FML Assignment – 1 Report

Q1.

The training and test dataset were loaded, and an additional dummy feature was added to form the feature matrix to facilitate the intercept term in linear regression. The features were combined and transposed for matrix operations.

i. **Least Squares Solution:**
Initially, I implemented the least squares solution W_{ML} to the regression problem using the analytical solution.
The analytical solution is:

$$(W_{\mathrm{ML}}) = (XX^T)^{-1} X\, y$$

ii. **Standard Gradient Descent Algorithm:**
Now, w is obtained using gradient descent algorithm starting from the initial w = [0 0 0], with a learning rate of 0.01 for 1000 iterations.
To solve the least squares problem using Gradient Descent, the update rule for the weights is given by:

$$W^{(t+1)} = W^{(t)} - \alpha \nabla J\left(W^{(t)}\right)$$

The cost function for gradient descent algorithm is given by:

$$J\left(W^{(t)}\right) = ||X^T W - Y||^2$$

Through multiple iterations, the norm difference $||W_t - W_{ML}||^2$ was plotted to visualize the convergence of the algorithm.

iii. **Stochastic Gradient Descent (SGD):**
In SGD, weights are updated using small random batches of data, enhancing efficiency. The update rule remains similar:

$$W^{(t+1)} = W^{(t)} - \alpha \nabla J\left(W^{(t)}\right)$$

Here, α is the learning rate. By selecting a batch size of 100, the algorithm was executed over several iterations, yielding a plot of $||W_t - W_{ML}||^2$ to track convergence.

iv. Ridge Regression

Ridge Regression is a regularized form of linear regression that introduces a penalty on large weights to prevent overfitting. The objective function is:

$$J_{\text{ridge}}(w) = ||X^T W - y||^2 + \lambda ||w||^2$$

where λ is the regularization parameter. The best value for λ was determined through cross-validation, and validation error was plotted against λ.

Upon comparison, the test errors for Ridge Regression and Least Squares were computed, revealing that the Least Squares model outperformed Ridge Regression, suggesting that regularization was unnecessary for this dataset.

v. Kernel Regression

To enhance model performance, a polynomial kernel regression was implemented. The polynomial kernel is defined as:

$$K(x, x_i) = (x^T x_i + c)^d$$

where:

c is a constant,

d is the degree of the polynomial.

Kernel regression solves for the coefficients α using:

$$\alpha = K_{\text{train}}^{-1} y_{\text{train}}$$

Predictions on the test dataset are made using:

$$y_{\text{pred}} = K_{\text{test}}\, \alpha, \text{ where } K_{\text{test}} = \text{K}(x_{\text{test}}, x_{\text{train}})$$

After evaluating different polynomial degrees, it was found that the polynomial kernel of degree 3 provided the best performance, achieving a test error significantly lower than that of the Least Squares method.

Conclusion

The analysis demonstrated that while Least Squares regression provided a solid baseline, the use of polynomial kernel regression yielded improved predictive performance, highlighting the importance of model selection based on the underlying data distribution. This study emphasizes the need for careful evaluation of various regression techniques to achieve optimal results in predictive modeling tasks.