Viva Questions with Short Answers (Programs 1 to 9)

1. Correlation, Heatmap, Pairplot (California Housing)

1. Q: What is the purpose of calculating a correlation matrix?

   A: To identify relationships between numerical variables.

2. Q: Why is a heatmap useful in visualizing correlations?

   A: It visually represents correlation strengths with color gradients.

3. Q: What does a value close to 1 or -1 in a correlation matrix signify?

   A: A strong positive or negative correlation respectively.

4. Q: Why do we use a sample in the pairplot?

   A: To reduce computation and make the plot faster to render.

5. Q: How can pairplots help in feature selection?

   A: By revealing feature relationships and redundancy.

2. PCA on Iris Dataset

1. Q: What is PCA and why is it used?

   A: Principal Component Analysis is used for dimensionality reduction.

2. Q: How many principal components were used in this code?

   A: Two principal components were used.

3. Q: What do PC1 and PC2 represent in the plot?

   A: They are the first two principal components capturing most variance.

4. Q: Why is dimensionality reduction important?

   A: It simplifies data, reduces overfitting, and improves performance.

5. Q: How does PCA differ from feature selection?

   A: PCA creates new features, while feature selection chooses existing ones.

3. Find-S Algorithm

1. Q: What is the Find-S algorithm used for?

A: To find the most specific hypothesis from positive training examples.

2. Q: Why do we consider only positive examples in Find-S?

A: Find-S ignores negative examples and generalizes from positives only.

3. Q: What does '?' signify in the final hypothesis?

A: It represents a wildcard, meaning 'any value'.

4. Q: Can Find-S handle noisy data?

A: No, it's sensitive to noise and assumes error-free data.

5. Q: What is the output of the Find-S algorithm?

A: A hypothesis consistent with all positive examples.

4. KNN from Scratch

1. Q: What does KNN stand for and how does it work?

A: K-Nearest Neighbors; it predicts based on the majority class of k closest points.

2. Q: How is distance measured in this implementation?

A: Using the absolute difference (L1 norm).

3. Q: What happens if k is too small or too large?

A: Too small: sensitive to noise; too large: may miss local patterns.

4. Q: How are ties handled in KNN?

A: By selecting the most frequent class among neighbors.

5. Q: What is the time complexity of KNN during prediction?

A: O(n) per query, where n is the number of training samples.

5. Locally Weighted Regression (LWR)

1. Q: What is Locally Weighted Regression?

A: A non-parametric regression method that weights data points based on proximity.

2. Q: How is the weight calculated in LWR?

A: Using a Gaussian kernel based on the distance to the query point.

3. Q: What is the role of the 'tau' parameter?

A: Tau controls the width of the weighting function, affecting locality.

4. Q: What does the red line in the plot represent?

A: The LWR model's predicted fit line.

5. Q: Is LWR a parametric or non-parametric method?

A: It is a non-parametric method.

6. Linear & Polynomial Regression

1. Q: What is the difference between linear and polynomial regression?

A: Linear fits straight lines, polynomial fits curves.

2. Q: What does the red line indicate in the linear regression plot?

A: It shows the best fit line predicted by the model.

3. Q: How is PolynomialFeatures used in sklearn?

A: It generates polynomial and interaction features.

4. Q: What is overfitting in polynomial regression?

A: Model fits noise in training data, harming generalization.

5. Q: Why do we transform features for polynomial regression?

A: To allow linear models to fit nonlinear data.

7. Decision Tree - Breast Cancer

1. Q: What dataset is used in this program?

A: The Breast Cancer Wisconsin dataset from sklearn.

2. Q: What does `max_depth=3` do in the DecisionTreeClassifier?

A: It limits the tree depth to prevent overfitting.

3. Q: What does the `plot_tree` function show?

A: It visualizes the structure and splits of the decision tree.

4. Q: How do we evaluate the decision tree model?

A: Using accuracy score on the test dataset.

5. Q: What are the advantages of decision trees?

A: They are easy to interpret, handle both numerical and categorical data.

8. Naive Bayes - Olivetti Faces

1. Q: What is the Olivetti Faces dataset?

   A: A collection of face images used for classification tasks.

2. Q: What is the Gaussian Naive Bayes classifier?

   A: A probabilistic classifier assuming feature independence and Gaussian distribution.

3. Q: How is the model evaluated in this code?

   A: Using the accuracy score on the test set.

4. Q: What are the assumptions of Naive Bayes?

   A: Features are conditionally independent given the class.

5. Q: Why is Naive Bayes suitable for high-dimensional data?

   A: It performs well with many features and small datasets.

9. KMeans Clustering - Breast Cancer

1. Q: What is the goal of KMeans clustering?

   A: To partition data into k groups based on similarity.

2. Q: How many clusters are used in this example?

   A: Two clusters.

3. Q: Why is PCA used before plotting?

   A: To reduce dimensions for visualizing high-dimensional data.

4. Q: What do the red X markers represent?

   A: They are the centroids of the clusters.

5. Q: Is clustering supervised or unsupervised learning?

   A: Unsupervised learning.