 1 \end{bmatrix}$$

Q13. **X** is a *n*-by-*v* data matrix with *n* number of samples and *v* number of variables. Here $n > v$. We have performed PCA for this data. $\mathbf{p}_i$ is the i-th principal component, with eigenvalue λ_i. Prove that the variance of the data projected on $\mathbf{p}_i$ is equal to λ_i.

Q14. We created a binary classifier by Logistic regression. The classifier is tested on test data set with different probability cut-offs. The confusion matrices for these cut-offs are shown.

	p cut-off = 0.1	
	Predicted	
Actual	Positive	Negative
Positive	40	10
Negative	50	100

	p cut-off = 0.3	
	Predicted	
Actual	Positive	Negative
Positive	30	20
Negative	30	120

	p cut-off = 0.5	
	Predicted	
Actual	Positive	Negative
Positive	20	30
Negative	10	140

	p cut-off = 0.7	
	Predicted	
Actual	Positive	Negative
Positive	10	40
Negative	0	150

Draw the ROC curve for the classifier using this data. You must show all steps in your calculations. You don't need a graph paper. Roughly position the data point and mark their coordinates. Try to maintain the scale in the plot for the ROC curve.

Q15. We performed linear regression to fit data to an equation $Y = aX + b$. R^2 is the coefficient of determination for this regression. ρ is the Pearson correlation coefficient between *X* and *Y*. Prove that $R^2 = \rho^2$