B.Sc. Engg. CSE 7th Semester

21 May 2019

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

SEMESTER FINAL EXAMINATION — **WINTER SEMESTER, 2018-2019**

DURATION: 3 Hours — **FULL MARKS: 150**

CSE 4709: 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. a) Briefly explain the factors that you need to consider while choosing the appropriate machine learning algorithms. 12

   b) Write short note on the followings: 3×3

      i. Overfitting and under-fitting problem

      ii. Curse of dimensionality

      iii. Bias-variance tradeoff

   c) What is transfer learning? How can you transfer what is learned for one task to improve the learning in other related task? 2+2

2. A set of reasonably clean sample records was extracted by Barry Becker from the 1994 Census database. We are interested in predicting whether a person makes over 50K a year. For simplicity suppose we model the two features with two Boolean variables, x_1; $x_2 \in \{0, 1\}$; and label $y \in \{0, 1\}$; where $y = 1$ indicates a person makes over 50K. In Figure 1, there are three positive samples ("+" for $y = 1$) and one negative samples ("-" for $y = 0$).

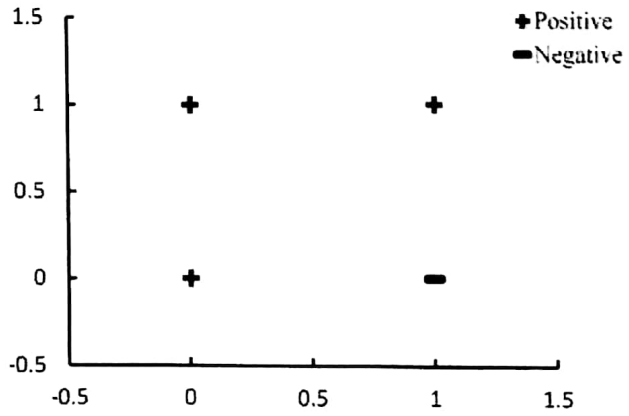

Figure 1: Sample dataset

Answer the followings:

   a) If we train a KNN classifier (K=1) based on data in Figure 1, and then try to classify the same data. Which sample(s) must be misclassified by this classifier? 5

   b) For predicting samples in Figure 1, which model is better: Logistic Regression or Linear Regression? Explain why. 5

   c) Is there any logistic regression classifier using x_1 and x_2 that can perfectly classify the examples in Figure 1? How about if we change label of point (0; 1) from "+" to "-"? 5

   d) Suppose we have trained a linear regression model $y = ax + b$ where $a = 0.5$ and $b = 1.0$, on a set of training data points $D = \{(1,0;\ 1,6);\ (1,5;\ 1,5);\ (3,0;\ 2,4)\}$. Calculate the mean squared errors of this model on D. 5

   e) Suppose we learn a Naïve Bayes classifier from the examples in Figure 1, using MLE (maximum likelihood estimation) as the training rule. Determine the prior probabilities $P(Y)$, and conditional probabilities $P(X_i|Y = 1)$. 5

3. a) Write the differences between generative and discriminative models of probabilistic classifiers. 5

b) Write the algorithm of Naïve Bayes classifier for discrete valued feature. 7

c) Consider a set of training examples given in Table 1 to train a robot, 'RecycleBot' to predict whether or not an office contains a recycling bin.

Table 1: Training dataset

	Status	Floor	Department	Office size	Recycle bin
1.	Faculty	Four	CSE	Medium	Yes
2.	Student	Four	EEE	Large	Yes
3.	Staff	Five	CSE	Medium	No
4.	Student	Three	EEE	Small	Yes
5.	Staff	Four	CSE	Medium	No

i. Construct a Bayesian network that represents all attributes in the 'RecycleBot' example, assuming that the predicting attributes are pairwise independent. Provide the probability table for each of the predicting attributes. 5

ii. Show how a Naïve Bayesian classifier would classify the following instance: *Status='Student', Floor='Four', Department='CSE', Office size='Small', Recycle bin=?* 8

4. a) Create a neural network with only one hidden layer (of any number of units) that implements $(A \vee \neg B) \otimes (\neg C \vee \neg D)$. Draw your network, and show all weights of each unit. 10

b) Figure 2 below is a small convolutional neural network that converts a 13x13 image into 4 output values. The network has the following layers/operations from input to output: convolution with 3 filters, max pooling, ReLU, and finally a fully-connected layer. For this network we will not be using any bias/o set parameters (b).

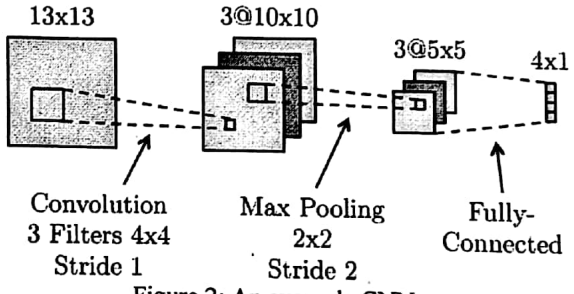

Figure 2: An example CNN

Answer the followings:

i. How many weights in the convolutional layer do we need to learn? 2

ii. How many ReLU operations are performed on the forward pass? 2

iii. How many weights do we need to learn for the entire network? 2

iv. What is the disadvantage of a fully-connected neural network compared to a convolutional neural network with the same size layers? 4

c) What is vanishing gradient problem? How do you solve this problem? 5

5. a) Write the mathematical derivations of `the objective function of support vector machine (SVM). 11

b) Suppose that you have a linear SVM binary classifier. Consider a point that is currently classified correctly, and is far away from the decision boundary. If you remove the point from the training set, and re-train the classifier, will the decision boundary change or stay the same? Explain your answer. 5

c) Consider the following training samples given in Figure 3.

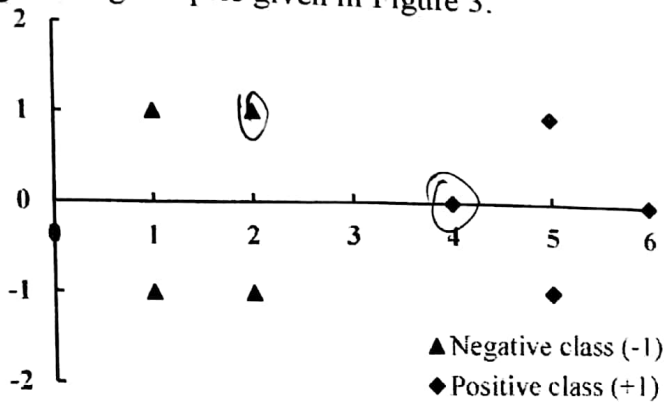

Figure 3: SVM training examples

Answer the followings:

i. Write the augmented feature vectors of the training samples. 2

ii. Redraw the points in $\Re^2$ space indicating the support vectors. 2

iii. Mathematically determine the discriminative hyperplane and draw the line. 5

6. a) Write differences between bootstrapping and cross-validation. 5

b) Briefly explain the steps of Stacking as an ensemble classifier. 10

c) Suppose in a classification problem, you have the probabilities of the three models: M1, M2, M3 (shown in Table 2) for five observations of test data set. 5

Table 2: Output of three machine learning models

M1	M2	M3	Output
.70	.80	.75	
.50	.64	.80	
.30	.20	.35	
.49	.51	.50	
.60	.80	.60	

What will be the predicted category for these observations if you apply probability threshold greater than or equals to 0.5 for category "1" or less than 0.5 for category "0"? Fill the table with the category you have determined.

d) Briefly explain the working principle of random forest algorithm. 5

7. a) Mrs. X has lost gender information of one of her customers, and does not know whether to make a skirt or trousers. The customer who is missing gender information has only the measurement of waist and hip which are 28 and 34 respectively. Using K-NN classifier with K=3, find the missing gender information. The training set is given in Table 3. 12

Table 3: Training set

S/N	Waist (cm)	Hip (cm)	Gender
1	28	32	Male
2	33	35	Male
3	27	33	Female
4	31	36	Female
5	28	34	?

b) Imagine you are dealing with text data. To represent the words you are using word embedding, i.e. representing words as vector of tokens. In word embedding, you will end up with 1000 dimensions. Now, you want to reduce the dimensionality of this high dimensional data such that, similar words should have a similar meaning. In such case, which algorithm are you most likely choose? Explain mathematically how you are going to reduce dimensions. 8

c) Given the covariance matrix, 5

$$\begin{bmatrix} 2.0 & 8.0 \\ 8.0 & 0.6 \end{bmatrix}$$

Find out the first two principal vectors.

8. a) For the distance matrix (Table 4), perform the iterations of agglomerative clustering (single linkage) and draw the corresponding Dendrogram. 12

Table 4: Distance Matrix

	A	B	C	D
A	0	1	4	5
B		0	2	6
C			0	3
D				0

b) Suppose you have got the result (shown in Table 5) of the confusion matrix of IRIS dataset after applying a clustering algorithm.

Table 5: Confusion Matrix

		Predicted Class		
		1	2	3
Actual Class	1	46	1	3
	2	3	45	2
	3	0	0	50

i. Find out the precision and recall from the confusion matrix. 4
ii. Find out the F-Measure. 2
iii. Find the cluster purity of individual clusters. 3

c) "One of the problems with hierarchical clustering is that there is no objective way to say how many clusters are there." – Explain the assertion with example. 4