Roll No: ----------

UPES

UNIVERSITY OF PETROLEUM AND ENERGY STUDIES

End Semester Examination, December 2017

Program: B.Tech CSE +BAO | **Semester – V**
Subject (Course): Data Mining & Prediction Modeling | **Max. Marks : 100**
Course Code : CSIB 338 | **Duration : 3 Hrs**
No. of page/s: 2

Section A
(All questions are compulsory and each carry 5 marks)

1. Differentiate between: (2.5x4=10)
   a) Eager Learners and Lazy Learners
   b) Text Mining and Web Mining
   c) Classification and Clustering
   d) Nominal and Ordinal Attributes
2. Define a pattern in data mining. Discuss application of Pattern Mining. (5)
3. What do you mean by "Data Migration Tool"? Discuss any one tool of such type. (5)

Section B
(All questions are compulsory and carry 10 marks)

4. How do you tackle noisy and inconsistent data? Briefly explain the steps of Data Preprocessing. (10)
5. Explain the basis of Model Evaluation and selection. Suppose there are two models M1 and M2.
   For M1: TP=6954, FN=46, FP=412 and TN=2588
   For M2: TP=6800, FN=134, FP=566 and TN=2500
   Among M1 and M2 which one is more preferable model? (10)
6. Write an algorithm for k-nearest-neighbor classification given k, the nearest number of neighbors and n, the number of attributes describing each tuples. Classify the person of age=48 for loan amount= $142,000. (10)

Age	Loan	Default	House Price Index	Distance
25	$40,000	N	135	102000
35	$60,000	N	256	82000
45	$80,000	N	231	62000
20	$20,000	N	267	122000
35	$120,000	N	139	22000
52	$18,000	N	150	124000
23	$95,000	Y	127	47000
40	$62,000	Y	216	80000
60	$100,000	Y	139	42000
48	$220,000	Y	250	78000
33	$150,000	Y	264	8000
48	$142,000	?		

7. Why is tree pruning useful in decision tree induction? Briefly outline the major steps of decision tree classification. (10)

OR

Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33,33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. (5+3+2)

(a) Use smoothing by bin means to smooth the data, using a bin depth of 3.Illustrate your steps. Comment on the effect of this technique for the given data.

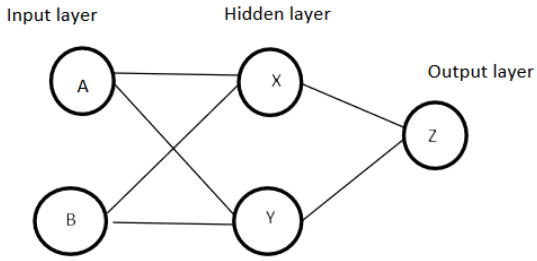
(b) How might you determine outliers in the data?

(c) What other methods are there for data smoothing?

Section C
(All questions are compulsory and each carry 20 marks)

8. "The support vector machine is highly accurate classification method", justify the statement. SVM classifier suffers from slow processing when training with a large data set, why? How we can solve this problem and make the SVM scalable. Categorize the types of hyperplane, if any. Explain with the concept of projection (orthonormal). (3+3+3+2+9)

9.

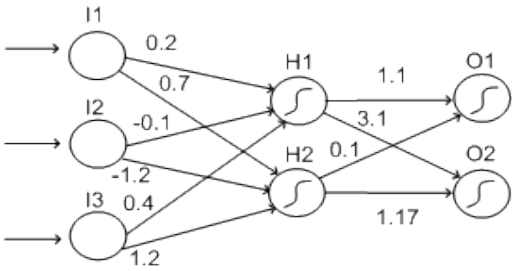


Input		Output
A	B	Z
0	0	0
0	1	1
1	0	1
1	1	1

Learning rate=0.35

Biases are $\theta_x = \theta_y = \theta_z = 0$. Neural Network of above diagram has two nodes (A,B) in the input layer, two nodes in the hidden layer (X,Y)and one node in the output layer (Z). The values given to weights are taken randomly and will be changed during back propagation iterations. Initial weights of the top input nodes taken at random are 0.4, 0.1 .Weights of bottom input node are 0.8 and 0.6. Weights of top hidden node is 0.3 and that of bottom hidden node is 0.9. (20)

OR



It is given that**:**

I1=10, I2=30, I3=10, learning rate(l)=0.1, t1(E)=target value for O1=1, t2(E)=target value for O2=0, θ = 0 for all nodes. Implement feed forward NN to minimize the error. (20)

Roll No: ----------

UPES

UNIVERSITY OF PETROLEUM AND ENERGY STUDIES

End Semester Examination, December 2017

Program: B.Tech CSE +BAO **Semester – V**
Subject (Course): Data Mining & Prediction Modeling **Max. Marks : 100**
Course Code : CSIB 338 **Duration : 3 Hrs**
No. of page/s: 2

Section A
(All questions are compulsory and each carry 5 marks)

1. Define the terms: (2x5=10)
   a) KDD b) Binary Attributes c) Data Cleaning d) Coverage e) Support
2. What is data mining? Briefly explain major issues of data mining. (2+3)
3. Why there is need of preprocessing the data? Write down the different phases of pre-processing of data. (2.5+2.5=5)

Section B
(All questions are compulsory and carry 10 marks)

4. Why is Naïve Baysian classification is called Naïve. Briefly explain the outline of major ideas of Baysian Algorithm. (2+8)
5. Briefly explain the SVM Classification algorithm with advantages and limitations. (10)
6.

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
human	warm	yes	no	no	mammals
python	cold	no	no	no	reptiles
salmon	cold	no	no	yes	fishes
whale	warm	yes	no	yes	mammals
frog	cold	no	no	sometimes	amphibians
komodo	cold	no	no	no	reptiles
bat	warm	yes	yes	no	mammals
pigeon	warm	no	yes	no	birds
cat	warm	yes	no	no	mammals
leopard shark	cold	yes	no	yes	fishes
turtle	cold	no	no	sometimes	reptiles
penguin	warm	no	no	sometimes	birds
porcupine	warm	yes	no	no	mammals
eel	cold	no	no	yes	fishes
salamander	cold	no	no	sometimes	amphibians
gila monster	cold	no	no	no	reptiles
platypus	warm	no	no	no	mammals
owl	warm	no	yes	no	birds
dolphin	warm	yes	no	yes	mammals
eagle	warm	no	yes	no	birds

Write down the rule based algorithm with parameters that can affect the results. Write down limitation of this algorithm, if any. Find the class label of given species: (6+2+2)

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
hawk	warm	no	yes	no	?
grizzly bear	warm	yes	no	no	?

7. Differentiate between Classification and Clustering. Briefly describe the major steps of Feed forward Back Propagation algorithm for classification. (3+7)

OR

Explain the basis of Model Evaluation and selection. Suppose there are two models M1 and M2.
For M1: TP=6954, FN=46, FP=412 and TN=2588
For M2: TP=6800, FN=134, FP=566 and TN=2500
Among M1 and M2 which one is more preferable model? (10)

Section C
(All questions are compulsory and each carry 20 marks)

8. Suppose that the data mining task is to cluster the following eight points (with (x, y) representing location into three clusters: (8+6+6)

A1(2, 10), A2(2, 5), A3(8, 4), B1(5, 8), B2(7, 5), B3(6, 4), C1(1, 2), C2(4, 9):

The distance function is Euclidean distance. Suppose initially we assign A1, B1, and C1 as the center of each cluster, respectively.
a) Write down k-means algorithm
b) Use k-means algorithm for the three cluster centers after the first round execution
c) Find the final three clusters

9. Differentiate between: (4x5=20)
a) Differentiate between Lazy and Eager Learners
b) Lift and Correlation
c) Web Mining and Text Mining
d) Active Learning and Transfer Learning

OR

Write short notes on:
a) Spatial Mining b) Machine Learning c) Outlier Detection d) Weka Tool