G.K. GUJAR MEMORIAL CHARITABLE TRUST'S

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

UNIT TEST EXAMINATION - I / II / III

Name of Student: Pratik Satish More

Batch ~~Class~~: No.:- FST 2.0 D9 Division:

Roll No. Date: 14-2-24

Subject: Class Assissment -3 Signature of Supervisor:

Que. No.	1	2	3	4	5	6	7	8	9	10	Total Marks
Marks Obtained											32/100

(Sign. of Subject Teacher)

Q.1] What is the difference between a list and a tuple in python? Provide an example of when you would use it.

→ Lists are mutable, meaning you can modify their elements after creation, while tuples are immutable, meaning their elements cannot be changed after creation.

```
List = ["task1", "task2", "task3"]
List.append("task4")
print(List)
["task1", "task2", "task3", "task4"]
```

Tuple

```
a = (3, 5)
point[0] = 2
```

This will raise a type error

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

→

```
def factorial(n):
    if n == 0:
        return 1
    else:
        return n * factorial(n-1)
```

Q3] Explain concept of list comprehesion in python Provide an example of how it can be used to create list.

→ List comprehesion is allows you to generate a new list by applying an expression to each item in an existing iterable and optionally filtering the items based on a condition.

new_list = [expression for item in iterable if condition]

2

Q4] Briefly explain the purpose of the following python libraries: Numpy, Pandas and Matplotlib.

→ ① Numpy : Numpy is a powerful library for numerical computing in python. It provides support for large, multi-dimensional arrays and matrices along with a collection of mathematical functions to operate on these arrays efficiently.

② Pandas : Pandas is a library built on top of Numpy that provides high-level data structures and functions designed to make working with structured data easy and intuitive.

③ Matplotlib : Matplotlib is a comprehensive plotting library for creating static, animated, and interactive visualizations in python. It provide a MATLAB-like interface and supports a wide variety of plots, bar plots, histograms and more.

4

Q5] Write a python script that reads a text file, counts the frequency of each word and print the top three most frequent words.

4

→

```
from collections import Counter
def count_words (file_path):
```

```
with open (file_path, 'r') as file:
    text = file.read().lower()
words = text.split()
word_counts = counter (words)
print ("Top three most frequent words:")
for word, count in word_counts.most common(3):
    print (f"{word}: {count} times")
```

Section B

Q1] Define supervised learning and unsupervised learning Provide an example of each.

→ Supervised learning: Supervised learning is a type of machine learning where the model is trained on a labeled dataset.

Example - Classification is a common task in supervised learning. For instance, consider a dataset containing images of fruits along with their corresponding labels (e.g. "apple", "banana", "orange"). In this scenario the task is to train a model to classify images of fruits into their respective catagories based on the features extracted from the images.

Unsupervised learning: Unsupervised learning is a type of machine learning where the model is trained on an unlabeled dataset.

Example: Clustering is a common task in unsupervised learning. For instance, consider a dataset containing customer purchase data, where each data point represents a customer and their purchasing behavior In this scenario, the task is to group similar customers together based on their purchasing behavior.

Q2] Explain the baias - variance tradeoff in the context of machine learning models. How does it impact model performance?

→ Bias: Bias refers to the error introduced by approximating a real-world problem with a simplified model. A high bias model tends to oversimplify the underlying relationships in the data and may underfit the training data.

4 variance: Variance refers to the error due to the model's sensitivity to small fluctuations in the training data

The tradeoff arieses because reducing bias often increases variance, and vice versa. A model with high bias tends to have low variance and a model with high variance tends to have low bias. The goal is to find the right balance between baias and variance that minimizes the total error and results good generalization performance on unseen data.

Q3] Describe the steps involved in the machine learning pipeline. Provide a brief explanation of each step.

→ 1] Data collection: This step involves gathering the relevant data from various sources, such as databases, APIs, files or external datasets.

2] Data Preprocessing: Data preprocessing involve cleaning, transforming and preparing the raw 6 data for further analysias.

3] Feature Engineering: Feature engineering is the process of creating new features or transforming existing features to improve the performance of the machine learning model.

4] Model selection :- In this step, you choose the appropriate machine learning algorithms to solve the problem at hand.

5] Model Training :- Once the model is selected it is trained on the training data to learn the underlying patterns and relationships.

6] Model Evaluation : After training the models performance is evaluated using the data or crosss validation techniques.

7] Model Tuning : Model tuning involves optimizing the hyperparameters of the model to improve its performance

8] Model Deployment : Once the model is trained and evaluated it can be deployed into production for making predictions on new unseen data

9] Monitoring and maintenance :- After deployment the model's performance is monitoring over time to ensure that it continues to perform well in real-world scenarios.

Q4] What is cross-validation and why is it important in machine learning ? Provide an example of a cross-validation technique.

→ Crosss-validation is a technique used to assess the performance and generalization ability of machine learning model.

Cross-validation is important in machine learning for several reasons

① Better estimate of Performance : Cross-validation provides a more reliable estimate of the model's performance compared to a single train-test split.

② Reduced overfitting :- Cross-validation helps in detecting overfitting by assessing the model's performance on multiple subsets of the data.

③ Optimization of Hyperparameters :- Cross-validation is often used in hyperparameter tuning to find the best combination of hyperparameters that maximizes the models performance.

Example.

K-Fold Cross-Validation is widely used cross-validation technique that involves splitting the data into k equal-sized folds. The model is then trained and evaluated k times, each time using a different fold as the test set and the remaining folds as the ~~test~~ training set.

Q5] Differentiate between regression and classification in the context of machine learning. provide example

→

Regression	Classification
① Regression is a type of supervised learning task where the goal is to predict a continuous ~~##~~ numerical values or quantity based on input features.	① Classification is another type of supervised learning task where the goal is to predict the catagory or class label of input data based on its features.
② The objective of regression is to learn the relationship between the ~~out~~ input feature and the target variable, which is typically represented as a mathematical function	② The objective of classification ~~is~~ ~~The~~ is to learn a decision boundry that separates the different classes in the ~~sp~~ separates the different class feature space.

5

③ In regression the output variable is continuous and can take any real values within specific range.

④ Example :- It include predicting house prices based on features such as size, location and number of bedrooms.

③ In classification the output variable is categorical the output belongs to a finite set of predefined classes or categories.

④ Example :- It include spam email detection, image classification, sentiment analysis.

Q6] Briefly explain the k-nearest neighbors (kNN) algorithm. How does it work and what are its main parameters?

→ The k-nearest neighbors (kNN) algorithm is a simple and intutive machine learning algorithm used for both classification and regression tasks. It works based on the principle of similarity where the prediction for a new data point is made by considering the majority class or averaging the values of its k nearest neighbors in the training dataset.

KNN algorithm working :-

① Calculate Distance :- For each new data point to be classified or predicted, calculate its distance to all other data points in the training dataset.

② Find Nearest Neighbors :- Identify the KNN to the new data point based on the calculated distance. These are the data points with the smallest distance to the new point.

③ Majority vote :- For classification tasks assign the class label that appears most frequently among the KNN to the new data point.

④ Make prediction: Assign the predicted class label or predicted value to the new data point based on the majority vote or average calculated in the previous step

Parameters of the KNN.

(i) k: The number of nearest neighbors to consider when making predictions. The choice of k is a crucial parameter that significantly affects the performance of the algorithm.

(ii) Distance Metric: The measure used to calculate the distance between data points.

(iii) Weighting scheme: In some implementations of KNN a weighting scheme used to give more weight to the predictions of closer neighbors.

Section C

Q1] Define the term mean, median and mode. Explain when each measure of central tendency is most appropriate.

→ (i) Mean: The mean is the average of a set of numbers. It is calculated by adding up all the numbers in the set and then dividing by the total count of numbers.

- The mean is most appropriate when dealing with data that is symmetrically distributed and does not contain outliers.

(ii) Median: The median is the middle value of a set of numbers when they are arranged in ascending or descending order. If there is an even number of values the median is the average of the two middle values.

- The median is most appropriate when dealing with skewed data or data with outliers as it is not affected by extreme values. It gives a better representation of the central value in such cases.

(iii) Mode :- The mode is the value that appears most frequently in a set of numbers.
- The mode is most appropriate for categorical data or when identifying the most commonly occurring value in a dataset.

Q2] A dataset has a standard deviation of 10, if a data point 2 standard deviations above the mean, what percentage of the data is below this point in a normal distribution?

→ In a normal distribution:
- About 68% of the data falls within one standard deviation above and below the mean.
- About 95% of the data falls within two standard deviations above and below the mean.

4 Since the data point in question is 2 standard deviations above the mean, it falls within the top 2.5% of the distribution (half of the remaining 5% above the mean). Therefore, the percentage of the data below this point is 100% - 2.5% = 97.5%.

Q3] Explain the concept of p-value in hypothesis testing. How is it used to make decision in statistical analysis?

→ The p-value in hypothesis testing represents the probability of observing a test statistic as extreme as or more extreme than the one calculated from the sample data assuming that the null hypothesis is true.

5 It is used is statistical analysis:
(1) Formulate Hypotheses: First, you formulate a null hypothesis (H0), which represents the default assumption that there is no effect or no difference. Then you have an

alternative hypothesis (H1), which typically represents what you're trying to prove or find evidence for.

(ii) Choose significance level: You choose a significance level (alpha) which is the threshold for deciding wether the p-value provides enough evidence to reject the null hypothesis. Commonly used significance levels 0.05 & 0.01.

(iii) Calculate test statistic and p-value: You collect sample data and calculate a test statistic based on your chosen hypothesis test.

(iv) Compare p-value to significance level: If the p-value is less than or equal to the chosen significance level (alpha), you reject the null hypothesis in favor of the alternative hypothesis. This indicates that the observed effect is statistically significant.

Q5] Describe the

Q5] A random variable x follows a normal distribution with a mean of 50 and a standard deviation of 8. Calculate the z-score for a value of $x = 58$.

→ To calculate the z-score for a given value $x = 58$ in a normal distribution with a mean $\mu = 50$ and a standard deviation $\sigma = 8$ you can use the formula for z-score:

$$z = \frac{x - \mu}{\sigma} = \frac{58 - 50}{8}$$

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

$$z = 1$$

So, the z-score for $x = 58$ is 1.

Page No.
Date

Q4] Describe the difference between correlation & causation. Provide an example to illustrate your explanation.

Correlation	Causation
(i) It refers to a statistical relationship between two variables where they tend to move together in a consistent manner.	(i) It refers to a relationship where one variable directly influences or causes changes in
(ii) The weather gets warmer with increase in sales of icecreams.	(ii) Warmer weather caused more sales of icecreams
(iii) Coffee drinkers are at high risk of lung cancer.	(iii) Smokers like to drink coffee and smokers develop lung cancer.
(iv) stress causes rumination ~~with~~ which causes depression.	(iv) stress causes depression

4

Section - D

Q4] Explain the concept of overfitting in machine learning. How can it be mitigated ?.

→ • Overfiting : Overfitting occurs when a model becomes too complex relative to the amount and noise level of the training data. The model essentially memorizes the training data instead of learning the underlying patterns, leading to poor performance or new data.

5

• Mitigation of overfitting :

(i) Crooss - validation : Cross - validation techniques such as k - fold cross - validation can help evaluate the model's performance on multiple subsets of the

data, providing a more reliable estimate of its generalization performance.

(ii) Regularization :- Techniques like L1 and L2 regularization add a penalty term to the model's loss function discouraging overly complex models and promoting simpler models that generalize better to new data.

(iii) Feature section :- Choosing relevant features and removing irrelevant or redundant ones can help reduce the complexity of the model and mitigate overfitting.

(iv) Simpler model :- Using simpler models with fewer parameters, such as decision trees with limited depth or linear models can reduce the risk of overfitting, especially when the dataset is small or noisy.

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

→ Support vector Machine (SVM) is a supervised learning algorithm used for classification and regression task.

SVM working :

(i) Linear Separation :- In the case of linearly separable data SVM aims to find the hyperplane that maximizes the margin which is the distance between the hyperplane and nearest data points.

(ii) Margin maximization : SVM aims to find hyperplane that not only separates the data points but also maximizes the margin between the hyperplane & the support vector.

The role of the kernel in SVM is to implicity map the input data into a higher dimensional space where it may become linearly separable. The choice of kernel function determines the mapping used to transform the data.

The kernal Function computes the dot product bet^n data points are ~~more~~ in higher dimensional space without explicity computting the transformation. This is known as the kernel trick.

1) Linear kernel : $k(x, x') :- x^T x'$
2) Polynomial kernal :- $k(x, x') = (x^T x' + c)^d$
3) Radial basis F^n (RBF) kernel :- $k(x, x') = e^{-\gamma \| x - x' \|^2}$
4) Sigmoid kernel : $k(x, x') = \tanh(a x^T x' + c)$

Q3] Deep learning :- It is subset of machine learning that uses artificial neural networks with multiple layer (hence the term "deep") to learn from large amount of data. Deep learning algorithms are designed to automatically learn hierarchical representation of data through the composition of multiple nonlinear transformation.

diff :-

i) Feature Representation :- Traditional machine learning algorithm ~~a~~ftenely on hand crafted feature engineering where domain expects manually design feture from row data. It constant deep learning model automatically learn feature representation directly from the data

ii) Model Complexity :- Deep learning model are typically more complex than traditional machine learning models with many layer of interconnected ~~Adr~~ neurals.

iii) Scalability :- Deep learning algorithms scale well with large data sets than to parallel computing & distributed traning techniques.

iv) Generalization :- Deep learning model often have high ~~co~~ capacity & can learn from diverse & heterogeneous data sourses.

e.g. convolutional neural Network (CNN)

CNN are type of deep neural network commonly used for image ergonition & computer vision. They consist of multiple layer ~~induvi~~ including convolution layers pooling layers & Fall connected key.

Q4) i) Importance of feature scaling

(i) Improves convergence :- Feature scaling can help algorithms converge faster during traning especially for gradient based optimization algorithm like gradient descent.

(ii) Prevent Dominance of feature :- Algorithms that uses distance based matrics (e.g. k-nearest neighbours, support vector machines with radial ations & influence the model disproportionately.

(iii) Enhances Regularization :- Regularization techniques like L_1 & L_2 regularization penalize large coefficients in linear models.

6

2) Affected by scaling

(i) Gradient Descent - Based Algorithm :- Algorithm like linear regression logistic regression neural networks & support vector machines benefit from feature scaling to improve convergence & stability during traning.

(ii) Distance - Based Algorithms - k nearest neighbors, support vector machines with radial basis F^n (RBF) kernel, & clustering algorithms like k-means & sensitive to the scale of input fetures.

iii) Principal component Analysis (PCA) :-

PCA is dimensionallize reduction technique that involves finding the principal componentes of data.