SRN | | | | | | | | | | | |

UE15CS322
Instructor – Chidambaran Iyer

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

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

UE15CS322 – DATA ANALYTICS

Time: 3 hrs | Answer All Questions | Max Marks: 100

General Instructions: All questions are compulsory

I
A)

(i) Explain in words Principal Component Analysis (2 marks)

(ii) Suppose that the first and second central moment of a random variable can be classified as either as high or low. Now as a data analytics student, define a random variable for each of the professional below and the professional's desired classification (with respect to the first and second moment) for the random variable.

a) Shop keeper (2 marks)
b) Product Quality Control officer (2 marks)

(iii) In a throw of two dice, compute and draw the probability density function for the random variable X. X is defined as the sum of the numbers shown on two dice. (4 marks)

B) Using appropriate mathematical notations, define and explain

(i) Discrete joint probability density function (2 marks)
(ii) Conditional probability density function (2 marks)
(iii) Normal distribution (2 marks)
(iv) Student's t-distribution (2 marks)
(v) F- distribution (2 marks)

II
A)

(i) Define using appropriate mathematical notations, an unbiased estimator and a minimum variance estimator? What are these properties called? (3 marks)

(ii) Write down any five assumptions of the classical linear regression model. (5 marks)

(iii) Write down the formula for the coefficient of determination (R^2). What does it measure? Ensure you explain all the terms used. (2 marks)

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)

III

(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)

SRN

B)

(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 describe the k-means clustering method. (2+3 marks)

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 Bag of Words? Assume that you have a sample phrase and a small repository of information, say S, both on the same topic. Write down all the steps involved before using cosine similarity clearly. (6 marks)

(ii) What can Named Entity Extraction do, which Bag of Words or even N-grams sequence cannot? Explain. (2 marks)

(iii) What are the disadvantages of N-grams sequence? (2 marks)