CS40003
Data Analytics

End-Autumn Semester Test
(Session 2016-2017)

Full Marks: 100   Time: 180 minutes

Instructions

- *This is a question-cum-answer booklet. No separate sheet is required for solving any problem and answering.*
- *There are two parts in the question paper. Answer to both parts.*
- *To give your answers, use the ANSWER SHEET given in the **Page 12-16** of the booklet. Don't give answer anywhere else. Put a CIRCLE on the option you have chosen as correct (for Part-A) and only write the answers (for Part-B). You are advised to record your all calculations in the rough space clearly showing question numbers.*

Part A

All questions in this part are of multiple choice type questions.

For a question, there may be one or more option(s) is(are) correct.

For question with more than one correct options, credit will be given on pro-rata basis.

No credit will be given, if wrong options(s) is(are) chosen.

There is NO NEGATIVE marking.

Each correct answer to a question carries 2 marks only.

Give answer on the ANSWER SHEET (Page 12) only.

1. Consider a regression problem, where we want to predict variable y from a single feature x given a training data. Consider two possible regression models as follows.

    $Model\ 1: y = \beta'_0 + \beta'_1 x$
    $Model\ 2: y = \beta'_0 + \beta'_1 x + \beta'_2 x^2$

    Which of the following statement is possibly the best answer?

    (a) Model 1 fits the training data better than Model 2.
    (b) Model 2 fits the training data better than Model 1.
    (c) Model 1 and model 2 both fit equally well the training data.
    (d) It is impossible to say which model fits better than the other.

2. Which of the following is an unsupervised techniques?

    (a) k-Means clustering.
    (b) k-NN classification technique.
    (c) Bayes' Naïve probabilistic prediction method.
    (d) Market basket analysis.

3. Following plots (that is, A and B) represent variation between regression variable (X) and response variable (Y).

Figure Q. 4.
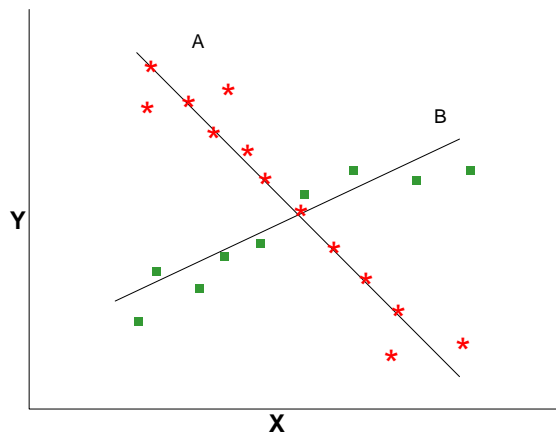
With reference to the above graph, which of the following statement is true?

(a) A is with positive and B is with negative correlation.
(b) A is with negative and B is with positive correlation.
(c) Both A and B are with high degrees of correlation.
(d) Both A and B are with low degrees of correlation.

4. Which of the following is not true in the case of Bayes' classification technique?

(a) Predicts class membership probabilities.
(b) Based on Bayes' theorem on probability.
(c) All classes are mutually exclusive and exhaustive.
(d) All attributes are independent given a class.

5. Naïve Bayesian classifier cannot be applied to training data with

(a) categorical attribute.
(b) continuous attribute.
(c) a null occurrence of some attributes' values for some classes.
(d) None of the above

6. The maximum value of entropy of a training set of size n with k number of class labels such that each tuple is defined with m number of attributes is

(a) $\log_k n$
(b) $\log_2 k$
(b) $\log_2 m$
(c) $\log_m n$

7. Which of the following splitting criteria is used in CART algorithm?

(a) Information gain.
(b) Gain ratio.
(c) Gini index.
(d) Weighted average entropy.

8. Which of the following statement is true about building a decision tree?

(a) ID3 always results a binary decision tree.
(b) CART is applicable to any type of attribute.

(c)   C4.5 is only applicable when record in the training set is defined in terms of categorical attributes.

(d)   CART results an *n*-ry decision tree always.

9.   Building an SVM for a given training data becomes

(a)   an equality constraint convex optimization problem.

(b)   an inequality constraint convex optimization problem.

(b)   a top down, divide and conquer, recursive problem.

(c)   a mapping problem.

10.   If L_P and L_D denote the dual form of Lagrangian while considering the training an SVM, then which of the following statement(s) is(are) correct?

(a) L_D involves the calculation of less number of parameters than L_P.

(b) L_P is maximization problem, whereas L_D is a minimization problem.

(c) L_P involves the calculation of $W.x$ whereas L_D involves the calculation of $x_i.x_j$ (all symbols are with their usual meanings).

(d) L_P is computationally faster than L_D.

11.   Mark the statements, which is(are) not necessarily correct, when we talk about Google's solution to handle Big data.

(a) HDFS is based on the concept of "Scale Out" architecture.

(b) HDFS is a file system designed for storing very large files reliably with streaming access patterns.

(c) HDFS runs on clusters of commodity hardware to process data locally.

(d) HDFS has a "Read once, Write often" model of data access.

12.   MapReduce is a distributed programming models meant for large clusters. Which of the following is(are) the important characteristic(s) of MapReduce programming?

(a) All data from the data servers are fetched into master node where parallel computing takes place.

(b) Two user defined functions namely *map()* and *reduce()* are transferred to all the data nodes.

(c) Job trackers and Task trackers are JVMs residing in Namenode and Datanode, respectively.

(d) MapReduce allows multiprogramming in an interactive fashion.

13.   Which of the following estimation strategies comes under the category of "Leave-one-out" cross-validation?

(a) Bootstrap method.

(b) Random subsampling.

(c) k-fold cross-validation.

(d) N-fold cross-validation.

14.   Which of the following parameter(s) is(are) essential in order to measure the "true accuracy"?

(a) Predictive accuracy.

(b) Confidence level.

(c) Size of the test data set.

(d) Size of the training data set

15. Following measurements are known while testing a binary classifier with classes $+$ and $-$

f_{++}, f_{+-}, f_{-+} and f_{--}, where f_{xy} denotes the number of instances that were with class "x" and classified as "y". Which of the following metrics represent "Recall" and "Precision"?

(a) $\dfrac{f_{++}}{f_{++}+f_{+-}}$

(b) $\dfrac{f_{++}}{f_{++}+f_{-+}}$

(c) $\dfrac{f_{-+}}{f_{-+}+f_{--}}$

(d) $\dfrac{f_{+-}}{f_{++}+f_{+-}}$

16. The distance d according to Minkowski matric between two objects x and y in n-dimensional space where x_i and y_i denote the values of ith attribute of the object x and y, respectively is

$$d(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^r \right)^{\frac{1}{r}}$$

The distance d becomes "City-block metric" when

(a) $r = -1$
(b) $r = 1$
(c) $r = 2$
(d) $r \in R, R$ is the set of all real numbers

17. We are to measure the similarity between two text documents. Which of the following metric(s) is(are) best suitable for the purpose?

(a) Jaccard coefficient.
(b) Eucledian distance.
(c) Cosine similarity.
(d) Set difference.

18. Identify the correct statement(s) in the following

(a) The best centroid for minimizing SSE (Sum of Square of Errors) of a cluster is the mean of the object in the clusters.
(b) The best centroid for minimizing SAE (Sum of Absolute Errors) of a cluster is a median of the objects in the cluster.
(c) SSE is used to measure the performance of k-Means clustering algorithm when L_1 norm is used to measure the proximity of two objects.
(d) SAE is used to measure the performance of k-Means clustering algorithm when L_2 norm is used to measure the proximity of two objects.

19. Which of the following statements is/are not correct?

(a) k-Means clustering algorithm does not work on categorical data.
(b) PAM algorithm does not require the value of k (the number of clusters) a priori.
(c) Agglomerative clustering techniques do not require the number of clusters to be specified.
(d) DIANA is a partition based, whereas AGNES is a hierarchical-based clustering

methods.

20. Given a database of transaction, a rule in "Market basket analysis" (MBA) is defined as $r: x \to y$ where x and y denote two sets of items in a transaction. Which of the following represents the "completeness" to measure the rule strength of r (symbols bear their usual meaning)?

(a) $\dfrac{\sigma(x \to y)}{\sigma(y)}$

(b) $\dfrac{\sigma(x \to y)}{\sigma(x)}$

(c) $\dfrac{\sigma(x \to y)}{|D|}$

(d) $\dfrac{\sigma(x \to y)}{|x|}$

Part B

This part includes 20 concept/problem solving type questions.

You should solve each question in the space provided in the booklet. Clearly mention the question number there.

Don't use any extra sheet for problem solving.

Write down your answers on the ANSWER SHEET **(Page 12-16).**

Each correct answer to a question carries 3 marks only.

Do not give answer elsewhere.

21. For a database of 2500 transactions, five association rules chosen at random are shown in Table Q. 21. Decide the rule(s), which is/are rejectable? Assume, *minsup* (μ) = 1% and *minconf*(τ) = 80%

Table Q. 21

Rule $X \to Y$	$\sigma(X \to Y)$	$\sigma(X)$	$\sigma(Y)$
r_1	700	720	800
r_2	140	150	650
r_3	1000	1000	2000
r_4	200	400	250
r_5	295	300	700

22. With reference to an arbitrary experiment, a set of clusters with different k values and *SSE* (Sum of Square of Errors) measure are obtained, which is shown in Table Q.22.

Table Q. 22

k	1	2	3	4	5	6	7	8
SSE	62.8	12.3	9.4	9.3	9.2	9.1	9.05	9.0

(a) For which value of k, you should choose the partition in this case?

(b) What is the expected value of SSE if $k = n$, n being the number of objects under clustering?

23. Write down the clustering objectives with the proximity measures
    (a) L_2 norm
    (b) L_1 norm
    (c) Cosine similarity

24. Consider the following set of records (see Table Q. 24), where each record is defined by two ordinal attributes $size$ ={S, M, L} and $quality$ = {EX, A, B, C} such that $S < M < L$ and $EX > A > B > C$.

Table Q. 24

Object	Size	Quality
A	S(0.0)	A(0.66)
B	L(1.0)	EX(1.0)
C	L(1.0)	C(0.0)
D	M(0.5)	B(0.32)

Write down the similarity matrix.

25. You are given two documents with frequency count of 10 words in each document as shown in the form of X and Y below.

$$X = [3, \quad 2, \quad 0, \quad 5, \quad 0, \quad 0, \quad 0, \quad 2, \quad 0, \quad 0]$$
$$Y = [1, \quad 0, \quad 0, \quad 0, \quad 0, \quad 0, \quad 0, \quad 1, \quad 0, \quad 2]$$

Calculate the similarity measure between X and Y. Also, mention the metric used.

26. A classifier is tested with a test set of size 100. The classifier predicts 80 test tuples correctly. Calculate the following.

    (a) Observed frequency
    (b) Standard error rate
    (c) True accuracy. Assume T_α with confidence level $\alpha = 95\%$ is 1.96.

27. The ROC curves of three classifiers C_1, C_2 and C_3 are shown in Figure Q.27.
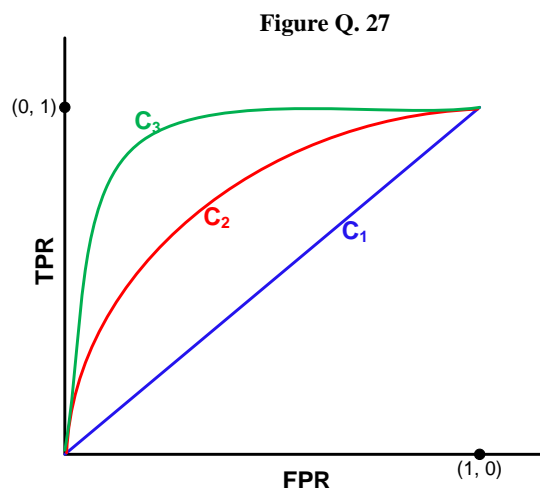
Figure Q. 27

With reference to the ROC curves in Figure Q. 27 answer the following.

    (a) Which classifier performs random guessing?
    (b) Which classifier is very close to the perfect classifier?
    (c) What is the TPR and FPR of the worst classifier?

28. Table Q.28 shows the confusion matrix of a classification problem with six classes labelled as C_1, C_2, C_3, C_4, C_5 and C_6.

Table Q. 28

Class	C_1	C_2	C_3	C_4	C_5	C_6
C_1	52	10	7	0	0	1
C_2	15	50	6	2	1	2
C_3	5	6	6	0	0	0
C_4	0	2	0	10	0	1
C_5	0	1	0	0	7	1
C_6	1	3	0	1	0	24

Calculate the mean error rate of the classification to classify a record belongs to class C_1.

29. Write the three differences between RDBMS and Hadoop ways of data processing. You should choose the parameters as mentioned in Table Q. 29.

Table Q. 29

Parameter	RDBMS	Hadoop
Storage Framework		
Access		
Update		

30. Write down the 3 advantages and 3 limitations of MapReduce programing paradigm.

31. Consider a training data set as shown in Table Q. 31.

Table Q. 31

Person	Gender	Height	Class
1	F	1.6	S
2	M	2.0	M
3	F	1.9	M
4	F	1.88	M
5	F	1.7	S
6	M	1.85	M
7	F	1.6	S
8	M	1.7	S
9	M	2.2	T
10	M	2.1	T
11	F	1.8	M
12	M	1.95	M
13	F	1.9	M
14	F	1.8	M
15	F	1.75	S

Obtain the decision trees with the splitting order (a) Gender-Height and (b) Height – Gender.

32. Write down the (a) learning and (b) classifier equations of SVM classifying linearly not separable data.

33. A set of training data with attributes A_1 and A_2 and two classes problem is given in Table Q. 33. The Lagrange multiplier is also known, which are also given in Table Q.33.

   a) Write down the model specification.
   b) Which vectors are support vectors according to the model?

Table Q. 33

A_1	A_2	Y	λ_i
0.38	0.47	+	65.52
0.49	0.61	-	65.52
0.92	0.41	-	0
0.74	0.89	-	0
0.18	0.58	+	0
0.41	0.35	+	0
0.93	0.81	-	0
0.21	0.10	+	0

34. For the following dataset (say *D*), (a) obtain the frequency table and hence (b) calculate the average entropy $E_{Age}(D)$ of the dataset.

Table Q. 34

Age	Eye sight	Astigmatic	Use Type	Class
1	1	1	1	3
1	1	1	2	2
1	1	2	1	3
1	1	2	2	1
1	2	1	1	3
1	2	1	2	2
1	2	2	1	3
1	2	2	2	1
2	1	1	1	3
2	1	1	2	2
2	1	2	1	3
2	1	2	2	1
2	2	1	1	3
2	2	1	2	2
2	2	2	1	3
2	2	2	2	3
3	1	1	1	3
3	1	1	2	3
3	1	2	1	3
3	1	2	2	1
3	2	1	1	3
3	2	1	2	2
3	2	2	1	3
3	2	2	2	3

35. Calculate the entropy of the dataset as shown in Table Q. 35.

Table Q.35

X	Y	Class
1	1	A
1	2	B
2	1	A
2	2	B
3	2	A
3	1	B
4	2	A
4	1	B

36. A set of records is shown in Table Q. 36.

Table Q. 36

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

With reference to the above-mentioned table, calculate the following.

a) $P(C_i)$ for each class C_i in the table
b) Given the following test tuple
$X = (age \leq 30, income = medium, student = yes, credit\ rating = fair)$, predict the class of X (You should show your all calculations in the rough space in this regard).

37. For the following Table Q. 37 calculate (a) Prior probability *P(Y=A)* and (b) Posterior probability $P(Y=A \mid x=x_2)$

Table Q. 37

X	Y
x_1	A
x_2	A
x_3	B
x_3	A
x_2	B
x_1	A
x_1	B
x_3	B
x_2	B
x_2	A

38. For different types of data, calculation of different correlation coefficients is known. Give the type of data on which the following correlation coefficients are applicable.

(a) Charles' Spearman's correlation coefficient (r)

(b) Kart Pearson's coefficient of correlation (r^*)

(c) Chi-square coefficient of correlation (χ^2)

39. A survey on *Gender* and *Hobby* of a population of size 1500 is recorded as under.

Table Q.39

Hobby	Gender		
		Male	Female
	Book	250	200
	Computer	50	1000

With reference to data in Table Q. 39 answer the following.

(a) Observed Frequency

(b) Expected frequency

(c) χ^2 correlation coefficient

40. Given a time series data set for T observations on a time series random variable $Y = [Y_1, Y_2, \ldots \ldots \ldots, Y_{T-1}, Y_T]:$. Write down the formula for calculating autocorrelation coefficient ρ_j. State all the symbols used in your expression precisely.

--- * ---

ANSWER SHEET
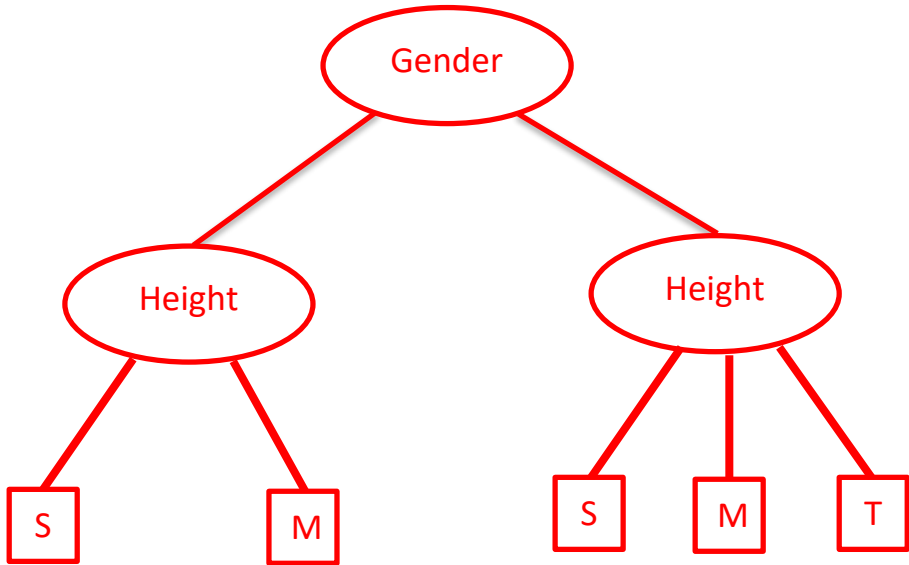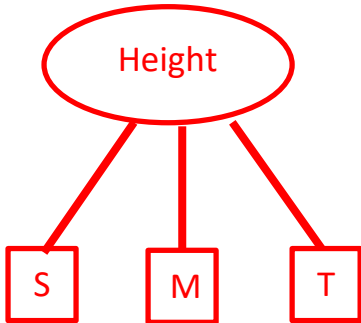
Part A

1.	A	B	C	(D)
2.	(A)	B	C	D
3.	A	(B)	C	(D)
4.	A	B	(C)	(D)
5.	A	B	C	(D)
6.	A	(B)	C	D
7.	A	B	(C)	D
8.	A	(B)	C	D
9.	A	(B)	C	D
10.	(A)	B	(C)	D
11.	A	B	C	(D)
12.	A	B	(C)	D
13.	A	B	C	(D)
14.	(A)	(B)	(C)	D
15.	(A)	(B)	C	D
16.	A	(B)	C	D
17.	A	B	(C)	D
18.	(A)	(B)	C	D
19.	A	(B)	C	(D)
20.	(A)	B	C	D

Part B

21.	Rejectable rule(s) : r_4 [Confidence is below the minconf (r) = 80%]

22.		
	a)	$k = 3$
	b)	$SSE = 0$

23.		
	a)	To minimize the SSE
	b)	To minimize the SAE
	c)	To maximize the TC

24.		A	B	C	D
	A	0.0	1.056	1.0	1.599
	B	1.056	0.0	1.0	0.599
	C	1.0		0.0	0.599
	D				0.0

25.		
	a)	*Similarity* = 0.31
	b)	*The metric used* = Cosine similarity

26.		
	a)	Observed frequency = 0.80
	b)	Standard error rate = 0.04
	c)	True accuracy = 0.7216

27.		
	a)	Random guessing classifier : C_1
	b)	Perfect classifier : C_3
	c)	Worst classifier : FPR =1, TPR = 0

28.	Mean error rate = 0.18 ≈ 18%

29.	Parameter	RDBMS	Hadoop
	Storage Framework	Scale Up	Scale Out
	Access	Interactive & Batch	Batch only Not Interactive
	Update	Read/Write many times	Write once, Read many times

30. Advantages
1. Allow parallel computing

2. No overhead on data transfer while program in execution

3. Open source

	Limitations 1. Batch processing not interactive 2. Only designed for a specific problem domain 3. Follows functional programming paradigm, which is not easy for the programmers
31.	(a) Decision tree according to splitting order **Gender-Height** (b) Decision tree according to splitting order **Height-Gender**

32.	a)	Learning an SVM:
		Maximize $\sum \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j Y_i Y_j K(X_i \cdot X_j)$
		Subject to $\lambda_i \geq 0$, $\sum_{i=1}^{n} \lambda_i \cdot Y_i = 0$
	b)	Classifier: $\delta(\mathbf{x}) = \sum_{i=1}^{n} \lambda_i \cdot Y_i K(X_i \cdot X) + b$

| 33. | a) | Model : $-6.64 \, X_1 - 9.32 \, X_2 + 7.93 = 0$ |
| | b) | Support vectors: Objects with non-zero Lagrange multiplier values |

34. **a)** Frequency table

	Age=1	Age=2	Age=3	Row Sum
Class 1	2	1	1	4
Class 2	2	2	1	5
Class 3	4	5	6	15
Column Sum	8	8	8	24

N=24

Column Sums

b) Average Entropy $E_{Age}(D)$: **1.2867**

35. Entropy : 1

36.	a)	$P(buys\_computer = \text{"Yes"}) = \frac{9}{14} = 0.643$ $P(buys\_computer = \text{"No"}) = \frac{5}{14} = 0.357$
	b)	The tuple X belong to class, "buys_computer" = "Yes"

37.	a)	Prior probability : P(Y=A) = 0.5
	b)	Posterior probability : $P(Y=A \mid X = X_2) = 0.2$

38.	a)	Charles Spearman's correlation coefficient: r = Ordinal attribute
	b)	Karl Pearson's correlation coefficient: $r^* =$ Numerical attribute
	c)	χ^2 correlation coefficient : $\chi^2 =$ Categorical attribute

39.	a)	Observed frequency :	

Hobby		*Gender*	
		Male	Female
	Book	250	200
	Computer	50	1000

.

b) Expected frequency :

Hobby		*Gender*	
		Male	Female
	Book	90	360
	Computer	210	840

.

c) χ^2 correlation coefficient $= 507.93$

40.

Auto correlation coefficient : $P_j = \dfrac{Cov\ (Y_t, Y_{t-j})}{\sqrt{\sigma_{Y_t}\ \sigma_{Y_{t-j}}}}$