INDIAN STATISTICAL INSTITUTE

MS QMS

TEST ON MULTIVARIATE DATA ANALYSIS

Date: 06 May 2022 Time: 3 hours Maximum Marks: 50

Answer as many questions as you can. The maximum you can score is only 50 marks

1. Explain multicollinearity? How is it measured? Why collinear explanatory variables are not desirable in regression? Describe how variance inflation factor (*VIF*) can detect multicollinearity? Why *VIF* is a better measure to detect multicollinearity than correlation matrix? Explain how dropping some of the collinear explanatory features or usage of shrinkage methods can tackle the multicollinearity problem?

[12]

2.

   a. Explain the logistic regression and discriminant analysis methodologies? Explain the similarities and differences between logistic regression and discriminant analysis? Compare linear and quadratic discriminant analysis?

   b. A mobile service provider wants to win back sleeping customers by offering various credit percentages. A logistic regression model is developed to classify whether win back will be successful or not (*y*) for various % of the credit (*x*). The model coefficients are β_0 = -2.7117 and β_1 = 0.6513, Estimate the % of the credit to be offered so that at least 75% of sleeping customers can be won back?

   c. The win back for eight credit % is given in the table below. Estimate the class probabilities and classify the cases as success or failure using the aforementioned logistic regression model? Compute the accuracy & misclassification % and provide your comment on the performance of the model? Compute specificity & sensitivity and list down the insights you can provide to the mobile company on the usage of the model?

% Credit	Win back	P(x/y = success)	P(x/y = failure)
4.0	Failure	0.0003	0.7979
5.5	Failure	0.4839	0.0089
3.0	Failure	0.0000	0.1080
4.5	Success	0.0089	0.4839
5.2	Success	0.2218	0.0448
6.0	Success	0.7979	0.0003
3.5	Failure	0.0000	0.4839
6.0	Success	0.7979	0.0003

   d. Based on P(x/ y= success) and P(x/ y = failure) given in the table above, classify the cases into success or failure using discriminant analysis. Compute the accuracy & misclassification % and provide your comment on the performance of the model? Compute specificity & sensitivity and list down the insights you can provide to the mobile company on the usage of the model? Compare the performance of the logistic regression and discriminant models?

[4+2+4+6]

3.

a. Provide two uses of ridge and lasso regressions? Explain the ridge and lasso regression methodologies and highlight their major differences? How optimum value of λ is computed in ridge and lasso regressions?

b. An application support process of an Information Technology (*IT*) company has developed ridge and lasso regression models to estimate the resolution time of tickets using skill and effort as explanatory variables. The model coefficients are given in the table below:

Model	Intercept	Skill	Effort
Ridge	-57.9611	104.459	0.1071
Lasso	- 86.5223	112.716	0.115

Apply the ridge and lasso regression models to the data given in the table below. Compute R^2, adjusted R^2, mean square error (*MSE*) and root mean square error (*RMSE*). Comment on the performance of the models.

Skill	Effort	Resolution Time
3.33	30	294
3.05	245	291
3.1	220	287
3.25	300	309
3.19	275	302
3.1	185	280
3.28	180	303

[6+8]

4. Write short notes on the following:

   a. Multidimensional scaling

   b. Correspondence analysis

[6+6]