	Unit 1
1	What is data mining and data warehouse?
2	Compare data base processing Vs. data mining processing.
3	Explain applications of data mining in detail.
4	Explain all data mining models and tasks.
5	What is KDD? Explain with diagram.
6	Write a short note on visualization.
7	Discuss the issues in Data mining.
8	What is fuzzy logic? Explain in brief with example.
9	Define following terms: 1. Information retrieval 2. Precision 3. Recall 4. Similarity 5. Granularities 6. Facts 7. Roll ups 8. Drill down
10	Define cube and explain with example.
11	Write a short note on star schema.
12	Explain characteristics of data warehouse.
13	Discuss the ways to improve the performance of data warehouse applications.
14	Write a short note on OLAP operations.
15	Write a short note on point estimation.
16	Compute mean, variance and standard deviation for (1,3,4,6,5).
17	Define 1. Mean 2. Median 3. Mode 4. Variance 5. Standard deviation (Sample/Population) 6. Bias 7. MSE
18	Compute mean, median and mode for (15, 10, 18, 20, 28, 32).
19	What is Jackknife estimate technique?
20	Find out Jackknife estimate for variance X={1, 5, 6} Mean X= {5, 6, 6}.
21	Estimate P that maximizes the likelihood that the given sequence of heads and tails would occur for {H, H, H, T, T} Note: Assume coin with H and T equally likely.

22	Estimate the missing data and continues until convergence using Expectation Maximization {1, 5, 10, 4,*, *}. (Guess μ^0=3)
23	Prove that X11 belongs to class h2 using Bayes theorem.

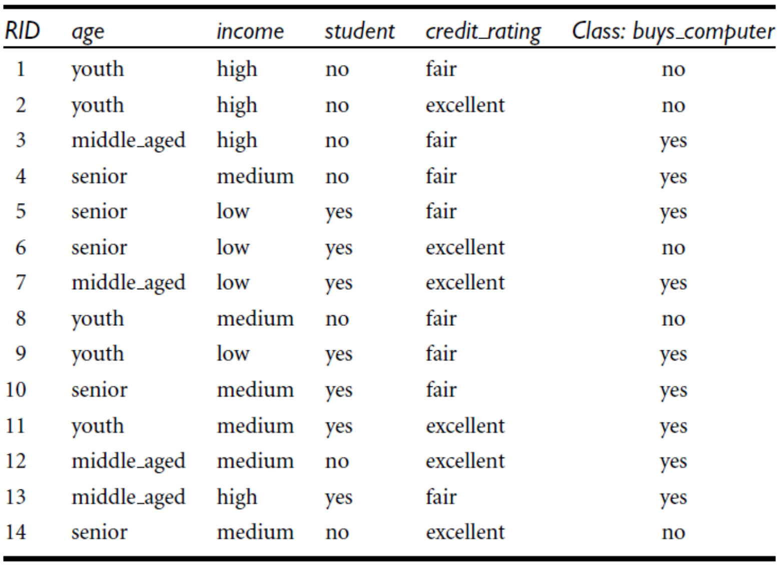
ID	Income	Credit	Class	x_i
1	4	Excellent	h_1	x_4
2	3	Good	h_1	x_7
3	2	Excellent	h_1	x_2
4	3	Good	h_1	x_7
5	4	Good	h_1	x_8
6	2	Excellent	h_1	x_2
7	3	Bad	h_2	x_{11}
8	2	Bad	h_2	x_{10}
9	3	Bad	h_3	x_{11}
10	1	Bad	h_4	x_9

24	Write a short note on Hypothesis testing.
25	Find Chi square statistics for Observed value = {51, 95, 67, 78, 88} Expected value=76
26	Write a short note on linear regression.
27	Write a short note on non-linear regression.
28	Explain correlation in detail.
29	Find correlation between Ice cream sales Vs temperature

Temperature	Ice Cream Sales (in
14.2	215
16.4	325
11.9	185
15.2	332
18.5	406
22.1	522

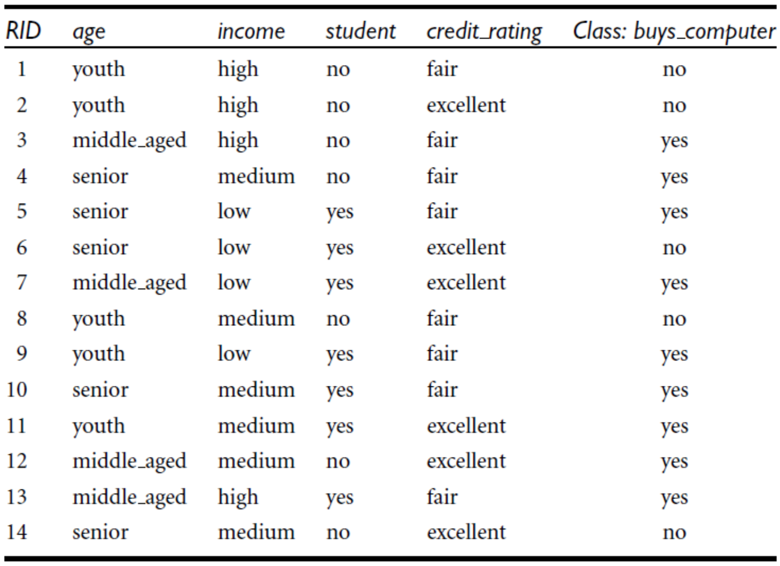
30	Write a short on similarity measures.
	Unit 2
31	Explain the need of data pre-processing.
32	List and explain major tasks in data processing.
33	Explain terms Quartile and Inter-Quartile range.
34	What are Box plot and Quantile plot?
35	What is histogram and scatter plot?
36	Write a short note on data cleaning tasks.
37	Explain Binning with example.
38	Explain Data aggregation, generalization and smoothing.
39	Write a short note on data transformation.
40	Write a short note on data normalization.
41	Define 1. Association rule

	2. Support 3. Confidence	
w) 42	Explain apriori algorithm with example.	
42a	Write a short note on Association rule mining.	
	Unit 3	
43	What is classification? Discuss the issues.	
44	What is prediction? Discuss the issues.	
45	Write a short note on decision tree.	
46	Write a short note on Bayesian classifier.	
47	Write a short note on Rule based classifier.	
48	Write a short note on Neural network classifier.	
49	Write a short note on Support Vector Machine.	
50	Define coverage and accuracy in rule based classifier.	
51	Explain triggering and firing of rules.	
52	Explain rule based and class based ordering.	
53	Discuss “The accuracy on its own is not a reliable estimate of rule	
54	Consider a training set that contains 100 positive examples and 400 negative examples for each of the following candidate rule. R1 : A + (covers 4 positive and one negative examples) R2 : B + (covers 30 positive and 10 negative examples) R3 : C +	
55	Consider a training set that contains 100 positive examples and 400 negative examples for each of the following candidate rule. R1 : A + (covers 4 positive and one negative examples) R2 : B + (covers 30 positive and 10 negative examples) R3 : C +	
56	Consider a training set that contains 100 positive examples and 400 negative examples for each of the following candidate rule. R1 : A + (covers 4 positive and one negative examples) R2 : B + (covers 30 positive and 10 negative examples) R3 : C +	
57	Write the difference between classification and clustering.	
58	Explain supervised and unsupervised learning.	
59	Explain term pruning and overfitting.	
60	Find information gain for “income” in following data	



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

61	Write a short note on Gini index.	
62	Classify the following tuple using Naïve Bayesian classifier. X=(age=youth, income=low , student=yes, credit_rating=fair) using following training data.	



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

63 Find out the population for the year 2013 using linear regression.

2005	2006	2007	2008	2009	2013
12	19	28	35	45	?

64 Write a short note on confusion matrix.

65 Classify X_1=4, X_2=7 using K-nearest neighbour (assume k=3).

X	Y	Class
7	7	B
7	4	B
3	4	G
1	4	G

Unit 4

66 List all the requirements of clustering Data mining.

67 Write a short note on type of data in clustering analysis.

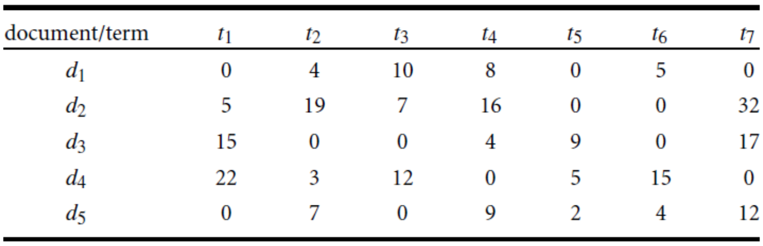
68 Compute Euclidean and Manhattan distance for X_1 (1, 2) and X_2 (3,6).

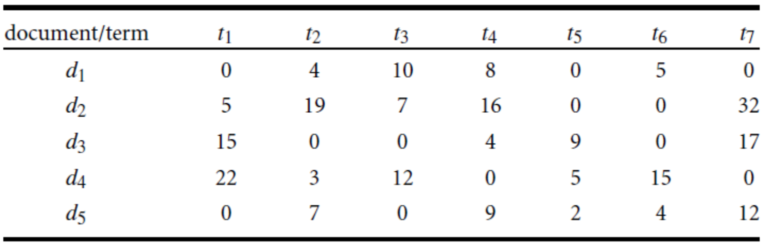
68 a Compute Euclidean and Manhattan distance for X_1 (1, 2) and X_2 (4,6).

69 Compute

1. Similarity between A and B
2. Similarity between C and B
3. Similarity between A and C and comment on the most similar tuples.

Name	Gender	F	C	T1	T2	T3
A	F	Y	N	P	P	N
B	M	Y	N	Y	N	P
C	F	Y	P	P	N	N

<table>
<tr><td>70</td><td>Write a short note on K-means clustering.</td><td></td></tr>
<tr><td>71</td><td>Write a short note on K-medoids clustering.</td><td></td></tr>
<tr><td>72</td><td>Write a short note on partitioning approach.</td><td></td></tr>
<tr><td>73</td><td>Write a short note on Hierarchical approach.</td><td></td></tr>
<tr><td>74</td><td>Write a short note on DBSCAN.</td><td></td></tr>
<tr><td>75</td><td>List and discuss major clustering approaches.</td><td></td></tr>
<tr><td>76</td><td>Write a short note on ROCK.</td><td></td></tr>
<tr><td>77</td><td>Explain agglomeration and divisive approach.</td><td></td></tr>
<tr><td>78</td><td>Apply hierarchical clustering using single linkage to following data.
A (1,1), B(1.5,1.5), C(3,4), D(4,4), E(3,3.5)</td><td></td></tr>
<tr><td>79</td><td>What are outliers? How to find out? Write the applications.</td><td></td></tr>
<tr><td></td><td><b>Unit5</b></td><td></td></tr>
<tr><td>80</td><td>What is graph mining and social network?</td><td></td></tr>
<tr><td>81</td><td>What are multimedia and spatial databases?</td><td></td></tr>
<tr><td>82</td><td>Explain set and listed valued attribute with example.</td><td></td></tr>
<tr><td>83</td><td>Explain set and complex structure valued attribute.</td><td></td></tr>
<tr><td>84</td><td>What is spatial aggregation and approximation? Explain with example.</td><td></td></tr>
<tr><td>85</td><td>Define plan, plan database and plan mining.</td><td></td></tr>
<tr><td>86</td><td>Explain the types of dimensions in spatial data cube.</td><td></td></tr>
<tr><td>87</td><td>Explain measures in spatial data cube.</td><td></td></tr>
<tr><td>88</td><td>Discuss approaches for similarity based retrieval in image database.</td><td></td></tr>
<tr><td>89</td><td>Write a short note on mining association in multimedia data.</td><td></td></tr>
<tr><td>90</td><td>Write a short note on text mining.</td><td></td></tr>
<tr><td>91</td><td>Define
1. Term frequency
2. Term frequency matrix
3. Relative term frequency
4. Inverse document frequency</td><td></td></tr>
<tr><td>92</td><td>Compute TF, IDF and TF-IDF for t2 in d2 for following data.

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

<table>
<tr><td>92 (a)</td><td>Compute TF, IDF and TF-IDF for t3 in d4 for following data.
<table><tr><th>document/term</th><th>t_1</th><th>t_2</th><th>t_3</th><th>t_4</th><th>t_5</th><th>t_6</th><th>t_7</th></tr><tr><td>d_1</td><td>0</td><td>4</td><td>10</td><td>8</td><td>0</td><td>5</td><td>0</td></tr><tr><td>d_2</td><td>5</td><td>19</td><td>7</td><td>16</td><td>0</td><td>0</td><td>32</td></tr><tr><td>d_3</td><td>15</td><td>0</td><td>0</td><td>4</td><td>9</td><td>0</td><td>17</td></tr><tr><td>d_4</td><td>22</td><td>3</td><td>12</td><td>0</td><td>5</td><td>15</td><td>0</td></tr><tr><td>d_5</td><td>0</td><td>7</td><td>0</td><td>9</td><td>2</td><td>4</td><td>12</td></tr></table></td><td></td></tr>
<tr><td>93</td><td>Discuss “Web poses grate challenges for effective resources and knowledge discovery”</td><td></td></tr>
</table>