UNIT I – CHAPTER 1,2,3 MH DUNHAM

1. Define
    a) data mining
    b) data mining algorithm
2. Explain the types of data mining tasks
3. Write a short note on KDD
4. Write a short note on visualization
5. Explain issues in data mining
6. Write short notes on the following:
    a) Database/ OLTP systems
    b) Fuzzy systems
    c) Information Retrieval
    d) Dimensional Modeling
    e) Data Warehousing
    f) OLAP/ DSS
    g) Statistics, Machine Learning & Pattern Matching
7. Explain Point estimation
8. Write short notes on:
    h) MSE
    i) RMSE
    j) Jack Knife Estimate
    k) MLE
    l) EM
9. Write short note on Box plot and scatter diagram.
10. Explain Bayes Theorem
11. Write short note on Hypothesis testing
12. Explain linear regression with an example
13. Explain Correlation analysis
14. Write short notes on Similarity measures.
15. What is a Decision tree? Explain in detail
16. Write a short note on Neural networks
17. Explain the concept of Genetic algorithms in detail

Unit 1 ch3 , 5 C J Cios

1. What is data? State the relationship between values, features, objects, data sets, database and datawarehouse
2. Explain in brief about datawarehouse and its components.
3. Write a note on Multidimensional data cube.
4. Explain the various reasons that affect the amount and quality of data.
5. Explain any one technique that speeds up the algorithm.
6. Explain the following categories of knowledge representation.
    a. Rules
    b. Graphs
    c. Trees
    d. Networks
7. Explain relationship between Knowledge Representation Schemes & different levels of granularity
8. Explain concept of Granularity in rules.

UNIT II – CHAPTER Ch 2 (han & Kamber)

1. Why do we need data pre-processing ? Why is it important? Explain the major tasks in data pre-processing.
2. Define the following terms with respect to central tendency of data :
   a. Distributive measure, Algebraic Measure & Holistic measure
   b. Mean & Weighted Arithmetic mean
   c. Median, Mode, Range, Mid-range

3. Explain box-plot and all terms involved.

4. With respect to data cleaning explain how are the following problems treated:
   a. Missing values
   b. Noisy data

5. Co-relation analysis

   A 2 × 2 contingency table for the data of Example 2.1. Are *gender* and *preferred_Reading* correlated?

	male	*female*	Total
fiction	250 (90)	200 (360)	450
non_fiction	50 (210)	1000 (840)	1050
Total	300	1200	1500

   Are gender & Preferred reading co-related? Find chi-square value

6. Explain the following with respect to data transformation:
   a. Smoothing
   b. Aggregation
   c. Generalization
   d. Normalization
   e. Attribute construction

7. What are data reduction techniques? List the different strategies used for it. Explain reduction using Attribute subset selection.

8. Explain numerosity reduction using sampling methods.

UNIT II – CHAPTER Ch 5 (han & Kamber)

1. Define the following with respect to association mining:
   a. Association Rule
   b. Support & Confidence of a rule
   c. Frequent Itemset
   d. Closed Frequent Itemset
   e. Maximal Frequent Itemset

2. Explain Apriori Algorithm with an example

3. What do you mean by frequent it m set? Use Apriori algorithm to generate frequent item sets for the following by taking support threshold as 60%:

TID	Items
1	{ Egg , Milk}
2	{ Egg, Chips, Butter, Beer}
3	{ Milk, Chips, Butter, popcorn}
4	{ Egg, Milk, Chips, Butter}
5	{Egg, Milk, Chips, popcorn}

4. Define support and confidence to measure the strength of association rule. Calculate support and confidence for the rules jam => (butter, bread) and butter => bread, from the following:

Basket 1: bread, butter, jam
Basket 2: bread, butter
Basket 3: bread, butter, milk
Basket 4: beer, bread
Basket 5: beer, milk

5. Explain how to generate Frequent pattern (FP) tree for the following data sets with minimum support 3:

TID	***Items bought***	***(ordered) frequent items***
100	***{f, a, c, d, g, i, m, p}***	***{f, c, a, m, p}***
200	***{a, b, c, f, l, m, o}***	***{f, c, a, b, m}***
300	***{b, f, h, j, o, w}***	***{f, b}***
400	***{b, c, k, s, p}***	***{c, b, p}***
500	***{a, f, c, e, l, p, m, n}***	***{f, c, a, m, p}***

6. What do you mean by frequent item set? Use Apriori algorithm to generate frequent item sets for the following by taking support threshold as 60%:

TID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{ Bread, Milk, Diapers, Beer}
5	{ Bread, Milk, Diapers, Cola}

Define support and confidence in association rule .Following is the frequent item set table explain how you can generate association rule for {2, 3, 5} with minimum confidence = 75%

x	{1}	{2}	{3}	{5}	{1,3}	{2,3}	{2,5}	{3,5}	{2,3,5}
f	2	3	3	3	2	2	3	2	2

A database has five transactions. Let $min\_sup = 60\%$ and $min\_conf = 80\%$.

TID	*items_bought*
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y }
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I ,E}

7. (a) Find all frequent itemsets using Apriori and FP-growth, respectively. Compare the efficiency of the two mining processes.

8. Explain the following terms:

   a. Correlation rule
   b. Lift
   c. All confidence
   d. Cosine

9. What is constraint based association mining? What are the different types of constraints?
10. Write a short note on Association rule mining

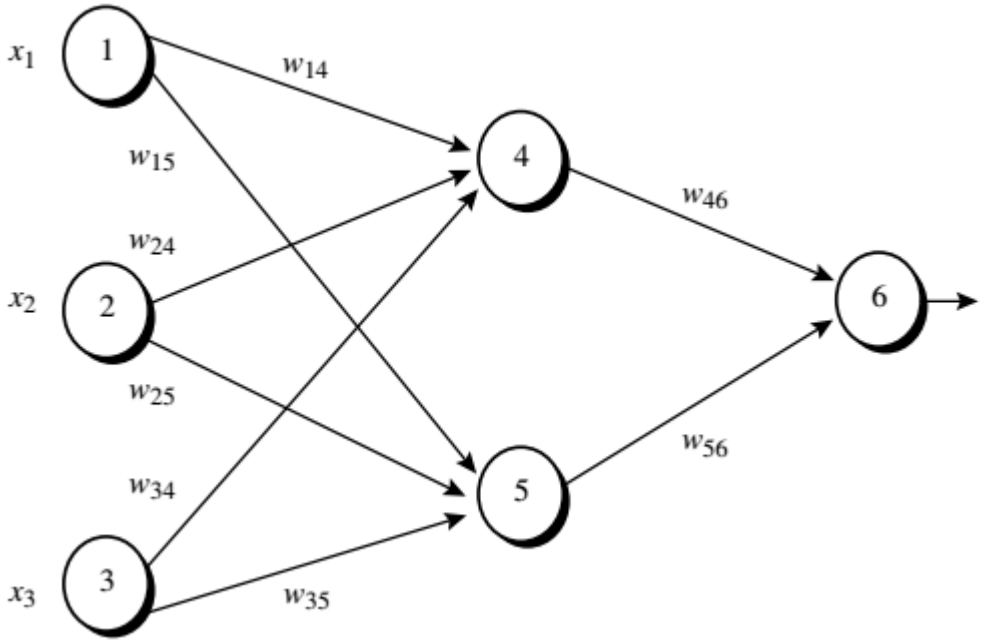
UNIT III – CHAPTER Ch 6 (han & Kamber)

1. Explain in brief classification and prediction with an example.
2. State the issues regarding classification process.
3. Explain algorithm for decision tree induction
4. Using Attribute selection: For the data given below compute
   a. Information Gain (age, income, student, creditrating)
   b. Gain Ration (income)
   c. Gini Index {low, medium} {high}

Class-labeled training tuples from the *AllElectronics* customer database.

RID	*age*	*income*	*student*	*credit_rating*	*Class: buys_computer*
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

5. Write a short note on Bayesian classifier
6. For the table given in Q.4 predict the class lable of tuple X using naïve bayes classifier ***X*** = (*age = youth, income = medium, student = yes, credit rating = fair*)
7. Write a short note on Rule based classifier or Explain in brief the IF – THEN Rules for classification.
8. Explain Classification using backpropagation algorithm
9. For the given neural network, give the calculations in learning using Backpropagation algorithm for predicting the class of tuple X with a learning rate = 0.9



Initial input, weight, and bias values.

x_1	x_2	x_3	w_{14}	w_{15}	w_{24}	w_{25}	w_{34}	w_{35}	w_{46}	w_{56}	θ_4	θ_5	θ_6
1	0	1	0.2	−0.3	0.4	0.1	−0.5	0.2	−0.3	−0.2	−0.4	0.2	0.1

10. Write a short note on classification using Support Vector Machine.
11. Explain the following terms with respect to Classifier Accuracy Measures:
    a. Classifier Error Rate
    b. Confusion matrix
    c. Sensitivity
    d. Specificity
    e. Precision
    f. Accuracy

12. Explain the following ensemble methods:
    a. Bagging
    b. Boosting

UNIT IV – CHAPTER Ch 6 (han & Kamber)

1. What is cluster analysis? Explain the types of data in cluster analysis
2. List and discuss major clustering approaches.
3. Derive the equations for Manhattan distance, Euclidean distance with an example.
4. Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8):

   (a) Compute the *Euclidean distance* between the two objects.

   (b) Compute the *Manhattan distance* between the two objects.

   (c) Compute the *Minkowski distance* between the two objects, using $q = 3$.
5. Compute Euclidean and Manhattan distance for X1 (1, 2) and X2 (3,6).
6. For the given Binary data, Compute the dissimilarity for
   a. Jack and Mary
   b. Mary and Jim
   c. Jack and Jim

name	gender	fever	cough	test-1	test-2	test-3	test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	Y	N	N	N	N

7. Give the outline of k-means partitioning method.
8. Give the outline of k-mediods partitioning method.
9. Explain agglomerative hierarchical clustering method.
10. Explain Divisive hierarchical clustering method
11. Write a short note on DBSCAN.
12. Write a short note on ROCK.
13. Apply hierarchical clustering using single linkage to follow ing data.
    A (1,1), B(1.5,1.5), C(3,4), D(4,4), E(3,3.5)
14. Explain the following with reference to density based clustering (give proper diagram):
    a. ε-neighborhood of an object.
    b. core object
    c. directly density-reachable object
    d. density-reachable object
    e. density-connected object
15. What are outliers? How to find out? Write the applications.

UNIT V – CHAPTER Ch 9 (han & Kamber)

1. What is graph mining and social network analysis?
2. Explain Characteristics of Social Networks
3. Explain Tasks in Link Mining
4. Explain Challenges in Link Mining
5. Explain multi relational data mining
6. What are multimedia and spatial databases?
7. Explain set and listed valued attribute with example.
8. Explain set and complex structure valued attribute.
9. What is spatial aggregation and approximation? Explain with example.
10. Explain Generalization of Structured Data
11. Explain the types of dimensions & measures in a spatial data cube

12. Explain the different approaches for similarity Search in Multimedia Data
13. Explain concept of mining Associations in Multimedia Data
14. Explain the following with respect to text mining:
    a. Precision and Recall, *F score*
    b. Document ranking
    c. Tokenization
    d. Stop list
    e. Word stem
    f. term-frequency matrix
    g. Relative term frequency
    h. inverse document frequency

15. ExplainText Indexing Techniques
16. Compute TF, IDF and TF-IDF for t2 in d2 for follow ing data.

document/term	t_1	t_2	t_3	t_4	t_5	t_6	t_7
d_1	0	4	10	8	0	5	0
d_2	5	19	7	16	0	0	32
d_3	15	0	0	4	9	0	17
d_4	22	3	12	0	5	15	0
d_5	0	7	0	9	2	4	12

17. Compute TF, IDF and TF-IDF for t3 in d4 for follow ing data.

document/term	t_1	t_2	t_3	t_4	t_5	t_6	t_7
d_1	0	4	10	8	0	5	0
d_2	5	19	7	16	0	0	32
d_3	15	0	0	4	9	0	17
d_4	22	3	12	0	5	15	0
d_5	0	7	0	9	2	4	12