GUJAR MEMORIAL CHA...

DR. ASHOK GUJAR TECHNICAL INSTITUTE'S
DR. DAULATRAO AHER COLLEGE OF ENGINEERING, KARAD

UNIT TEST EXAMINATION - I / II / III

Name of Student: Bagwan Soheb Jabbar

Class: Batch No - D9 Division:

Roll No. Date: 14/2/24

Subject: Assesment 3. Signature of Supervisor:

Que. No.	1	2	3	4	5	6	7	8	9	10	Total Marks
Marks Obtained											51/100

(Sign. of Subject Teacher)

Section A

Q.1] What is the difference a list & a tupple in python?

→

list :- ① lists are mutable : you can modify, add or remove element after creation

② list defined using square brackets

e.g. list = [1, 2, 3]

Tuple :- ① Tuple are immutable : once created, their element cannot be changed.

② Tuple use parentheses

e.g. Tuple = (1, 2, 3)

Example use cases

1] Use list when mutability matter

e.g. - Managing a dynamic collection of user input

User data = [' John, 25]

User data = 26 # modify age

2] Use tuple for Immutable data.

coordinates : (10, 20)

Choose list when you need a collection that can be modified & tuple when you want a fixed unchangable set of value.

Q2] Write a python function to calculate the factorial of given number.

→ def Factorial (n):
    If (n) == 0 or n == 1:
        return 1
    else:
        return n* factorial (n-1)

2

* Example usage:
    number s
    result = factorial (number)
    print (f" The factorial of {number}: {result}

Q.3] Explain the concept of list comprehension in python. provide an example of how it can be used to create list

→ List comprehension is a concise way to create lists a python. It provides a more readable & compact syntax for generating lists by specifying the element to include & the condition to meet. The basic structure is

3 [expression for item in trable if condition]

e.g:-
Here an example that generates a list of squares for even number betn 0 & 9 using list compression

Q.4 Briefly explain the purpose of the following Python libraries: Numpy, pandas & Matplotlib

→ 1] Numpy:-

- Numpy is a powerful library for numerical computing in python It provides support for large, multi-diamtional arrays & matrices along with a collection of mathematical functions to operate on these arrays efficiency.

2] Pandas:-

- Pandas is a data manipulation & analysis library that provides easy-to-use data structures, such as Dataframes, for working with structured data.
- It simplifies task like cleaning, filtering & aggregating data.
- widly used in data analysis & preparation, handling missing data, & performing operations on structured dataset.

3/2

3] Matplotlib:-

- Matplotlib is a 2D plotting library for creating static animated & interactive visualization in python. It allows user to create a wide variety of charts & graphs to visualize data.
- Ideal for data visualization in scientific computing & statistical analysis.

Section B

Q.17 Define supervised learning & unsupervised learning provide an example of each.

→ Supervised learning:-

- supervised learning is a type of machine learning where the algorithm is trained on a labled dataset. In this approach the model learns to map input data to corresponding output lables.

Example:-

- Task - Image classification [distinguishing bet cats & dogs]
- Dataset - collection of images with labels
- Training - The algorithm learns the patterns & features associated with each class from image
- Testing - Once trained the model can predict the lables of new images.

Unsupervised learning:-

- Unsupervised learning involves training a algorithm on an unlabeled dataset.

Example:-

- Task - clustering customer preferences in an e-commerce platform.
- Dataset:- Purchasing history data without label. Categories
- Algorithm - The model idenifies natural grouping or clusters based on similarities in purchasing ~~beh~~ behaviour.
- output - clusters or segments of customers with similar preferences emerge from analysis.

Q.3] Describe the steps involved in the machine learning pipeline provide a brief explanation of each step

→ Steps:-

① Data collection:-

Gather relevent data for the problem at hand the quality and quantity of data significantly impact model performance

② Data cleaning & processing:-

clean & preprocess the data to handle missing values outliers, & in consistencies convert data into a suitable format for training model

③ Feature Engineering:-

Create new features or transform existing ones to enhance the model's ability to capture patterns in the data. This step helps improve model performance.

④ Data spliting:-

Divide the dataset into training & testing sets.

⑤ Model selection:-

chose a machinelearning algorithm & model The selection nd upon nature & characteristics of data.

⑥ Model training:-

Train the selected model using the training data.

⑦ Hyperoparameter Tuning:-

fine-tune the hyperparameter of the model to optimize its performance.

⑧ Model evaluation:-

Assess the model performance

(9) Model deployment:-
integrate model into production environments.
(10) Monitoring & maintenance.

Q.4] What is cross-validation & why is it important in machine learning provide an example of a cross validation.

→ Cross-validation:-
- Cross validation is a resampling technique used in a machine learning to assess the performance of a model & to reduce the risk of overfitting or underfitting.
- It involves partitioning the dataset into subsets training the model on some of these subsets & evaluting on the remaing subsets

Importance of cross validation.
1) Better performance Estimation:- Cross-validation provides a more realible estimate of how well a model will generlize to unseen data compared to a single train-test split.
2) Reduced Overfitting or underfitting Risk:- By evaluation the model on multiple subsets, cross validation helps in identifying if model is overfitting & underfitting

example:-
from sklearn.model_selection import kfold from sklearn model-selection import cross-val-score from sklearn ensemble import RandomForest classifier

Example using kfold cross-validation with Random forest classifier

```
model = Random-forest classifier()
kf = KFold (n-splits = 5, shuffle = True,
            random. state = 42)
```

* performance cross-valitation & calculate accuracy

```
accaracy scores = cross_val_score(model, x, y, cv=kf,
                  scoring = 'accuracy')
```

Display average accuracy

```
print ("Average Accuracy", accuracy_scores.mean())
```

Q.5] Differentiate betn regression & classification in the context of machine learning provide an example for each.

→

Regression ~~algorithm helps~~	classification
1] The output variable has to be real value or continuous in nature.	1] The output variable in discrete in the classification SML probelm
2) Regression algorithm helps to map the input value & the continous output variable.	2] classification algorithm helps in mapping the input value with the output variable which is discrete in nature
3) Regression algorithms are only used for data that is continous	3) classification algorithm are only used for data is decrete
4) linear regression, decision trees & netural network are common used for tasks. example:- predicting house prices based on features like square footage, number of bedrooms location.	4) Logistic regression, decision trees, support vector machine & neural network common task examples- Identifying whether anemail is spam or not based on its content & characteristic.

Q.6] Briefly explain the k-nearest neighbors (kNN) algorithm how does it work, & what are its main parameters &

→ The k-nearest neighbors (kNN) algorithm a simple & versatile supervised machine learning algorithm used for both classification & regression tasks. It make predictions based on the majority class or average of the k-nearest data points in the future space

How kNN works:-

1] Training:-

The algorithm stores the entire training dataset in memory.

2) prediction (classification):-

- for a new data point, the algorithm identifies the k-nearest neighbors from the training set based on a distance metric.

- The majority class among these neighbors is assigned to the data point.

3) prediction:- (Regression)

- for regression task, the algorithm calculates the average of the target values of the k-nearest neighbors & assigns it to the new data point

Main parameter.

① Number of Neighbors (k):-

② Distance Metric:-

Measure similarity or dissimilarity betn data points

③ Weighting of Neighbors:-

Defines the contribution of each neighbor to the prediction

PAGE NO.:
DATE / /

1) mean :- i) It is the average of a set of values calculates by summing all the values & dividing by the total number of value.

i) mean = $\Sigma_i = \frac{1}{n} x_i$ ii) The mean is most appropriate when the data is symmetrically distributed & does not contain extreme outliers. It sensitive to extreme values (outliers) & can be influenced by them, which may lead to a biased estimate of central tendency.

2) median :- i) It is the middle value of a sorted list of values. If the Num of values is odd, the median is even & the values is even median average of the middle values.

3) Mode :- i) It is the value that appears most frequently in a dataset. 2) A data can have one mode (unimodal), two modes (bimodal) or more than two modes (multi modal)

iii) The mode is most appropriate for categorical or nominal data, where calculating the mean & median may not be meaningful It is also useful for idenifying the most common value in dataset.

Q.2] A dataset has a standard deviation of 10. If a data point is 2 standard deviations above the mean what percentage of the data is below this point in a normal distribution?

→ In a normal distrubution, approximatly 68% of the data falls within one standard deviation of the mean. 95% falls within three standard deviation.

If a data point is 2 standard deviations above the mean, we are looking at the top 2.5% of the distribution (because 95% is within two std deviation, leaving 5% in the tails, & we're interested in one tail

Page No :
Date : \ \

So, the percentage of data below this point is
100% - 2.5% = 97.5%.

Therefore, approximately 97.5% of the data is below a data point that is 2 standard deviations above the mean in a normal distribution.

Q.4) Describe the diffrence betn correclation & causation provide an example to illustrate your explanation

→ Correlation :-

Defination → Correlation measures the statistical relationship betn two variables. If two variables are correlated, it means that a change in one variable is assosiated with a change in ~~one variable~~ the other but it doesn't imply a cause & effect relationship.

Example - Ice cream sales & the number of drowings are positively correlated. However, it doesn't mean that buying more icecream causes more drowings or vice versa

2. Causation :-

Defination → Causation implies a cause & effect relationship betn two variable. A change in one variable directly causes a change in order

Example → The statement "smoking causes lung cancer. asserts a causal relationship betn smoking & the development of lung cancer

PAGE NO.:
DATE / /

Q.5) A random variable X follows a normal distribution with a mean of 50 & a standard deviation of 8, calculate the z-score for a value of $x = 58$

→ The z-score (or standard score) for a particular value in a normal distribution is a measure of how many standard deviation that value is from the mean

$$z = \frac{x - \mu}{\sigma}$$

In this case

$x = 58, \ \mu = 50, \ \sigma = 8$

Now plug these values into formula.

$$z = \frac{58 - 50}{8}$$

$$= \frac{8}{8}$$

$$z = 1$$

So, the z-score for $x = 58$ in this normal distribution is 1

3

Section D

Q.1] Explain the concept of overfitting in machine learning How can it be mitigated?

→ Overfitting occurs when a machine learning model learns the training data too well, capturing noise & fluctuations in the data rather than the underlying patterns. This can lead to poor generalization performance on new unseen data. Overfitting is characterized by a model that performs well on the training set but poorly on test & validation data.

Mitigation Strategies for overfitting:-

1) Cross Validation :- Use techniques like cross-validation to assess the model's performs on different subsets of the data.

2) Train with more data :-
Increasing the size of the training dataset can help the model learn more robust patterns & reduce the chance of memorizing noise.

3) Simplify the model :-
choose a simpler model with fewer parameters. This helps prevent the model from being too flexible & capturing noise.

4) Regularization :-
Regularization adds a penalty term to the loss function, discouraging extreme parameter values.

5) Feature selection :-
carefully choose relevant features & eliminate unnecessary ones.

PAGE NO.:
DATE / /

6] Ensemble methods:-

Random forests or gardient boosting these methods combine predictions from multiple models often reducing overfitting.

7) Dropout

8) Early stopping:-

Moniton the model's performance on a validation set during training.

Q.2) Briefly describe the support vector machine (SVM) algorithm. What is the role of the kernel in SVM

→

support vector machine (SVM)

Defination:- SVM is a supervised machine learning algorithm used for classification & regression tasks. It works by finding the hyperplane that best separates different classes in the feature space.

Role of the kernel in SVM

- Linear SVM:-
  - In its simplest form, SVM uses a linear kernel & the decision boudary is a straight line.
  - The linear kernel is effictive when the data is already linearly separable in the feature space

Non-linear SVM:-

- for non-linearly seperable data SVM can use non linear kernels to map the input features

Page No :
Date : \ \

~~Non~~ linear into higher-dimensional space
- This transformation allows for finding a hyperplane in the transformed space that corresponds to a non-linear decision boundary in original space