csd3186s25-a.sg

Started on	Monday, 10 February 2025, 10:22 AM
State	Finished
Completed on	Monday, 10 February 2025, 10:55 AM
Time taken	32 mins 53 secs
Grade	**17.00** out of 20.00 (**85**%)

Quiz navigation

1 ✔ 2 ✔ 3 ✔ 4 ✔ 5 ✔ 6 ✔ 7 ● 8 ✔ 9 ✔ 10 ✖

Show one page at a time

Finish review

Question 1 Correct. Mark 2.00 out of 2.00. Flag question

What does "Overfitting" refer to in Machine Learning?

Select one:

- A model that performs badly on both training and test data
- A model that generalizes well to new data
- A model that ignores patterns in the data
- A model that performs well on training data but poorly on new data ✔

Your answer is correct.

The correct answer is: A model that performs well on training data but poorly on new data

Question 2 Correct. Mark 2.00 out of 2.00. Flag question

What is the primary objective of Linear Regression?

Select one:

- Identify clusters in a dataset
- Classify data into predefined categories
- Minimize false positives in classification
- Find the relationship between independent and dependent variables ✔

Your answer is correct.

The correct answer is: Find the relationship between independent and dependent variables

Question 3 Correct. Mark 2.00 out of 2.00. Flag question

What type of problems is Logistic Regression commonly used for?

Select one:

- Clustering problems
- Regression problems
- Classification problems ✔
- Dimensionality reduction

Your answer is correct.

The correct answer is: Classification problems

Question 4 Correct. Mark 2.00 out of 2.00. Flag question

What is the key idea behind K-Nearest Neighbours (KNN)?

Select one:

- It applies linear transformations to reduce dimensions
- It builds a mathematical function to predict outputs
- It constructs a tree structure to make decisions
- It classifies data points based on the majority class among their nearest neighbours ✔

Your answer is correct.

The correct answer is: It classifies data points based on the majority class among their nearest neighbours

Question 5 Correct. Mark 2.00 out of 2.00. Flag question

What is the main advantage of Decision Trees?

Select one:

- They always outperform other models
- They do not require labelled data
- They are resistant to overfitting
- They are easy to interpret and visualize ✔

Your answer is correct.

The correct answer is: They are easy to interpret and visualize

Question 6 Correct. Mark 2.00 out of 2.00. Flag question

A simple linear regression model is given by the equation:

$$Y = 41.5 + 7.5X$$

where:

- **41.5** is the intercept (b)
- **7.5** is the slope (W1)

(a) Using the given regression equation, predict the exam score when a student studies for **6 hours**:

Predicted score : 86.5 ✔ (Please round your answer to 1 decimal)

(b) The residual for each data point is calculated as:

Residual = Y_actual - Y_predicted

If a student studied **3 hours** and the actual exam score was **65**, calculate the residual using the given regression equation.

Residual : 1.0 ✔ (Please round your answer to 1 decimal)

Question 7 Partially correct. Mark 1.00 out of 2.00. Flag question

The following code snippet is incomplete for obtaining the accuracy of a logistic regression model. Complete the code by replacing "..." with your solution.

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
import numpy as np

X_train = np.array([[0], [1], [2], [3], [4]])
y_train = np.array([1, 0, 1, 1, 1])

lgr = LogisticRegression()
lgr.fit(X_train, y_train)

# TODO: get the accuracy for training data
y_pred = ...
accuracy = ...

print("Accuracy on training data is", accuracy)
```

ANSWER:

y_pred = lgr.predict(X_train) ✖

accuracy = accuracy_score(y_train,y_pred) ✔

Question 8 Correct. Mark 2.00 out of 2.00. Flag question

You are given the following training dataset with two features:

Point	X_1	X_2	Class
A	2	3	0
B	5	6	0
C	8	9	1
D	4	7	1
E	6	2	0

A new test point **T(4,5)** needs to be classified using the **KNN algorithm**.

Compute the **Euclidean distance** between the test point **T(4,5)** for the first TWO points in the dataset only. (round to 2 decimal points)

Distance to A : 2.83 ✔

Distance to B : 1.41 ✔

Question 9 Correct. Mark 2.00 out of 2.00. Flag question

You are given a dataset used for binary classification. The root node of the decision tree contains all the data, and its entropy is **0.94** . You are considering two candidate features, **Feature A** and **Feature B** , to split the data at the root node. The split results in the following child nodes:

Split on Feature A:

- **Left Child:** 40 samples with an entropy of **0.00**
- **Right Child:** 60 samples with an entropy of **0.97**

Split on Feature B:

- **Left Child:** 50 samples with an entropy of **0.72**
- **Right Child:** 50 samples with an entropy of **0.72**

Compute the Information Gain (IG) for each feature using the formula:

$$IG = Entropy(parent) - Weighted\ sum\ of\ Entropy(children)$$

And also decide which feature provides the better split based on the IG.

IG (Feature A) = 0.36 ✔ (round to 2 decimal place)

IG (Feature B) = 0.22 ✔ (round to 2 decimal place)

Best split = A ✔

Question 10 Incorrect. Mark 0.00 out of 2.00. Flag question

Given the following dataset:

```
# Training data
X_train = [
    [1, 3],
    [2, 2],
    [5, 5],
    [8, 10]
]
Y_train = [0, 0, 0, 1]
# Test data
X_test = [5, 6]]
```

And a partially implemented KNN class:

```
import numpy as np
from collections import Counter

class KNearestNeighbors:
    def __init__(self, k=3):
        self.k = k

    def fit(self, X, y):
        self.X_train = X
        self.y_train = y

    def predict(self, X):
        X = np.array(X)
        y_pred = [self._predict(x) for x in X]
        return np.array(y_pred)

    def _euclidean_distance(self, x):
        # YOUR CODE HERE

    def _predict(self, x):
        distances = self._euclidean_distance(x)

        k_indices = np.argsort(distances, kind='mergesort')[:self.k]
        print(f"for {x}, indices of {self.k} nearest neighbours in training dat

        k_nearest_labels = [self.y_train[i] for i in k_indices]
        most_common = Counter(k_nearest_labels).most_common(1)
        return most_common[0][0]

# Will be used to train a KNN model as per the following
knn = KNearestNeighbors()
knn.fit(X_train, Y_train)

knn.predict(X_test)
```

Study the **KNearestNeighbors** class above and define the **_euclidean_distance** method inside the class.

_euclidean_distance(self, x) : Computes the Euclidean distance between a test sample and all training samples.

- **Inputs:**
- x: An 'ndarray' containing the test sample

- **Outputs:** Returns the list of Euclidean distance between x and training samples.

```
def _euclidean_distance(self, x):
    i dunno
```

Incorrect

Evaluation details:

Evaluation:

-Test 1: 1 (-1.000)

```
Incorrect program output
 --- Input ---

 --- Program output ---
>  File "submit.py", line 32
>    i dunno
>          ^
>SyntaxError: invalid syntax

 --- Expected output (exact test)---
>for [5 6], indices of 3 nearest neighbours in training data:
>[2 0 1]
>

-Summary of tests

>+----------------------------+
>|  1  Test run/ 0 tests passed |
>+----------------------------+
```

Possible solution:

```
def _euclidean_distance(self, x):
    return [np.sqrt(np.sum((np.array(x_train) - x) ** 2)) for x_train in self.X_train]
```

Finish review