Team Members:
Kashish Sehgal(kashishsehgal73@gmail.com)
Manasi Dogra
Rakshit Joshi(rakshitjoshi97@gmail.com)
Soumya Kumar
(Bharati Vidyapeeth's College of Engineering)

CELESTINI PROJECT INDIA 2018
TAKE-HOME EXAM
March 19, 2018

1. **This exam has 5 questions. Write the answers in the space provided in the questions. Return your solutions in PDF format by 11:59PM (IST) on Mar 31, 2018 to email <celestiniprizeindia@gmail.com>.**
2. **Submit all the code solutions in a single zip file or using a GitHub link. Provide a readme file to the code solution for each question.**

1. Multiple-choice questions (10 points)

Select one or more correct solutions. Please write your answer next to **Solution:**

A.) What types of learning, if any, best describe the following three scenarios:
(i) A coin classification system is created for a vending machine. In order to do this, the developers obtain exact coin specifications from the U.S. Mint and derive a statistical model of the size, weight, and denomination, which the vending machine then uses to classify its coins.
(ii) Instead of calling the U.S. Mint to obtain coin information, an algorithm is presented with a large set of labeled coins. The algorithm uses this data to infer decision boundaries which the vending machine then uses to classify its coins.
(iii) A computer develops a strategy for playing Tic-Tac-Toe by playing repeatedly and adjusting its strategy by penalizing moves that eventually lead to losing.

[a] (i) Supervised Learning, (ii) Unsupervised Learning, (iii) Reinforcement Learning
[b] (i) Supervised Learning, (ii) Not learning, (iii) Unsupervised Learning
[c] (i) Not learning, (ii) Reinforcement Learning, (iii) Supervised Learning
[d] (i) Not learning, (ii) Supervised Learning, (iii) Reinforcement Learning
[e] (i) Supervised Learning, (ii) Reinforcement Learning, (iii) Unsupervised Learning

Solution: [d] (i) Not learning, (ii) Supervised Learning, (iii) Reinforcement Learning

---

B.) For an imbalanced dataset, which of the following metric/tool is not that useful?
[a] F1 measure
[b] Accuracy
[c] Confusion Matrix
[d] Precision
Solution: [b] Accuracy

C.) Consider the following implementation of a function mysteryFunction (pseudocode), where x is a positive integer:

```
mysteryFunction (x)
    xs = str(x)
    if len(xs) == 1
        return int(xs)
    n = int(xs[0]) + int(xs[1])
    if len(xs) == 2
        return n
    else
```

```
        return n + mysteryFunction(xs[2:])
```

What does mysteryFunction(3223) return

[a] 0

[b] 10

[c] 5

[d] 1

Solution: [b] 10

D.) What is the output of the following program (in C) for input "Celestini Project"

```
#include "stdio.h"
int main()
{
   char arr[100];
   printf("%d", scanf("%s", arr));
   return 2;
}
```

[a] 0

[b] -1

[c] 1

[d] 2

Solution:[c] 1

E.) Which of the following options suggest the best approach to fix the high bias and high variance in a machine learning model? (Assume model has been trained on at least 1000 samples)

[a] To fix high bias, we can add more training samples; to fix high variance, we can reduce the number of training examples so it fits on them less

[b] To fix high bias, we can reduce our model's complexity; to fix high variance, we can increase our model's complexity

[c] To fix high bias, we can increase our model's complexity; to fix high variance, we can try reducing the number of features in the dataset

[d] To fix high bias, we can decrease the number of training samples; to fix high variance, we can increase the number of features in the dataset

Solution:[c] To fix high bias, we can increase our model's complexity; to fix high variance, we can try reducing the number of features in the dataset

---

F.) The major advantage(s) of prototyping over a Raspberry Pi over prototyping on a personal computer are
[a] cost
[b] faster processing speed
[c] small form factor
[d] low power consumption
Solution:[a],[c],[d]

G.) Which of the following statement(s) are correct?
[a] A machine learning model with higher accuracy will always indicate a better classifier.
[b] When we increase the complexity of a model, it will always decrease the test error.
[c] When we increase the complexity of a model, it will always decrease the train error.
Solution:[a] and [c]

H.) What is the output of the program (in C)?

```
#include <stdio.h>
int main()
{
int celestini[6] = {6,5,4,3,2,1};
int *ptr = (int*)(&celestini+1);
printf("%d %d", *(celestini+1), *(ptr-1));
return 0;
}
```

[a] 5 1
[b] 4 3
[c] 6 4
[d] 5 3
Solution:[a] 5 1

I.) A poor binary classification model for detecting a **rare** cancer disease *always* predicts positive for presence of the disease. What can we infer about the model's performance?
[a] The model has high accuracy, maximum precision but low recall.
[b] The model has poor accuracy, poor precision but maximum recall.
[c] The model has poor accuracy, maximum precision and minimum recall.
[d] The model has maximum accuracy, maximum precision but minimum recall.
Solution:[b] The model has poor accuracy, poor precision but maximum recall.

J.) Which of the following problems are best suited for a machine learning approach?
(i) Classifying numbers into primes and non-primes.
(ii) Detecting potential fraud in credit card charges.
(iii) Determining the time it would take a falling object to hit the ground.
(iv) Determining the optimal cycle for traffic lights in a busy intersection.
[a] (ii) and (iv)
[b] (i) and (ii)
[c] (i), (ii), and (iii).
[d] (iii)
Solution:[a] (ii) and (iv)
---

2. Programming (10 points)

Given two sparse matrices A and B, perform multiply and convolution operation of the matrices in their sparse form itself. The result should consist of two sparse matrices, one obtained by multiplying the two input matrices, and the other obtained by convolution of the two matrices.

Recall that a sparse matrix is a matrix in which most of the elements are zero. Assume both the matrices are of size NxN. Assume the number of non-zero elements in A and B are m1 and m2 respectively. Note that other entries of matrices will be zero as matrices are sparse.
Note: You may use any data-structure to represent the sparse matrix. The solution approach should not use in-built libraries for the multiplication or convolution of matrices.

(i) Write code to solve the above problem in Python, Java or C++
Also given in the github link provided

```
// Java code to perform add,
// multiply and transpose on sparse matrices
import java.util.Scanner;


public class sparse_matrix {


	int MAX = 100;
	int data[][] = new int[MAX][3];
	int row, col;
	int len;
	public sparse_matrix(int r, int c)
	{
		row = r;
		col = c;
		len = 0;
	}

	public void insert(int r, int c, int val)
	{
		if (r > row || c > col) {
			System.out.println("Wrong entry");
		}

		else {
			data[len][0] = r;
```

```
                data[len][1] = c;
                data[len][2] = val;
                len++;
        }
}

public void Convolution(sparse_matrix b)
{

        if (row != b.row || col != b.col) {
                System.out.println("Matrices can't be Convoluted");
        }

        else {

                int apos = 0, bpos = 0;
                sparse_matrix result = new sparse_matrix(row, col);

                while (apos < len && bpos < b.len) {

                        if (data[apos][0] > b.data[bpos][0] ||
                        (data[apos][0] == b.data[bpos][0] &&
                        data[apos][1] > b.data[bpos][1]))

                        {



                                bpos++;
                        }

                        else if (data[apos][0] < b.data[bpos][0] ||
                        (data[apos][0] == b.data[bpos][0] &&
                        data[apos][1] < b.data[bpos][1]))

                        {

                                apos++;
                        }

                        else {
```

```
                                int addedval = data[apos][2] * b.data[bpos][2];


                                result.insert(data[apos][0],
                                data[apos][1],
                                addedval);
                                apos++;
                                bpos++;
                        }
                }
                result.print();
        }
}

public sparse_matrix transpose()
{

        sparse_matrix result = new sparse_matrix(col, row);
        result.len = len;
        int count[] = new int[col + 1];
        for (int i = 1; i <= col; i++)
        count[i] = 0;

        for (int i = 0; i < len; i++)
        count[data[i][1]]++;

        int[] index = new int[col + 1];
        index[1] = 0;
        for (int i = 2; i <= col; i++)

        index[i] = index[i - 1] + count[i - 1];

        for (int i = 0; i < len; i++) {

                int rpos = index[data[i][1]]++;
                result.data[rpos][0] = data[i][1];
                result.data[rpos][1] = data[i][0];
                result.data[rpos][2] = data[i][2];
        }
        return result;
}
```

```
public void multiply(sparse_matrix b)
{

	if (col != b.row) {
		System.out.println("Can't multiply, "+ "Invalid dimensions");

		return;
	}
	b = b.transpose();
	int apos, bpos;
	sparse_matrix result = new sparse_matrix(row, b.row);

	for (apos = 0; apos < len;) {

		int r = data[apos][0];
		for (bpos = 0; bpos < b.len;) {

			int c = b.data[bpos][0];

			int tempa = apos;
			int tempb = bpos;

			int sum = 0;

			while (tempa < len && data[tempa][0] == r
			&& tempb < b.len && b.data[tempb][0] == c) {

				if (data[tempa][1] < b.data[tempb][1])
				tempa++;

				else if (data[tempa][1] > b.data[tempb][1])
					tempb++;
				else
				sum += data[tempa++][2] * b.data[tempb++][2];
			}
			if (sum != 0)
			result.insert(r, c, sum);

			while (bpos < b.len && b.data[bpos][0] == c)

			bpos++;
		}
```

```
                while (apos < len && data[apos][0] == r)

                apos++;
        }

        result.print();
}

// printing matrix
public void print()
{
        System.out.println("Dimension: " + row + "x" + col);
        System.out.println("Sparse Matrix: \nRow Column Value");

        for (int i = 0; i < len; i++) {

                System.out.println(data[i][0] + " "
                + data[i][1] + " " + data[i][2]);
        }
}

public static void main(String args[])
{
        Scanner sc=new Scanner(System.in);
        System.out.println("Enter the value of n");
        int n = sc.nextInt();
        // create two sparse matrices and insert values
        sparse_matrix a = new sparse_matrix(n, n);
        sparse_matrix b = new sparse_matrix(n, n);

        System.out.println("Enter the value of m1 ");
        int m1 = sc.nextInt();
        System.out.println("Enter the value of m2");
        int m2 = sc.nextInt();

        System.out.println("Enter data for matrix 1");
        for(int i=0;i<m1;i++)
        {
                System.out.println("Enter row,col,data");
                int r,c,v;
                r= sc.nextInt();
```

```
                c= sc.nextInt();
                v= sc.nextInt();
                a.insert(r,c,v);
            }
            System.out.println("Enter data for matrix 2");
            for(int i=0;i<m2;i++)
            {
                System.out.println("Enter row,col,data");
                int r,c,v;
                r= sc.nextInt();
                c= sc.nextInt();
                v= sc.nextInt();
                b.insert(r,c,v);
            }

            // Output result
            System.out.println("Convolution ");
            a.Convolution(b);
            System.out.println("\nMultiplication: ");
            a.multiply(b);

        }
}
```

(ii) What is the best time complexity of your solution (in terms of m1,m2,N)?
O(m1+m2)
(iii) What is the best space complexity of your solution (in terms of m1,m2,N)?
O(m1+m2)

3. Programming II (10 points)

Write an efficient algorithm that searches for a value in an m x n matrix. This matrix has the following properties:

- Integers in each row are sorted in ascending from left to right.
- Integers in each column are sorted in ascending from top to bottom.
  For example,
  Consider the following matrix:

```
[
  [1,   4,  7, 11, 15],
  [2,   5,  8, 12, 19],
  [3,   6,  9, 16, 22],
  [10, 13, 14, 17, 24],
  [18, 21, 23, 26, 30]
]
```

Given target = 5, return true.
Given target = 20, return false.

(i) Write code to solve the above problem in Python, Java or C++.

```
#include<iostream>
using namespace std;

int search(int mat[100][100],int m, int n, int x)
{
   int i=0, j=n-1;//set indexes for top right element
   while ( i < m && j >= 0 )
   {
      if ( mat[i][j] == x )
      {
         cout<<"\nElement found at "<<i<<" "<<j;
         return 1;
      }
      if ( mat[i][j] > x )
      j--;
      else //  if mat[i][j] < x
      i++;
   }
   cout<<"\n Element not found";
   return 0;// if ( i==n || j== -1 )
}
```

```
int main()
{
   int r,c,e,mat[100][100];
   cout<<"\nEnter number of rows and colomns";
   cin>>r>>c;
   for(int i=0; i<r ; i++)
   {
      for(int j=0; j<c; j++)
      {
         cin >>mat[i][j];
      }
   }
   cout<<"\nEnter element to be searched";
   cin >> e;
   search(mat,r,c,e);
   return 1;
}
```

(ii) What is the best time complexity of your solution (in terms of m, n)?

O(n) or O(m)

(iii) What is the best space complexity of your solution (in terms of m, n)?

O(n*m)

4. Problem Solving (20 points)

Please select either problem 4A or 4B and provide your solution in detail. You may solve both problems for extra credit though it is not required.

4A. Cryptosystem Identifier (select either 4A or 4B)

Cryptography is associated with the process of converting plain text into unintelligible text and vice versa. The goal of problem is to identify the cryptosystem used in encrypting a given cryptogram using Support Vector Machine (SVM) and Back propagation Neural Networks (BPNN). We consider that the cryptogram are derived using Simple substitution or Vigenere.

[a] Simple substitution (SS) ciphers work by replacing each plaintext character by another one character. To decode cipher text letters, one should use reverse substitution and change the letters back.

[b] Vigenere cipher is a kind of polyalphabetic substitution cipher. It is about replacing plaintext letters by other letters. Parties have to agree on a common shared keyword (which may also be a sentence), which is used during encryption algorithm.

Data generation approach: Create 50 cryptograms by Simple Substitution (Key size: 26) and 50 cryptograms by Vigenere cryptosystems (key size: 3). Each of the cryptograms should be of size 200 characters consisting of only upper case alphabets and white spaces (i.e. total 27 characters).

You can use the following links for encoding

- Vigenere: https://www.mathworks.com/matl-abcentral/mlc-downloads/downloads/submissions/29443/versions/1/previews/VigenereDetails.html
- Simple substitution: https://in.mathworks.com/matlabcentral/fileexchange/31522-substitution-cipher-encoder-and-decoder

We are providing you with a dataset of ten plaintext, ten cryptograms by Vigenere, and ten cryptograms by simple substitution for testing your solution in the attachment (dataset_cryptosystem.doc)

Hint: You may consider using frequency pattern of the cryptograms for training the dataset.

(i) Write the solution for implementing Cryptosystem Identifier in MATLAB or Python. Give a brief description of what feature vectors you have used, how you designed the

machine-learning model for SVM and BPNN, and what loss function did you use in each case.

According to our study of the problem, the two ciphers mentioned in the problem have a characteristic difference in the character frequency distribution in the encrypted text.
The english language follows certain character distributions in general.For example ‘e’ is used maximum number of times(aroung 12.7%) in our speech or text followed by ‘t’ and other characters in order.

In simple substitution, the trend of frequency diatribution remains same with only difference being in the shuffled letters. That is, in encrypted text, some character would have a distribution of about 12.7% but that would not be ‘e’, but the trend of distribution would remain the same.

In vignere cipher, this trend of distribution gets disturbed which was originally the purpose of this cipher.Here the frequencies of characters would even out for all 26 letters.

So, as features we used the frequency distribution(all 26 characters) of encrypted texts in sorted order as the features of our model.

(ii) Compare the performance of the classifiers based on SVM and BPNN using test samples. Did you use a validation approach on the dataset? What performance metric did you use to compare the performance? Why is this a good metric?

The performances are:
SVM :100%
BPNN: 100%

and on our self generated test data it was about :
SVM: 100% or 99.75%
BPNN: 98.75 to 100%

Validation appraoches:

Due to less data to train, we first decided to use leave one out cross validation so as to not lose any samples.
But afterwards we used random subsampling. But results were almost the same in each case, and we got about 98-100% accuracy in both the cases.

Performance metric:
First, we used accuracy(correctly classified samples / total samples) as our performance metric. **As our accuracy was quite good so, we didnt feel the need for using precision/recall method** which is otherwise a better metric as it gives us a better idea our model and predictions(TP.TN,FP,FN) which can further help even analysis of out training data(Whether there were more samples of a particular type) and further different aspects.

(iii) Plot the performance of your system for SVM and BPNN by varying parameters in your model.

SVM:

1)Kernels
- linear: 96.5% accuracy
- rbf: 98.75-100% accuracy

2)Max iterations:
- <5: 70-90% accuracy
- 5-10: 90-96% accuracy
- >10 : 98 -100% accuracy

BPNN:

1)hidden layers:
- 1-3: 42-48% accuracy
- 3-5: 48-60% accuracy
- >5: >95% accuracy

2)activation function:
- tanh: 48-50% (occasionaly 90%) accuracy
- relu: >97% accuracy
- logistic: 90-97% accuracy

3)solver:
- lbfgs:95-99% accuracy
- sgd:60-90% accuracy
- adam:50-60% accuracy

4)Learning rate:

1e-5: >99% accuracy
1e-4: 97-99% accuracy
1e-3: <95% accuracy

You will be graded based on what you have submitted as well as your ability to explain your code.

4B. Designing IoT system (select either 4A or 4B)

Many applications such as robot navigation (wheeled robot for instance) require an estimate of where the obstacle is relative to the robot.

(i) Design a SONAR system using Arduino UNO that records the distance of the obstacle and the angle by which the sensor has rotated on the console.
Things you will require:

- Arduino UNO kit (https://www.amazon.in/Arduino-ATmega328P-ATMEGA16U2-Compatible-Cable/dp/B06XB81X82)
- jumpwires
- breadboard/PCB boards
- ultrasonic sensor HC-SR04 (https://www.amazon.in/Adraxx-HC-SR04-Ultrasonic-Distance-Measuring/dp/B01LXFUAFV)

(ii) Discuss the system you have designed with the following specifications:
[a] Explain the working principle behind the transceiver and how it measures the distance and angle.
The ultrasonic sensor uses sonar to determine the distance to an object.
the transmitter (trig pin) sends a signal: a high-frequency sound
when the signal finds an object, it is reflected and
the transmitter (echo pin) receives it.
By measuring the time required for the echo to reach to the receiver, we can calculate the distance.

[b] Plot a graph between the estimated distance (y-axis) and actual distance (x-axis)
Given in github repository.
[c] Discuss any parameter which affects the performance of the system in the plot obtained
Ambient Temperature(Although as our testing area was same , we didntobserve any significant changes)
[d] Find the workable ranges of obstacle resolution (minimum and maximum size of the objects which can be detected)
The workable range of of object resolution are: min size of object-0.5 cm(radius)

Submit this along with code files and readme in a .zip format or Github link. Also provide a demo video showing the results clearly.

(iii) Optional Part: Additional credits for novelty in circuit design (customised circuitry). Provide a blueprint of the circuit diagram using easyEDA (https://easyeda.com/) in case of customized circuitry. Can you construct a touch detection system using the same system which would convert it to give back the {x,y} coordinates of the point where touch is performed knowing the distance of the obstacle (finger in this case) and angle at which the sensor rotates. In case you give this a try include all necessary documentation and code files in .zip format.

5. Solving socio-economic problems using technology (10 points)

Select one of the two problems below:
(i) Analytics and alerts on road safety using car mounted dashboard cameras
(ii) Analytics and alerts on air pollution in Delhi using vision and IoT sensors

Discuss in about 500-600 words how you would design a solution for the problem you selected above. Your solution approach needs to consider the following parts:
a) datasets or data acquisition for training
b) choice of machine learning algorithm to run online or offline
c) what platform can be used to run machine learning algorithm (for e.g. Raspberry Pi, smartphone, cloud)
d) sending alerts over the network via peer-to-peer methods or cloud architecture.

This question is open-ended so you need to outline the design choices you will make. Include an architecture diagram and how you would measure the performance of the system you design. What demo can you show and what key challenges do you expect. (Note: Additional credits on out-of the box feasible and interesting ideas)

We would like to propose some points on how we would like to help towards raod safety.

Our cameras as proposed below would be connected to raspberry pi installed in the car which in turn would be conneced to an andriod mobile of the user. Our algorithms and computing would be done offline on the raspberry module(To speed up the response time) and alerts(in-car) would be in the form of sound signals or a small screen mounted on the dasahboard itself.
For other alerts, we would use the connected android mobile to send messages to preferred contacts.The connected mobile would also help to update our algorithms or make updates on the rapberry installed.

1 .Dashboard mounted camera facing the driver:
We would record the drivers face at regular intervals of time and using opencv and other algorithms do an expression analysis on the driver on wether he looks weary or not.
Another point being we would scan the eyes of driver for redness which would help us identify if he is drunk or is too tired to drive.This would also help us to calculate the percentage of time driver pays attention to the road.

We would propose a scoring system based on these results with parameters being:
(i).Face expression whether drunk or not.
(ii).Redness in the eyes.
(iii)Percentage of time driver was paying attention on the road.
(iv).The speed of car.(This would be in combination with above point...If spped is more, driver would need to pay more attention on the road)

When the threshold scoring value is crossed we would alert the driver with sound signals and preferably a message to emergency contacts submiited initially by the driver.(**Special focus would be made on removing false positives and incorporating tottaly obvious cases**)

For even better results we have API(s) from different organisations which have done work in computer vision. For Example Kairos has a face expression analysis API and using open cv inbuilt modules or by using deep learing methods we can scan and analyse the eyes of the driver.

2.Android application connected to our raspberry pi in the car:
The drivers mobile would have to be connected with the raspberry in the car and a specialized application would be made which would serve the following points:
(i)An automated voicemail would be sent to the caller(in case a someone calls the driver during the drive) saying that the person is driving and cant pickup your call.If the person calls again, then we would move on to the next point
(ii)The driver cant pickup a call unless the mobile is connected to bluetooth system in the car or the car is not moving.

3.Road facing cameras(2) or a camera combined with a ultrasonic sensor(Kind of a parking sensor) to measure distance of objects in front of thr car:
Here, we would use the cameras to identify objects in front of the car.Special focus on identifying people and other cars most accurately and calculating the distance from the car.

Here, we would have to collect data from the car companies on the stopping distance of different cars according to their current speed.

If the margin of distance and stopping time for tha speed is reducing we would first flash alerts on the dashboard mounted screen(so as to not scare the driver) but after its completely necessary sound alerts would be given for the driver to apply the breaks.

For even more accuracy we can look into the auto breaking algorithms used by companies like Mercedes and Ford in their cars(Various Research papers have also been published on the subject) for our alerts to be more accurate and real time.

4.Black box for cars(Recordings,speed and location data)

The raspberry would store the above mentioned data and in case of an accident, we could use this to analyse what went wrong .Not only in case of an accident, but this would also help in our data collection for further uses and also would help the police in analysis of a crime (if it was a case).

The only drawback for this method would be that storage on a raspberry is limited and we would have to upload the data to cloud in regular intervals which would require data services from the connected mobile mentioned above.