**ICDS Midterm Examination**

Machine Learning Focus

**Instructions:**

* This exam consists of three sections: Multiple Choice Questions , Short Answer Questions , and Programming Questions .
* Total time: 120 minutes.
* Please answer all questions.

**Section 1: Multiple Choice Questions (2 points each)**

Question 1.1 | K-Nearest Neighbors (KNN)  
Which of the following tasks can be solved using the K-Nearest Neighbors (KNN) algorithm?  
A: Partitioning genomic data into clusters  
B: Predicting travel time based on traffic and weather conditions  
C: Segmenting satellite images into land cover types  
D: Recommending movies to a new user based on similar users' ratings

Question 1.2 | Supervised vs. Unsupervised Learning  
Which of the following is an example of supervised learning?  
A: Grouping students based on their height, weight, and hair color  
B: Clustering genes into groups based on expression patterns  
C: Predicting whether a user will purchase a product based on their demographic data  
D: Reducing the dimensionality of high-dimensional data

Question 1.3 | Linear Regression  
What is the purpose of the loss function in linear regression?  
A: To compute the distance between clusters in K-Means  
B: To minimize the prediction error between the predicted and actual values  
C: To calculate the gradient of the activation function in neural networks  
D: To determine the optimal number of clusters in K-Means

Question 1.4 | Gradient Descent  
What happens if the learning rate in gradient descent is set too low?  
A: The algorithm converges too quickly to a suboptimal solution  
B: The algorithm may take too long to converge or get stuck in local minima  
C: The algorithm overshoots the global minimum  
D: The algorithm ignores the gradient entirely

Question 1.5 | Feature Engineering  
Why is feature engineering important in machine learning?  
A: It reduces the computational complexity of clustering algorithms  
B: It transforms raw data into meaningful representations that improve model performance  
C: It eliminates the need for labeled data in supervised learning  
D: It ensures that all features have the same scale

Question 1.6 | K-Means Clustering  
What does the "K" in K-Means clustering represent?  
A: The number of iterations required for convergence  
B: The number of clusters to form  
C: The learning rate used in gradient descent  
D: The distance metric used to measure similarity

Question 1.7 | Activation Functions  
Which activation function is commonly used in the hidden layers of neural networks to introduce non-linearity?  
A: ReLU  
B: Sigmoid  
C: Tanh  
D: Softmax

Question 1.8 | Bias-Variance Tradeoff  
What is the primary cause of overfitting in machine learning models?  
A: High bias  
B: Low variance  
C: High variance  
D: Low bias

Question 1.9 | Clustering Metrics  
Which of the following is a common objective in clustering analysis?  
A: Minimize intra-cluster variance  
B: Maximize the number of clusters  
C: Minimize the learning rate  
D: Maximize the distance between centroids

Question 1.10 | Inductive Bias  
What is inductive bias in machine learning?  
A: A set of assumptions that the learner uses to generalize unseen data  
B: The process of reducing the dimensionality of data  
C: The use of labeled data to train supervised learning models  
D: The calculation of gradients in gradient descent

Question 1.11 | Types of Machine Learning Tasks  
Which of the following is a classification task?  
A: Predicting the price of a house based on its features  
B: Estimating the commute time based on traffic and weather  
C: Determining whether an email is spam or not  
D: Grouping customers into segments based on purchasing behavior

Question 1.12 | Distance Metrics  
Which distance metric is most intuitive for measuring similarity between two points in Euclidean space?  
A: Manhattan distance  
B: Minkowski distance  
C: Euclidean distance  
D: Geodesic distance

Question 1.13 | Neural Networks  
What is the role of the activation function in a neural network?  
A: To compute the loss during training  
B: To introduce non-linearity into the model  
C: To reduce the dimensionality of the input data  
D: To cluster similar data points together

Question 1.14 | Model Evaluation  
Which of the following metrics is used to evaluate the performance of a classification model?  
A: Mean squared error  
B: Precision and recall  
C: Intra-cluster variance  
D: Learning rate

Question 1.15 | Lazy Learners  
Which of the following algorithms is considered a "lazy learner"?  
A: Linear regression  
B: K-Nearest Neighbors (KNN)  
C: K-Means clustering  
D: Neural networks

**Section 2: Short Answer Questions (5 points each)**

Question 2.1 | Feature Engineering  
Explain why feature engineering is crucial in machine learning. Provide an example of how poor feature selection can lead to suboptimal model performance.

Question 2.2 | Bias-Variance Tradeoff  
Define the bias-variance tradeoff in machine learning. How does it affect the choice of models?

Question 2.3 | Time Complexity  
Consider the K-Means clustering algorithm. Given **n** data points, **k** clusters, **d** dimensions, and **i** iterations, derive and explain the time complexity of the algorithm.

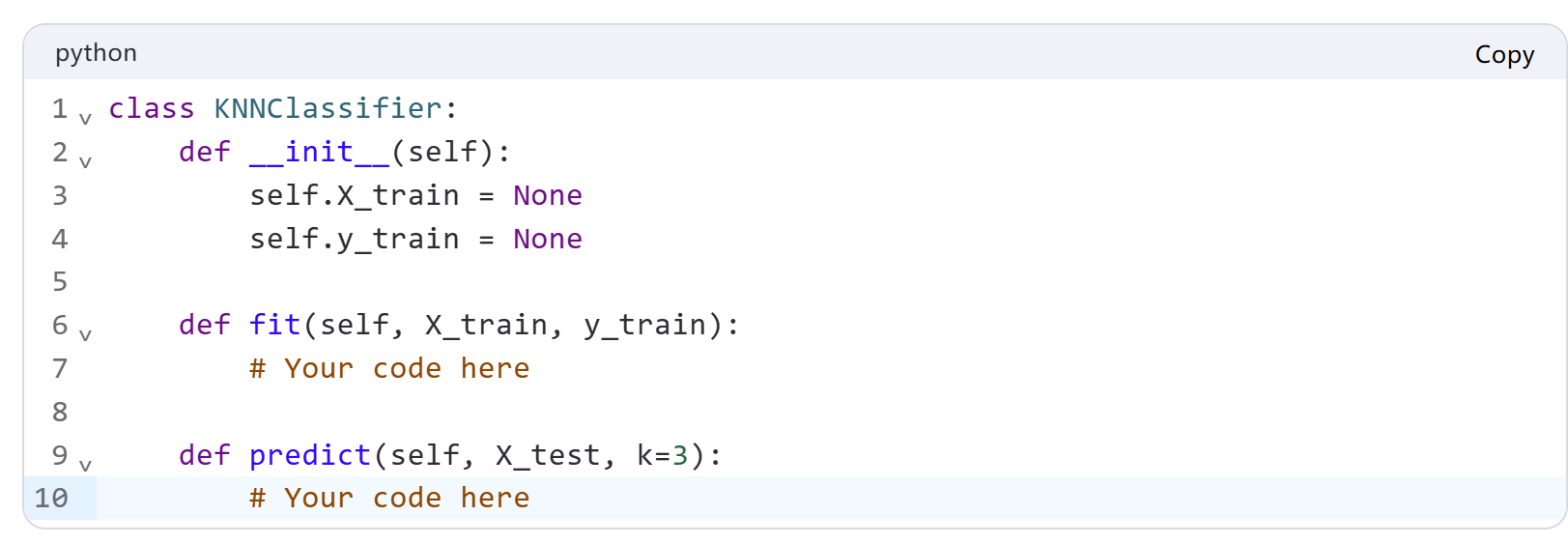
Question 2.4 | Activation Functions  
Compare and contrast the ReLU and Sigmoid activation functions. In which scenarios would you prefer one over the other?

Question 2.5 | Gradient Descent  
Explain the concept of gradient descent in the context of linear regression. What challenges might arise during its implementation?

**Section 3: Programming Questions (10 points each)**

Question 3.1 | Implementing K-NN  
Write a Python class **KNNClassifier** that implements the K-Nearest Neighbors algorithm. The class should have the following methods:

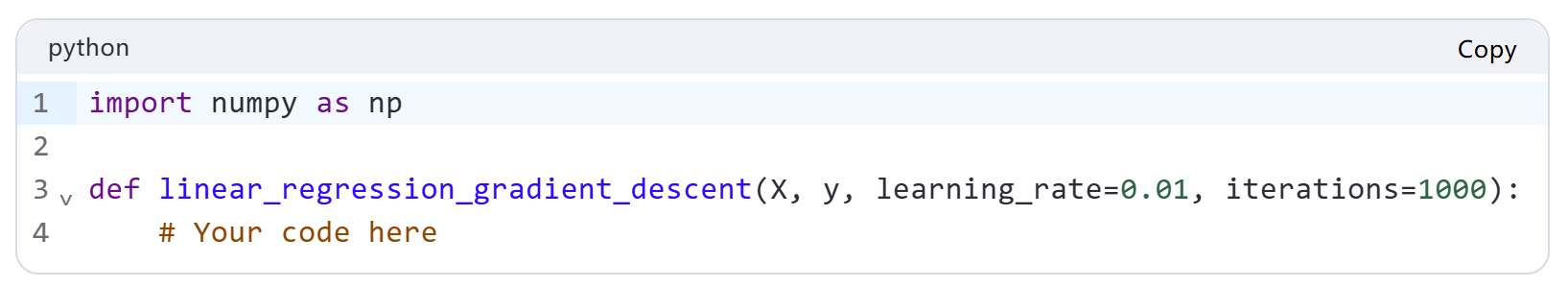
1. **fit(X\_train, y\_train)**: Trains the model using the training data.
2. **predict(X\_test, k=3)**: Predicts the labels for the test data using the K-NN algorithm. Use Euclidean distance as the distance metric.



Question 3.2 | Gradient Descent for Linear Regression  
Implement a Python function **linear\_regression\_gradient\_descent** that performs linear regression using gradient descent. The function should take the following arguments:

1. **X**: A numpy array of shape **(n\_samples, n\_features)** representing the input data.
2. **y**: A numpy array of shape **(n\_samples,)** representing the target values.
3. **learning\_rate**: The learning rate for gradient descent.
4. **iterations**: The number of iterations for gradient descent.

The function should return the learned weights **w**.



Question 3.3 | K-Means Clustering  
Implement a Python function **kmeans** that performs K-Means clustering. The function should take the following arguments:

1. **X**: A numpy array of shape **(n\_samples, n\_features)** representing the input data.
2. **k**: The number of clusters.
3. **max\_iterations**: The maximum number of iterations for convergence.

The function should return the cluster centroids and the labels for each data point.

