B.Sc. Engg. CSE 6th Semester

28 May 2024

ISLAMIC UNIVERSITY OF TECHNOLOGY (IUT)

ORGANISATION OF ISLAMIC COOPERATION (OIC)

Department of Computer Science and Engineering (CSE)

SEMESTER FINAL EXAMINATION

DURATION: 3 HOURS

SUMMER SEMESTER, 2022-2023

FULL MARKS: 150

CSE 4621: Machine Learning

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

Answer **all 6 (six)** questions. Figures in the right margin indicate full marks of questions with corresponding COs and POs in parentheses.

---

1. a) What are the benefits of using the convolution operation in Convolutional Neural Networks (ConvNets) over regular Neural Networks (NNs)? — 6 (CO1) (PO1)

   b) If we initiate all weights in a Neural Network to a constant c, we suffer from the weight symmetry problem after each update. Again, following the standard approach of random initialization with $W_{ij} \simeq N(0, \sigma^2)$ will cause the output to converge to zero or diverge depending on whether the σ value is too small or too large. In this scenario, how can you initialize the weights? — 4 (CO1) (PO1)

   c) Consider the following hypothetical ConvNet architecture defined by series of different layers (i.e. Convolutional layer, RELU function, Max Pooling layer, Flattening, FC layer, Softmax layer). Calculate the shape of the output volume and the number of parameters (weights and bias) at each layer. — 12 (CO1) (PO1)

   You should write the output shapes in the format $(H,\ W,\ C)$, where $H,\ W,\ C$ are the *height*, *width* and *channel* dimensions, respectively. Unless specified, assume valid convolution where appropriate.

   Starting with an input sized $64 \times 64 \times 3$:

   CONV3-16 > RELU > CONV3-32 > RELU > POOL2 > CONV5-16 > CONV5-32 > POOL-2 > Flatten > FC28 > FC14 > Softmax-5

   Here the notations follow the convention given below:

   i. CONVf-N denotes a convolutional layer with N filters, each them of size $f \times f$.
   ii. POOL-n denotes a $n \times n$ max-pooling layer with stride of n and 0 padding.
   iii. FLATTEN denotes to flattenning the feature vector.
   iv. FC-N denotes a fully-connected layer with N neurons.
   v. Softmax-N denotes Softmax function with N neurons.

   d) Why are the residual connections often used in deep neural networks? — 3 (CO1) (PO1)

2. a) During decision tree generation for classification, instead of taking a binary split for the numeric attribute, we use ternary split using two thresholds w_{ma} and w_{mb}. In other words, we use three potential branches where samples can take j_{th} branch according to the following conditions: — 8 (CO3) (PO3)

$$x_k < w_{ma}; \quad w_{ma} \leq x_k \leq w_{mb}; \quad x > w_{mb}$$

   Propose a modification of the tree induction method along with impurity measure to learn those two thresholds.

   b) Compare between Discriminative model and Discriminant function with examples. — 7 (CO2) (PO2)

c) From the training data in Table 1 classify the tuple $x = [youth, low, no, fair]^T$ using Naïve Bayes classifier. Calculate individual class-conditional probability for each class. 10 (CO1) (PO1)

3. a) What properties does an impurity test function need to satisfy? Give an example of impurity test function and draw its response graph for a two-class problem. 5 (CO1) (PO1)

b) Consider the Table 1 which presents a training set of class-labeled tuples from the **AllElectronics** customer database. In this example, each attribute is discrete valued. The class label attribute,*buys_computer* has two distinct values namely, {yes, no}. Therefore, there are two distinct classes. Let class C_1 correspond to *yes* and class C_2 correspond to *no*. There are nine tuples of class *yes* and five tuples of class *no*. RID represents the record/tuple number, which is not useful for our task. Generate a decision tree from this database along with all calculations for split generation. 20 (CO1) (PO1)

Table 1: Training samples from the **AllElectronics** customer database for Question 3.b

RID	*age*	*income*	*student*	*credit_rating*	*Class: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
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

4. a) Imagine you are working on a project to analyze customer purchasing behavior for a large retail chain with millions of transactions recorded over several years. The dataset contains a wide range of features such as customer demographics, purchase history, time of purchase, product categories, and more. However, the dataset is extremely high-dimensional, with thousands of features, making it challenging to extract meaningful insights and build predictive models efficiently. As a machine learning expert, you have been tasked with developing a customer categorization model to identify distinct customer groups based on their purchasing patterns. However, due to the high dimensionality of the dataset, traditional machine learning algorithms like clustering or classification models struggle to effectively process and extract insights from the data. How can you leverage Principal Component Analysis (PCA) to address the challenges posed by the high-dimensional nature of the dataset? 8 (CO1) (PO1)

b) What are the implications of having a *positive, negative,* or *zero* covariance between variables in terms of their linear relationship and impact on statistical analysis? 3 × 3 (CO2) (PO2)

c) Let an orthonormal transformation $y = \Phi^T x$, where the matrix Φ contains all eigenvectors. Show that for orthonormal transformations, Euclidean distances are preserved, i.e., . $||y||^2 = ||x||^2$ 8 (CO1) (PO1)

5. a) Explain why k-means clustering algorithm may not find the best solution? How can you choose the right value of k? 5 + 3 (CO1) (PO1)

b) Consider the data set as given in Table 2 consisting of the scores of two variables on each of six individuals. 6 + 6 (CO1) (PO1)

Table 2: Dataset for Question 5.b

Sample	x_1	x_2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	7.0	7.0
5	3.5	5.0
6	4.5	5.0

Apply the *k-medoids* clustering algorithm with the value $k = 2$. Provide all required calculations up to two cluster-center updates.

c) "Both *k-means* and *k-medoids* clustering algorithms produce convex shaped clusters" — Do you agree or disagree? Justify your answer. 5 (CO2) (PO2)

6. a) The Support Vector Machine (SVM) is a highly accurate classification method. However, SVM classifiers suffer from slow processing when training with a large set of data tuples. Discuss how to overcome this difficulty and develop a scalable SVM algorithm for efficient SVM classification in large data sets. 7 (CO1) (PO1)

b) Consider a Support Vector Machine to classifier the following training data for a two-class problem given in Table 3:

Table 3: Training data for Question 6.b

Class	x_1	x_2
+	1	1
+	2	2
+	2	0
-	0	0
-	1	0
-	0	1

i. After plotting these six training points (**use a separate graph paper**), construct the weight vector for the optimal hyperplane, and calculate the optimal margin width. 7 (CO1) (PO1)

ii. If you remove one of the support vectors does the size of the optimal margin decrease, or stay the same, or increase? Mention explicitly which one you have removed. 3 (CO2) (PO2)

[Note: You do not need to calculate the solutions by solving, rather find the answers by inspecting the graph.]

c) How does the kernel trick work in SVM classifier? Give an example with the polynomial kernel that can be effectively utilized to handle non-linearly separable data sets. 4 + 4 (CO1) (PO1)