Ex/PG/CRSE/T/128A/2024

M. TECH IT (COURSEWARE ENGG) EXAMINATION
First Year Second Semester - 2024
LEARNING ANALYTICS AND EDUCATIONAL DATA MINING

TIME: 3 Hours — Full Marks: 100

Answer any ***five*** questions.
[All the question parts must be answered together]

1. a) Briefly explain with an example why normalisation is required in data pre-processing? — 5+5+5+5
   b) What are the different methods of handling noisy data in machine learning?
   c) Briefly illustrate with an example which data mining method is used to determine the sequential dependencies among different events.
   d) Which plot is used to detect outliers in machine learning?

2. a) How can the overfitting issue be addressed if the classification accuracy of the training data is 80% and the accuracy of the test data is 60%? — 5+5+5+5
   b) A data scientist develops a binary classifier to predict whether a patient has a particular disease based on a series of test results. The data scientist randomly selected 400 patients' data from the population. The disease affects 3% of the population. Which cross-validation strategy should the data scientist adopt, and why?
   c) Why is Adjusted R-square preferred over R-square in regression problems?
   d) How is the ROC curve used for multiclass classification problems?

3. a) Describe the methods used to avoid overfitting in decision trees. — 5+5+10
   b) Briefly explain the properties of Gini Impurity for the following 3 cases.
   Case 1: When 100% of observations belong to the positive class
   Case 2: When 50% of observations belong to the positive class
   Case 3: When 0% of observations belong to the positive class
   c) Construct a decision tree based on the given training data. For splitting, use information gain as a measure for impurity. Build a separate branch for each attribute. The decision tree shall stop when all instances in the branch have the same class.

[Turn over

ID	time	gender	area	risk
1	1-2	m	urban	low
2	2-7	m	rural	high
3	>7	f	rural	low
4	1-2	f	rural	high
5	>7	m	rural	high
6	1-2	m	rural	high
7	2-7	f	urban	low
8	2-7	m	urban	low

4. a) Write a Python program to nest 3 dictionaries, dict1, dict2, and dict3, under a parent dictionary called dictparent. 5+5+10

b) Consider the following Series object, 'company,' and its profit in 'Crores.'

Company	Crores
TCS	350
Reliance	200
Infosys	800
Wipro	150

Write the Python code to display the company's name with a profit>250. Add 100 to all the elements and display the Series object 'company'.

c) Write the output of the following Python code snippets.

I.

```
List1 = [1, 2, 3]
List1 = list1 + 2
print(list1)
```

II.

```
x = ("apple", "banana", "cherry")
y = list(x)
y[1] = "kiwi"
x = tuple(y)
print(x)
```

III.

```
import numpy as np
a=np.array ([[11,2,3,4],[10,20,30,40]])
print (a)
print (a[1][2])
print (a[1, 2])
```

IV.
```
def printWords(s):
s = s.split(' ')
for word in s:
if len(word)%2==0:
          print(word)
s = "i am GRABY"
          printWords(s)
```

V.
```
import pandas as pd
import numpy as np
s = pd.Series(np.arange(10,50,10))
print(s)
print (s.ndim)
print(s.shape)
print(len(s))
```

5. a) How does the cost function help to determine the minimum values for θ_0 and θ_1 in linear regression? Why is a partial derivative used in a linear regression model? 10+8+2

b) Consider a classification problem of predicting whether a photograph contains a man or a woman. A test dataset contains 10 records, their expected outcomes, and predictions from a classification algorithm. (i) Compute the confusion matrix for the data. (ii) Calculate the accuracy, precision, recall, and specificity of the prediction.

	Actual	Predicted
1	Man	Woman
2	Man	Man
3	Woman	Woman
4	Man	Man
5	Woman	Man
6	Woman	Woman
7	Woman	Woman
8	Man	Man
9	Man	Woman
10	Woman	Woman

c) Why is linear regression not suitable for the classification of binary classes?

6. a) What are the criteria for terminating the K-means clustering algorithm? Why is the silhouette method preferred over the elbow method in finding the optimal number of clusters? 5+5+10

b) For the given data, compute two clusters using the K-means algorithm for clustering where initial cluster centroids are (1.0, 1.0) and (5.0, 7.0). Show the cluster calculations for two iterations using Euclidean distance.

S. No	A	B
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

7. a) In Logistic Regression, $h_\theta(x)$ is a nonlinear function, and the hypothesis $h_\theta(x) = \sigma(z) = \frac{1}{1+e^{-(\theta_0 + \theta_1 x)}}$ gives a non-convex function- Justify the statement. 10+5+5

b) Explain each of the terms in the following equations.

i. $\log loss = J(\theta) - \frac{1}{m}\sum_{i=1}^{m}[(y^{(i)}\log(\boldsymbol{h_\theta}(\mathbf{x}^{(i)})) + (1 - y^{(i)})\log(1 - \boldsymbol{h_\theta}(\mathbf{x}^{(i)}))]$

ii. $\theta_j := \theta_j - \propto \frac{1}{m}\sum_{i=1}^{m}(h_\theta(\mathrm{x}^{(i)}) - \mathrm{y}^{(i)})\, \mathrm{x}_j{}^{(i)}$

8. Write short notes on *(ANY FOUR)* 5x4

a) Association rule mining

b) Transactional data

c) Heat map

d) Posterior probability

e) Adaptive and Intelligent Hypermedia System