# 

Assignment#5: Machine Learning: Evaluation Report: Regularized Linear and Logistic Regression with Cross-Validation (Task 1-5).

Swapnali Dashrath

Oakland University

CSI6160 - Machine Learning

Prof. Mohammad Wardat.

November 21, 2024.

**Evaluation Report: Regularized Linear and Logistic Regression with Cross-Validation (Task 1-5)**

This report presents the implementation and evaluation on Regularized Linear and Logistic Regression on Diabetes and Breast Cancer dataset from sklearn.datasets with 5-fold validation using ML libraries (numpy, pandas, seaborn, matplotlib, sklearn etc). From Scratch Implementation is using Python (code file submitted along with report) and Evaluation is based evaluation metrics MSE for Ridge and accuracy for Logistic Regression.

**Task 1: Dataset selection and exploration.**

1. **Dataset Loading**: The dataset is loaded from sklearn.datasets for diabetic and breast cancer.
2. **Exploration and Preprocessing**:
   * Normalized all features to ensure uniform scaling and better model performance.
   * Added a bias term for intercept handling in linear models.
   * Visualized key features and relationships to understand correlations and patterns.
3. **Summary**: Dataset contains continuous features and target values suitable for regression tasks.

**Task 2: Implementation with Libraries**

* **Linear Regression with Ridge Regularization (L2):**
  + Implemented using Ridge from sklearn.linear\_model.
  + Grid search and 5-fold cross-validation were performed to identify the optimal regularization coefficient (alpha).
  + Performance Metric: Mean Squared Error (MSE).
* **Logistic Regression with L2 Regularization:**
  + Implemented using LogisticRegression from sklearn.linear\_model.
  + Grid search and 5-fold cross-validation were used to tune the regularization coefficient.
  + Performance Metrics: Accuracy and F1 Score.
* **Results:**
  + Optimal alpha for Ridge Regression minimized MSE significantly.
  + Logistic Regression achieved the best accuracy with balanced regularization.

**Part 3: Implementation WITHOUT Libraries**

* **Approach:**
  + Ridge Regression and Logistic Regression were implemented from scratch:
    - Ridge Regression: Gradient Descent with L2 penalty term.
    - Logistic Regression: Sigmoid activation with L2 penalty.
  + Cross-validation was performed using KFold from sklearn.model\_selection.
* **Challenges:**
  + Efficient implementation of the gradient descent for large datasets.
  + Ensuring convergence for all alpha values.
* **Performance:**
  + Results aligned closely with library-based implementation.
  + Ridge Regression: Computed MSE for each alpha.
  + Logistic Regression: Computed accuracy and F1 scores.

#### Part 4: Evaluation and Comparison

* **Metrics Used:**
  + Linear Regression: **Mean Squared Error (MSE).**
  + **Logistic Regression:** Accuracy.
* **Performance Visualization:**
  + Ridge Regression: MSE decreased as regularization increased but increased again for very high alpha values, showing underfitting.
  + Logistic Regression: Accuracy peaked at an optimal regularization coefficient and dropped for too small or too large values of alpha.
* **Bar Charts:**
  + Visualized MSE for Ridge Regression and accuracy for Logistic Regression across different regularization coefficients.

#### Part 5: Analysis and Reflection

* **Impact of Regularization:**
  + Ridge Regression:
    - Regularization helped prevent overfitting for smaller alpha.
    - Excessive regularization (large alpha) led to underfitting, increasing the MSE.
  + Logistic Regression:
    - Moderate regularization balanced bias and variance, improving accuracy.
    - Over-regularization reduced model flexibility, leading to lower accuracy.
* **Key Insights:**
  + Regularization is crucial for balancing model complexity and generalization.
  + Optimal regularization coefficients significantly improve performance, as observed in both regression and classification tasks.
* **Visualizations:**
  + Plots of MSE for Ridge Regression and accuracy for Logistic Regression highlight the impact of different regularization coefficients.

### Conclusion :

This exercise demonstrated the importance of regularization in both linear and logistic regression. Implementing these algorithms from scratch enhanced understanding of the underlying mechanics, while library-based implementations ensured scalability and ease of experimentation. Cross-validation was invaluable in identifying optimal hyperparameters and ensuring robust model evaluation.

**References**

Scikit-Learn. (n.d.). *Ridge regression*. Retrieved from <https://scikit->learn.org/stable/modules/linear\_model.html#ridge-regression

Scikit-Learn. (n.d.). *Logistic regression*. Retrieved from <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html>

*Starmer, J. (n.d.). StatQuest: Regularization (Ridge and Lasso Regression) [YouTube channel]. Retrieved from* [*https://www.youtube.com/watch?v=Q81RR3yKn30*](https://www.youtube.com/watch?v=Q81RR3yKn30)

*Simplilearn. (n.d.). Logistic regression in Python. [YouTube video]. Retrieved from* [*https://www.youtube.com/watch?v=zM4VZR0px8E*](https://www.youtube.com/watch?v=zM4VZR0px8E)