In-Class Assessment for Machine Learning

Instructions:

This assessment is designed to evaluate your knowledge of Linear Regression, Multiple Regression, and Polynomial Regression.
The assessment consists of multiple-choice questions and code completion questions.
You have 2 hours to complete the assessment.

Multiple-Choice Questions:

1. What is Linear Regression?
   a. A supervised learning algorithm used for classification problems.
   b. A supervised learning algorithm used for regression problems.
   c. An unsupervised learning algorithm used for clustering problems.
   d. An unsupervised learning algorithm used for dimensionality reduction.

2. Which of the following is NOT an assumption of Linear Regression?
   a. Linearity
   b. Homoscedasticity
   c. Multicollinearity
   d. Normality

3. What is Multiple Regression?
   a. A supervised learning algorithm used for classification problems.
   b. A supervised learning algorithm used for regression problems involving multiple independent variables.
   c. An unsupervised learning algorithm used for clustering problems.
   d. An unsupervised learning algorithm used for dimensionality reduction.

4. Which of the following is an advantage of Multiple Regression over Simple Linear Regression?
   a. More accurate predictions
   b. Simpler model interpretation
   c. Easier to implement
   d. None of the above

5. What is Polynomial Regression?
   a. A supervised learning algorithm used for classification problems.
   b. A supervised learning algorithm used for regression problems that model the relationship between the response variable and the independent variable as an nth degree polynomial.

c. An unsupervised learning algorithm used for clustering problems.
d. An unsupervised learning algorithm used for dimensionality reduction.

6. Which of the following is NOT an assumption of Polynomial Regression?
   a. Linearity
   b. Homoscedasticity
   c. Normality
   d. Independence

7. What is the coefficient of determination (R-squared) used for in Linear Regression?
   a. To measure the accuracy of the predictions made by the model.
   b. To measure the significance of the independent variables.
   c. To measure the degree of multicollinearity between the independent variables.
   d. To measure the goodness of fit of the model.

8. Which of the following statements is true about Multicollinearity in Multiple Regression?
   a. It is desirable to have high levels of multicollinearity between the independent variables.
   b. It can lead to unstable estimates of the regression coefficients.
   c. It has no effect on the accuracy of the predictions made by the model.
   d. It increases the interpretability of the model.

9. Which of the following statements is true about Overfitting in Polynomial Regression?
   a. It occurs when the model is too simple and fails to capture the complexity of the data.
   b. It occurs when the model is too complex and fits the noise in the data.
   c. It has no effect on the accuracy of the predictions made by the model.
   d. It occurs only in Linear Regression.

10. Which of the following statements is true about Regularization in Linear Regression?
    a. It is used to reduce the bias of the model.
    b. It is used to reduce the variance of the model.
    c. It has no effect on the accuracy of the predictions made by the model.
    d. It increases the complexity of the model.

11. Which of the following is an example of Linear Regression?
    a. Predicting the likelihood of a customer to buy a product based on their demographic information.
    b. Predicting the category of a news article based on its headline.
    c. Predicting the price of a house based on its size and location.
    d. Predicting the sentiment of a tweet based on its content.

Case Studies based MCQs(more than one option could be correct):

1. A car rental company wants to predict the rental price of its cars based on the age of the car and the number of miles driven. Which type of regression would be most appropriate for this problem?
   a. Linear Regression
   b. Multiple Regression
   c. Polynomial Regression
   d. Logistic Regression

2. A clothing retailer wants to predict the sales of its products based on the price of the product and the marketing spend on the product. However, the retailer suspects that there might be a non-linear relationship between the price and the sales. Which type of regression would be most appropriate for this problem?
   a. Linear Regression
   b. Multiple Regression
   c. Polynomial Regression
   d. Logistic Regression

3. A healthcare provider wants to predict the length of hospital stay for patients based on their age, gender, medical history, and the severity of their illness. However, the provider suspects that there might be a strong correlation between some of the independent variables. Which technique can be used to address this issue?
   a. Multicollinearity
   b. Regularization
   c. Principal Component Analysis
   d. Cross-Validation

4. A real estate agent wants to predict the selling price of a house based on its location, size, number of bedrooms, and age. However, the agent suspects that the relationship between the independent variables and the dependent variable might not be linear. Which type of regression would be most appropriate for this problem?
   a. Linear Regression
   b. Multiple Regression
   c. Polynomial Regression
   d. Logistic Regression

5. A marketing agency wants to predict the conversion rate of a digital advertising campaign based on the target audience, the ad creative, and the ad spend. However, the agency suspects that there might be interactions between the independent variables. Which technique can be used to address this issue?

a. Multicollinearity
b. Regularization
c. Principal Component Analysis
d. Interaction Terms

Coding Question:

Case Study: Predicting House Prices using Multiple Polynomial Regression
In this case study, you will build a multiple polynomial regression model to predict the prices of houses based on their characteristics. You will use the "California Housing" dataset from sklearn, which contains information about houses in California.

Dataset Description

The dataset contains the following columns:

- MedInc: Median income of the block.
- HouseAge: Median age of the houses in the block.
- AveRooms: Average number of rooms per household in the block.
- AveBedrms: Average number of bedrooms per household in the block.
- Population: Number of people residing in the block.
- AveOccup: Average household occupancy in the block.
- Latitude: Latitude of the block in decimal degrees.
- Longitude: Longitude of the block in decimal degrees.
- Target: Median house value of the block in units of 100,000.

Use the following code to import the dataset from sklearn:
from sklearn.datasets import fetch_california_housing (for sklearn version 1.2 or above)

Or download from:
https://www.kaggle.com/datasets/camnugent/california-housing-prices

Tasks

Your tasks are as follows:

1. Load the California Housing dataset into a Pandas DataFrame.

2. Preprocess the dataset. Remove any rows with missing values. Drop the Latitude and Longitude columns.

3. Visualize the dataset. Plot histograms of all the features.

4. Split the dataset into training and test sets with a 80/20 split.

5. Train a multiple polynomial regression model on the training set. Use only the MedInc, HouseAge, AveRooms, AveBedrms, and Population columns. Use polynomial features up to degree 2.

6. Evaluate the performance of the multiple polynomial regression model on the test set. Compute the Mean Squared Error (MSE) for the model.

7. Fine-tune the parameters of the multiple polynomial regression model to improve its performance.

8. Create a function that takes the trained multiple polynomial regression model as input and a set of features, and returns the predicted house price for those features.

9. Repeat steps 5 to 8 using polynomial features up to degree 3 and compare the performance of the models.

10. Repeat steps 5 to 8 using only the MedInc and Population columns and compare the performance of the models.

11. Choose the best-performing model and create a final version of the model using the entire dataset.