Data Mining and Machine Learning

Final Examination, II Semester, 2023–2024

Date : 3 May, 2024 | Marks : 40
Duration : 3 hours | Weightage : 40%

1. There are three biased coins c_1, c_2, and c_3. You are given a sequence of 1000 coin tosses, where each outcome corresponds to tossing one of $\{c_1, c_2, c_3\}$, chosen uniformly at random. Let $\{p_1, p_2, p_3\}$ be the probabilities of heads for the coins $\{c_1, c_2, c_3\}$, respectively. You have prior information that p_1 is less than 0.5 and p_2 and p_3 are greater than 0.5. Describe, in algorithmic pseudocode, an iterative procedure to estimate $\{p_1, p_2, p_3\}$. *(5 marks)*

2. Explain how to cluster points using a mixture of Gaussians. Can this also be used to detect outliers? *(5 marks)*

3. Explain how clustering can be used for image segmentation — that is, to identify objects in an image. *(5 marks)*

4. The 0−1 loss function assigns a cost of 1 to every misclassified input and a cost of 0 to every correctly classified input. This loss function is minimized when the model makes no errors on the training data. Explain with respect to the perceptron algorithm why the 0−1 loss function is not always adequate to learn a good model. *(5 marks)*

5. (a) For $z = wx + b$, how does the shape of the sigmoid function $\sigma(z) = (1 + e^{-z})^{-1}$ vary with w and b?

   (b) Given two input features x_1, x_2, explain how to construct a neural network to approximate a "rectangular box" function $g(x_1, x_2)$ with height h for $\ell_1 \leq x_1 \leq r_1$ and $\ell_2 \leq x_2 \leq r_2$. In other words, the function to be approximated is the following.

$$g(x_1, x_2) = \begin{cases} h & \text{if } \ell_i \leq x_i \leq r_i, i \in \{1, 2\} \\ 0 & \text{otherwise} \end{cases}$$

(5 marks)

6. Consider a neural network that is layered and completely connected. Suppose we initialize two nodes n_1 and n_2 from the same layer with the same biases and same weights on incoming and outgoing edges. What can you say about the final weights and biases that will be learned for n_1 and n_2 through backpropagation? What can you conclude about initialization strategies for such networks? *(5 marks)*

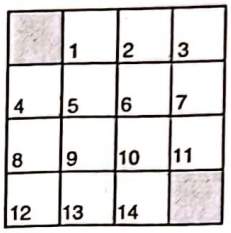
7. Two astronomers independently count stars in the same region of the sky using their telescopes. The region has N stars. The counts reported by the astronomers are M_1 and M_2, respectively. Each astronomer has a small probability of miscounting the stars by ± 1. It is also possible that their telescopes are faulty and do not focus properly, denoted by boolean events F_1 and F_2, respectively. With a faulty telescope, an astronomer may undercount by as many as 3 stars.

   (a) Draw a Bayesian network to represent the relationship between N, M_1, M_2, F_1 and F_2.

   (b) Suppose $M_1 = 12$ and $M_2 = 14$. What are the possible values of N for each of the different combinations of F_1 and F_2?

(5 marks)

8. Consider the 4×4 grid-world to the right. The non-terminal states are $\{1, 2, \ldots, 14\}$ and the terminal states are the shaded squares. There are four actions, $\{\text{up}, \text{down}, \text{left}, \text{right}\}$, which result in a deterministic move in the given direction. A move that would take the agent off the grid leaves the position unchanged. The reward is -2 for any transition that results in a change of position. A move off the grid that does not change the position has a reward of -1. Formally, $r(s, a, s') = -2$ if $s \neq s'$ and $r(s, a, s') = -1$ if $s = s'$.



(a) Consider the uniformly random policy π that chooses each of the four directions with equal probability. Assume we start with an initial value $v(s) = 0$ for each state s. Compute one iteration of v_π.

(b) Describe the new policy after applying policy improvement based on this one step computation of v_π.

(5 marks)

CMI/02/05/2024/Final Exam May 2024

Chennai Mathematical Institute

Distributed Computing and Big Data

DURATION: 2 HOURS MAX MARKS: 35.

Instructions

- Please remember to mention your name and roll number in your answer sheet.
- This is an individual task. Do not discuss with anyone.
- This is a closed book exam. You are not allowed to carry books or cheatsheets.
- No electronic devices (calculators, laptops, etc) are allowed in the exam hall. Wherever heavy calculation is involved, you need not evaluate it to the final number unless it is explicitly asked for. For example, it is acceptable to leave the answer as $\frac{1}{1+\frac{5}{32}}$. You need not evaluate it to 0.865.
- First section has negative marks. No negative marks for the rest of the sections.

Section 1: All questions carry one mark each. -0.5 for wrong answers. Answer in True/False.

Question 1. One Petabyte space is enough to store 4 Million ebooks of size 1 MB each.

Question 2. Based on Amdahl's law, we can expect a linear increase in speed-up for a specific job as we increase the number of processors.

Question 3. Yet Another Resource Negotiator, is a resource management and job scheduling technology for the Hadoop distributed processing framework

Question 4. One of the items in pig philosophy is that pigs live anywhere.

Question 5. As per the principles of object-oriented programming, an object has an identity while a class does not.

Question 6. An impedance mismatch occurs in relational databases when a relational database needs to be transformed into an object-oriented model.

Question 7. BSON is a binary serialization format used to store documents in MongoDB.

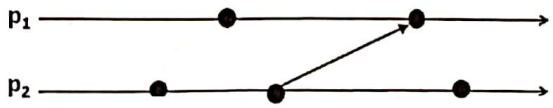
Question 8. NoSQL datastores are not ACID compliant.

Question 9. Changing a block in a blockchain makes all the following blocks invalid.

Question 10. A carrier hotel is a facility strategically based in a location closer to users that houses networks and cloud services.

Section 2: All questions carry 2 marks each.

Consider the following space-time execution diagram while answering the questions in this Section.



Question 11. List all the happens-before relationships.

Question 12. Annotate the events using scalar time.

Question 13. Annotate the events using vector time.

Question 14. Annotate the events using matrix time.

Question 15. Identify an inconsistent cut.

Section 3: All questions carry 2 marks each.

Question 16. Consider the following Pig script.

```
Lines = LOAD 'file1' USING PigStorage() as (line:chararray);
Words = FOREACH Lines GENERATE FLATTEN(TOKENIZE(line)) AS word;
Groups = GROUP Words BY word;
Counts = FOREACH Groups GENERATE group, COUNT(Words) as Cnt;
Results = ORDER Counts BY Cnt ASC;
Dump Results;
```

Assume that the input file 'file1' contains the following two lines:

```
cmi is the best
the best in chennai is cmi
```

What does the pig script output?

Question 17. The following pig script was written to find the most expensive iphone. However, it has errors. Identify the errors and correct them.

Pig Script:

```
A = LOAD 'file2' USING PigStorage(',') AS (year:int,product:chararray,cost:int);
B = GROUP A BY $2; -> $1
C = FOREACH B GENERATE MIN(A.cost);
DUMP C;
```

Input File ('file2' contains year,product,cost):

```
2022, iphone, 50000
2023, iphone, 65000
2024, iphone, 72000
```

Expected output is 72000.

Question 18. How many nodes are created when the following three statements are executed by Neo4j?

```
1. CREATE (p:Person{name:'Venkatesh'})-[:Teaches]->(c:Course{name:'BigData'})
2. CREATE (p:Person {name:'Raj'})-[:StudentOf]->(o:Org{name:'CMI'})
3. MATCH (a:Person),(b:Org) WHERE a.name = 'Venkatesh' AND b.name = 'CMI'
   CREATE (a)-[:FacultyAt]->(b)
```

Section 4: All questions carry 3 marks each.

Question 19. Describe a map-reduce design for computing median of a large list of numbers. Assume that the input file contains 2 Million lines. Each line contains an integer ranging between 1 and 1000.

Question 20. Assume that Indian Railways wants to store the train running schedule (past and live status) information in MongoDB. Provide a database design along with at least one or two queries as example to indicate how you would query the data.

Question 21. Design a RESTful web service for a learning management system such as moodle. Include at least three items in your object model.

Linear Algebra and its Applications
Final Examination
29/04/2024

(**Note:** You may use a calculator, but not your phone. For Questions 1 to 4, no justification is needed.)

1. In each of the following provide an appropriate example of a 2×2 matrix. [5 points]

   (a) A cannot be diagonalized and it is invertible.
   (b) A cannot be diagonalized but it is singular.
   (c) A is diagonalizable and it is non-singular.
   (d) A has orthogonal columns but it is not invertible.
   (e) A has orthogonal columns and it is diagonalizable.

2. Determine the truth value of the following statements. Just write **T** or **F**. [5 points]

   (a) If U is a matrix with orthonormal columns then $UU^T = I$.
   (b) A square matrix with orthonormal columns has real eigenvalues.
   (c) The singular values of an orthonormal matrix are all equal to 1.
   (d) The eigenvalues of an orthogonal matrix need not be all 1.
   (e) For any matrix A, the eigenvalues of A^TA are positive.

3. Consider the following SVD factorization of a movie-ratings matrix A:

$$\begin{bmatrix} 2 & 2 & 2 & 0 \\ 4 & 4 & 4 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix} = \begin{bmatrix} 0.44 & 0 \\ 0.99 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 7.74 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

   Here the rows of A are the users and the columns are the movies labeled *Drama1, Drama2, Drama3, Comedy1*. The matrix U connects users to movie geners (present as latent features of A), the matrix V connects movies to geners and finally, Σ describes the strength of each genere. In this synthetic example it is clear that the number of geners is equal to the rank of the rating matrix. Answer the following questions in order to use the above matrix factorization for recommending movies. [10 points]

   (a) Suppose a new user has watched only *Drama2* movie and rated it 3. Determine the vector $x \in \mathbb{R}^4$ which encodes this information and can be used to determine recommendations.
   (b) Which of the above matrix has to be used to map the ratings vector x to the 2-dimensional "genere space"?
   (c) Determine the representaiton of new user's ratings in the genere space.
   (d) Appropriately map the above representation back into the "movie space" in oder to interprete the genere each movie partakes. Conclude by recommending movies to the new user.
   (e) What is the potential use of the map given by the matrix UU^T?

4. Let A be an $n \times n$, real matrix and $b \in \mathbb{R}^n$ be a nonzero vector. Consider the following algorithm and answer the questions based on it: [10 points]

$q_1 = b/\|b\|_2$
for $k = 1, 2, \ldots$ do
  $v = Aq_k$
  for $j = 1$ to k do
    $h_{jk} = \langle q_j, v \rangle$
    $v = v - h_{jk} q_j$
  end for
  $h_{k+1,k} = \|v\|_2$
  $q_{k+1} = v/h_{k+1,k}$
end for

(a) The vectors $\{q_1, \ldots, q_k\}$ form an orthonormal basis of the subspace spanned by which vectors? (Hint: try to answer in terms of A, b).

(b) For each k, denote by Q_k the $n \times k$ matrix whose columnns are $q_1, \ldots, q_k$ and by $H_{k+1,k}$ the $(k+1) \times k$ matrix whose entries are h_{jk}'s. Express the above algorithm in the matrix language using A, Q_k and $H_{k+1,k}$.

(c) What is $Q_k^T A Q_k$ for each $k = 1, \ldots, n$?

(d) Which two matrices have the exact same eigenvalues?

(e) If A is symmetric then each H_k is ... and .... Fill in the blanks.

5. Use QR factorization (Gram-Schmiddt or Householder) method to find x that minimizes $\|Ax - b\|^2$, where [5 points]

$$A = \begin{bmatrix} 3 & -6 \\ 4 & -8 \\ 0 & 1 \end{bmatrix}, b = \begin{bmatrix} -1 \\ 7 \\ 2 \end{bmatrix}.$$

6. Find the (approximate) dominant eigenvector of the matrix [10 points]

$$A = \begin{bmatrix} 4 & 5 \\ 6 & 5 \end{bmatrix}$$

using power iteration. Start with the vector $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and perform 4 iterations, with scaling. Use this approximate eigenvector to calculate the dominant eigenvalue.

7. Consider the matrix [15 points]

$$A = \begin{bmatrix} 1 & -1 \\ -2 & 2 \\ 2 & -2 \end{bmatrix}.$$

(a) Find the full singular value decomposition of A.

(b) Write down an orthonormal basis for the following four fundamental spaces: row space of A, column space of A, null space of A and null space of A^T.

(c) What is the pseudoinverse of A?

- This exam has 3 questions for a total of 100 marks.
- Warning: CMI's academic policy regarding cheating applies to this exam.

All arrays in this question paper have their indices starting at 0. If you wish to use 1-based indexing, you must clearly state this in each such answer. The arrays in these questions are objects whose sizes are fixed, in the sense that the size cannot be changed after the array is created. In particular, these are *not* Python's lists, whose sizes can be changed using, say, `append()`. Also: note that Python's `list.append()` does *not* run in *worst-case* constant time; be mindful of this when writing your pseudocode.

Unstated assumptions and lack of clarity in solutions can and will be used against you during evaluation. Please ask the invigilators if you have questions about the questions.

1. The input to this problem is an array A of length $n \geq 1$. Each element of array A is a pair of the form (StudentID, marks) where StudentID is an alphanumeric string, and marks is a non-negative integer. The pair (StudentID, marks) being present in A means that the student whose ID is StudentID, got marks marks in some exam. There may be more than one pair with the same StudentID; each such pair denotes the marks that this particular student got in a different exam. All pairs with the same StudentID occur *consecutively* in the array. The goal is to find the *maximum* marks that each student scored, among all the exams whose marks for this student are present in A. **[30]**

   Write the *complete pseudocode* for an algorithm MaxMarks(n, A) that **prints out** the highest marks corresponding to each StudentID that is present in array A. For each distinct StudentID present in array A, the output should have *exactly one line* that lists this StudentID and the corresponding maximum marks. The algorithm must run in $O(n)$ time and take at most *constant* extra space, in the worst case. Assume that two StudentIDs can be compared for equality in constant time, using the == operator. Assume also that you can use

   - A[i][0] to access the StudentID in location i of array A in constant time,
   - A[i][1] to access the marks in location i of array A in constant time, and
   - A Python-like function `print()` that prints out its arguments, and then a newline, in constant time. You may use Python-like—or any other sensible—string interpolation to include values of variables in the printed-out string.

   You will get the credit for this question *only if* your solution *correctly* solves *all* valid instances of the stated problem *within the required time and space bounds.*

   You do *not* have to explain why your pseudocode is correct. You do *not* have to provide an analysis of its running time and space.

2. The input to this problem consists of an array A of $n \geq 1$ integers, and an integer x. Array A is *sorted* in non-decreasing order. The goal is to find the *number of times* that integer x appears in A. Note that this number could be zero. **[30]**

   Write the *complete pseudocode* for an algorithm Counter(A, x) that returns the number of times that x appears in A. The algorithm must run in $O(\log_2 n)$ time in the worst case. Assume that two

numbers can be compared using the operators $\{<, >, ==\}$ in constant time. Assume also that you can use A[i] to access the number in location i of array A, in constant time.

You will get the credit for this question *only if* your solution *correctly* solves *all* valid instances of the stated problem *within the required time bound.*

You do *not* have to explain why your pseudocode is correct. You do *not* have to provide an analysis of its running time.

3. Recall that a palindrome is a string that reads the same in either direction. A *non-trivial palindrome* is a palindrome with length (number of characters) at least two. Consider the following problem: **[40]**

Palindrome Sequence

- Input: An integer $n \geq 2$ and a string S of length n. String S is an array indexed from 0; its elements are thus $S[0], S[1], \ldots, S[(n-1)]$.
- Output: True if S can be obtained by concatenating one or more non-trivial palindromes, and False otherwise. Equivalently: True if S can be partitioned (that is: cut up, without dropping any element, and without rearranging the pieces) into one or more non-trivial palindromes, and False otherwise.

Some examples with $n = 10$:

- True instances: `abacabaddd`, `fjbubjfhuh`, `mttmzizzcz`, `fjfyspsyqq`, `abcdeedcba`
- False instances: `daabbbcccd`, `azatznnzth`, `ummnurlxmv`, `xqeppajynx`, `tyjglnvmaa`

Write the *complete* pseudocode for a *non-recursive* algorithm IsNTPSequence(n, S) that solves Palindrome Sequence using *dynamic programming.* The algorithm should have a worst-case running time of $O(n^c)$ for some small constant c.

You will get the credit for this question *only if* your pseudocode is *complete*, and implements a *non-recursive DP algorithm* which *correctly* solves *all* valid instances of the problem *within the required time bound.*

Make sure that you correctly initialize your DP table, that you check for sentinel values wherever required, and that you always compute and store the value in any cell of the table *before* you access the value in that cell for further computation.

You do *not* have to explain why your pseudocode is correct. You do *not* have to provide an analysis of its running time.