

50
100

Class Assessment - 3

Section A :- Python Basics.

What is the difference a list and a tuple in python? Provide an example of when you would use each.

List

Tuple

- | | |
|--|---------------------------------------|
| 1) List is mutable | 2) Tuple is immutable. |
| 2) List iteration is slower and is time consuming. | 2) Tuple iteration is faster. |
| 3) List consumes more memory. | 3) Tuples consumes less memory. |
| 4) List operations are more error prone | 4) Tuples operations are safe. |
| 5) List provides many in-built methods | 5) Tuples have less in-built methods. |

6) e.g :-

```
list_num = [1, 2, 3, 4]  
print(list_num)
```

6) e.g :-

```
tup_num = (1, 2, 3, 4)  
print(tup_num)
```

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

```
import math  
def fact(n),
```

```
return math.factorial(n))\nnum = int(input("Enter the number: "))\nf = fact(num)\nprint("Factorial of", num, "is", f).
```

o/p : Enter the number : 6
Factorial of 6 is 720.

Q1. Explain the concept of list comprehension in Python. Provide an example how it can be used to create a list.

List comprehension:-

- 1) When you want to create a new list based on the values of an existing list
- 2) E.g. Based on a list of fruits, you want a new list, containing only the fruits with the letter 'a' in the name.
- 3) With list comprehension you can do all with only one line of code

```
Fruits = ["apple", "banana",  
          "cherry", "kiwi", "mango"]
```

```
newlist = [x for x in Fruits if "an" in x]
```

```
print(newlist)
```

Briefly explain the purpose of the following python libraries : Numpy, Pandas, and matplotlib

Numpy :- Numpy can be used to perform a wide variety of mathematical operation on array

- 2) Numpy aims to provide an array objects that is up to $50 \times$ faster than traditional python lists
- 3] i) Pandas :- Pandas is most commonly used for data wrangling & data manipulation purpose
ii) Pandas is used for creating heterogeneous two-dimensional data objects
- c) Matplotlib: Matplotlib is a comprehensive library for creating static, animated & interactive visualization in python.
2) It provides an object-oriented API for embedding plots into applications using general purpose GUI toolkits like Tkinter.

Section B: Machine Learning Basics

Define supervised learning and unsupervised learning. Provide an example of each.

A] supervised learning:

- i) In supervised learning, the machine is trained on a set of labeled data, which means that the input data is paired with the desired o/p.
ii) The machine then learns to provide the o/p for new input data.

A) Unsupervised learning:-

- 1) In this machine is trained on a set of unlabeled data, which means that the input data is not paired with the desired o/p.
- 2) The machine then learns to find patterns & relationships in the data.
- 3) It is used for tasks such as clustering, dimensionality reduction & anomaly detection.

Q. 2.

Explain the bias-variance tradeoff in the context of machine learning models. How does it impact model performance?

- 1) In machine learning, as you try to maintain one component of the error, the other component tends to increase & vice versa.
- 2) Finally, the right balance of bias & variance is key to creating an effective & accurate model.
- 3) A high level of bias can lead to underfitting, which occurs when the algorithm is unable to capture relevant relations b/w features & target output.
- 4) A high bias model typically includes more assumptions about the target function or end results.

Q. 3

Describe the steps involved in the machine learning pipeline.



① The steps in a machine learning pipelines include Data ingestion, Data validation, Data pre-processing, model training & tuning, model analysis, model versioning, model deployment, Feedback loop.

② Data Ingestion: In this step the data is processed into a well-organized format which would be suitable to apply for further steps.

③ Data validation: It is required to perform before training a new model. It focuses on statistics of new data.

④ Data Pre-Processing: The pre-processing step involves preprocessing the raw data & making it suitable for the ML model. The process includes different substeps, such as data cleaning, feature scaling, etc.

⑤ Model Training & tuning: In this step, the model is trained to take the input & predicts an output with highest possible accuracy.

⑥ Model Analysis: After model training we need to determine the optimal set of parameters by using loss of accuracy metrics.

⑦ Model Deployment: After training & analyzing the model, it's time to deploy the model. A model can be deployed in three ways which are

- using the model server
- In a browser
- on edge device

Q.4 A) Cross-validation:

1) Cross-validation is a technique for evaluating ML models by training several ML models subsets of the available input data & evaluating them on the complementary subset of the data.

2) It is important in ML because it is to detect overfitting i.e. failing to generalize patterns.

3) It is used in ML to evaluate the performance of a model on unseen data.

4) E.g. - k-fold cross-validation.

In this technique, the whole dataset is partitioned in k parts of equal size & each partition is called a fold.

It's known as k -fold since there are k parts where k can be any integer - 3, 4, 5, etc. One fold is used for validation & other $k-1$ folds are used for training the model.

Q.5

Regression

Classification

1) In this problem statement the target variables are continuous.

2) In this problem statement, the target variable is discrete.

2) Problem like House Price prediction, Rainfall Prediction like problems are solved using regression algorithm.

2) Problems like spam Email, classification, Disease prediction like problems are solved using classification algo.

numerical values

labels

- 8) Eg. we uses are stock price prediction, house price prediction.
- 9) Eg. we classification are spam detection, image recognition.

Section C:- statistics & probability.

- 1) Measures of central tendency help you find the middle or the average of a dataset.
- 2) The 3 most common measures of central tendency are the mode, median & mean.
- 3) mode: the most frequent value.
- 4) median: the middle number in an ordered dataset.
- 5) mean: The sum of all values divide by the total number of values
- 6) In addition to central tendency, the variability & distribution of your dataset is import to understand when performing descriptive statistics

- 1) The P-value is defined as the probability under the assumption of no effect or no difference of obtaining a result equal to or more extreme than what was actually observed.
- 2) The P stands for probability & measures how likely it is that any observed difference group is due to chance.
- 3) In statistical hypothesis testing P-value can be defined as the measure of probability that real valued test statistic at least as extreme as the value actually obtained.
- 4) A P-value indicate the probability of getting an effect rates no less than that actually observed in the effect + sample data.
- 5) P-values are used in statistical hypothesis testing to determine whether to reject the null hypothesis.
- 6) The smaller p-value, the stronger the likelihood that you should rejects null hypothesis.

- 1) Understanding the diff between correlation and causation is essential in data science
- 2) Statistics
- 3) Correlation refers to the statistical relation ship betw two variables.
- 4) While causation is the relationship between cause & effect.
- 5) While correlation can help identify patterns

it does not imply causation

5) correlation is when two things happen together while causation is when one thing causes another thing to happen

e) for e.g. you might say that there is a correlation b/w ice cream sales & crime rate because between notice that they seem to rise & fall together

Section D :- Advanced topics.

a. 1

- 1) Overfitting is an undesirable machine learning behavior that occurs when the ML model gives accurate prediction for training data but not for new data.
- 2) When data scientists use ML models for making on a known dataset.
- 3) Overfitting is modeling error which occurs when function too closely fit to limited set of data pts.
- 4) In this case, the ML model learns the details & noise in the training data such that it negatively affects the performance of on test data
- 5) It can be mitigated by using diff techniques such as train with more data, data feature selection, cross-validation etc.

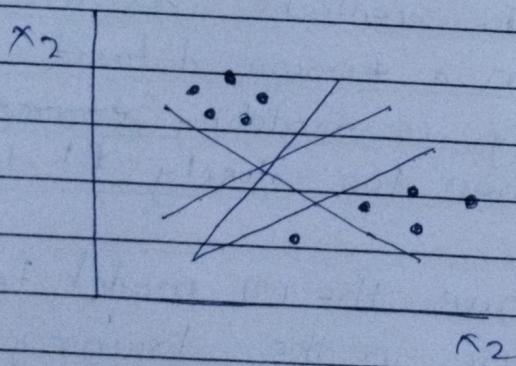
Q.2

D SVM:- SVM is a ML algorithm that uses supervised learning models to solve complex classification, regression & outliers detection problems by performing optimal dual transformation that determine boundaries b/w data pts based on predefined class labels or off.

2) SVM is a supervised ML algo. used for both classification & regression

3) The main objective of SVM algo. is to find the optimal hyperplane in N-dimensional space that can separate data pts in different classes in the feature spaces.

a) Let's consider two independent variables x_1, x_2 & one dependent variable which is either a block or white circle.



Q.3

Deep learning:-

i) Deep learning, on the other hand is a subset of machine learning that uses neural networks with multiple layers to analyze nested patterns and relationships in data.

- ⑦ It is inspired by the structure and function of the human brain and has been successful in variety of tasks, such as computer vision, natural language processing and speech recognition.
- ⑧ Deep learning algorithms trains machines by learning from examples.
- ⑨ Industries such as healthcare, ecommerce, entertainment and advertising commonly use deep learning.

- Q. 4
- ① Feature scaling is the process of normalizing the range of features in dataset.
 - ② Real-world datasets often contain features that are varying in degrees of magnitude, range and units.
 - ③ Therefore, in order for machine learning models to interpret these features on the same scale, we need to perform feature scaling.
 - ④ Feature scaling also improve the convergence & stability of some machine learning algorithm, particularly those that use gradient based optimization.
 - ⑤ It preserves the relationship between the minimum & maximum values of each feature, which can be important for some algorithm.

→ It is in

- a) By scaling features to a similar range the algorithm can take steps more uniformly across diff dimensions, speeding up the learning process.
- b) When using gradient descent based optimization algorithm feature scaling can help speed up convergence and improve model performance.