A

Course File Report

On

“Data Mining”

Department of

Computer Science & Engineering

CMR ENGINEERING COLLEGE

(Affiliated to J.N.T.U, HYDERABAD)

Kandlakoya (v), Medchal -501 401

(2024-2025)

CMR ENGINEERING COLLEGE
UGC AUTONOMOUS
(Approved by AICTE - New Delhi. Affiliated to JNTUH and Accredited by NAAC & NBA)

COURSE INSTRUCTOR NAME: Mr. Siliveri Kiran Kumar ACADEMIC YEAR: 2024-25

SUBJECT NAME: DATA MINING

EMAIL-ID: siliverikirankumar1990@cmrec.ac.in CLASS ROOM NO: B-216

CONTACT NO: 8919850117

SEM START DATE AND END DATE: 29-07-24 TO 30-11-2024

CONTENTS OF COURSE FILE

1. **Department vision & mission**
2. **List of PEOs, POs, PSOs**
3. **List of Cos (Action verbs as per blooms with BTL)**
4. **Syllabus copy and suggested or reference books**
5. **Individual Time Table**
6. **Session plan/ lesson plan**
7. **Session execution log**
8. **Lecture notes (handwritten or softcopy printout-5 units)**
9. **Assignment Questions with (original or Xerox of mid 1 and mid 2 assignment samples)**
10. **Mid exam question papers with (Xerox of mid 1 and mid 2 script samples)**
11. **Scheme of evaluation**
12. **Mapping of Cos with Pos and PSOs**
13. **COs, POs, PSOs Justification**
14. **Attainment of Cos, Pos and PSOs (Excel sheet)**
15. **Previous year question papers**
16. **Power point presentations (PPTs)**
17. **Innovative Teaching method**
18. **References (Textbook/Websites/Journals)**

HOD

1. DEPARTMENT VISION & MISSION

Vision:

To produce globally competent and industry-ready graduates in Computer Science & Engineering by imparting quality education with the know-how of cutting-edge technology and holistic personality.

Mission:

1. To offer high-quality education in Computer Science & Engineering in order to build core competence for the graduates by laying a solid foundation in Applied Mathematics and program framework with a focus on concept building.
2. The department promotes excellence in teaching, research, and collaborative activities to prepare graduates for a professional career or higher studies.
3. Creating an intellectual environment for developing logical skills and problem-solving strategies, thus developing, an able and proficient computer engineer to compete in the current global scenario.

2. Program Outcomes (POs):

Engineering Graduates will be able to satisfy these NBA graduate attributes:

1. **Engineering knowledge:** Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
2. **Problem analysis:** Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.

3. **Design/development of solutions:** Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.

4. **Conduct investigations of complex problems:** Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.

5. **Modern tool usage:** Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.

6. **The engineer and society:** Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.

7. **Ethics:** Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

8. **Individual and team work:** Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.

9. **Communication:** Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.

10. **Project management and finance:** Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.

11. **Life-long learning:** Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

2.3 Program Specific Outcomes (PSOs):

PSO1: Professional Skills and Foundations of Software development: Ability to analyze, design and develop applications by adopting the dynamic nature of Software developments.
PSO2: Applications of Computing and Research Ability: Ability to use knowledge in cutting edge technologies in identifying research gaps and to render solutions with innovative ideas.

Course Outcomes:

S. No	Course Out Come
CO1	**Explain** the types of the data to be mined and present a general classification of tasks and primitives to integrate a data mining system (Understanding).
CO2	**Apply** preprocessing methods for any given raw data and extract interesting pattern from large amounts of data. (Applying)
CO3	**Discover** the role played by data mining in various fields. (Analyzing)
CO4	**Choose** and employ suitable data mining algorithms to build analytical applications.(Evaluating)
CO5	**Evaluate** the accuracy of supervised and unsupervised model and algorithms. (Evaluating)

Action Words for Bloom's Taxonomy					
Knowledge	**Understand**	**Apply**	**Analyze**	**Evaluate**	**Create**
define	explain	solve	analyze	reframe	design
identify	describe	apply	compare	criticize	compose
describe	interpret	illustrate	classify	evaluate	create
label	paraphrase	modify	contrast	order	plan
list	summarize	use	distinguish	appraise	combine
name	classify	calculate	infer	judge	formulate
state	compare	change	separate	support	invent
match	differentiate	choose	explain	compare	hypothesize
recognize	discuss	demonstrate	select	decide	substitute
select	distinguish	discover	categorize	discriminate	write
examine	extend	experiment	connect	recommend	compile
locate	predict	relate	differentiate	summarize	construct
memorize	associate	show	discriminate	assess	develop
quote	contrast	sketch	divide	choose	generalize
recall	convert	complete	order	convince	integrate
reproduce	demonstrate	construct	point out	defend	modify
tabulate	estimate	dramatize	prioritize	estimate	organize
tell	express	interpret	subdivide	find errors	prepare
copy	identify	manipulate	survey	grade	produce
discover	indicate	paint	advertise	measure	rearrange
duplicate	infer	prepare	appraise	predict	rewrite
enumerate	relate	produce	break down	rank	role-play
listen	restate	report	calculate	score	adapt
observe	select	teach	conclude	select	anticipate
omit	translate	act	correlate	test	arrange
read	ask	administer	criticize	argue	assemble
recite	cite	articulate	deduce	conclude	choose
record	discover	chart	devise	consider	collaborate
repeat	generalize	collect	diagram	critique	collect
retell	give examples	compute	dissect	debate	devise
visualize	group	determine	estimate	distinguish	express
	illustrate	develop	evaluate	editorialize	facilitate
	judge	employ	experiment	justify	imagine
	observe	establish	focus	persuade	infer
	order	examine	illustrate	rate	intervene
	report	explain	organize	weigh	justify
	represent	interview	outline		make
	research	judge	plan		manage
	review	list	question		negotiate
	rewrite	operate	test		originate
	show	practice			propose
	trace	predict			reorganize
	transform	record			report
		schedule			revise
		simulate			schematize
		transfer			simulate
		write			solve
					speculate
					structure
					support
					test
					validate

4. Syllabus Copy

UNIT – I

Introduction to Data Mining: Introduction, What is Data Mining, Definition, KDD, Challenges, Data Mining Tasks, Data Preprocessing, Data Cleaning, Missing data, Dimensionality Reduction, Feature Subset Selection, Discretization and Binaryzation, Data Transformation; Measures of Similarity and Dissimilarity- Basics.

UNIT – II

Association Rules: Problem Definition, Frequent Item Set Generation, Mining Frequent Patterns– Associations and correlations – Mining Methods– Mining Various kinds of Association Rules– Correlation Analysis– Constraint based Association mining. Graph Pattern Mining, SPM.

UNIT – III

Classification: Problem Definition, General Approaches to solving a classification problem , Evaluation of Classifiers , Classification techniques, Decision Trees-Decision tree Construction , Methods for Expressing attribute test conditions, Basic concepts–Decision tree induction–Bayesian classification, Rule–based classification, Lazy learner.

UNIT – IV

Clustering and Applications: Problem Definition, Clustering Overview, Cluster analysis–Types of Data in Cluster Analysis–Categorization of Major Clustering Methods– Partitioning Methods, Hierarchical Methods– Density–Based Methods, Grid–Based Methods, Outlier Analysis.

UNIT – V

Web and Text Mining: Introduction, Web Mining, Web Content Mining, Web Structure Mining, Web Usage Mining, Text Mining-Unstructured Text, Episode Rule Discovery for Texts, Hierarchy of Categories, Text Clustering.

TEXT BOOKS:

1. Data Mining- Concepts and Techniques- Jiawei Han, MorganKaufmann Publishers, Elsevier, 2 Edition, 2006.

2. Introduction to Data Mining, Pang-Ning Tan, Vipin Kumar, Michael Steinbanch,Pearson Education.

3. Data mining Techniques and Applications, Hongbo Du Cengage India Publishing

REFERENCES:

1.Data Mining Techniques, Arun K Pujari, 3rd Edition, Universities Press.

2. Data Mining Principles & Applications – T.V Sveresh Kumar, B. Esware Reddy,Jagadish S Kalimani, Elsevier.

3. Data Mining, Vikaram Pudi, P Radha Krishna, Oxford University Press

4. Ian H. Witten and Eibe Frank, Data Mining: Practical Machine Learning Tools and Techniques (Second Edition), Morgan Kaufmann, 2005.

5. INDIVIDUAL TIME TABLE

	I	II	III	IV		V	VI	VII
MON		**DM-3C**	**AI LAB-3A**				**DM-3A**	
TUE	**AI LAB-3A**			**DM-3A**			**AI LAB-3A**	
WED	**DM-3C**					**DM-3A**		
THU	**DM-3C**						**DM-3A**	
FRI				**DM-3C**			**DM-3C**	
SAT		**DM-3A**		**DM-3C**		**DM-3C**		**DM-3A**

6. Session plan/Lesson Plan.

CMR ENGINEERING COLLEGE

UGC AUTONOMOUS

(Approved by AICTE - New Delhi. Affiliated to JNTUH and Accredited by NAAC & NBA)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ACADEMIC YEAR: 2024-25

SESSION PLAN

FACULTY NAME: SILIVERI KIRAN KUMAT

SUBJECT: DATA MINING

YEAR & SEM: III B.TECH I SEM **SECTION: C**

S.NO	Topic (JNTU syllabus)	Sub-Topic	No. Of Lectures Required	Planned Date	Conducted Date	Methods of Teaching	Remarks
	UNIT - I						
1	Introduction to Data Mining	Introduction	L1	29/07/24	29/07/24	M1	
2		What is Data Mining Definition	L2	1/08/24	1/08/24	M1	
3		KDD, Challenges	L3	2/08/24	2/08/24	M1	
4		Data Mining Tasks	L4	4/08/24	4/08/24	M1	
5		Data Pre Processing	L5	5/08/24	5/08/24	M1	
6		Data Cleaning	L6	5/08/24	5/08/24	M1	

7		**Missing Data**	**L7**	**07/08/24**	**08/08/24**	**M1**	
8		**Dimensionality Reduction**	**L8**	**12/08/24**	**14/08/24**	**M1**	
9		**Feature Subset Selection**	**L9**	**19/08/24**	**21/08/24**	**M1**	
10		**Discrimination and Binarization**	**L10**	**22/08/24**	**22/08/24**	**M1**	
11		**Data Transformation**	**L11**	**28/08/24**	**28/08/24**	**M1**	
12	**ASSOCIATION RULES**	**Measures of Similarity and Dissimilarity-Basis**	**L12**	**29/08/24**	**29/08/24**	**M1**	
		UNIT- II					
13		**Problem Definition**	**L13**	**04/09/24**	**04/09/24**	**M1**	
14		**Frequent Item Set Generation**	**L14**	**05/09/24**	**05/09/24**	**M1**	
15		**The APRIORI Principle**	**L15**	**18/09/24**	**18/09/24**	**M1**	
16		**Support and Confidence Measures**	**L16**	**19/09/24**	**19/09/24**	**M1**	

17		**Association Rule Generation**	**L17**	**20/09/24**	**20/09/24**	**M1**	
18		**APRIORI Algorithm**	**L18**	**24/09/24**	**24/09/24**	**M1**	
19		**The Partition Algorithms**	**L19**	**25/09/24**	**25/09/24**	**M1**	
20		**FP-Growth Algorithms**	**L20**	**25/09/24**	**26/09/24**	**M1**	
21		**Compact Representation of Frequent Item Set**	**L21**	**26/09/24**	**27/09/24**	**M1**	
22		**Maximal Frequent Item Set**	**L22**	**27/09/24**	**28/09/24**	**M1**	
23		**Closed Frequent Item Set**	**L23**	**30/09/24**	**30/09/24**	**M1**	
		UNIT-III					
24		**Classification Problem Definition**	**L24**	**01/10/24**	**01/10/24**	**M1**	
25	**CLASSIFICATION**	**General Approach to solving a classification Problem**	**L25**	**03/10/24**	**03/10/24**	**M1**	
26		**Evaluation of Classifiers**	**L26**	**04/10/24**	**04/10/24**	**M1**	

27		**Decision Trees**	**L27**	**14/10/24**	**14/10/24**	**M1,M4**	
28		**Classification Techniques**	**L28**	**14/10/24**	**14/10/24**	**M1,M4**	
29		**Decision Tree Construction**	**L29**	**15/10/24**	**15/10/24**	**M1,M4**	
30		**Methods for Expressing attribute test conditions**	**L30**	**16/10/24**	**16/10/24**	**M1,M4**	
31		**Measures for Selecting the Best Split**	**L31**	**16/10/24**	**17/10/24**	**M1,M4**	
32		**Algorithm for Decision Tree Induction**	**L32**	**17/10/24**	**18/10/24**	**M1,M4**	
33		**Naïve Bayes Classifier**	**L33**	**18/10/24**	**19/10/24**	**M1,M4**	
34		**Bayesian Belief ?Networks**	**L34**	**21/10/24**	**21/10/24**	**M1,M4**	
35		**K-Nearest Neighbour Classification Algorithm and Characterstics**	**L35**	**22/10/24**	**22/10/24**	**M1,M4**	
		UNIT-IV					
36		**Problem Definition,Clustering Overview**	**L36**	**23/10/24**	**23/10/24**	**M1,M4**	
37		**Evaluation of Clustering Algorithms**	**L37**	**23/10/24**	**23/10/24**	**M1,M4**	

38		**Partitioniong Clustering=K-Means Algorithm Algorithm**	**L38**	**24/10/24**	**24/10/24**	**M1,M4**	
3	**CLUSTERING**	**K-Means Additional issues, PAM Algorithm**	**L39**	**25/10/24**	**25/10/24**	**M1,M4**	
40		**Hierarchical Clustering-Aggloramative Methods**	**L40**	**28/10/24**	**28/10/24**	**M1,M4**	
41		**Divisive Methods**	**L41**	**29/10/24**	**29/10/24**	**M1,M4**	
42		**Basic Aggloramative Hierarchical Clustering Algorithm**	**L42**	**30/10/24**	**30/10/24**	**M1,M4**	
43		**Specific Techniques, Key issues in Hierarchical Clustering**	**L43**	**30/10/24**	**30/10/24**	**M1,M4**	
44		**Strengths and Weakness, Outlier Detection**	**L44**	**01/11/24**	**01/11/24**	**M1,M4**	
		UNIT-V					
45		**Introduction**	**L45**	**04/11/24**	**04/11/24**	**M1,M4**	
46		**Web Mining**	**L46**	**05/11/24**	**05/11/24**	**M1,M4**	
47		**Web Content**	**L47**	**06/11/24**	**06/11/24**	**M1,M4**	

		Mining					
48		**Web Structure Mining**	**L48**	**06/11/24**	**06/11/24**	**M1,M4**	
49	**Wb and Text Mining**	**Web Usage Mining**	**L49**	**07/11/24**	**07/11/24**	**M1,M4**	
50		**Text Mining-Unstructured Text**	**L50**	**08/11/24**	**08/11/24**	**M1,M4**	
51		**Episode rule discovery for Text**	**L51**	**08/11/24**	**09/11/24**	**M1,M4**	
52		**Hierarchy of Categories**	**L52**	**09/11/24**	**09/11/24**	**M1,M4**	
53		**Text Clustering**	**L53**	**09/11/24**	**09/11/24**	**M1,M4**	

METHODS OF TEACHING

M1 : Lecture Method	**M6 : Tutorial**
M2 : Demo Method	**M7 : Assignment**
M3 : Guest Lecture	**M8 : Industry Visit**
M4 : Presentation /PPT	**M9 : Project Based**
M5 : Lab/Practical	**M10 : Charts / OHP**

NOTE:

1. Any Subject in a Semester is supposed to be completed in 55 to 65 periods.
2. Each Period is of 50 minutes.
3. Each unit duration & completion should be mentioned in the Remarks Column.
4. List of Suggested books can be marked with Codes like T1, T2, R1,R2 etc.

7. SESSION EXECUTION LOG

S no	Units	Scheduled started date	Completed date	Remarks
1	I	29-7-24	15-8-24	COMPLETED
2	II	16-8-24	30-8-24	COMPLETED
3	III	01-9-24	29-9-24	COMPLETED
4	IV	16-10-24	28-10-24	COMPLETED
5	V	2-11-24	16-11-24	COMPLETED

8. Lecture notes (Hand Written or softcopy printout 5 units) –attached

9. Assignment Questions along with sample Assignments Scripts

Data Mining

Mid -I Assignment Questions

1)Explain concept hierarchy generation for nominal data.

a) Describe feature subset selection.

2)Explain various data preprocessing techniques , how data reduction helps in data preprocessing.

a) Illustrate Apriori algorithm for the following data set TidList

1	Milk ,Dal ,Sugar ,Bread
2	Dal ,Sugar ,Wheat ,Jam
3	Milk ,Bread ,Curd ,Paneer
4	Wheat ,Paneer ,Dal ,Sugar
5	Milk ,Paneer ,Bread
6	Wheat ,Dal ,Paneer ,Bread

3) Explain how can you improve the performance of Apriori Algorithm?

4) Considering support threshold (50%) & Confidence(60%) , Illustrate FP Growth Algorithm for the following transactional dataset.

T1	I1 I2 I3
T2	I2 I3 I4
T3	I4 I5
T4	I1 I2 I4
T5	I1 I2 I3 I5
T6	I1 I2 I3 I4

5) Discuss about the classification of 2 step model and discuss about the issues regarding classification.

Mid –II Assignment Questions

1. a. What is a decision tree? Explain decision tree induction algorithm. (CO3)
   b. How to solve a classification problem using k-nearest neighbour algorithm? (CO3)
2. Explain the Naive Bayesian Classification algorithm. (CO3)
3. a. Describe the different types of hierarchical methods? (CO4)
   b. Explain the key issues, strengths and weaknesses of hierarchical clustering algorithms. (CO4)
4. Explain Partitioning Clustering - K -Means Algorithm with an example. (CO4)
5. a. Give a brief note on PageRank algorithm used in web structure mining? (CO5)
   b. Compare and contrast text mining with web content mining using lucid examples. (CO5)

10. Mid exam Question Papers along with sample Answers Scripts

Data Mining

I MID Question Paper for the A.Y 2024-25,I-SEM

CLASS: B.Tech III CSE A, B, C & D

SUB: **DATA MINING (DM)**

SET 1:

PART-A

Answer all Questions: **Marks: 2X5=10M**

1. List out the applications of data mining.(CO1)
2. Define binaryzation.?(CO1)
3. Define closed frequent itemset. (CO2)
4. What is the need of confidence measure in association rule mining?(CO2)
5. Define decision tree ? (CO3)

PART B **Marks: 3X5=15M**

6. Explain Data Mining Task Primitives?(CO1)

OR

7. Explain Data -preprocessing approaches?(CO1)

8. Explain FP growth algorithm with an example? (CO2)

Transaction 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, O}

Minimum support be 3. OR

9. Explain partition algorithm for the following dataset ? (CO2)

Transaction	Itemset
T1	I1,I5
T2	I2,I4,
T3	I4,I5
T4	I2,I3
T5	I5
T6	I2,I3,I4

10. Explain the Classification techniques? (CO3)

OR

11. Write the General Approaches to solving a classification problem .(CO3)

II MID Question Paper for the A.Y 2024-25,I-SEM

PART-A

Answer all Questions: Marks: 2X5=10M

1. List the characteristics of k-nearest neighbor algorithm. (CO3)
2. Give Brief Discussion about Clustering Problem Definition? (CO4)
3. What is the need of outlier detection? List two applications of it. (CO4)
1. Define web mining. (CO5)
2. Define text clustering. (CO5)

PART B **Marks: 3X5=15M**

3. Illustrate the Measures for Selecting the Best Split? (CO3)

OR

7. Explain about Bayesian Belief Networks? (CO3)

8. Give Specific techniques, Key Issues in Hierarchical Clustering. (CO4)

OR

9. Explain about Partitioning Clustering-K-Means Algorithm. (CO4)

10. Give brief explanation about unstructured text. (CO5)

OR

11. Describe episode rule discovery for texts. (CO5)

11. Scheme of Evaluation

MID-I

SCHEME OF EVALUATION

<table>
<tr><th>S.NO</th><th>THEORY</th><th>MARKS</th><th>TOTAL MARKS</th></tr>
<tr><td colspan="4">PART-A</td></tr>
<tr><td>1</td><td>applications of data mining.</td><td>2</td><td>2</td></tr>
<tr><td>2</td><td>Define binaryzation</td><td>2</td><td>2</td></tr>
<tr><td>3</td><td>Define closed frequent itemset.</td><td>2</td><td>2</td></tr>
<tr><td>4</td><td>measure in association rule mining</td><td>2</td><td>2</td></tr>
<tr><td>5</td><td>Define decision tree</td><td>2</td><td>2</td></tr>
<tr><td></td><td colspan="3">PART-B</td></tr>
<tr><td>6</td><td>Explain Data Mining Task</td><td>2</td><td rowspan="2">5</td></tr>
<tr><td></td><td>Data Mining Task Types</td><td>3</td></tr>
<tr><td rowspan="2">7</td><td>Explain Data -preprocessing approaches?</td><td>2</td><td rowspan="2">5</td></tr>
<tr><td>Data preprocessing steps</td><td>3</td></tr>
<tr><td rowspan="2">8</td><td>Explain FP growth algorithm with an example? (CO2)
<table><tr><th>Transaction ID</th><th>Items</th></tr><tr><td>T1</td><td>{E, K, M, N, O, Y}</td></tr><tr><td>T2</td><td>{D, E, K, N, O, Y}</td></tr><tr><td>T3</td><td>{A, E, K, M}</td></tr><tr><td>T4</td><td>{C, K, M, U, Y}</td></tr><tr><td>T5</td><td>{C, E, I, K, O, O}</td></tr></table></td><td>2</td><td rowspan="2">5</td></tr>
<tr><td>Minimum support be 3. Solve Problem</td><td>3</td></tr>
<tr><td>9</td><td>Explain partition algorithm for the following dataset ? (CO2)
<table><tr><td>Transaction</td><td>Itemset</td></tr><tr><td>T1</td><td>I1,I5</td></tr><tr><td>T2</td><td>I2,I4,</td></tr><tr><td>T3</td><td>I4,I5</td></tr><tr><td>T4</td><td>I2,I3</td></tr><tr><td>T5</td><td>I5</td></tr><tr><td>T6</td><td>I2,I3,I4</td></tr></table></td><td>2</td><td>5</td></tr>
</table>

	Solution problem(find final result)	3	
10	Explain the Classification	2	5
	techniques	3	
11	Write Explain the Classification	2	5
	the General Approaches to solving	3	

MID-II

S.NO	THEORY	MARKS	TOTAL MARKS
	PART-A		
1	List the characteristics	2	2
2	Discussion about Clustering	2	2
3	Need of outlier detection	2	2
4	Define web mining.	2	2
5	Define text clustering.	2	2
	PART-B		
6	Illustrate the Measures	3	5
	Selecting the Best Split	2	
7	Explanation	2	5
	About Bayesian Belief Networks?	3	
8	Give Specific techniques,	3	5
	Key Issues in Hierarchical Clustering.	2	
9	Explanation about Partitioning Clustering-K-Means Algorithm.	2	5
	Example K-Means	3	
10	Give brief explanation about	2	5
	Unstructured text. Diagram	3	
11	Describe episode rule discovery	2	5
	for texts. Examples	3	

12. Mapping of COs with POs and PSO's

Course Outcomes	**Relationship of Course Outcomes (CO) to Program Outcomes (PO)**											
CO/PO	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12
CO1	3	-								1	1	
CO2	3	3	2	2					2	1		1
CO3	3	3	2	2					2	1		1
CO4	3	2	2	2					2	1		1
CO5	3	2	2	2					2	1		1

13. CO's, PO's, PSO's Justification

S. No	**Course Out Come**
CO1	Explain the types of the data to be mined and present a general classification of tasks and primitives to integrate a data mining system (Understanding).
CO2	Apply preprocessing methods for any given raw data and extract interesting pattern from large amounts of data. (Applying)
CO3	Discover the role played by data mining in various fields. (Analyzing)
CO4	Choose and employ suitable data mining algorithms to build analytical applications.(Evaluating)
CO5	Evaluate the accuracy of supervised and unsupervised model and algorithms. (Evaluating)

Justification:

CO1.: Explain the types of the data to be mined and present a general classification of tasks and primitives to integrate a data mining system
Correlated with PO1 moderately: Because it contributes the knowledge on fundamentals of Data Mining which makes students get engineering knowledge and student can categorize different utilities. So, overall the correlation of CO1 to PO1 is good.
Correlated with PO10 moderately: Because it provides communication in complex activities with effective reports and design documentation. So Correlation of CO1 with PO10 is low.
Correlated with PO11 moderately: Because it demonstrates knowledge and understanding of the Engineering and management Principles. So Correlation of CO1 with PO11 is low.

CO2.: Apply preprocessing methods for any given raw data and extract interesting pattern from large amounts of data
Correlated with PO1 moderately: Because it provides fundamentals of computer science. So, correlation is good.
Correlated with PO2 moderately: Because it Apply preprocessing methods for any given raw data. So, correlation is good.
Correlated with PO3 moderately: Because it provides solutions for complex engineering problems. So, correlation is good.
Correlated with PO4 moderately: Because it provides Analyses and Interpretation of data. So, correlation is good.
Correlated with PO9 moderately: Because it provides function effectively as an individual for data. So, correlation is average.
Correlated with PO10 moderately: An ability to communicate effectively with a range of audiences
Correlated with PO12 moderately: Recognition of the need for and an ability to engage in continuing professional development.

CO3.: Discover the role played by data mining in various fields.
Correlated with PO1 moderately: Because it provides an engineering specialization to the solution of complex engineering problems. So, correlation is good.

Correlated with PO2 moderately: An ability to analyze a problem, and identify and formulate the computing requirements appropriate to its solution.
Correlated with PO3 moderately: An ability to design, implement, and evaluate a computer-based system, process, component, or program to meet desired needs with appropriate consideration for public health and safety, cultural, societal and environmental considerations.
Correlated with PO4 moderately: An ability to design and conduct experiments, as well as to analyze and interpret data.
Correlated with PO9 moderately: An ability to function effectively individually and on teams, including diverse and multidisciplinary, to accomplish a common goal.
Correlated with PO10 moderately: An ability to communicate effectively with a range of audiences
Correlated with PO12 moderately: Recognition of the need for and an ability to engage in continuing professional development.

CO4.: Choose and employ suitable data mining algorithms to build analytical applications
Correlated with PO1 moderately: An ability to apply knowledge of computing, mathematics, science and engineering fundamentals appropriate to the discipline.
Correlated with PO2 moderately: An ability to analyze a problem, and identify and formulate the computing requirements appropriate to its solution.
Correlated with PO3 moderately: An ability to design, implement, and evaluate a computer-based system, process, component, or program to meet desired needs with appropriate consideration for public health and safety, cultural, societal and environmental considerations.
Correlated with PO4 moderately: An ability to design and conduct experiments, as well as to analyze and interpret data.
Correlated with PO9 moderately: An ability to function effectively individually and on teams, including diverse and multidisciplinary, to accomplish a common goal.
Correlated with PO10 moderately: It is an ability to communicate effectively with a range of audiences.
Correlated with PO12 moderately: Recognition of the need for and an ability to engage in

continuing professional development.

CO5.: Evaluate the accuracy of supervised and unsupervised model and algorithms
Correlated with PO1 moderately: To apply knowledge of computing, mathematics, science and engineering fundamentals appropriate to the discipline.
Correlated with PO2 moderately: To analyze a problem, and identify and formulate the computing requirements appropriate to its solution.
Correlated with PO3 moderately: To design, implement, and evaluate a computer-based system, process, component, or program to meet desired needs with appropriate consideration for public health and safety, cultural, societal and environmental considerations.
Correlated with PO4 moderately: To design and conduct experiments, as well as to analyze and interpret data.
Correlated with PO9 moderately: to function effectively individually and on teams, including diverse and multidisciplinary, to accomplish a common goal.
Correlated with PO10 moderately: To communicate effectively with a range of audiences.
Correlated with PO12 moderately: An ability to engage in continuing professional development.

14. Attainment of CO's, PO's and PSO's (Excel Sheet)

After Result

15. Previous Year Question Papers/ Question Bank

Unit -1

1. What is Data Mining? Explain the process of knowledge discovery in database?

2. Write a short note on Data Mining Functionalities?

3. Describe Classification of Data Mining Systems?

4. Explain Data Mining Task Primitives?

5. What are the various forms of data pre-processing techniques?

6. Mention Data mining functionality, classification, prediction, clustering & evolution analysis?

7. What are the challenges in methodology of Data Mining technology?

8. Discuss issues to consider during Data Mining?

9. What defines a Data Mining Task Explain at least 5 primitives?

10. What is knowledge discovery?

11. Explain the motivating challenges in development of data mining.

12. Explain with example the data mining tasks

13.What is a data? What do you mean by quality of data?

14.What is a data set? Explain the various types of data sets

15. What is data preprocessing?

Unit –II

1. What is Apriori algorithm?
2. Explain the association rule Mining?
3. What is more efficient method for Generalizing association rule explain?
4. What is meant by association rule?
5. Explain the Partition Algorithms ?
6. State and explain FP-Growth Algorithms?
7. What is meant by Frequent itemset. ?
8. What is meant by Maximal Frequent Item Set?
9. What is meant by Closed Frequent Item Set?

Unit –III

1. Give Brief discussion about classifiers Problem Definition,
2. General Approaches to solving a classification problem ,
3. Write short note on Evaluation of Classifiers ,
4. List Classification techniques,
5. State and explain Decision Trees-Decision tree Construction ,
6. Explain Methods for Expressing attribute test conditions,
7. Explain the Measures for Selecting the Best Split,
8. Algorithm for Decision tree Induction ;
9. Explain about Naive-Bayes Classifier,
10. Explain about Bayesian Belief Networks
11. Explain K- Nearest neighbor classification-Algorithm and Characteristics.

Unit –IV

1. Give Brief Discussion about Clustering Problem Definition?
2. Give Clustering Overview
3. Explain Evaluation of Clustering Algorithms
4. Explain about Partitioning Clustering-K-Means Algorithm
5. Write a short note on K-Means Additional issues
6. Explain about PAM Algorithm.
7. Write a short note on Hierarchical Clustering.
8. Explain Agglomerative Methods and divisive methods
9. Explain Basic Agglomerative Hierarchical Clustering Algorithm.
10. Give Specific techniques, Key Issues in Hierarchical Clustering
11. Explain Hierarchical Clustering Strengths and Weakness
12. Explain about Outlier Detection.

Unit –V

1. Write a short note on Web and Text Mining?

2. Explain about web mining.
3. Explain about web content mining.
4. Explain web structure mining?
5. State and explain we usage mining?
6. Explain about Text mining
7. Give brief explanation about unstructured text
8. Explain episode rule discovery for texts.
9. Write about hierarchy of categories, text clustering.

Code No: 137BQ **R16**

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY HYDERABAD
B. Tech IV Year I Semester Examinations, March - 2021
DATA MINING
(Common to CSE, IT)

Time: 3 Hours **Max. Marks: 75**

Answer any Five Questions
All Questions Carry Equal Marks

- -

1.a) How to handle redundancy in data integration?
b) Explain principal component analysis as a method of dimensionality reduction. [7+8]

2. How can we mine closed frequent item sets? Explain. [15]

3. Explain market basket analysis and its relevance to association rule. Explain the Apriori algorithm using the following transactional data assuming that the support count is 22%. Illustrate with an example.

TID	LIST OF ITEMS
001	milk, dal, sugar, bread
002	Dal, sugar, wheat ,jam
003	Milk, bread, curd, paneer
004	Wheat, paneer, dal, sugar
005	Milk, paneer, bread
006	Wheat, dal, paneer, bread.

[15]

4. Discuss K- Nearest neighbor classification-Algorithm and Characteristics. [15]

5. How Neural Networks can be used for Data classification? Which algorithm is suitable? Explain them with example? [15]

6. Explain various issues and challenges in data mining. [15]

7.a) Describe web usage mining.
b) Explain about Text Clustering with an illustrative example. [7+8]

8.a) Write and explain about the k-medoids algorithm.
b) Describe distance based outlier detection. [8+7]

--ooOoo--

1.a) Define data mining. [2]
b) List the methods of filling missing values. [3]
c) Define closed frequent itemset. [2]
d) What is the need of confidence measure in association rule mining? [3]
e) List the measures for selecting best split in decision tree construction. [2]
f) Quote an example for Bayesian belief network. [3]
g) What are the limitations of single linkage algorithm? [2]
h) List the typical requirements of clustering in data mining. [3]
i) What is meant by stop words? [2]
j) Give the taxonomy of web mining [3]

PART – B

(50 Marks)

2. Discuss data mining as a step in knowledge discovery process and various challenges associated. [10]

OR

3. Use a flowchart to summarize the following procedures for attribute subset selection:
a) Stepwise forward selection
b) Stepwise backward elimination. [10]

4. Classify frequent pattern mining methods and explain the criteria followed for classification. [10]

OR

5. Apply apriori algorithm to find frequent itemsets from the following transactional database. Let min_sup = 30%. [10]

TID	Items_bought
1	Pen, notebook, ruler
2	Pencil, eraser, sharpener
3	Pen, ruler, chart, sharpener
4	Pencil, clip, eraser
5	Ruler, pin, story book, pen
6	Marker, chart, sketchpens

6. State classification problem and briefly explain general approaches to solve it. [10]

OR

7. Apply Naïve-Bayesian classifier to identify class label(campus_placement) to the new sample/student < 7 to 8, 'Fair', 'Excellent','No'>. [10]

SID	CGPA	Coding Skills	Soft Skills	Hackathon Participation	Campus_placement
1	7 to 8	Excellent	Fair	Yes	Yes
2	8 to 9	Fair	Excellent	Yes	Yes
3	9 to10	Poor	Fair	No	Yes
4	5 to 6	Poor	Excellent	No	No
5	7 to 8	Excellent	Poor	No	No
6	8 to 9	Fair	Fair	Yes	Yes
7	9 to 10	Poor	Poor	No	No

8. Suppose that the data mining task is to cluster the following eight students into three clusters, the distance function is Manhattan.Assign record 1,2,3 as the centroid of each cluster respectively.Use the k-means algorithm to show the final three clusters. [10]

RecordID	Height(cms)	Weight(kgs)
1	145	35
2	165	55
3	170	90
4	135	60
5	140	50
6	160	75
7	150	40
8	155	65

OR

9. Appraise the importance of outlier detection and its application. Explain any one approach

R16

Code No: 137BQ

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY HYDERABAD
B. Tech IV Year I Semester Examinations, February/March - 2022
DATA MINING
(Common to CSE, IT)

Time: 3 Hours **Max. Marks: 75**

Answer any five questions
All questions carry equal marks

1.	a)	Explain Various Data Mining Functionalities with an example.	[8+7]
	b)	Illustrate about Data Mining Task Primitives.	
2.	a)	What is Data Cleaning? Describe various methods of Data Cleaning.	[7+8]
	b)	Discuss about the Issues to be considered during Data Integration.	
3.		Write a note on Maximal Frequent Item Set and Closed Frequent Item Set.	[15]
4.		Explain about the Apriori algorithm for finding frequent item sets with an example.	[15]
5.		Discuss about Decision tree induction algorithm with an example.	[15]
6.		Discuss about Naïve-Bayes classification algorithm with an example.	[15]
7.		Write partitioning around mediods algorithm.	[15]
8.		Explain about hierarchy of categories in text mining.	[15]

---ooOoo---

Code No: 157BC **R18**

JAWAHARLAL NE... ...IVERSITY HYDERABAD

B. Tech IV Year ... ... Examinations, February/March - 2022

DATA MINING

(Common to CSE, IT)

Time: 3 Hours Max. Marks: 75

Answer any Five Questions
All Questions Carry Equal Marks

---

1. Explain the need of data preprocessing and various forms of preprocessing. [15]

2. What is a data warehouse? Demonstrate integrating data mining system with a data warehouse with a neat diagram. [15]

3. Apply FP-Growth algorithm to the following data for finding frequent item sets, consider support threshold as 30%. [15]

TID	List of ItemIDs
1	I1,i2, i4, i5
2	I2, i4, i7
3	I2,i3,i4,i5
4	I1,i3,i4,i7
5	I1,i2,i3,i4,i5
6	I3,i4,i5,i6

4.a) How to identify sub-graphs in a graph?
b) Give an overview of correlation analysis. [8+7]

5.a) Explain classification as a two step process.
b) State Bayes theorem. How this concept is used in classification. [8+7]

6. What is a decision tree? Explain decision tree induction algorithm. [15]

7.a) Contrast k-means clustering with k-medoids clustering approach.
b) Discuss the merits and demerits of hierarchical approaches for clustering. [8+7]

8. How to apply mining techniques to unstructured text database? Explain with example. [15]

—ooOoo—

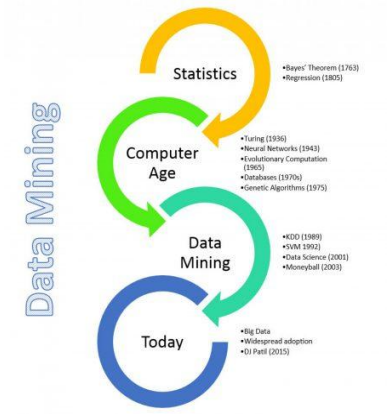
16. Power Point Presentations (PPTs)

Unit-1
INTRODUCTION TO DATA MINING

By,
MRUTYUNJAYA S Y
Assistant Professor,
CSE dept,
CMREC, Hyderabad

HISTORY

- The history of Data Mining started very recently as it is commonly considered with new technology.
- However data is a discipline with a long history.
- It starts with the early Data Mining methods **Bayes' Theorem (1700`s)** and **Regression analysis (1800`s)** which were mostly identifying patterns in data.



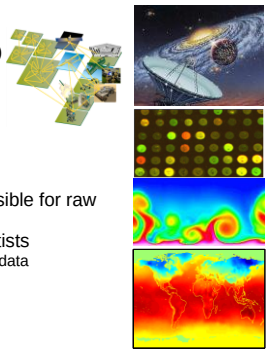
Why Mine Data? Commercial Viewpoint

- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - purchases at department/ grocery stores
  - Bank/Credit Card transactions

- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an *edge* (e.g. in Customer Relationship Management)

Why Mine Data? Scientific Viewpoint

- Data collected and stored at enormous speeds (GB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data
- Traditional techniques infeasible for raw data
- Data mining may help scientists
  - in classifying and segmenting data
  - in Hypothesis Formation



What Is Data Mining?

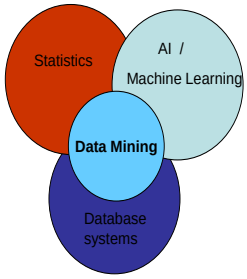
- Data mining (knowledge discovery in databases):
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases
  - Knowledge discovery(mining) in databases (KDD), knowledge extraction, data/pattern analysis.
  - **Data mining** is a process used by companies to turn raw **data** into useful information. By using software to look for patterns in large batches of **data**, businesses can learn more about their customers and develop more effective marketing strategies as well as increase sales and decrease costs.

????Questions????

- Where exactly the DATA can be stored?
- How and from where to extract the DATA?
- From where the DATA we can be MINE?
- Answer is **DATA WAREHOUSE**.....

Origins of Data Mining

- **Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems**
- **Must address:**
  - **Enormity of data**
  - **High dimensionality of data**
  - **Heterogeneous, distributed nature of data**



Database Processing vs. Data Mining Processing

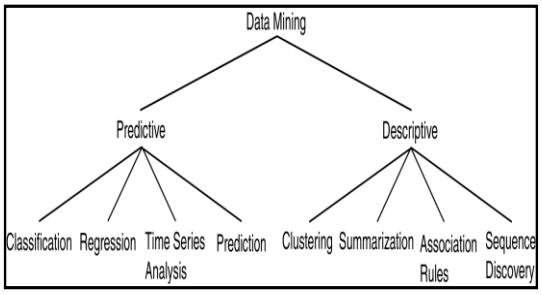
- Query
  - Well defined
  - SQL
- Query
  - Poorly defined
  - No precise query language

- Output
  - Precise
  - Subset of database
- Output
  - Fuzzy
  - Not a subset of database

Query Examples

- Database
  - Find all credit applicants with last name of Smith.
  - Identify customers who have purchased more than $10,000 in the last month.
  - Find all customers who have purchased milk
- Data Mining
  - Find all credit applicants who are poor credit risks. (classification)
  - Identify customers with similar buying habits. (Clustering)
  - Find all items which are frequently purchased with milk. (association rules)

Data Mining Models and Tasks



Data Mining: Classification Schemes

- Decisions in data mining
  - Kinds of databases to be mined
  - Kinds of knowledge to be discovered
  - Kinds of techniques utilized
  - Kinds of applications adapted
- Data mining tasks
  - Descriptive data mining
  - Predictive data mining

Decisions in Data Mining

- **Databases to be mined**
  - Relational, transactional, object-oriented, object-relational, time-series, text, multi-media, heterogeneous, legacy, WWW, etc.
- **Knowledge to be mined**
  - Characterization, discrimination, association, classification, clustering, trend, deviation and outlier analysis, etc.
  - Multiple/integrated functions and mining at multiple levels
- **Techniques utilized**
  - Database-oriented, data warehouse (OLAP), machine learning, statistics, visualization, neural network, etc.
- **Applications adapted**
  - Retail, telecommunication, banking, fraud analysis, DNA mining, stock market analysis, Web mining, Weblog analysis, etc.

Data Mining Tasks

- Prediction Tasks
  - Use some variables to predict unknown or future values of other variables
- Description Tasks
  - Find human-interpretable patterns that describe the data.

Common data mining tasks
- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]

Classification: Definition

- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - **A *test set* is used to determine the accuracy of the model.**
  - Usually, the given data set is divided into ***training and test sets,*** with training set used to build the model and test set used to validate it.

Classification: Application 1

- Direct Marketing
  - Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.
  - Approach:
    - Use the data for a similar product introduced before.
    - We know which customers decided to buy and which decided otherwise. This *{buy, don't buy}* decision forms the *class attribute*.
    - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
      - Type of business, where they stay, how much they earn, etc.
    - Use this information as input attributes to learn a classifier model.

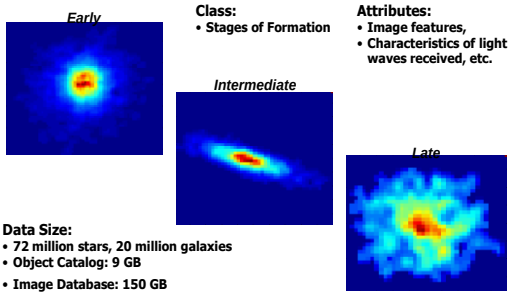
Classification: Application 2

- Fraud Detection
  - Goal: Predict fraudulent cases in credit card transactions.
  - Approach:
    - Use credit card transactions and the information on its account-holder as attributes.
      - When does a customer buy, what does he buy, how often he pays on time, etc
    - Label past transactions as fraud or fair transactions. This forms the class attribute.
    - Learn a model for the class of the transactions.
    - Use this model to detect fraud by observing credit card transactions on an account.

Classification: Application 3

- Sky Survey Cataloging
  - Goal: To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory).
    - 3000 images with 23,040 x 23,040 pixels per image.
  - Approach:
    - Segment the image.
    - Measure image attributes (features) - 40 of them per object.
    - Model the class based on these features.
    - Success Story: Could find 16 new high red-shift quasars, some of the farthest objects that are difficult to find!

Classifying Galaxies

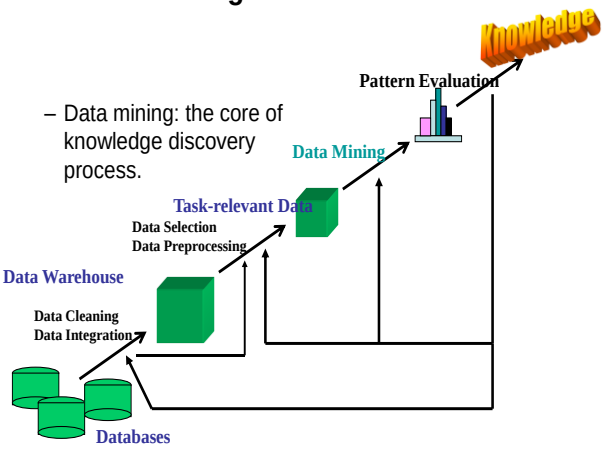


Class:
- Stages of Formation

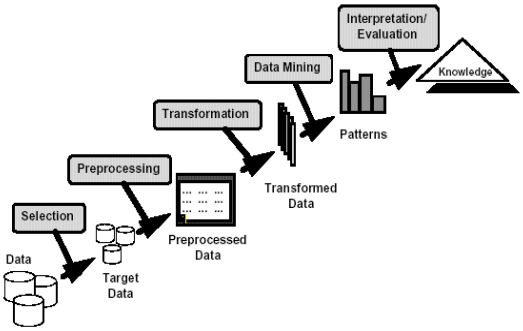
Attributes:
- Image features,
- Characteristics of light waves received, etc.

Data Size:
- 72 million stars, 20 million galaxies
- Object Catalog: 9 GB
- Image Database: 150 GB

Data Mining: A KDD Process



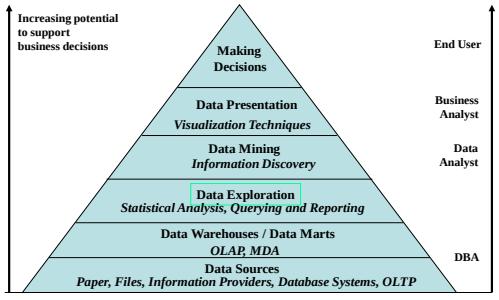
– Data mining: the core of knowledge discovery process.



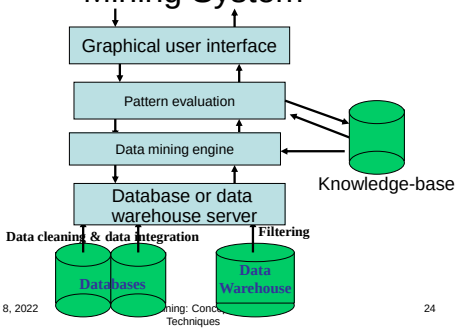
Steps of a KDD Process

- Learning the application domain:
  - relevant prior knowledge and goals of application
- Creating a target data set: data selection
- **Data cleaning** and preprocessing: (may take 60% of effort!)
- **Data reduction and transformation**:
  - Find useful features, dimensionality/variable reduction, invariant representation.
- Choosing functions of data mining
  - summarization, classification, regression, association, clustering.
- Choosing the mining algorithm(s)
- **Data mining**: search for patterns of interest
- **Pattern evaluation and knowledge presentation**
  - visualization, transformation, removing redundant patterns, etc.
- Use of discovered knowledge

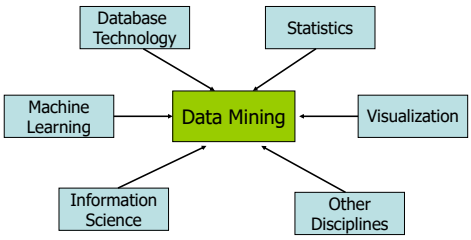
Data Mining and Business Intelligence



Architecture of a Typical Data Mining System



Data Mining: Confluence of Multiple Disciplines



Examples of Large Datasets

- Government: IRS, NGA, …
- Large corporations
  - WALMART: 20M transactions per day
  - MOBIL: 100 TB geological databases
  - AT&T 300 M calls per day
  - Credit card companies
- Scientific
  - NASA, EOS project: 50 GB per hour
  - Environmental datasets

DATA MINING APPLICATIONS

- Areas of Use (Huge usage in All Fields)
  - Internet – Discover needs of customers
  - Economics – Predict stock prices
  - Science – Predict environmental change
  - Medicine – Match patients with similar problems → cure
- Credit Card Company wants to discover information about clients from databases. Want to find:
  - Clients who respond to promotions in "Junk Mail"
  - Clients that are likely to change to another competitor

Data Preprocessing

- Why preprocess the data?
- Descriptive data summarization
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

Why Data Preprocessing?

- Data in the real world is dirty
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., occupation=" "
  - noisy: containing errors or outliers
    - e.g., Salary="-10"
  - inconsistent: containing discrepancies in codes or names
    - e.g., Age="42" Birthday="03/07/1997"
    - e.g., Was rating "1,2,3", now rating "A, B, C"
    - e.g., discrepancy between duplicate records

Why Is Data Dirty?

- Incomplete data may come from
  - "Not applicable" data value when collected
  - Different considerations between the time when the data was collected and when it is analyzed.
  - Human/hardware/software problems
- Noisy data (incorrect values) may come from
  - Faulty data collection instruments
  - Human or computer error at data entry
  - Errors in data transmission
- Inconsistent data may come from
  - Different data sources
  - Functional dependency violation (e.g., modify some linked data)
- Duplicate records also need data cleaning

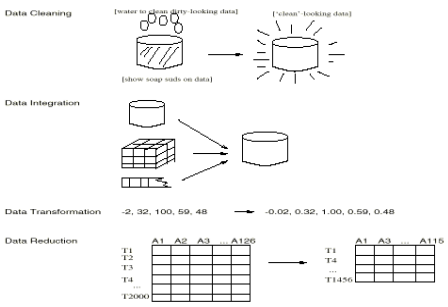
Why Is Data Preprocessing Important?

- No quality data, no quality mining results!
  - Quality decisions must be based on quality data
    - e.g., duplicate or missing data may cause incorrect or even misleading statistics.
  - Data warehouse needs consistent integration of quality data
- Data extraction, cleaning, and transformation comprises the majority of the work of building a data warehouse

Major Tasks in Data Preprocessing

- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data integration
  - Integration of multiple databases, data cubes, or files
- Data transformation
  - Normalization and aggregation
- Data reduction
  - Obtains reduced representation in volume but produces the same or similar analytical results
- Data discretization
  - Part of data reduction but with particular importance, especially for numerical data

Forms of Data Preprocessing



Data Cleaning

- Importance
  - "Data cleaning is one of the three biggest problems in data warehousing"—Ralph Kimball
  - "Data cleaning is the number one problem in data warehousing"—DCI survey
- Data cleaning tasks
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data
  - Resolve redundancy caused by data integration

Missing Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data
- Missing data may need to be inferred.

Noisy Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention
- Other data problems which requires data cleaning
  - duplicate records
  - incomplete data
  - inconsistent data

How to Handle Noisy Data?

- Binning
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Regression
  - smooth by fitting the data into regression functions
- Clustering
  - detect and remove outliers
- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)

Data Cleaning as a Process

- Data discrepancy detection
  - Use metadata (e.g., domain, range, dependency, distribution)
  - Check field overloading
  - Check uniqueness rule, consecutive rule and null rule
  - Use commercial tools
    - Data scrubbing: use simple domain knowledge (e.g., postal code, spell-check) to detect errors and make corrections
    - Data auditing: by analyzing data to discover rules and relationship to detect violators (e.g., correlation and clustering to find outliers)
- Data migration and integration
  - Data migration tools: allow transformations to be specified
  - ETL (Extraction/Transformation/Loading) tools: allow users to specify transformations through a graphical user interface
- Integration of the two processes
  - Iterative and interactive (e.g., Potter's Wheels)

17. Innovative Teaching method if any(Attached Innovative Assignment)

QUESTIONS

1. What are five challenges when conducting data mining?

2. What are the different problems that data mining solve?

18. References (Tesxt Book/Websites/ Journals)

1. https://www.tutorialspoint.com/data_mining/index.htm
2. https://www.javatpoint.com/data-mining
3. https://data-flair.training/blogs/data-mining-tutorial/
4. https://www.guru99.com/data-mining-tutorial.html
5. http://ir.inflibnet.ac.in:8080/ir/bitstream/1944/435/1/04Planner_22.pdf
6. https://www.talend.com/resources/data-mining-techniques/

Text Book Link:

T1 :

http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf