Python for data science

Week 4 assignment

1. Which of the following are regression problems? Assume that appropriate data is given. [1 mark]

   (a) Predicting the house price.

   (b) Predicting whether it will rain or not on a given day.

   (c) Predicting the maximum temperature on a given day.

   (d) Predicting the sales of the ice-creams.

   Answer: a, c, d

2. Which of the followings are binary classification problems? [1 mark]

   (a) Predicting whether a patient is diagnosed with cancer or not.

   (b) Predicting whether a team will win a tournament or not.

   (c) Predicting the price of a second-hand car.

   (d) Classify web text into one of the following categories: Sports, Entertainment, or Technology.

   Answer: a, b

3. If a linear regression model achieves zero training error, can we say that all the data points lie on a hyperplane in the (d+1)-dimensional space? Here, d is the number of features. [1 mark]

   (a) Yes

   (b) No

   Answer: a

Read the information given below and answer the questions from 4 to 6:

Data Description:

An automotive service chain is launching its new grand service station this weekend. They offer to
service a wide variety of cars. The current capacity of the station is to check 315 cars thoroughly per

day. As an inaugural offer, they claim to freely check all cars that arrive on their launch day, and
report whether they need servicing or not!

Unexpectedly, they get 450 cars. The servicemen will not work longer than the working hours, but the data analysts have to!

Can you save the day for the new service station?

How can a data scientist save the day for them?

He has been given a data set, **'ServiceTrain.csv'** that contains some attributes of the car that can be easily measured and a conclusion that if a service is needed or not.

Now for the cars they cannot check in detail, they measure those attributes and store them in **'ServiceTest.csv'**

Problem Statement:

Use machine learning techniques to identify whether the cars require service or not

Read the given datasets 'ServiceTrain.csv' and 'ServiceTest.csv' as train_data and test_data respectively and import all the required packages for analysis.

4. Which of the following machine learning techniques would NOT be appropriate to solve the problem given in the problem statement? [1 mark]

   (a) kNN

   (b) Random Forest

   (c) Logistic Regression

   (d) Linear regression

Answer: d

Prepare the data by following the steps given below, and answer questions 6 and 7.

- Encode categorical variable, Service - Yes as 1 and No as 0 for both the train and test datasets.

- Split the set of independent features and the dependent feature on both the train and test datasets.
- Set **random_state** for the instance of the logistic regression class as 0.

5. After applying logistic regression, what is/are the correct observations from the resultant confusion matrix? [1 mark]

   (a) True Positive = 29, True Negative = 94

   (b) True Positive = 94, True Negative = 29

   (c) False Positive = 5, True Negative = 94

   (d) None of the above

   Answer: a, c

6. The logistic regression model built between the input and output variables is checked for its prediction accuracy of the test data. What is the accuracy range (in %) of the predictions made over test data? [1 mark]

   (a) 60 - 79

   (b) 90 - 95

   (c) 30 – 59

   (d) 80 – 89

   Answer: b

7. How are categorical variables preprocessed before model building? [1 mark]

   (a) Standardization

   (b) Dummy variables

   (c) Correlation

   (d) None of the above

   Answer: b

The Global Happiness Index report contains the Happiness Score data with multiple features (namely the Economy, Family, Health, and Freedom) that could affect the target variable value.

Prepare the data by following the steps given below, and answer question 8

- Split the set of independent features and the dependent feature on the given dataset
- Create training and testing data from the set of independent features and dependent feature by splitting the original data in the ratio 3:1 respectively, and set the value for **random_state** of the training/test split method's instance as 1

8. A multiple linear regression model is built on the Global Happiness Index dataset **'GHI_Report.csv'**. What is the RMSE of the baseline model? [1 mark]

   (a) 2.00

   (b) 0.50

   (c) 1.06

   (d) 0.75

   Answer: c

9. A regression model with the following function $y = 60 + 5.2x$ was built to understand the impact of humidity (x) on rainfall (y). The humidity this week is 30 more than the previous week. What is the predicted difference in rainfall? [1 mark]

   (a) 156 mm

   (b) 15.6 mm

   (c) -156 mm

   (d) None of the above

   Answer: a

10. X and Y are two variables that have a strong linear relationship. Which of the following statements are incorrect? [1 mark]

    (a) There cannot be a negative relationship between the two variables.

    (b) The relationship between the two variables is purely causal.

    (c) One variable may or may not cause a change in the other variable.

    (d) The variables can be positively or negatively correlated with each other.

    Answer: a, b