Linear Regression

This assignment consists of several exercises designed to enhance your understanding of linear regression and its extensions. Follow the instructions carefully and provide detailed explanations, derivations, and code where applicable.

Question 1: Predicting Car Prices

Problem Statement:

You are provided with a dataset, data.csv, that contains features such as the number of cylinders, engine size, and the number of doors, along with car prices. The goal is to build a predictive model using linear regression.

Tasks:

1 . **Train a Linear Regression Model:** Develop a model to predict car prices based on the given features. 2 . **Model Evaluation:** Calculate the R^2 score to assess the model's performance. Discuss whether the correlation between the predicted and actual prices is strong ($R^2 > 0.8$). 3 . *(Optional)* **Experimentation with Basis Functions:** Explore techniques like basis functions or Gaussian features to enhance predictions and report your findings.

Hints:

- Use libraries such as pandas for preprocessing and scikit-learn for linear regression.
- Consider feature scaling or polynomial feature transformations to improve results.

Question 2: Data Splitting

Problem Statement:

Implement linear regression using gradient descent, incorporating a basis function of your choice. You may use the provided dataset in question 1 or choose one of your own.

Tasks:

1. Data Split

Perform a train/validation/test split for the dataset using the following configurations:

- **Experiment 1:** Train: 80%, Validation: 10%, Test: 10%
- **Experiment 2:** Train: 70%, Validation: 10%, Test: 20%
- **Experiment 3:** Train: 60%, Validation: 10%, Test: 30%

2. Basis Function

Choose a basis function of your choice for transforming the input features. Try experimenting with different basis functions, including more complex ones.

3. Gradient Descent

Implement gradient descent to optimize the linear regression model.

4. Early Stopping & Overfitting Prevention

Implement early stopping to prevent overfitting during training. Ensure the model stops training once the validation error does not improve for a set number of iterations (patience).

5. Evaluation

- Evaluate the model's performance on the test set for each experiment.
- Track training loss, validation loss, and test loss for each configuration.

6. Additional Tasks

- Compare the results of using simpler basis functions versus more complex ones.
- Discuss how different splits impact the model's performance and generalization.
- Comment on the effects of early stopping in your results.

Question 3: Weighted Linear Regression

Problem Statement:

This problem explores weighted linear regression, where each training sample is assigned a different weight. The cost function for weighted regression is given as:

$$J(w) = \frac{1}{2} \sum_{i=1}^{m} u^{(i)} \left(w^T x^{(i)} - t^{(i)} \right)^2$$

Here:

- w: Parameter vector of the model.
- $x^{(i)}$: Feature vector of the i-th sample.
- $t^{(i)}$: Target value of the i-th sample.
- $u^{(i)}$: Weight for the i-th sample.

Tasks:

1 . **Reformulate the Cost Function:** Show that $J(w)$ can be rewritten as:

$$J(w) = (Xw - \tilde{t})^T U (Xw - \tilde{t})$$

- Explain the role of U (the diagonal weight matrix) in the model and its influence on regression.

2 . **Derive the Normal Equation:** Derive the closed-form solution for w that minimizes $J(w)$.

3 . **Weighted Linear Regression and Maximum Likelihood Estimation:** Assume that the training set $\{(x^{(i)}, t^{(i)}); i = 1, \ldots, m\}$ consists of m independent training samples, where the $t^{(i)}$ values are observed with different variances. Assume that the following holds:

$$p(t^{(i)} \mid x^{(i)}, w) = \frac{1}{\sqrt{2\pi}\,\sigma^{(i)}} \exp\left(-\frac{(t^{(i)} - w^T x^{(i)})^2}{2(\sigma^{(i)})^2}\right)$$

where $t^{(i)}$ has mean $w^T x^{(i)}$ and variance $(\sigma^{(i)})^2$. Assume that the values of $\sigma^{(i)}$ are known. Show that finding the maximum likelihood estimate for w from the above equation is equivalent to solving the weighted linear regression problem. Also, express the relationship between the weights $w^{(i)}$ and the $\sigma^{(i)}$ values.

۴. **Implementation:**

- Implement simple linear regression using the provided dataset: features.csv and labels.csv.
- Implement weighted linear regression. For weights, use the formula:

$$u^{(i)} = \exp\left(-\frac{2 \cdot (x_i - x_i')^2}{\tau^2}\right)$$

- x_i: Feature vector of the i-th data point.
- x_i': Reference feature vector (e.g., current test point).
- τ: Parameter controlling the influence region.
  - Plot the resulting hypothesis (line) for different values of τ (0.1, 0.3, 2, 10) and explain how the parameter affects predictions.

Question 4: Noise and Regularization

Problem Statement:

Consider a linear model:

$$y(x, \mathbf{w}) = w_0 + \sum_{i=1}^{D} w_i x_i$$

with the error function:

$$E_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 .$$

Now assume Gaussian noise ϵ_i with zero mean and variance σ^2 is added to input variables x_i. Show that minimizing E_D over the noise distribution is equivalent to minimizing the sum-of-squares error for noise-free inputs with an added weight-decay regularization term (excluding w_0 from regularization).

Question 5: Hierarchical Linear Models

Problem Statement:

Hierarchical Linear Models (HLM) address nested data structures (e.g., students within schools). The model is expressed as:

$$y_{ij} = w_0 + w_1 x_{ij} + u_j + \epsilon_{ij}$$

- u_j: Random effect for group j.
- ϵ_{ij}: Residual error for observation i in group j.

Tasks:

1 . **Explain the Model:** Describe the structure and cost function of HLM, emphasizing its suitability for hierarchical data.

2 . **Synthetic Data:** Create synthetic hierarchical data and fit an HLM to the dataset.

3 . **Analysis:** Evaluate the effects of intra-group and inter-group variations on predictions.

Hints:

- Use libraries such as statsmodels or PyMC for hierarchical modeling.
- Visualize the nested structure to understand the model.

Question 6: Penalized Spline Regression Using the Car Prices Dataset(Points practice)

Problem Statement:

One of the constraints of basis functions is that they should be piecewise continuous. This means that any small change in the input space does not affect the result significantly. To solve this limitation, the input space can be divided into smaller regions, and in each region, a separate model can be used. This approach leads to the concept of Penalized Splines. In this context, the concept of splines and penalization will be investigated, and the findings should be summarized in 2 to 3 pages.

First, implement the **Penalized Spline** on a given dataset and report the results. The cost function for Penalized Spline is defined as follows:

$$||\mathbf{t} - \phi(\mathbf{X})\beta||^2 + \lambda\beta^T D\beta$$

where $y(x)$ can be defined as:

$$y(x) = \beta_0 + \beta_1 x + ... + \beta_p x^p + \sum_{j=1}^{q} \beta_{pj}(x - k_j)_+^p = \sum_{j=0}^{p+q} \beta_j \phi_j(x)$$

with:

$$(x-k)_+ = \begin{cases} x-k & \text{if } x > k \\ 0 & \text{otherwise} \end{cases}$$

Here, p represents the degree of the polynomial, and k_j values represent the knots, which divide the input space into different regions. The functions in each region may differ from those in other regions (Figure 1). Functions in each region must be continuous at the boundaries to ensure a smooth output (Figure 1). To determine these regions, the input space $[a, b]$ can be divided into q equal subintervals.

Also, in Equation (1), the matrix D is defined as:

$$D = \int_a^b \phi''(x)\phi''(x)^T dx$$

or:

$$D = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$

To implement this, divide the data randomly into two parts: **training** and **testing**, consisting of 70 and 30 samples, respectively. Use Penalized Spline to fit the training data. Then, calculate the error for the test data for both Penalized Spline and linear regression methods. Compare the results by plotting the mean squared error for both methods.

Task:

1 . Fit the **Penalized Spline** and compare it with linear regression in terms of performance.

2 . Vary the number of knots q and polynomial degree p, and plot the performance in a single graph.

3 . Provide the code, analysis, and results as part of the final report.

Submission Guidelines

- Submit all derivations, code, and visualizations in a single ZIP file.
- Ensure all plots are labeled, explanations are detailed, and code is well-documented.
- IMPORTANT: For each question, cite the resources you used.