Sr. No..... 015605.....

May 2024

B.Tech(ENC/CE/CE(HINDI)/IT/CSE(AIML))- VI SEMESTER

Data Mining (PEC-CS-D601)

Time: 3 Hours | Max. Marks:75

Instructions:
1. *It is compulsory to answer all the questions (1.5 marks each) of Part -A in short.*
2. *Answer any four questions from Part -B in detail.*
3. *Different sub-parts of a question are to be attempted adjacent to each other.*
4. *Any other specific instructions*

PART -A

Q1 (a) How can the confidence of an association rule $X \rightarrow Y$ be calculated? (1.5)
एसोसिएशन नियम $X \rightarrow Y$ के आत्मविश्वास की गणना कैसे की जा सकती है?

(b) Define classifier accuracy. (1.5)
क्लासिफायर एक्यूरेसी को परिभाषित करें।

(c) Explain any two methods for filling up the missing values during data preprocessing. (1.5)
डेटा प्रीप्रोसेसिंग के दौरान लुप्त मानों को भरने की किन्हीं दो विधियों की व्याख्या करें।

(d) Differentiate between Classification and Clustering. (1.5)
वर्गीकरण और क्लस्टरिंग के बीच अंतर बताएं।

(e) Explain the importance of Web Mining. (1.5)
वेब माइनिंग का महत्व समझाइये।

(f) Give the limitations of Hierarchical Clustering. (1.5)
हिरार्चिकाल क्लस्टरिंग की सीमाएँ बताइए।

(g) Define the term Outlier. (1.5)
आउटलायर शब्द को परिभाषित करें।

(h) What is the basic idea behind Histogram method of sampling. (1.5)
नमूनाकरण की हिस्टोग्राम विधि के पीछे मूल विचार क्या है?

(i) Name any three properties of data streams. (1.5)
डेटा स्ट्रीम के किन्हीं तीन गुणों के नाम बताइए।

(j) Generate the Clustering Feature for point (3,5). (1.5)
बिंदु (3,5) के लिए क्लस्टरिंग सुविधा उत्पन्न करें।

PART -B

Q2 (a) For the following transaction dataset create the FP tree and also find out the conditional Pattern Base (10)

निम्नलिखित ट्रांसक्शन डेटासेट के लिए एफपी ट्री बनाएं और कंडीशनल पैटर्न बेस का भी पता लगाएं।

T_Id	Items
T1	{E,K,M,N,O,Y}
T2	{D,E,K,N,O,Y}
T3	{A,E,K,M}
T4	{C,K,M,U,Y}
T5	{C,E,I,K,O}

(b) Write down the algorithm for K-mean algorithm (5)

K-मीन एल्गोरिथम लिखिए।

Q3 (a) What is time series analysis? Explain four major component of Time Series Data. (5)

समय श्रृंखला विश्लेषण क्या है? टाइम सीरीज़ डेटा के चार प्रमुख घटकों की व्याख्या करें।

(b) Following are the points on positive and negative plane respectively: (10)

Positive labelled {(4,0) (5,1) (5,-1) (6,0)}

Negatively Labelled {(1,1) (1,-1) (2,1)(2,-1)}

Find the best fit line or hyperplane to classify the point

सकारात्मक और नकारात्मक तल पर क्रमशः निम्नलिखित बिंदु हैं:

सकारात्मक लेबल {(4,0) (5,1) (5,-1) (6,0)}

नकारात्मक लेबल {(1,1) (1,-1) (2,1)(2,-1)}

बिंदुओं को वर्गीकृत करने के लिए सबसे उपयुक्त रेखा या हाइपरप्लेन ढूंढें।

Q4 (a) What are the parameters on the basis of which Classification and Prediction methods can be evaluated? (5)

वे कौन से पैरामीटर हैं जिनके आधार पर वर्गीकरण और भविष्यवाणी विधियों का मूल्यांकन किया जा सकता है?

(b) Explain Decision tree induction algorithm for classification. Discuss the usage of information gain in this. (10)

वर्गीकरण के लिए डिसीजन ट्री इंडक्शन एल्गोरिथम की व्याख्या करें। इसमें सूचना लाभ के उपयोग पर चर्चा करें।

Q5 (a) Explain the difference between Euclidian and Manhattan Distance. For the following 1-D points generate the distance matrix using Euclidian distance method. (5)

Points : [3, 5, 1, 10, 8]

यूक्लिडियन और मैनहट्टन दूरी के बीच अंतर स्पष्ट करें। निम्नलिखित 1-डी बिंदुओं के लिए यूक्लिडियन दूरी विधि का उपयोग करके दूरी मैट्रिक्स उत्पन्न करें।

अंक : [3, 5, 1, 10, 8]

(b) Differentiate between: (10)

- OLAP vs OLTP
- Web Mining vs Data Mining

अंतर करो:

- ओएलएपी बनाम ओएलटीपी
- वेब माइनिंग बनाम डेटा माइनिंग

Q6 (a) A database has following four sequences of transactions: (15)

SNo	SID	Items_bought
01	S1	<a {a,b} {a,c} d{c,e,f}>
02	S2	<{a,d} c {b,c,d} {a,b,e}>
03	S3	<{e,f} {a,b} {d,e,f}c b>
04	S4	<e g {a,d,f} c b>

Let min sup = 2, Find all frequent sub-sequences using GSP approach.

एक डेटाबेस में लेनदेन के निम्नलिखित चार क्रम होते हैं:

SNo	SID	Items_bought
01	S1	<a {a,b} {a,c} d{c,e,f}>
02	S2	<{a,d} c {b,c,d} {a,b,e}>
03	S3	<{e,f} {a,b} {d,e,f}c b>
04	S4	<e g {a,d,f} c b>

मान लीजिए न्यूनतम समर्थन = 2, जीएसपी दृष्टिकोण का उपयोग करके सभी