PES University, Bengaluru

(Established under Karnataka Act No. 16 of 2013)

UE18CS312

OCTOBER 2020: IN SEMESTER ASSESSMENT B Tech FIFTH SEMESTER TEST – 1

UE18CS312 (4 credit subject) - Data Analytics
Scheme and Solutions

Time: 2 Hrs	Answer All Questions	Max Marks: 60

1. a) An online certification course has been offered to students in the fifth and seventh semesters of Computer Science and Engineering. The number of registrations and number of successful certifications across the country at the end of each month as recorded by the course is provided below: **4 (2+2)**

Month	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct
Number of registrations	44	101	386	4,904	12,106	74,696	1,02,458	12,524
Successful certifications	6	59	174	359	18,036	72,599	96,239	6,980

(i) If we use a Cox-comb plot to visualize this data, how many sectors would this plot have and how would we represent the data provided in the table?

(ii) A potential registrant wants to answer the question: "Is the increase in the number of certifications between the fifth and seventh semester students statistically significant?" Assuming the detailed data for fifth and seventh semesters is available, suggest an approach one might take to answer this question.

2 marks each

(i) (Either a schematic diagram or an explanation in words) Eight sectors for the eight months for which data is available. The size of the sector would be proportional to number of registrations. We would use different shades for registrations versus certifications (or a band in one color (with the area proportionate to the number of certifications) around each sector representing the registrations for that month)

(ii) Design a t-test or a z-test (since the number is rather large)

For the t test:

- Calculate t
  - by finding the difference between mean certifications of one semester and the other (let us call this num)
  - for each group calculate the variance divided by the number of observations-1 (let us call the variance divided by degrees of freedom, σ^2_1 and σ^2_2)
  - compute sqrt($\sigma^2_1 + \sigma^2_2$) (let us call this denom)
  - compute num/denom
  - calculate degrees of freedom (add the number of observations from each group -2)
  - look up the value in the table and interpret the value of t

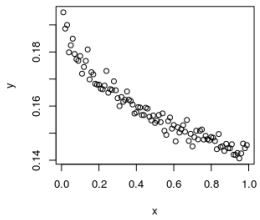
Similarly for the z-score (the p-value compared to the value in the table for a given significance will determine whether the difference is statistically significant or not)

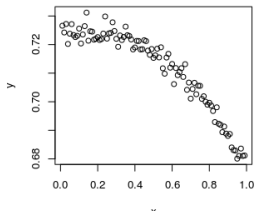
b) In the data shown in Question 1(a) above, **3**

(i) Is there any anomaly? Briefly explain your answer.

		(ii) The organizers of the course have realized that data has not been recorded during weekends in April. Suggest a method to fill in the missing values. 1.5 marks each (i) In the month of July, the number of certifications seems to exceed the number of registrations. This appears anomalous. (It cannot be argued as not anomalous because it includes registrants from previous months, given the numbers do not tally.) (ii) [open ended] Data for weekends can be modeled based on weekend data in March and May as a function of the data during the week and that can be used to predict values for weekends in April based on the cases during the week in April (any other reasonable approach that is suitably justified can be considered)	
	c)	The range of scores of students for various components of their project submission is (3,8). How much will a student who has scored 7 on this scale get, if the marks are rescaled to a new range, (16,25)? (((7-3)/(8-3))*(25-16)) + 16 = **23.2** (rounded to 23)	3
2.	a)	Briefly explain the sampling technique(s) used in each of the following cases: (i) A Kitkat factory produces ten different flavours of the chocolates and has twenty assembly lines (two for each flavour). A taste tester selects a random chocolate bar from every other line. (ii) A restaurant has placed a feedback card on every table and allows diners to choose whether they would like to provide a feedback on behalf of their group or not. Solution (2+2): (i) Systematic sampling (every other line) followed by simple random sampling (ii) Convenience (cluster) sampling, since the customers can choose to give feedback or not and the feedback, if given, would be of the entire table (including elders, adults, youth and children) (Convenience sampling – can be given complete credit; cluster based sampling can be given 1 mark)	4
	b)	For the following examples, identify the datatypes as numeric/ categorical, ordinal/ interval/ ratio and discrete/ continuous (as applicable) (i) Movie rating on a scale of 1 to 5 (ii) When booking a flight ticket, response to whether the wheelchair service would be required for the passenger (yes/ no) (iii) Temperature (recorded in Centigrade from various regions around the world) Solution (1 mark each): (i) Numeric, Ordinal, Discrete (ii) Numeric, Discrete, Binary (or nominal, discrete) (iii) Numeric, interval, continuous (considered interval as temperature can be negative and holds no true zero)	3
	c)	Twenty engineers and twenty pilots were subject to tests and scores were measured for the following six features: (i) Intelligence (ii) Conformance to procedure (iii) Eyesight (iv) Hearing (v) Sensory motor coordination and (vi) Perseverance. Briefly outline the steps to extract two principal components from this data to visualize the two groups of twenty points in the 2-dimensional rectangular plane. Solution (1 mark for each step): (i) Subtract the mean of each feature (or subtract the mean and divide by the standard deviation for each feature). Compute the outer product for each mean adjusted feature vector and add this to obtain the the covariance matrix for the 40 points (or two separate covariance matrices for each of the 20 points) (ii) Perform eigen analysis to obtain the Eigen values; select the two Eigen vectors corresponding to the largest two Eigen values	3 (2+1)

		(iii) Project the data (6x1) on to the selected Eigen vectors (6x2) to obtain (2x1) vectors that are approximations for each point. Plot this on a 2D graph using two different colors (one color for the twenty points representing engineers and one color representing twenty points for pilots)	
3.	a)	Taste testers Aman and Mani have rated the quality of food at a restaurant on six days in the week as follows: **Day** | **M** | **Tu** | **W** | **Th** | **F** | **Sa** **Aman's rating** | 4 | 2 | 3 | 5 | 1 | 3 **Mani's rating** | 3 | 3 | 2 | 5 | 2 | 2 Given: mean and standard deviation of ratings: μ_{Aman}= 3, σ_{Aman} = 1.291, μ_{Mani} = 2.833, σ_{Mani} = 1.067, correlation coefficient(Aman, Mani) = 0.7258 (i) What are β_0 and β_1 if we must predict Aman's rating in terms of Mani's rating using simple linear regression with the following model? Rating(Aman) = β_0 + β_1Rating(Mani) (ii) What is the coefficient of determination for this model? (iii) How can we measure the influence Mani's rating of the food on Thursday has on the model? (Suggest the test or statistic that can be used for this.) Solution: (i) β_1 = rRating(Aman)/ Rating(Mani) = 1.927 $\beta_0 = \mu_{Aman} - \beta_1\mu_{Mani}$ = 3-1.927*2.833 = -2.459 (ii) Coefficient of determination = r*r = 0.5267 (iii) Cook's distance or DFBeta can be used to measure the influence that Mani's rating of Thursday's food has on the regression model	4
	b)	The correlation between two variables (#views for a video and average #videos posted per month) is found to be positively correlated. Answer the following questions with a line to justify your answer: (i) Is it necessarily true that the Pearson's correlation coefficient between #postings/ month and #views on a video would to be closer to 1 than it is to 0 for this data? (ii) Can we assume there is no cause-effect relationship between #postings per month and #views on a channel because correlation does not imply causation? 1.5 marks each (i) No, the relationship can be nonlinear and still positively correlated (ii) No, just because correlation does not imply causation does not mean there cannot be positive correlation when there is causation! We must do further tests to infer causation (Scheme 1 marks each, even if the answer is Yes, if the justification is a plausible one.)	3
	c)	For each of the following scatterplots suggest whether the data is suitable for linear regression and, if not, what transformation may be applied to make it amenable for modeling with linear regression. (i) (ii)	3





(i) Not suitable – x needs to be transformed to ln(x), sqrt(x), etc., and y to ln(y), sqrt(y) etc.
(ii) Not suitable – x could be transformed to x^2, x^3, etc., and y to y^2, y^3, etc.

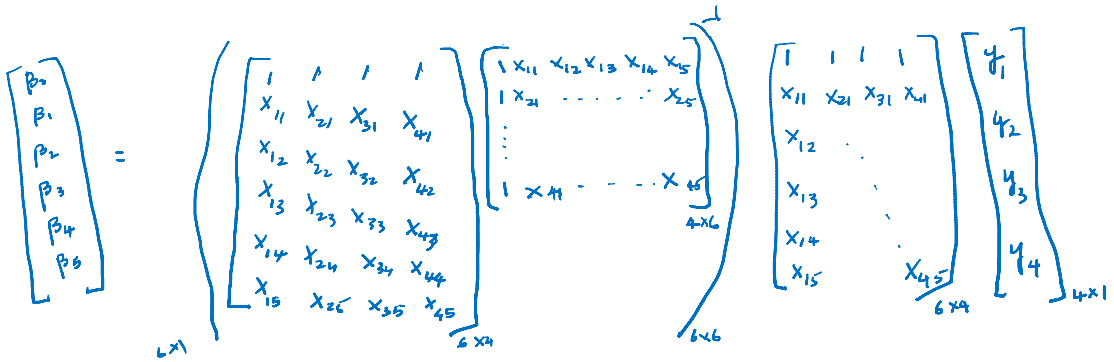
4. a) Write the linear algebraic expression for computing an estimate of the Beta vector in a multiple linear regression system to predict 4 dependent variables using 5 independent variables. — 4 (2+2)

In the table given below, identify the features that are significant (for an alpha = 0.01) and if there is insufficient data to do this, list out what other data is necessary to determine the significance of regression coefficients.

Term	Coef	SE Coef	T	P
Constant	389.166	66.0937	5.8881	0.000
X_1	2.125	1.2145	1.7495	0.092
X_2	5.318	0.9629	5.5232	0.000
X_3	4.22	0.3	14.06	0.043
X_4	-24.132	1.8685	-12.9153	0.000
X_5	-17.201	1.333	-12.9039	0.004

Solution (2 marks each):

$\beta\_hat = (X^TX)^{-1}X^TY$



(Either form is acceptable)

Variables X_1 and X_3 are not statistically significant; all others are statistically significant.

b) Rajesh has designed a logistic regression classifier to predict the likelihood of stars being visible in the night sky based on the humidity reported on any day: — 3 (1+2)

logit (p) = log(p/(1-p)) = $\beta_0 + \beta_1$*humidity, where p is the probability of a power cut.

Given that $\beta_0 = 1.8185$ and $\beta_1 = -0.0665$, answer the following questions:

(i) What does the value of β_0 mean?

(ii) If humidity on a day = 25, what is the probability with which stars are visible in the night sky according to this model?

Solution:

(i) When humidity = 0, the ln(odds) of sighting stars in the night sky = 1.8185 (OR the probability of stars being visible in the night sky = 0.86)

(ii) When humidity = 25, the probability with which stars are visible in the night sky according to this model

$\ln(p/(1-p)) = -1.8185-0.0665*25$

$(p/(1-p)) = \exp(0.156) = 1.17$

$p = 1.17/2.17 = 0.539$ (about 53% chance stars are visible)

c) In a collection of 1000 small rocks collected on a river bed, 100 happen to be precious stones. All those 100 precious stones along with 100 other rocks have been determined to be precious stones by a logistic regression classifier. What does the confusion matrix look like for this classifier? What should be done to obtain a receiver operator characteristics (RoC) plot for this logistic regression model? — 3 (2+1)

Solution:

	Precious stones	Other rocks
Precious stones	100	0
Rocks	100	800

If we vary the threshold that determines whether a rock is 'precious' or 'not precious', we can compute the TPR and FPR for each of these to plot the RoC graph.

5. a) With a schematic sketch, discuss the components of an additive time series data with level, trend and seasonality. What are cyclic components and, why are they usually not accounted for in models for time series data? — 4

(1 mark each (including the schematic diagram for what the component looks like (OE))

- Level (where the trend begins)
- Trend (upward or downward)
- Seasonality (repetitions within a calendar year)
- Cyclic (macro-economic changes or patterns that vary in a decade or several years and are not periodic and so cannot be predicted with most time series models)

b) For the data given below, use MAPE to compare forecast accuracy of single exponential smoothing (SES) with alpha = 0.7 with simple moving average (SMA) with a window size = 3 for time points t=5,6,7. [You can use the values of y available to make the forecasts for SMA and for SES assume the forecast, $F_4=y_4$] — 3

t	1	2	3	4	5	6	7
y_t	10	11	12	16	17	19	20
sMA					13	15	18.67
APE					4/17	4/19	1.34/20
SES					16	16	16
APE					1/17	3/19	4/19

(1 mark) MAPE (SMA) = 0.1709
(1 mark) MAPE (SES) = 0.1389
(1 mark) Single Exponential Smoothing is more accurate

c) Suggest an application for each of the following techniques to model time series data — 3

(i) Croston's method

(Open-ended) Forecasting the demand for any entity that has a sporadic demand (such as seasonal crops, fruits, etc., or winter (or summer) clothes, etc.)

(ii) Holt-Winter's method

(Open-ended) Forecasting a dependent variable where the underlying process has both trend and seasonality (such as the demand for flu medicine or a specific type of pesticide or school supplies or avionic parts or the price of stocks of companies that have seasonal highs or lows, with a general upward or downward trend)

(iii) ARIMA

(Open-ended) Forecasting the demand for any entity that has a sporadic demand (such as seasonal crops, fruits, etc., or winter (or summer) clothes, etc.)

6 a) Equation for ARIMA models (2+2) — 4

(i) ARIMA(0,1,0)

$F_{t+1} = e_t$

OR

$\Delta Y_{t+1} = e_t + e_{t+1}$

(ii) ARIMA(1,0,1)

$F_{t+1} = \beta y_t + \alpha e_t$

OR

$Y_{t+1} = \beta y_t + \alpha e_t + e_{t+1}$

b) Which model is better and why? (3+3) — 6

	Statistic	Model A	Model B	A or B?	Why?
1	AIC	258.24	251.42	Model B	AIC is the negative log likelihood the sample will fit/ estimate future values – lower the better
2	R^2	0.98	0.91	Model A	R^2 is expected to be higher for a better model designed for time series data
3	RMSE	0.048	0.051	Model A	RMSE is expected to be lower