Department of Computer Science
Banaras Hindu University, Varanasi 221005.

Mid-term Examination September 2018
CS304 – Machine Learning

Date: 22-09-2018 Duration: 60 minutes Total Marks: 20

Answer ALL questions

1. What is machine learning? Explain how supervised learning different from unsupervised learning. (3)
2. What is decision theory? List out the steps involved in decision making. (3)
3. If E and F are independent, then show that, E and F^c are also independent. (3)
4. If the density function of X equals

$$f(x) = \begin{cases} ce^{-2x} & 0 < x < \infty \\ 0 & x \leq 0 \end{cases}$$

Find the value of c and $P\{X > 2\}$. (3)

5. Obtain the equations of the lines of regression from the following data and find an estimate of Y which should correspond on the average to X = 6.2 (4)

X:	1	2	3	4	5	6	7	8	9
Y:	9	8	10	12	11	13	14	16	15

6. In a partially destroyed laboratory record of an analysis of correlation data, the following results only are legible: Variance of X = 9.
Regression equations: $8X + 10Y + 66 = 0$ and $40X - 18Y = 214$.
What were (i) The mean values of X and Y.
(ii) The correlation coefficient between X and Y. and
(iii) The standard deviation of Y? (3)
7. Define the following distributions: (i) Binomial (ii) Normal and (iii) Uniform (3)

REFNO[illegible]

Roll No. 17m1DCMP07

M.C.A. V SEMESTER/M.Sc. III SEMESTER EXAMINATION 2018-19

COMPUTER SCIENCE/COMPUTER APPLICATIONS

CS - 304 : Machine Learning

Time : Three hours **Max. Marks : 70**

Note: Question 1 is compulsory, attempt any four questions from the remaining five questions.

1. a) Define machine learning. Why has machine learning important now a days? (2)
   b) What is cross validation? Where it is used in machine learning? (2)
   c) What is hidden variable? Why do we need to consider it? (2)
   d) What are the different ways in which data can be missing from a model? How missing data can be handled? (2)
   e) Define the terms *density-reachable* and *density-connected.* (2)
   f) List out the various requirements of good clustering algorithm. (2)
   g) What is the *Curse of Dimensionality*? (2)

2. a) Explain the steps involved in designing a learning machine with an example. (6)
   b) Explain supervised learning and unsupervised learning with suitable example(s). (4)
   c) What is confusion matrix? Explain its use in Machine Learning Classifier. (4)

3. a) The joint density function of X and Y is given by (4)

$$f(x,y) = \begin{cases} 2e^{-x}e^{-2y} & 0 < x < \infty, 0 < y < \infty \\ 0 & \text{otherwise} \end{cases}$$

   Compute (a) P{X > 1, Y < 1}; (b) P{X < Y }; and (c) P{X < a}.

   b) What is decision theory? Explain the various steps involved in decision making. (4)

   c) The Thompson Lumber Company must decide whether or not to expand its product line by manufacturing and marketing a new product, backyard storage sheds. The various actions and states of natures with corresponding rewards are given below: (6)

Actions	States of Nature	
	Favorable Market	Unfavorable Market
Large plant	$200,000	-$180,000
Small plant	$100,000	-$20,000
No plant	$0	$0

   (i) Suggest a decision to the company using *minimax* regret method.
   (ii) Suppose that the probability of a favorable market is 0.6 and probability of an unfavorable market 0.4. Which alternative would give the greatest expected monetary value (EMV)?
   (iii) Calculate the expected value of perfect information (EVPI).

4. a) Explain the maximum likelihood estimation method for parameter estimation. (4)
   b) Explain the k-nearest neighbor classification algorithm with its merits and demerits. (6)

P.T.O.

(4)

c) Calculate the coefficient of correlation from the following data :

X:	1	2	3	4	5	6	7	8	9
Y:	9	8	10	12	11	13	14	16	15

Also obtain the regression line Y on X and obtain an estimate of Y which should correspond on the average to X = 6.2. (5)

5. a) Find out the class label of the following tuple based on the given training data using naive bayes algorithm.
X = (Outlook = Sunny, Temperature = Hot, Humidity = Normal, Windy = False)

Outlook	Temperature	Humidity	Windy	Play golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

b) Compute the variance of the sum obtained when 10 independent rolls of a fair die are made. (4)

c) Explain Bayesian belief network and conditional independence with example. (5)

6. a) Cluster following points in three clusters. Take initially A1, B1, C1 as a center points. Use K –means algorithms to show only the three cluster centers after the final round of execution. A1(2,10) A2(2,5) A3(8,4) B1(5,8) B2(7,5) B3(6,4) C1(1,2) C2(4,9) (use Euclidean distance). (5)

b) Explain the k-medoids algorithm. How does it differ from k-means algorithm? (5)

c) Explain expectation maximization (EM) algorithm for clustering. (4)

REFNO022407

Roll Number: 17419MCA008

MCA SEMESTER V / M.Sc SEMESTER III EXAMINATION 2019-20
Computer Science/Computer Applications
Paper No.: CS304: Machine Learning

Time: 3 hours | Full Marks: 70

Note: Question 1 is compulsory, attempt any four questions from the remaining five questions.

a) Define machine learning. (2)
b) What is cross validation? Where it is used in machine learning? (2)
c) What is hidden variable? Why do we need to consider it? (2)
d) Define the terms *density-reachable* and *density-connected*. (2)
e) What is outlier? List out any four applications of outlier analysis. (2)
f) What is the *Curse of Dimensionality*? (2)
g) What is over fitting problem? Why we need to avoid it? (2)

a) Explain the steps involved in designing a learning machine with an example. (6)

b) You are stranded on a deserted island. Mushrooms of various types grow wildly all over the island, but no other food is anywhere to be found. Some of the mushrooms have been determined as poisonous and others as not (determined by your former companions' trial and error). You are the only one remaining on the island. You have the following data to consider. (8)

Example	IsHeavy	IsSmelly	IsSpotted	IsSmooth	IsPoisonous
A	0	0	0	0	0
B	0	0	1	0	0
C	1	1	0	1	0
D	1	0	0	1	1
E	0	1	1	0	1
F	0	0	1	1	1
G	0	0	0	1	1
H	1	1	0	0	1
U	1	1	1	1	?
V	0	1	0	1	?
W	1	1	0	0	?

You know whether or not mushrooms A through H are poisonous, but you don't know about U through W. For first couple of questions, consider mushrooms only A through H.

(i) What is the information gain (entropy) of IsPoisonous?
(ii) What attribute should you choose as the root of the decision tree? Hint: You can figure this out by looking at the data without explicitly computing the information gain of all four attributes.
(iii) What is the information gain of the attribute you chose in the previous question?
(iv) Build a decision tree to classify mushrooms as poisonous or not.
(v) Classify mushrooms U, V and W using this decision tree as poisonous or not poisonous.
(vi) If the mushrooms of A through H that you know are not poisonous suddenly become scarce, should you consider trying U, V and W? Which one(s) and why? Or if none of them, then why not?

3. a) The joint density function of X and Y is given by (4)

$$f(x,y) = \begin{cases} 2e^{-x}e^{-2y} & 0 < x < \infty, 0 < y < \infty \\ 0 & \text{otherwise} \end{cases}$$

Compute (a) P{X > 1, Y < 1}, (b) P{X < Y }, and (c) P{X < a}.

b) Write short notes on logistic regression. (4)

c) What is decision theory? Explain the steps involved in decision making with an example. (6)

4. a) Explain the maximum likelihood estimation method for parameter estimation. (5)

b) Define the following distributions: (i) Binomial (ii) Normal and (iii) Uniform (3)

c) In a partially destroyed laboratory record of an analysis of correlation data, the following results only are legible: (6)

Variance of X = 9.

Regression equations: 8X - 10Y + 66 = 0 and 40X – 18Y = 214.

What were (i) The mean values of X and Y.

(ii) The correlation coefficient between X and Y. and

(iii) The standard deviation of Y?

5. a) Find out the class label of the following tuple based on the given training data using naive bayes algorithm. (8)

X = (Outlook = Sunny, Temperature = Hot, Humidity = Normal, Windy = False)

Outlook	Temperature	Humidity	Windy	Play golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

b) Explain the k-nearest neighbor classification algorithm with its merits and demerits. (6)

6. a) What are the requirements of good clustering algorithm? Discuss the differences and working of K-means and Hierarchal clustering. (9)

b) Explain expectation maximization (EM) algorithm for clustering. (5)

UCI Dataset Probabilistic

Yaha pe Calculus

.csv one file nhi mil rhi

OBE

31755

MCA-V Semester/M.Sc.-III Semester Examinations 2020-2021

Computer Science

Paper: CS-304

(Machine Learning)

Time: Four hours] [**Full Marks**: 70

Instructions

1. **The Question Paper contains 08 questions out of which you are required to answer any 04 questions. The question paper is of 70 marks with each question carrying 17.5 marks.**

   प्रश्नपत्र में आठ प्रश्न पूँछे गये हैं जिनमें से 4 प्रश्नों का उत्तर देना है। प्रश्नपत्र 70 अंकों का है, जिसमें प्रत्येक प्रश्न 17.5 अंक का है।

2. The total duration of the examination will be **4 hours** (Four hours), which includes the time for downloading the question paper from the Portal, writing the answers by hand and uploading the hand-written answer sheets on the portal.

   परीक्षा का कुल समय 4 घंटे का है जिसमें प्रश्नपत्र को पोर्टल से डाउनलोड करना, हस्तलिखित प्रश्नों का उत्तर पोर्टल पर अपलोड करना है।

3. For the students with benchmark disability as per Persons with Disability Act, the total duration of examination shall be **6 hours** (six hours) to complete the examination process, which includes the time for downloading the question paper from the Portal, writing the answers by hand and uploading the hand-written answer sheets on the portal .

   दिव्यांग छात्रों के लिये परीक्षा का समय 6 घंटे निर्धारित है जिसमें प्रश्नपत्र को पोर्टल से डाउनलोड करना एवं हस्तलिखित उत्तर को पोर्टल पर अपलोड करना है।

4. Answers should be hand-written on a plain white A4 size paper using black or blue pen. Each question can be answered in upto 350 words on 3 (Three) plain A4 size paper (only one side is to be used).
   हस्तलिखित प्रश्नों का उत्तर एक सादे सफेद A4 साइज के पन्ने पर काले अथवा नीले कलम से लिखा होना चाहिये। प्रत्येक प्रश्न का उत्तर 350 शब्दों अथवा A4 साइज के तीन पृष्ठों का होना चाहिये। प्रश्नों का उत्तर कापी के केवल एक पृष्ठ पर ही लिखना है।

5. Answers to each question should start from a fresh page. All pages are required to be numbered. You should write your Course Name, Semester, Examination Roll Number, Paper Code, Paper title, Date and Time of Examination on the first sheet used for answers.
   प्रत्येक प्रश्न का उत्तर नये पृष्ठ से शुरू करना है। सभी पृष्ठों को पृष्ठांकित करना है। छात्र को प्रथम पृष्ठ पर प्रश्नपत्र का विषय, सेमेस्टर, परीक्षा अनुक्रमांक, प्रश्नपत्र कोड, प्रश्नपत्र का शीर्षक, दिनांक एवं समय लिखना है।

Questions

1. a) Define machine learning. Why is machine learning important? Discuss the areas/disciplines that are influencing the machine learning. (9)

   b) Describe in detail all the steps involved in designing a learning system with an example. (8.5)

2. a) The distribution function of the random variable X is given (5)

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{x}{2} & 0 \le x < 1 \\ \frac{2}{3} & 1 \le x < 2 \\ \frac{11}{12} & 2 \le x < 3 \\ 1 & 3 \le x \end{cases}$$

   (i) Plot this distribution function. (ii) What is P{X > 1 2}?
   (iii) What is P{2 < X ≤ 4}? (iv) What is P{X < 3}?
   (v) What is P{X = 1}?

b) What is meant by k-fold cross validation. Given a data set with 1200 instances, how k- fold cross validation is done with k=1200. (6.5)

c) Differentiate between supervised and unsupervised training. Explain with suitable examples. (6)

3. a) Company ABC would like to market a new product X. The manager is trying to decide whether to produce the product X in large quantities (A1), moderate quantities (A2) or small quantities (A3). The manager does not know the demand for his product, but asserts that three events could occur: strong demand (S1), moderate demand (S2) or weak demand (S3). The profit, in thousands of rupees, with regard to marketing the product X is given in the following payoff table. (10)

Actions	States of Nature		
	S1	S2	S3
A1	44	27	10
A2	38	33	16
A3	29	25	20

(i) Suggest decisions to the company using Maximax, Maximin, Equal likelihood, Criterion of Realism and Minimax regret methods.

(ii) Suppose that the probability of a S1, S2 and S3 are 0.25, 0.4 and 0.35 respectively. Which alternative would give the greatest expected monetary value (EMV)?

(iii) Calculate the expected value of perfect information (EVPI).

b) The joint density function of X and Y is given by (7.5)

$$f(x,y) = \begin{cases} 2e^{-x}e^{-2y} & 0 < x < \infty, 0 < y < \infty \\ 0 & \text{otherwise} \end{cases}$$

Compute (a) P{X > 1, Y < 1}; (b) P{Y < X}; and (c) P{X < a, Y<b}

4. a) Explain maximum likelihood estimation method for parameter estimation. (5)

b) Explain the k-nearest neighbor classification algorithm with its merits and demerits. (6)

c) The following table shows the midterm and final exam grades obtained for students in a machine learning course. Use method of least squares using regression to predict the final exam grade of a student who received 97 on the midterm exam. (6.5)

Midterm Exam (X)	Final Exam (Y)
72	83
50	63
81	77
74	78
95	90
86	75
59	49
83	79
65	77
33	52
88	74
81	91

5. a) Explain naïve bayes classifier. Find out the class label of the following tuple based on the given training data using naïve bayes algorithm. (10)

X = (age <=30, Income = medium, Student = yes, Credit_rating = Fair)

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	high	yes	fair	yes
<=30	medium	yes	excellent	no
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no

b) Suppose the random variable X has distribution function (2.5)

$$F(x) = \begin{cases} 0 & x \leq 0 \\ 1 - \exp\{-x^2\} & x > 0 \end{cases}$$

What is the probability that X exceeds 1?

c) Explain classification using Bayesian Belief Network with an example. (5)

6. a) Explain working of k-means clustering algorithm. Suppose that the data mining task is to cluster the following seven points (with (x,y) representing location) into two clusters A1(1,1), A2(1.5,2), A3(3,4), A4(5,7), A5(3.5,5), A6(4.5,5), A7(3.5,4.5). The distance function is Euclidean distance. Suppose initially we assign A1,A5 as the centre for each cluster respectively. Using the K-means algorithm to find the three clusters and their centres after two round of execution. (10)

b) Explain the concept of Expectation Maximization(EM) Algorithm. How it can be used for clustering? (7.5)

7. a) Define the terms *Directly Density-Reachable, Density-Reachable and Density-Connected*. Explain DBSCAN algorithm for density based clustering. List out its advantages compared to K-means. (12)

b) What is outlier? Discuss the applications of outlier analysis. (5.5)

8. a) Use the k-means clustering algorithm and Euclidean distance to cluster the following eight 8 examples into three clusters: (10)
A1= (2, 10), A2= (2, 5), A3= (8, 4), A4= (5, 8), A5= (7, 5), A6= (6, 4), A7= 1(1, 2), A8= (4, 9). Find the new centroid at every new point entry into the cluster group. Assume initial cluster centers as A1, A4 and A7.

b) Discuss with examples some useful applications of machine learning. (7.5)
