Indian Institute of Information Technology Design & Manufacturing Kancheepuram Chennai-127

COM 526T / 406A ANALYTICS & SYSTEMS OF BIG DATA - ONLINE ASSESSMENT TEST (OAT) Max: 40

Q1. [11 marks]

(A)The following database has five transactions. Use Apriori algorithm to find all the frequent item sets with minimum support 3. TID data items(bought) T1: {K, A, D, B} , T2: {D, A, C, E, B}, T3 : {C, A, D, B} , T4 : {A, B, D} and T5 :{C, D, E}.

(B) Assume you are asked to design an ARM algorithm finding all the association rules in a data to mine rules whose supports are between 30% and 50%, and the accuracy > 60% . You may refine the basic Apriori algorithm to generate rules satisfying the above criterion. You may test drive your algorithm over the same data set above.

(C)Analyse time and space complexity of Aclose, and FP growth algorithm making valid assumptions required.

(D) State and Prove the Downward Closure property of Apriori algorithm.

Q2[10 marks]

(A) In Naïve Bayesian Classification, probability of $p(H \mid X) = \frac{p(H)p(X \mid H)}{p(X)}$. Correlate each term in the equation to data classification notion of training and test data.

(B) Justify the terminology choice of Naïve for Bayes Classifier and represent the same naiveness in a graphical model assuming a class C and four task relevant attributes A1,A2,A3 and A4.

(C)For the data set given in Figure 1 apply Bayesian classifier to predict the class label for a test instance (35, medium, yes, fair)?. Apply Laplacian correction if required.

Q3[10 marks]

(A) Apply Hierarchichal clustering over the data set (2, 10), (2, 5), (8, 4), (5, 8), (7, 5), (6, 4), (1, 2), (4, 9),(8,6),(6,7) following (i) single link (ii) complete link strategies. Show the trace for each of these approaches separately.

(B) Prove or Disprove: Single link and Double link clustering yielding same results implies the same for average link clustering.

(C) Compare and contrast the Hierarchichal from K means clustering algorithm. The answer should clearly highlight the pros and cons of each algorithm. Please do not explain their working strategies for this question.

Q4[4 marks]

(A)Assume you are given the following data relating to exercise time and calories burnt:
{(0,0); (9,260),(13,320),(21,425),(30,452),(36,46), (42,550)}

(i)Draw the scatter plot for the above data. (ii) Identify the type of correlation. (iii) Draw the line of bet fit. (iv) Use this line to complete the question 400 calories = ---------------- exercise time.

(B) Fig 2 shows the hours spent on sleep by school children for a calendar year as part of their data analytics requirements. Arrive at at interpretations on sleep patterns v/s days of the week.

Q5[5 Marks]

(A) Neural Network models employ activation functions in their working. Justify the need for such functions from an ML point of view and also explore three activation functions and illustrate their working with a sample trace.

(B) Given a neural network model in Fig 3and it activation function as f(x) = 0 for x < 0 and 1 for x >=0. Compute the overall output from the neural network. Show the complete trace of how the inputs pass thru the hidden layers, activation and then the final output. Comment on the overall function that the given network simulates.

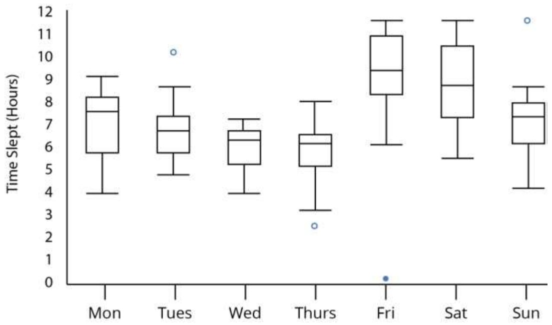

Fig.2

Age	Income	Marital Stat	Cred Rating	Sanction Loan
21	Low	No	Excellent	No
25	Low	No	Excellent	Yes
31	Medium	No	Excellent	Yes
32	High	?	Fair	No
36	High	Yes	Fair	Yes
41	Medium	Yes	Fair	Yes
45	Low	Yes	Fair	No
?	Low	Yes	Excellent	No
47	Medium	Yes	Fair	Yes

Fig.1

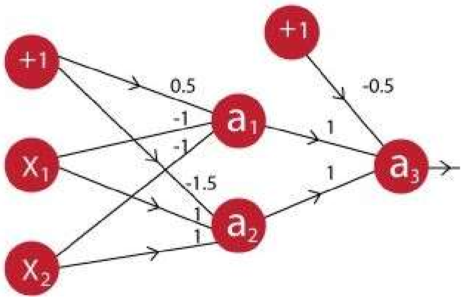

Fig.3