BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
I SEMESTER 2012-2013
CS C415/IS C415 – DATA MINING

Component: Mid Semester Test(closed book) **Date: 7th October, 2012**
Time: 90 Minutes **Weightage: 30%**

1. What are the problems associated with a rule set classifier if it is not mutually exclusive? How do you resolve these problems?
2. Sampling with replacement is used in bagging and boosting methods. A sample contains approximately 63% of the original data. Prove or disprove.
3. Compare the classifiers decision stump, decision tree and k-NN for bias and variance.

[2 each]

4. Pre-processing:
   a. Apply two methods of normalizations to the data 200, 300, 400, 600, 1000.
   b. Comment on the robustness of the Bayesian and rule based classifiers with respect to features.
   c. Discuss the type of the attributes of the TA evaluation dataset.
   d. Which similarity measures will you use for finding the similarity between the objects of TA evaluation dataset and why?
   e. Which attribute would you like to filter from the TA evaluation dataset and why?

[3+3+2.5+2.5+1]

5. Classification:
   a. Construct a decision stump for the data set given below.
   b. Use ensemble classifier to predict the class of the records in test dataset. Ensemble classifier consists of 3 base classifiers: Decision Stump (obtained in (a)), 3-NN and Naive Bayesian classifiers.

[2+10]

The following data consist of evaluations of teaching assistance's performance over two regular semesters. The scores were divided into 3 roughly equal-sized categories ("low", "medium", and "high") to form the class variable:

S No	English Speaker	Course category	Semester	Class Size	TA Evaluation	Data
1	1	3	1	19	high	Training
2	2	2	2	27	high	
3	2	3	2	46	medium	
4	2	2	2	31	medium	
5	2	2	2	24	medium	
6	2	1	2	37	low	
7	2	1	2	21	low	
8	1	3	1	25	high	
9	2	1	2	17	high	
10	1	1	2	32	low	
11	2	2	2	27	low	Testing
12	2	3	2	11	medium	

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
I SEMESTER 2012-2013 CS C415/IS C415 – DATA MINING
Comprehensive Exam

Part-A (Closed Book) Max Marks: **20** Date: 11th December 2012

Each Question caries 2 marks.

1. What are the factors that may affect the efficiency in frequent item set generation? How do you deal with these factors in the algorithms?
2. What are the two major steps in Apriori algorithm and Why are these necessary?
3. What is the best method for discritization of numerical attributes and why? Discritize the numerical attributes A and B of the dataset "2D" with this method.
4. Why do we need to eliminate instances in the sequential covering algorithm?
5. Suppose you are asked to analyze a classification problem. You have been given a set of classification rules generated by the decision tree. Further, you have also been given a choice of solving the same problem by rule based classifier. Will you analyze it with rule based classifier? Given that the set of rules generated by decision tree is mutually exclusive and exhaustive. Justify your answer?
6. Consider a credit card transaction dataset where each transaction is marked as legitimate or fraudulent. How do you evaluate a classifier for this dataset? Justify.
7. Prove or disprove that the k-means clustering algorithm will always produce a clustering.
8. Write steps to detect anomaly in a dataset using a data mining technique.
9. Give one example where k-means clustering produces only one cluster with any value of k=1 to 100.
10. Suppose you need to find frequent itemsets for the measure "Sum of Price" i.e. An itemset S will be considered frequent if Sum(S) ≤v . Prove or disprove that Apriori algorithm for finding frequent itemsets will work.

2D Dataset

#	A	B
1	0.61	17.52
2	15.26	8.19
3	1.06	126.22
4	3.02	9.74
5	0.26	21.22
6	0.18	5.33
7	0.46	123.44
8	1.28	23.85
9	0.46	6.33
10	1.14	13.22
11	0.24	137.52

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
I SEMESTER 2012-2013 CS C415/IS C415 – DATA MINING
Comprehensive Exam

Part-B (Open Book) Max Marks: **20** Date: 11th December 2012

Use Manhattan distance wherever needed.

1. Suppose we have market basket data consisting of 100 transactions and 20 items. If the support for item *a* is 25%, the support for item *b* is 90% and the support for itemset *{a, b}* is 20%. Let the support and confidence thresholds be 10% and 60%, respectively.
(a) Compute the confidence of the association rule *{a}* → *{b}*. Is the rule interesting according to the confidence measure?
(b) Compute the interest measure for the association pattern *{a, b}*. Describe the nature of the relationship between item *a* and item *b* in terms of the interest measure.
(c) What conclusions can you draw from the results of parts (a) and (b)?
(d) Prove that if the confidence of the rule *{a}* –→ *{b}* is less than the support of *{b}*, then:
   i. $c(\{\bar{a}\} \dashrightarrow \{b\}) > c(\{a\} \dashrightarrow \{b\})$,
   ii. $c(\{\bar{a}\} \dashrightarrow \{b\}) > s(\{b\})$,

[5]

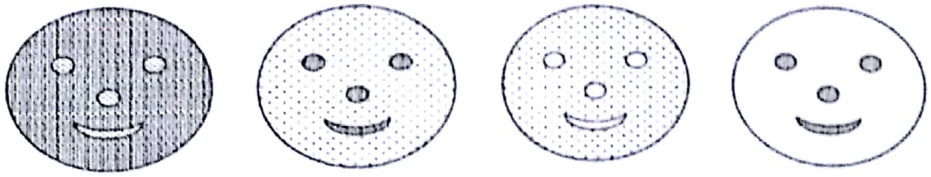
2. Consider the dataset given below:

A (.75, .75)	B (1.4, 1.3)	C (3, 1)	D (2.5, 3.5)	E (1.5, 3.75)
F (3, 4)	G (1, 2)	H (1.5, 2.5)	I (3, 2)	J (3.5, 3.75)

(a) Estimate the value of the parameter epsilon for dbscan with *Min_Pts*=3
(b) Find the dissimilarity matrix for the items A to F.
(c) Apply complete link algorithm on the data obtain in (b) and construct dendogram.
(d) Cut the dendogram appropriately to get optimal number of clusters using measures for cohesion and separation.

[2.5+1+2.5+2]

3. Consider the following four faces shown below. Darkness or number of dots represents density. Lines are used only to distinguish regions and do not represent points. For each figure, could you use single link, k-means, DBSCAN to find the patterns represented by the nose, eyes, and mouth? Explain.



(a) (b) (c) (d)

[4]

4. Consider a training set that contains 100 positive examples and 400 negative examples. For each of the following candidate rules,
*R*1: *A* –→ + (covers 4 positive and 1 negative examples),
*R*2: *B* –→ + (covers 30 positive and 10 negative examples),
*R*3: *C* –→ + (covers 100 positive and 90 negative examples)
Determine which is the best and worst candidate rule according to Rule accuracy and Foil's information gain.

[3]

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
I SEMESTER 2016-2017 CS F415 – DATA MINING

Mid Semester (closed book)

me: 90 Minutes | Weightage: 30% | Date: 7th Oct, 2016

se Gini index wherever applicable.

1. (a) What are the two major problems with one-vs-rest (1-*r*) approach for multi-class classification? How do you address these problems?
   (b) What is a class imbalance problem? Give an effective strategy to deal with it.
   (c) Give a situation where error correcting output coding method will be helpful.
   [2+2+2]

2. Compare the following classifiers on the basis of different characteristics:
   (a) Naïve Bayes or *k*-Nearest Neighbors
   (b) Decision tree and Rule based classifier
   [2+2]

3. (a) Discuss the type of the attributes in TA dataset.
   (b) Construct cost matrix for CL dataset to evaluate classification models.
   (c) Compare 11-NN and Naïve Bayes classifiers based on accuracy and cost matrix given in (b) for CL dataset.
   [2+2+8]

4. (a) Give all steps involved in an ensemble classifiers where base classifiers are constructed by manipulating the class labels and construct three base classifiers, decision stumps, for TA dataset using the same approach. Use equi-width binning to discretize the attribute "*class size*" into 2 bins.
   (b) Compare the accuracy of the ensemble classifier using base classifiers constructed in (a) with the decision stump constructed over the entire dataset.
   Note: Higher class level should be given to a node in case of a tie, e.g., if tie is between Low and Fair, class assigned will be Fair.
   [4+4]

Teaching Assistant (TA) Evaluation Dataset

The data consist of evaluations of teaching performance over regular and summer semesters of teaching assistant (TA) assignments at the Computer Science Department. The scores were divided into 4 roughly equal-sized categories ("low", "Fair", "Good" and "high") to form the class variable:

S. No	Native English speaker	Course	Summer /Regular	Class Size Range [1-50]	Evaluation	
1	2	1	R	21	Low	Training
2	2	1	R	48	Low	
3	2	1	R	33	Low	
4	1	3	S	20	Fair	
5	1	3	S	19	Fair	
6	1	2	R	15	Good	
7	2	2	S	37	Good	
8	2	2	R	10	Good	
9	2	3	S	17	High	
10	2	3	S	27	High	
11	2	2	R	15	Good	Testing
12	2	3	R	18	Fair	
13	1	1	R	12	Low	

				Training		Testing
S.No	Daily Exercise	Smoking	Retired	Cholesterol Level	Support Count	# of test records
1	Yes	Yes	Yes	High	1	3
2	Yes	Yes	No	High	42	
3	Yes	No	Yes	High	0	2
4	Yes	No	No	High	14	10
5	Yes	Yes	Yes	Low	3	
6	Yes	Yes	No	Low	8	
7	Yes	No	Yes	Low	6	
8	Yes	No	No	Low	16	
9	No	Yes	Yes	High	42	5
10	No	Yes	No	High	10	10
11	No	No	Yes	High	12	
12	No	No	No	High	6	
13	No	Yes	Yes	Low	12	
14	No	Yes	No	Low	1	
15	No	No	Yes	Low	24	10
16	No	No	No	Low	3	
Total					200	40

Name ID

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

I SEMESTER 2016-2017 CS F415 – DATA MINING Comprehensive Exam

Part-A (Closed Book) **Max Marks: 20** **Date: 12th Dec 2016**

Questions 1-20 carry 0.5 marks each. Questions 21-26 carry 1 mark each and question 27-28 carry 2 marks each. Use Manhattan distance wherever needed.

1. Data Mining is
   (a) searching a research paper on net
   (b) finding people with age above 25 and salary >40K drive sports Cars.
   (c) finding number of students who have got above 75% marks from a student database
   (d) studying that stocks of companies A and B perform similarly
2. Identify technique(s) to handle noisy data
   (a) Aggregation (b) Normalization
   (c) clustering (d) Association Relation
3. Which of the following comes under temporal data mining
   (a) summarization (b) time series analysis
   (c) Association Rule Mining (d) Sequence Mining
4. Identify the odd one
   (a) DBSCAN (b) Shared Nearest Neighbors
   (c) OPTICS (d) Chameleon
5. Identify correct statement(s) about Apriori and its variations
   (a) Apriori has better efficiency over Sampling and Partitioning for large databases for support>20%
   (b) Join step satisfies apriori property and anti monotone property
   (c) Items must be arranged in some order
   (d) No other measure can be used in place of support in Apriori algorithm
6. Identify correct statement(s)
   (a) High support => High confidence
   (b) High confidence => High support
   (c) All relations can be found using support and confidence measures
   (d) Both support and confidence are required to find frequent itemsets
7. Identify correct statement(s) about divisive hierarchical clustering algorithms
   (a) DIANA is a divisive hierarchical clustering algorithm
   (b) Divisive algorithms are computationally costlier than agglomerative clustering algorithms
   (c) Divisive algorithm starts with n clusters, n is the number of objects, then merges most similar clusters
   (d) CURE is a divisive hierarchical clustering method
8. Consider five points A, B, C, D and E, *Min_Pts* as 3 and ε as 0.5. A's distance to its 3rd nearest neighbor B is 0.35. Given B, C and D belongs to A's ε-neighborhood and E is density reachable to B. Identify correct statement(s) about these points
   (a) A's core distance is 0.5 (b) A is a core point
   (c) B is directly density reachable to A (d) Points A, B, C, and D belong to same DBSCAN clusters
9. Consider the data given in question 6. Identify the false statements from the following
   (a) A is density reachable to E (b) E is density reachable to A
   (c) B is directly density reachable to A (d) B's reachability distance with respect to A is 0.35
10. k-medoids takes 3 iterations to determine two clusters for 8 data points. The algorithm has to examine the cost of ___ no. of pairs
    (a) 48 (b) 36 (c) 16
    (d) 24 (e) none of these

11. Identify the linear classifier(s) from the following
    (a) k-nearest neighbor (b) Decision Tree
    (c) Rote Learner (d) Naïve Bayes
12. Identify eager learner(s) from the following
    (a) k-nearest neighbor (b) Decision Tree
    (c) Rote Learner (d) Naïve Bayes
13. Which of the clustering algorithm(s) find clusters of different sizes, shapes and densities
    (a) Chameleon (b) OPTICS
    (c) J P Algorithm (c) Shared Nearest Neighbors
14. Which of the clustering algorithm(s) find clusters of different sizes, shapes and densities in high dimensional data
    (a) Chameleon (b) OPTICS
    (c) J P Algorithm (c) Shared Nearest Neighbors
15. Which of the finding is the most suitable classifier for web document classification
    (a) k-nearest neighbor (b) Decision Tree
    (c) Rote Learner (d) Naïve Bayes
16. Identify true statement(s) about multi-level association rule mining
    (a) Multi-level association rules may lead to generation of redundant rules
    (b) uniform support is better than that of reduced support
    (c) level-by-level independent method of reduced support violets apriori property
    (d) Multi-level association rules can be obtained by Apriori algorithm
17. Identify internal measure(s) for cluster validation
    (a) information gain (b) purity
    (c) interest (d) silhouette coefficient
18. Multiple minimum support is
    (a) non-monotone (b) anti-monotone
    (c) can be converted into anti-monotone (d) cann't be converted into anti-monotone
19. Which of the following statement(s) are correct if itemset {A C D} is maximal
    (a) {A C} and {A E} are frequent (b) {A C E} is frequent
    (c) {A C} and {C D} are frequent (d) {A C D E} is frequent
20. Identify correct statement(s) about FP Growth
    (a) It is based on divide and conquer principle
    (b) It is faster than Apriori
    (c) It is not scalable because FP trees are constructed for all conditional base
    (d) It is an approximate solution
21. In the join step of the Apriori algorithm: the join, L_{k-1} X L_{k-1}, is performed, where members of L_{k-1} are joinable if their first (k-2) items are in common because

    _____

22. Name one association rule mining algorithm which can be used in incremental mining. Justify your answer.

    _____

23. What is D_3 if you apply DHP on the dataset 1. a,b,d, 2. abcde, 3. a,c,e, & 4. a,b,c,d

    _____

24. Consider A and B two decision tree classification models (having only binary splits) of Dataset D of 50 records. Model A and B have the following characteristics. Which model is better and why?

Model	Internal Nodes	External Nodes	#Misclassified Records
A	6	7	5
B	3	4	8

25. Give one example where k-means clustering produces only one cluster with any value of k=1 to 100.

26. Consider a credit card transaction dataset where each transaction is marked as legitimate or fraudulent. How do you evaluate a classifier for this dataset? Justify.

27. Find the appropriate range of ε for DBSCAN for the data (2,2), (3,3), (7,6), (6,7), (7,7), (1,2), (5,3), (9,9) using Min_Pts as 3.

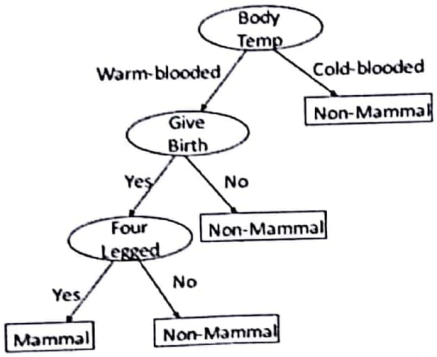
28. Draw a set of two or more clusters for which center-based, density-based and contiguity-based clustering algorithms give different number of clusters.

Part-B (Open Book) Max Marks: 20 Date: 12th Dec 2016

1. Consider

Name	Body Temp.	Gives Birth	Four legged	Hibernates	Class Mammal
Human	Warm-blooded	Yes	No	No	Yes
Pigeon	Warm-blooded	No	No	No	No
Elephant	Warm-blooded	Yes	Yes	No	Yes
Leopard Shark	Cold-blooded	Yes	No	No	No
Turtle	Cold-blooded	No	Yes	No	No
Penguin	Cold-blooded	No	No	No	No
Eel	Cold-blooded	No	No	No	No
Dolphin	Warm-blooded	Yes	No	No	Yes
Spiny anteater	Warm-blooded	No	Yes	Yes	Yes
Gila Monster	Cold-blooded	No	Yes	Yes	No

(a) Construct the confusion matrix for the above data using the following decision tree.



(b) Construct a dissimilarity matrix for the first five objects of the above data using four attributes (except name and class). Treat all four attributes as symmetric binary attributes.
(c) Find two clusters of first five objects using s-link and dissimilarity matrix obtained in (b).
(d) Evaluate this clustering using external measures purity and entropy.
(e) Find all the associations of attributes of the animal dataset with the class variable and comment on the classifier given in (a).

[2+1+2+2+2]

2. What is the main challenge in finding multi-level association rules? Write (in bullets) two ways two mine multilevel association rules. [4]

3. Show that bagging is a special case of AdaBoost. [2]

4. In Classification, the number of classes is fixed. But, in many real life problems, the number of classes associated with a problem may change with time (increase or decrease). Change in number of classes may be due to adding of new classes, removal of existing classes, merging of two classes, and splitting (n-way) of existing classes. Suggest a methodology to handle dynamic class classification problem.

Also write an application for the given case. [5]

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI
Second Semester 2016-17
COMPREHENSIVE EXAMINATION
PART A (OPEN BOOK)
CS F415 Data Mining

Date : 06-05-2017(FN) **Max.Duration : 90 minutes** **Max.Marks : 20 (20%)**

Instructions: Write precise and to the point answers.

1. Is the curse of dimensionality more problematic for a decision tree classifier or a naive Bayes classifier? Why? Compare and contrast with proper explanation. [3]

2. A database has 5 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

Find all frequent itemsets using Apriori and FP-growth, respectively. Compare the efficiency of the two mining processes. [4+4+1=9]

3. We want to cluster five documents into two clusters having inter-document distances shown below. Assume that at a certain step, documents C and D are selected as medoids. Which two documents will be selected as medoids in the next iteration using the PAM algorithm? You must explain the steps followed in arriving at your answer.

Document	**A**	**B**	**C**	**D**	**E**
A	0	1	2	2	3
B	1	0	2	4	3
C	2	2	0	1	5
D	2	4	1	0	3
E	3	3	5	3	0

[4]

4. Compute the Silhouette for the following clustering that consists of 2 clusters:
{(0,0), (0.1), (2,3)}, {(3,3), (3,4)}; use Manhattan distance for distance computations. Compute each point's silhouette; interpret the results (what do they say about the clustering of the 5 points; the overall clustering?)

[4]

****

IDNO		NAME	

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

Second Semester 2016-17

COMPREHENSIVE EXAMINATION

PART A (CLOSED BOOK)

CS F415 Data Mining

Date : 06-05-2017(FN) **Max.Duration : 90 minutes** **Max.Marks : 40 (20%)**

1. Which of the following is not a metric : [1]

(a) Euclidean distance

(b) Hamming distance

(c) $$\frac{S_{11}S_{00} - S_{10}S_{01}}{\sqrt{((S_{10} + S_{11}) + (S_{01} + S_{00}) + (S_{11} + S_{01}) + (S_{00} + S_{10}))}}$$

(d) $$\frac{S_{11}}{S_{11}+S_{10}+S_{01}}$$

ANS :

2. Which of the following agglomerative clustering algorithm is equivalent to MST clustering algorithm? [1]

A. Single Link
B. Complete Link
C. Group Average
D. Ward' s method

ANS :

3. Which of the following measure is equivalent to Geometric mean between the confidence of association rules? [1]

A. Interest Factor
B. Correlation coefficient
C. IS Measure
D. Kappa

ANS :

4. Give two other names to Anomaly/Outlier Detection. [1]

ANS :

5. Give one of the popular evaluation measure for hierarchical clustering. [1]

ANS :

6. Expand the following acronyms for the following clustering algorithms.

A. DBSCAN	
B. BIRCH	
C. CLIQUE	
D. ROCK	
E. OPOSSUM	

7. Give the definition of outliers with respect to the following

A. Probabilistic based Model	
B. Proximity based	
C. Density based	
D. Clustering based	
E. Hawkin's based	

8. Give two examples of measures which support the following properties

Inversion property	
Null Addition property	
Symmetry Under Variable Permutation	
Row/Column Scaling Invariance	
Antisymmetry Under Row/Column Permutation	

DNO		NAME	

. Agglomerative hierarchical clustering algorithms merge the pair of clusters that are the closest to each ther. Is this always a good approach? Why? Give an example. [4]

10 . Assume I run DBSCAN with MinPoints=5 and epsilon=0.2 for a dataset and I obtain 5 clusters and 7% of the objects in the dataset are classified as outliers. Now I run DBSCAN with MinPoints=10 and epsilon=0.2. How do expect the clustering results to change? [2]

11. Discuss the two problems masking and swamping with respect to Outliers. [2]

14. K-NN (k-nearest-neighbor) classifies are lazy classifiers. What does this mean? What is the disadvantage of using lazy classifiers? [2]

15. State the Space and Time Complexity for K-Means algorithm. [2]

16. What is the purpose of z-score? Assume a normalized attribute of an object has a z-score of −2; what does this s about the object's attribute value in relationship to the values of other objects? [2]

13. Assume we have an association rule

if Drink_Tea and Drink_Coffee then Smoke

that has a lift of 2. What does say about the relationship between smoking, and drinking coffee, and drinking tea? Moreover, the support of the above association rule is 1%. What does this mean? [2]

12. Assume the following dataset is given: (2,2), (4,4), (5,5), (6,6), (8,8),(9,9), (0,4), (4,0) . K-Means is used with k= to cluster the dataset. Moreover, Manhattan distance is used as the distance function (formula below) to compu distances between centroids and objects in the dataset. Moreover, K-Means's initial clusters C1, C2, C3, and C4 are follows:

C1: {(2,2), (4,4), (6,6)}
C2: {(0,4), (4,0)}
C3: {(5,5), (9,9)}
C4: {(8,8}}

Now K-means is run for a single iteration; what are the new clusters and what are their centroids? [4]