PES UNIVERSITY

SRN							

UE15CS322 Instructor -Chidambaran

lyer

PES University, Bangalore (Established under Karnataka Act No. 16 of 2013)

END SEMESTER ASSESSMENT (ESA) B.TECH. V SEMESTER- Dec 2017

UE150	S322 – DATA ANAL	YTICS	
Time: 3 hrs	Answer All Questions		Max Marks: 100
General Instructions: All questions are cor	npulsory		
I A)			
(i) Explain in words Principal Component	Analysis (2 marks)		
(ii) Suppose that the first and second centr low. Now as a data analytics student, defir professional's desired classification (with r	1e a random variable to	r each of the	professional balance
a) Shop keeper (2 marks) b) Product Quality Control officer			
(iii) In a throw of two dice, compute and d defined as the sum of the numbers shown of	raw the probability den on two dice.	sity function (4)	for the random variable X. X is marks)
B) Using appropriate mathematical notation	ns, define and explain		
 (i) Discrete joint probability density functio (ii) Conditional probability density functio (iii) Normal distribution (iv) Student's t-distribution (v) F- distribution 	on (2 marks) n (2 marks) (2 marks) (2 marks) (2 marks)		
II A)			
(i) Define using appropriate mathematical r What are these properties called?	notations, an unbíased e	stimator and (3 marks)	a minimum variance estimator?
(ii) Write down any five assumptions of the	classical linear regress	ion model.	(5 marks)
(iii) Write down the formula for the coefficiall the terms used.	ent of determination (R	t²). What doe (2 marks)	es it measure? Ensure you explain

SRN		

B)

(i) From a sample of 209 firms, the following regression results with the dependent variable being log (salary) were obtained:

Independent Variable	Coefficient	Std. Error
Intercept	4.32	0.32
Log (sales)	0.28	0.035
RoE	0.0174	0.0041
RoS	0.00024	0.00054

where salary = salary of CEO

- ; sales = annual firm sales
- ; RoE = return on equity in percent
- ; RoS = return on firm's stock.
 - a) Write down the population regression function. Ensure you explain the terms used.

(2 marks)

- b) Interpret the results, are the results as per your expectations? (2 marks)
- c) How many variables in the table are significant? Show your computations (4 marks)
- d) If $R^2 = 0.283$, How do you test do you the overall significance of the regression? What would be the null and the alternative hypothesis? Compute the F statistic. (2 marks)

Ш

- (i) For a time series y_t , mathematically derive the decomposition of y_t into its trend, seasonal, and error component. Derive the decomposition when y_t is additive & when y_t is multiplicative. (5+5 marks)
- (ii) Prove that in simple exponential smoothing, when α (alpha) is low, initial forecast values play more important role in forecast than when α (alpha) is high. (5 marks)
- (iii) Prove that MA(5) is equivalent to 2X4 MA.

(3 marks)

(iv) State the idea on which the method of moving average is based. Under what conditions does the moving average work best? (1+1 marks)

IV

A)

(i) In the context of recommendation systems, describe cold start.

(3 marks)

- (ii) Why is item based collaborative filtering more popular than user-based collaborative filtering. (3 marks)
- (iii) How does the concept of support help us identify a frequent itemset? Explain in detail. (4 marks)

В) :	SRN
(i) Define Confidence, Benchmark Confidence, and Lift Ratio.	(3 marks)
(ii) Why is Lift Ratio a better measure than Confidence?	(2 marks)
(iii) Mathematically define a cluster centroid. Then proceed to o	
V A)	
(i) With respect to business, why is text important?	(2 marks)
(ii) Compared to other forms of data, why is text so different?	(2 marks)
(iii) Define a document and corpus.	(2 marks)
(iv) Explain TFIDF in detail. In doing so also define TF & IDF.	(4 marks)
B)	
(i) Define cosine similarity. When is cosine similarity used in Ba phrase and a small repository of information, say S, both on the s pefore using cosine similarity clearly.	ng of Words? Assume that you have a sample name topic. Write down all the steps involved (6 marks)
ii) What can Named Entity Extraction do, which Bag of Words o	* *
iii) What are the disadvantages of N-grams sequence?	(2 marks)