189

B.Sc. Engg. CSE 5th Semester (24)

14 March 2016 (Afternoon)

ISLAMIC UNIVERSITY OF TECHNOLOGY (IUT)
ORGANISATION OF ISLAMIC COOPERATION (OIC)
Department of Computer Science and Engineering (CSE)

MID SEMESTER EXAMINATION — WINTER SEMESTER, 2015-2016

DURATION: 1 Hour 30 Minutes — FULL MARKS: 75

CSE 4541: Machine Learning

Programmable calculators are not allowed. Do not write anything on the question paper.

There are **4 (four)** questions. Answer any **3 (three)** of them.

Figures in the right margin indicate marks.

---

1. State the Task (T), Performance Measure (P) and Training Experience (E) of the following learning tasks. Moreover, explain which type of learning method we need for each case. 12
   i. Document clustering in news reports (Group like sports, politics, lifestyle, foreign news etc.)
   ii. Spam filtering in email
   iii. Computer Chess program

   Suppose you need to train a robot for navigating an environment to search for a goal. What are the challenges you are going to face? What type of machine learning method will be the best solution to solve the problems? Explain. 5

   To determine a relationship between the number of fish and the number of species of fish in samples taken for a portion of the Great Barrier Reef, P. Sale and R. Dybdahl fit a linear least squares polynomial to the following collection of data (Table 1), which were collected in samples over a 2-year period. Let x be the number of fish in the sample and y be the number of species in the sample. Determine the Linear least square polynomial for these data. 8

Table 1: Table for Question 1(c)

x	13	15	16	21	22	23	25	29	30	31
y	11	10	11	12	12	13	13	12	14	16

2. Explain some scenarios where FIND-S algorithm fails, in spite of its guarantee to output the most specific hypothesis. 4

   Consider the trading agent trying to infer which books or articles the user reads based on keywords supplied in the article. Suppose the learning agent has the following data. (Table 2) 12

Table 2: Table for Question 2(b)

Article	Crime	Academic	Local	Music	Reads
1	true	false	false	true	true
2	true	false	false	false	true
3	false	true	false	false	false
4	false	false	true	false	false
5	true	true	false	false	true

   Run the Candidate-Elimination algorithm on the above training examples and generate the sequence of S and G boundaries.

   What are the conditions for CANDIDATE-ELIMINATION algorithm to converge to the correct hypothesis? If we want to minimize the version space of the CANDIDATE-ELIMINATION algorithm, then what is the optimal way? Explain. 2+3

   What is the effect of the value of K in K-NN classifier? How will you choose a good value of K? 2+2

3. a) Briefly describe the attribute selection measures of ID3, C4.5 and CART algorithms to build a decision tree.

b) Consider the following Table 3. These are some training and test examples obtained from observing a user deciding whether to read articles posted to a threaded discussion board depending on whether the author is known or not, whether the article started a new thread or was a follow-up, the length of the article, and whether it is read at home or at work. $e_1,\ldots, e_{18}$ are the training examples. The aim is to make a prediction for the user action on e_{19}, e_{20}. Use Bayesian Classifier to classify e_{19} and e_{20}.

Table 3: Table for Question 3(b)

Ex.	Author	Thread	Length	Where Read	User Action	Ex.	Author	Thread	Length	Where Read	User Action
e_1	known	new	long	home	skips	e_{11}	unknown	follow Up	short	home	skips
e_2	unknown	new	short	work	reads	e_{12}	known	new	long	work	skips
e_3	unknown	follow Up	long	work	skips	e_{13}	known	follow Up	short	home	reads
e_4	known	follow Up	long	home	skips	e_{14}	known	new	short	work	reads
e_5	known	new	short	home	reads	e_{15}	known	new	short	home	reads
e_6	known	follow Up	long	work	skips	e_{16}	known	follow Up	short	work	reads
e_7	unknown	follow Up	short	work	skips	e_{17}	known	new	short	home	reads
e_8	unknown	new	short	work	reads	e_{18}	unknown	new	short	work	reads
e_9	known	follow Up	long	home	skips	e_{19}	unknown	new	long	work	?
e_{10}	known	new	long	work	skips	e_{20}	unknown	follow Up	long	home	?

c) How will you avoid the problem of probability zero in Bayesian classifier?

4. a) Briefly describe how SPRINT algorithm will follow the data structure to build the decision tree for the following data set (Table 4) where prediction class is whether the person will cheat or not.

Table 4: Table for Question 4(a)

Taxable Income	125K	100K	70K	120K	95K	60K	220K	85K	75K	90K
Marital Status (S = Single, M = Married, D = Divorced)	S	M	S	M	D	M	D	S	M	S
Cheat	No	No	No	No	Yes	No	No	Yes	No	Yes

b) Given a 5 GB data set with 50 attributes (each containing 100 distinct values) and 512 MB of main memory in your laptop, outline an efficient method that constructs decision trees in such large data sets. Justify your answer by rough calculation of your main memory usage.

c) How does the AdaBoost technique boost its final classifier? Explain mathematically.

d) 'Random forests are comparable in accuracy to AdaBoost, yet are more robust to errors and outliers.' - Explain why.

318

B.Sc. Engg. CSE 5th Semester (2-4)

20 May 2016

ISLAMIC UNIVERSITY OF TECHNOLOGY (IUT)
ORGANISATION OF ISLAMIC COOPERATION (OIC)
Department of Computer Science and Engineering (CSE)

SEMESTER FINAL EXAMINATION — WINTER SEMESTER, 2015-2016

DURATION: 3 Hours — FULL MARKS: 150

CSE 4541: Machine Learning

Programmable calculators are not allowed. Do not write anything on the question paper.

There are **8 (eight)** questions. Answer any **6 (six)** of them.

Figures in the right margin indicate marks.

---

1) Machine Learning intertwines with other fields of Computer Science. Explain an example mentioning the parts of the Data Mining, Machine Learning, Pattern Recognition, and Artificial Intelligence. 3

2) Assume you are given the task to build a system that can distinguish junk email. What is in a junk e-mail that lets us know that it is junk? How can the computer detect junk through a syntactic analysis? What would you like the computer to do if it detects a junk e-mail—delete it automatically, move it to a different file, or just highlight it on the screen? Explain your answer. 8

3) Consider the example task of learning the target concept whether your friend enjoy a sport or not. Each hypothesis is a vector of six constraints, specifying the values of the six attributes *Sky*, *AirTemp*, *Humidity*, *Wind*, *Water*, and *Forecast*. Consider the following data: 6

X_1 = <*Sunny, Warm, Normal, Strong, Warm, Same*> +
X_2 = <*Sunny, Warm, High, Strong, Warm, Same*> -
X_3 = <*Overcast, Cold, High, Strong, Warm, Same*> +
X_4 = <*Sunny, Warm, High, Strong, Cool, Change*> +

Apply FIND-S algorithm to generate the target hypothesis.

4) Define the following terms: 4×2

i. Cross-fold validation ii. Bootstrap iii. Boosting iv. Random Forest

5) Consider a learning agent who is trying to infer 'who will be the first this year in the sports'. Suppose the learning agent has the following data. 12

Table 1: Table for Question 2(a)

Examples	Was First last year	Male	Works hard	Plays	First this year?
1	yes	yes	no	yes	yes
2	yes	yes	yes	no	yes
3	no	no	yes	no	yes
4	no	yes	no	yes	no
5	yes	no	yes	yes	yes
6	no	yes	yes	yes	no

Run the Candidate-Elimination algorithm on the above training examples and generate the sequence of *S* and *G* boundaries.

6) Why naïve Bayesian classifier is called "naïve"? Briefly outline the major ideas of naïve Bayesian classification. 5

7) Briefly explain how SPRINT algorithm develops the data structure to build the decision tree. 4

8) How does BOAT algorithm work? What are the advantages of BOAT algorithm over the other scalable decision trees? 4

3. a) On the following (Table 2) apply the ID3 algorithm and show your calculation for selecting the first attribute where **Target Concept** = {buys_computer}, **Attributes** = {age, income, student, credit_rating}.

Table 2: Table for question 3(a)

examples	age	income	student	credit_rating	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

b) What is the problem with ID3 algorithm? How does C4.5 overcome this problem? Apply C4.5 algorithm on the same data of Question 3(a) and show your calculation for selecting the first attribute.

c) Generate the AVC-sets of RainForest algorithm for the data of Question 3(a). How will RainForest algorithm handle the memory in case of AVC-set and AVC-group, if these does not fit in the memory?

4. a) Least Square is the most popular method of parameter estimation for coefficients of regression models. But why do we minimize the sum of the square of the residuals? Explain with example.

b) How does the AdaBoost technique boost its final classifier? Explain mathematically.

c) What are the ways to measure the distance in case of nominal data and missing values when KNN classifier is used?

d) What is the problem of hard margin in the design of SVM? How can we solve it? Explain mathematically.

e) For the given classification problem in Figure 1, which SVM classifier will you use? Linear or Non-linear? How will you use it? Explain your answer.

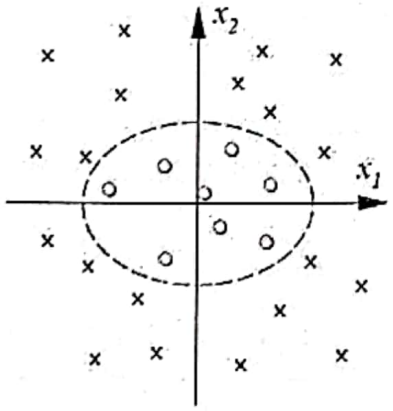

Figure 1: A classification problem for question 4(e)

Consider the following Boolean function shown in Table 3. From the table shown answer the following 4+6

i. Can this function be represented by a Perceptron? Explain your answer with appropriate figure.

ii. Construct a Perceptron that represents the function.

Table 3: Table for question 5(a)

A	B	~A v B
1	1	1
1	0	0
0	1	1
0	0	1

Weights are modified at each step according to the Perceptron training rule, which revises the weight w_i associated with input x_i according to the rule 3

$$w_i \leftarrow w_i + \Delta w_i$$
$$\Delta w_i = \eta\,(t - o)\,x_i$$

Why should this update rule converge toward successful weight values? 4

What is sigmoid function? How does it work for constructing multilayer networks? 8

In a learning task, you need to classify camera images of faces of various people in various poses. Images of 20 different people were collected, including approximately 32 images per person, varying the person's expression (happy, sad, angry, neutral), the direction in which they were looking (left, right, straight ahead, up), and whether or not they were wearing sunglasses. In total, 624 greyscale images were collected, each with a resolution of 120 x 128, with each image pixel described by a greyscale intensity value between 0 (black) and 255 (white). For applying Backpropagation algorithm, a number of design choices must be made. Describe your design choices.

Suppose a genetic algorithm uses chromosomes of the form $x = a\ b\ c\ d\ e\ f\ g\ h$ with a fixed length of eight genes. Each gene can be any digit between 0 and 9. Let the fitness of individual x be calculated as: $f(x) = a + b - c + d + e + f - g + h$ and let the initial population consist of four individuals with the following chromosomes: $x_1 = 65413532$, $x_2 = 87126601$, $x_3 = 23921285$, $x_4 = 41852094$ 12

Perform the Genetic algorithm using two point crossover and one point mutation for each pair set and select the best fitted sample. 5

Define the following fitness function selection methods: i. Roulette wheel selection ii. Tournament selection iii. Rank selection. 8

In a block stacking problem in Figure 2, your task is to discover a program using Genetic programming that can transform an arbitrary initial configuration of blocks into a stack that spells the word "**computer**". What should be the **primitive functions** and **terminal arguments** to formulate your task? Describe.

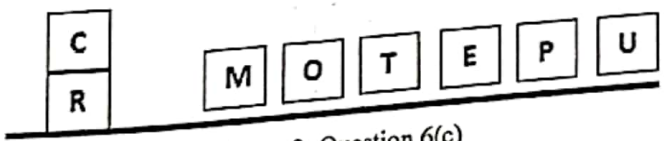

Figure 2: Question 6(c)

Suppose that the data mining task is to cluster points (with (x, y) representing location) into three clusters, where the points are $A_1(2, 10)$, $A_2(2, 5)$, $A_3(8, 4)$, $B_1(5, 8)$, $B_2(7, 5)$, $B_3(6, 4)$, $C_1(1, 2)$, $C_2(4, 9)$. The distance function is Euclidean distance. Suppose initially we assign A_1, B_1, and C_1 as the center of each cluster, respectively. Use the k-means algorithm to show the three cluster centers and the final clusters. (Use maximum three iterations) 10

b) Describe each of the following clustering algorithms in terms of the following criteria (1) shapes of clusters that can be determined; (2) input parameters that must be specified; and (3) limitations.
i. k-means ii. k-medoids iii. DBSCAN

c) In case of DBSCAN clustering algorithm define the following terms with appropriate figures:
i. Directly density reachable ii. Density reachable iii. Density connectivity

8. a) In a Latent Semantic Analysis problem, the following *word* × *documents* matrix is found.

$$\mathbf{A} = \begin{bmatrix} 2 & 3 & 5 \\ 1 & 0 & 1 \\ 3 & 4 & 1 \\ 2 & 1 & 3 \end{bmatrix}$$

Apply SVD algorithm and generate the **U**, **S** and **V** components of the matrix.

b) Write down the steps how PCA and SVD methods convert their axes to the projected axes using 2 diagrams.

c) After performing PCA, how do you reconstruct the original data from the reduced feature vector Explain with vector notations.