BT307 Biological Data Analysis

Total Marks: 35

End-semester Examination
Jan-May 2023

Duration: 3 hr

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

Section A: Five questions with three marks each

Q1. A gene expression study has been performed in three experimental conditions (E1, E2, and E3). The normalized fold changes in the expression of two genes (X and Y) in these conditions are given. Calculate Spearman's Correlation Coefficient between these two genes.

	E1	E2	E3
X	0	1	2
Y	5	3	2

Q2. 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 Euclidian distance between samples 3 and 4 in the PC1-PC2 space.

Data:

	Gene 1	Gene 2	Gene 3
Sample 1	1	2	3
Sample 2	4	5	6
Sample 3	1	2	3
Sample 4	2	4	6

Loading matrix:

$$\begin{bmatrix} 1 & 1 & 3 \\ 2 & 1 & 5 \\ 3 & 1 & 8 \end{bmatrix}$$

Q3. Find the constraints under which Mahalanobis distance is the same as Euclidian distance. You need to show all the steps for the derivation of these constraints.

Q4. We measured the expression of several genes in three experimental conditions. Data for two genes are shown here. Calculate the chord distance between these two genes.

Gene 1	Gene 2
1	-1
1	-1
1	1

Q5. We have four data points D1, D2, D3, and D4. The distance matrix for the data points is given. Perform hierarchical clustering of this data using single linkage. Show all steps in the clustering and draw the dendrogram showing the relations between clusters.

	D1	D2	D3	D4
D1	0	3	6	1
D2	3	0	8	8
D3	6	8	0	7
D4	1	8	7	0

Section B: Five questions with four marks each

Q6. Prove that for simple linear regression: TSS = RSS + ESS. Show all the steps of your proof.

Q7. Derive the relationship between the R-squared of linear regression and the Pearson correlation coefficient.

Q8. **X** is a data matrix with *n* number of samples and *v* number of variables. **S** is its covariance matrix. **u** is a unit vector. Prove that the variance of the data projected on **u** will be maximum if **u** is an eigenvector of **S**.

Q9. "The existence of multicollinearity in data leads to spurious results in the statistical test in multiple linear regression." — Explain this statement by showing the effect of multi-colinearity on the variance of parameters/coefficients in the regression model.

Q10. 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.

FP
TP
TN

	p cut-off = 0.2	
	Actual	
Predicted	Positive	Negative
Positive	150	25
Negative	25	0

	p cut-off = 0.4	
	Actual	
Predicted	Positive	Negative
Positive	120	8
Negative	55	17

	p cut-off = 0.6	
	Actual	
Predicted	Positive	Negative
Positive	80	2
Negative	95	23

	p cut-off = 0.8	
	Actual	
Predicted	Positive	Negative
Positive	60	1
Negative	115	24

Draw the ROC curve for the classifier using this data. You must show all steps in your calculations. Try t maintain the scale in the plot for the ROC curve.