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.**

Team Members:

1. Raghav Bansal (virgoraghavbansal@gmail.com)
2. Shreya Srivastava (shreyasrivastava2509@gmail.com)
3. Akshath Singhal (singhal.akshath97@gmail.com)
4. Raman Aggarwal (raman.dtu.a7@gmail.com)

Github Link:

https://github.com/bansalraghav/Project-Celestini

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: [a] (i) Supervised Learning, (ii) Unsupervised 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:[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
[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] cost
[c] small form factor
[d] low power consumption

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: [c] When we increase the complexity of a model, it will always decrease the train error.

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: [c] The model has poor accuracy, maximum precision and minimum 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++

Python code:

```
# function to multiply two sparse matrices
def sparseMultiply(mat1,mat2):
    for i in range(len(mat1)):
        for j in range(len(mat1)):
            for k in range(len(mat1)):
                c[i][j] += mat1[i][k]*mat2[k][j]
    return c



a = [[0,2,0],
     [1,0,0],
     [0,0,5]]

b = [[100,0,0],
     [0,100,0],
     [0,0,100]]

c = [[0,0,0],
     [0,0,0],
     [0,0,0]]

print(sparseMultiply(a,b))
```

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

O(N^3)

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

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++.

Python code:

```
# function to find the number from the array
def findNumber(b,element):
    m = 0
    n = len(b[0])-1
    # loop to find the number(starting from the top right corner)
    # Condition to not let m and n become out of bounds
    while(m < len(b) and n >= 0):
        if(b[m][n] == element):
            print("The element is found at position: ",m, ",", n)
            return 0
        elif(b[m][n] > element):
            n = n - 1
        else:
            m = m + 1

    print("The element is not found at any position.")
    return 0

# Input matrix
a = [[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]]
# Element to be searched
x = 50
```

```
# Matrix and element passed into the function
findNumber(a,x)
```

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

O(m + n)

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

O(1)

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. Crytosystem 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/matlabcentral/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.

Solution:

Frequency of the alphabets all the alphabets is found and sorted in plain text as well as cipher text

These frequency are the 26 features

Python's scikit library is used to train Support Vector Classifier with linear kernel

Tensorflow library is used to train a basic feed forward neural network which uses cross entropy error as the cost function

(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?

Solution:

Accuracy of SVM was found to be higher and to analyse the are classifier we plotted learning curves the check if the classifier is high biased or high variance exists

In case of Neural Network Tensorboard was to analyze training of the network by plotting histogram of weights, and learning rate

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

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:

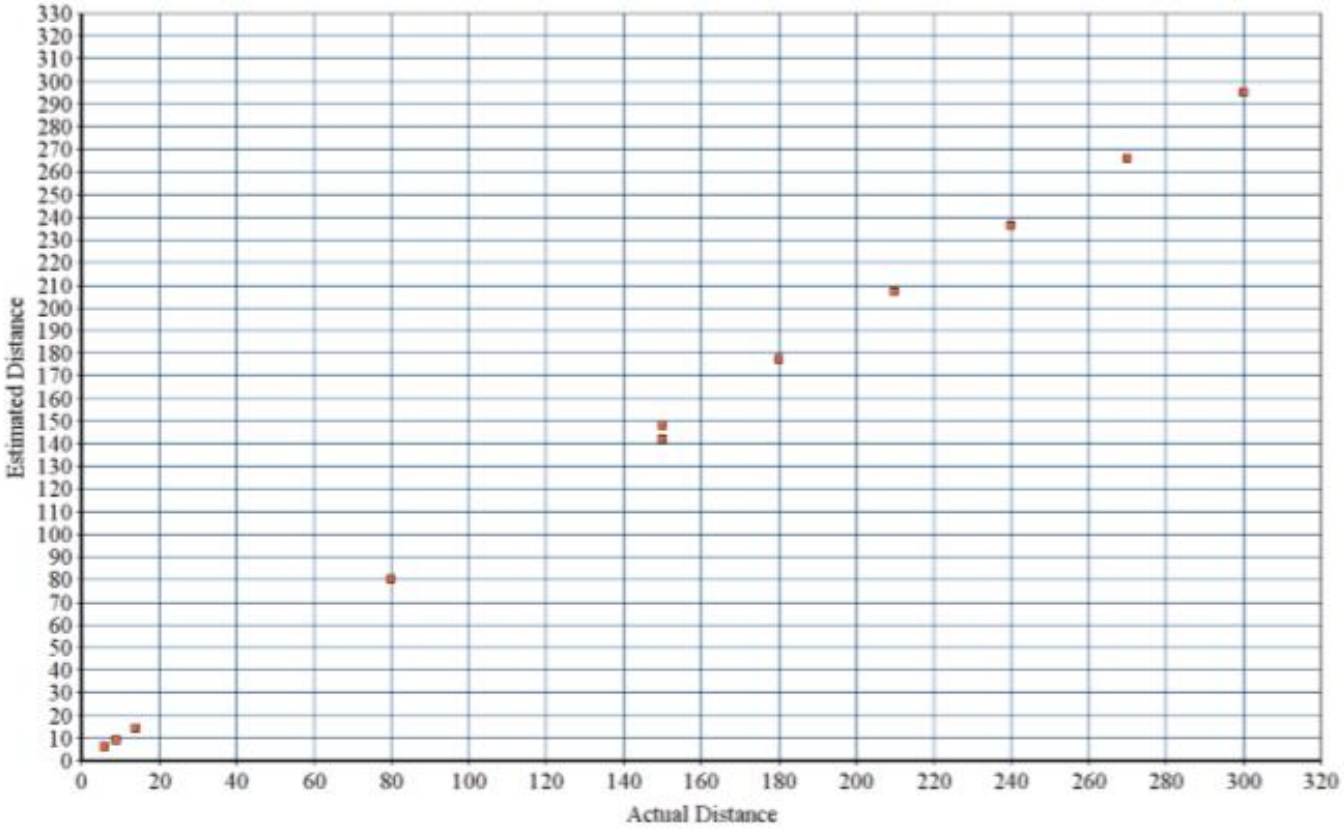
- Arduino UNO kit (https://www.amazon.in/Arduino-ATmega328P-ATMEGA16U2-Compatible-Cable/dp/B06XB81X82)
- jumpwires
- breadboard/PCB boards
- ultrasonic sensor HC-SR04 (https://www.amazon.in/Adraxx-HC-SR04-Ultrasonic-Distance-Measuring/dp/B01LXFUAFV)

(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
[b] Plot a graph between the estimated distance (y-axis) and actual distance (x-axis)
[c] Discuss any parameter which affects the performance of the system in the plot obtained in [b]
[d] Find the workable ranges of obstacle resolution (minimum and maximum size of the objects which can be detected)
Submit this along with code files and readme in a .zip format or Github link. Also provide a demo video showing the results clearly.

Answers:

[a] The working principle of the transceiver is similar to radar or sonar. It generates high frequency sound (40,000 Hz) and calculates the time interval between the sending of signal and receiving of echo.One ping of the sonar module consists of 8 pulses of 40KHz. To start a ping we need to provide a 10microseconds pulse on the trigger pin. The speed of sound is known to us(340 m/s). Hence we can calculate the distance between the sonar module and object by the formula:
Distance = (speed of sound * time interval)/2

Graph showing values measured by Ultrasonic Sensor



[b]

X(actual distance in cm)	6	9	14	80	150	180	210	240	270	300
Y(estimated distance in cm)	6.01cm	9.07	14.07	80.2	148	178	207	236	266	295

[c] Objects close to the sonar are measured with greater accuracy but objects further away have more error than the least count. To reduce this error, we may consider objects that are perpendicular to the line of sight of the sonar for greater accuracy.
The object should be stable and should be well within the working range of the sonar i.e should be within ~21 degrees (FOV) and ~4 degrees for spatial resolution.
Also signal from the sonar bounces from table or floor (or wherever it is put) if we measure larger distances. Hence it is better to keep the sonar on the edge so as to avoid the bouncing of signals.

[d] Upto 300cm:

Objects as big as a wall or pillar can be detected with error of 2-5 cm.

Beyond 300cm:

Large objects like pillars and guitar were detected but with an error of around 10cm.

(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.

Answers:

EasyEda design circuit added in the Question 4B folder.
We tried to make the touch detection system using a servo motor and LCD(picture included in the folder). We placed the sonar module on top of the servo which is rotating back and forth from 0 to 180 degrees. If the sonar has a large empty space in front of it which is greater than the workable range of the sonar then it will return junk values. As soon as we introduce a finger it will show us a value on the LCD. Using polar coordinates we can calculate the x and y coordinates as follows:

r = distance measured by the sonar
Theta = angle the servo has rotated
X coordinate = r*cos(theta)
Y coordinate = r*sin(theta)

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)

Solution:

The topic which we have chosen is
(i)Analytics and alerts on road safety using car mounted dashboard cameras

A rasberry pi or the users smartphone can be connected with the dashboard mounted camera and it will send alerts of the objects in front. Using machine learning we can train our model to handle different objects. Using image recognition we can also identify the type of object in front of the camera.(example cat,dog, traffic light etc).
If the car enters a speed limiting zone then we can also provide alerts on the speed the car has to maintain.

If the camera's range is x then we can also send alerts based upon the time the car takes to stop after applying brakes.(brakes applied early if car is moving fast).

Nowadays cars have charging ports inbuilt and that can be used to power the device or a power bank of sufficient mAh value can be attached which can bypass the charging in case the car is turned off.

(a)We will use dashboard mounted cameras for image acquisition and traffic cameras connected to online server.
(b)The machine learning model will run offline
(c)Any platform can be used to run the machine learning algorithm(Rasberry Pi, smartphone)