COMP551

Assignment 2: Linear Regression and Classification

Melody Mantegh - Junghoon Cho - Axel Refalo

March 2, 2025

Abstract

This study evaluates the performance of linear regression and logistic regression for binary and multiclass classification tasks. The models were implemented and evaluated on the Breast Cancer Wisconsin dataset for binary classification and the Wine Recognition dataset for multiclass classification. Following feature standardization and importance ranking, we evaluate model performance through accuracy for multiclass classification and AUROC for binary classification. Results show that logistic regression consistently outperforms linear regression in binary classification, particularly in cases where decision boundaries are nonlinear. Multiclass logistic regression achieves the highest accuracy across all models, surpassing K-Nearest Neighbors (KNN) and Decision Trees. These findings highlight the advantages of probabilistic modeling in classification tasks. They also underscore the importance of feature selection and model choice in optimizing predictive performance.

Introduction

This assignment compares the performance of linear regression, logistic regression, and multiclass classification models on two datasets: Breast Cancer Wisconsin and Wine Recognition. The Breast Cancer Wisconsin dataset, sourced from the UCI Machine Learning Repository, contains 30 numerical features extracted from fine needle aspirate (FNA) images, with the goal of distinguishing malignant from benign tumors (Wolberg & Mangasarian, 1993). Prior studies, such as Street et al. (1993), demonstrated that logistic regression and neural networks consistently outperform linear models, reinforcing the necessity of probabilistic modeling in medical diagnosis.

The Wine Recognition dataset, widely used in chemometrics, presents a multiclass classification task, predicting wine type based on 13 chemical composition features like alcohol content, flavonoids, and color intensity (Forina et al., 1991). Previous research (Bishop, 2006) suggests that multiclass logistic regression provides superior classification accuracy over linear methods, as it effectively models class probabilities.

We apply feature standardization and regression-based feature selection to refine our models. Using accuracy and AUROC scores, we assess the effectiveness of logistic regression over linear regression and examine the impact of feature importance on predictive performance.

1 Datasets

1.1 Breast Cancer Wisconsin

This dataset consists of 569 samples with 30 numerical features that are extracted from digitized images of fine needle aspirate (FNA) of breast masses. The features describe different cellular characteristics and the target variable "Diagnosis" classifies the tumor as benign (0) or malignant (1).
To assess feature importance, we compute the regression coefficients using the following formula: $\mathbf{w} = \mathbf{X}^\top \mathbf{y}/N \in \mathbb{R}^{D\times 1}$

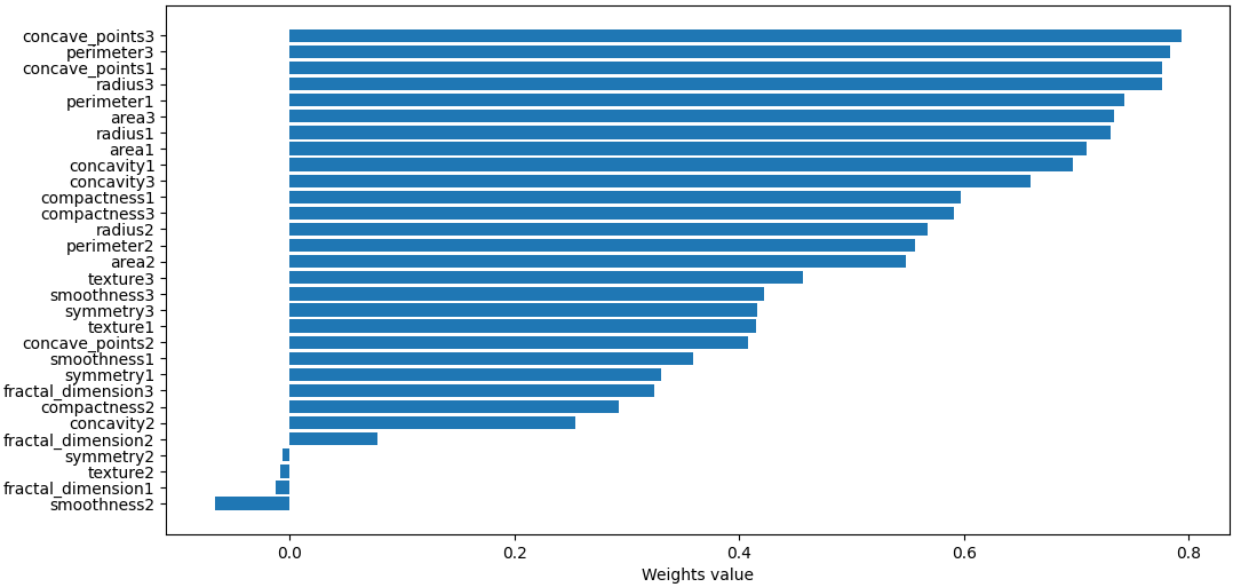

Figure 1: Barplot of the features weight for the breast cancer dataset

From the bar-plot the following features can be considered irrelevant: `fractal_dimension1`, `fractal_dimension2`, `texture2`, `symmetry2` and `smoothness2`

Furthermore, to analyze feature relevance, we examined pairwise plots (**Appendix 1**), which revealed substantial overlap between benign (0) and malignant (1) diagnoses in these features. The absence of clear separation or discernible trends suggests minimal correlation with the target variable, indicating that these features contribute little to predictive performance. Consequently, their removal is justified to enhance model efficiency and accuracy.

1.2 Wine Recognition

This dataset contains the chemical analysis of 13 different features from 178 wines. The wines all come from the same region in Italy but belong to three different cultivars. The goal of this dataset is to identify the cultivar using a supervised machine learning method. In our case, we will perform both multi-class linear regression and multi-class logistic regression on the dataset.

First, to highlight the relevant features of each cultivar or class, we compute the **Pearson Correlation Coefficient** for each class using: $\mathbf{W} = \tilde{\mathbf{X}}^\top \tilde{\mathbf{Y}}/N \in \mathbb{R}^{D\times C}$

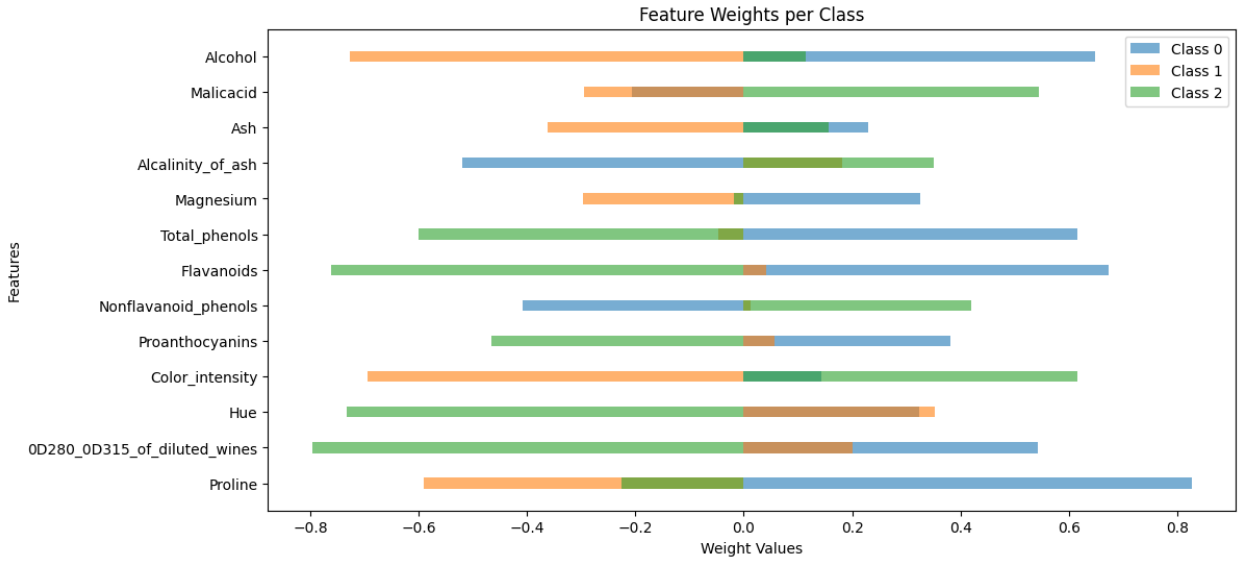

Figure 2: Bar plot of the feature weights for the wine recognition dataset

From the bar plot, we can see that certain features are more influential in distinguishing specific classes. However, we cannot disregard any features, as they all contribute to identifying one of the three classes.

2 Method

2.1 Linear Regression

Linear regression is a supervised learning method aimed at minimizing the Sum of Squared Error (SSE) **cost function**:

$$J(\mathbf{w}) = \sum_{n=1}^{N} (y^{(n)} - \hat{y}^{(n)})^2$$

As $J(w)$ is a convex function, we can find its global minimum by setting its derivative to zero. The solution to this equation is the closed-form solution:

- For **binary classification**, it is: $\mathbf{w} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$
- For **multi-class classification**, it is: $\mathbf{W} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$

Note: Computing multi-class linear regression is equivalent to performing a simple linear regression for each class.

2.2 Logistic Regression

Logistic regression is a classification method that uses the sigmoid function to transform a linear combination of features into a probability score. It is trained by minimizing the cross-entropy loss, ensuring that predicted probabilities align with actual class labels.

- **binary logistic regression** aims to minimize the cross-entropy $J(\mathbf{w}) = -y\log(\hat{y}) - (1-y)\log(1-\hat{y})$ where $\hat{y} = \frac{1}{1+e^{-a}}$, is the **sigmoid function** and $a = x \cdot w$
- **Multiclass logistic regression** extends this approach by minimizing the categorical cross-entropy loss: $J(\mathbf{W}) = -\sum_{c=1}^{C} y_c \log \hat{y}_c$ where $\hat{y}_c = \frac{e^{a_c}}{\sum_{c=1}^{C} e^{a_c}}$ is the **softmax function**, and $a_c = Xw_c$

The softmax function normalizes logits into a probability distribution across C classes.
To minimize the cost function, a **gradient descent** can be used as the cost function is concave. $\mathbf{w}^{(t)} = \mathbf{w}^{(t-1)} - \alpha \frac{\partial J(\mathbf{w}^{(t-1)})}{\partial \mathbf{w}^{(t-1)}}$ where $\alpha \in [0, 1]$ is the learning rate.

3 Results

3.1 Binary classification

3.1.1 Experiment on the learning rate and iterations on Gradient Descent

To determine the optimal hyperparameters for gradient descent, a series of manual experiments were conducted. The learning rate, α, was incrementally adjusted, and its impact on the cross-entropy loss was evaluated using the validation set. The optimal learning rate was selected based on its ability to produce a stable, smooth, and rapid decrease in loss without causing divergence.

In parallel, the maximum number of iterations was increased until the loss plateaued, indicating convergence. The convergence criterion was set to ensure that updates to the model weights became negligibly small, preventing unnecessary computations while maintaining high accuracy. These empirically derived hyperparameters were then applied to the test dataset to validate their effectiveness.

The final hyperparameter values, ensuring reliable convergence and optimal model performance, are summarized in Table 1.

learning rate	**max iterations**	**convergence criteria**	**threshold**
$\alpha = 0.005$	$max\_iter = 76000$	$\epsilon = 1e-8$	$threshold = 0.37$

Table 1: best hyper-parameters for the binary logistic regression

3.1.2 Comparative Analysis of Regression Coefficients

The feature importance ranking differs between simple regression and logistic regression due to their distinct optimization objectives. Simple regression minimizes squared errors, leading to coefficient estimates that capture overall variance in the dataset. In contrast, logistic regression prioritizes class separation, assigning greater weight to features that enhance discriminative performance.

As a result, the ranking of top features varies between the two models. While simple regression highlights features with strong linear correlations, logistic regression emphasizes those most

effective for binary classification. This difference underscores the importance of choosing feature selection methods tailored to the specific learning task.

The visual comparison in Figure 3 illustrates these differences, with simple regression favoring features contributing to overall variance, while logistic regression accentuates those that maximize class separability.

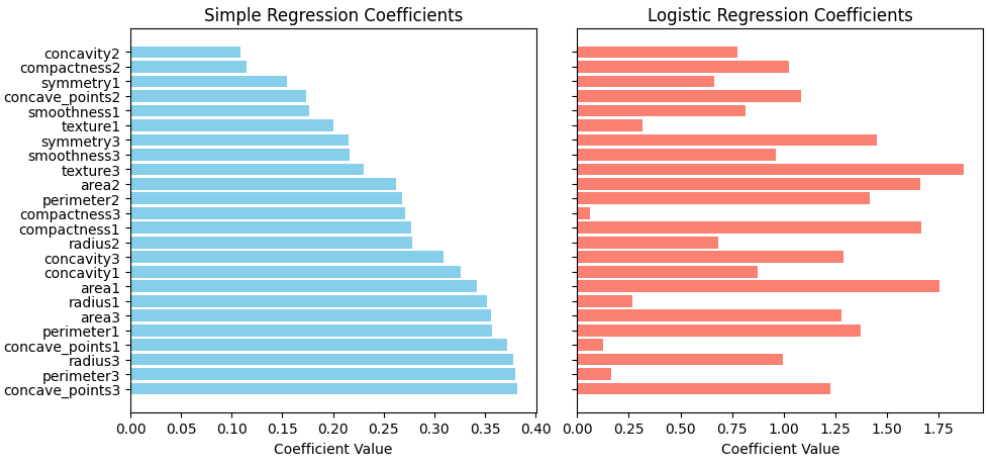

Figure 3: Top Features comparison between Simple Regression and Logistic Regression

3.1.3 Accuracy and AUROC evaluation

After experimenting with the hyper-parameters, we evaluate the **linear regression** and the **logistic regression** on binary classification. We first evaluate the model accuracy on the test data. The accuracy is evaluated by determining the threshold that gives the best accuracy on the test data. Then we evaluate, the AUROC. For reference, the same two tests are performed on the previous **KNN** and **DT** models.

Model	**accuracy**	**threshold**	**AUROC**
Linear Regression	97.7%	0.37	1.00
Logistic Regression	98.8%	0.37	0.998
K-Nearest Neighbors	96.5%	0.5	0.953
Decision Tree	93.3%	0.22	0.953

Table 2: Performance of models for binary classification on the breast cancer test data

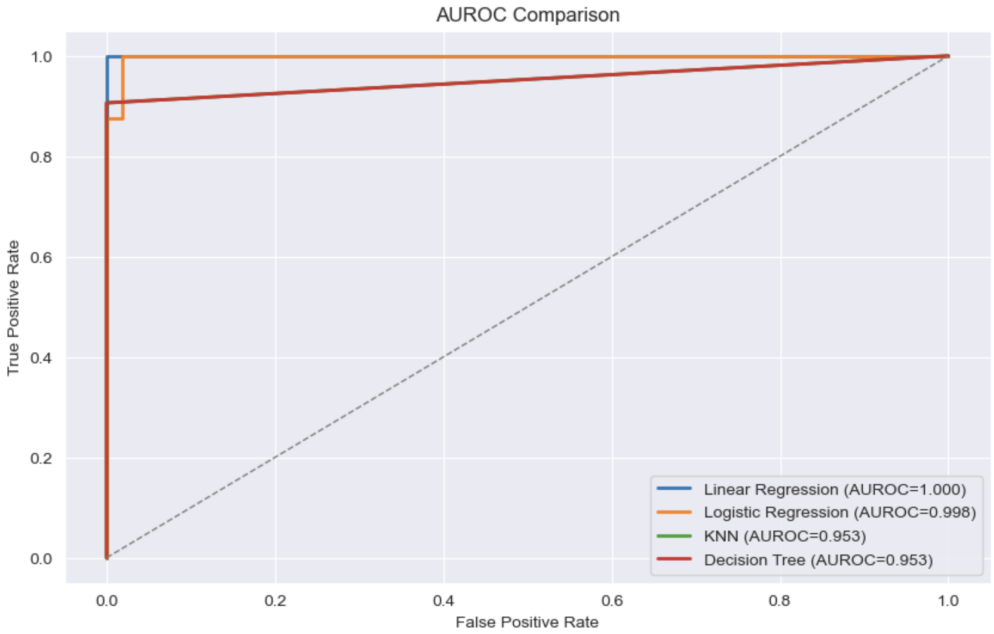

Figure 4: ROC models evaluation on the binary dataset

The results of these tests show us that the linear regression outperforms the logistic regression in terms of classification probabilities, whereas the logistic regression is the best model for the classification. These two models are also more performant than both KNN and DT models in terms of AUROC and accuracy.

3.2 Multi-class classification

3.2.1 Analytical gradients and the approximate gradients via numerical perturbation

To rigorously validate the implementation of our gradient function, we employed numerical gradient computation using a finite-difference approximation. Specifically, the analytical gradient was compared against the numerically computed gradient, yielding a discrepancy of approximately $3.66e-15$. This negligible difference confirms the accuracy of our gradient function implementation, thereby ensuring that the optimization procedures built on this foundation are based on the correct derivative calculation.

3.2.2 Experiment on the learning rate and iterations on Gradient Descent

A similar experimental procedure was performed for the multiclass logistic regression model to identify the optimal learning rate and maximum number of iterations. We adjusted these hyperparameters, observing the convergence behavior on the validation set, until we determined the combination that yielded the best performance on the test data.

learning rate	**max iterations**	**convergence criteria**
$\alpha = 0.001$	$max\_iter = 9500$	$\epsilon = 1e - 8$

Table 3: best hyper-parameters for the multi-class logistic regression

We monitored convergence by plotting both the training and validation losses on a single graph to assess the learning behavior and detect any potential overfitting. The overlapping trajectories of the two losses indicated that the model was generalizing well, and no overfitting occurred during training.

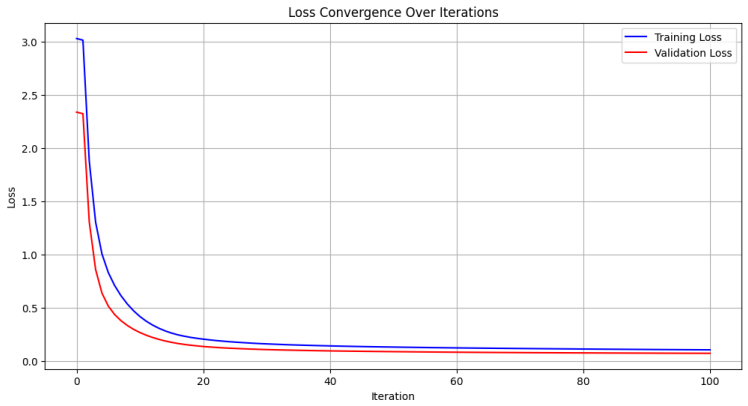

Figure 5: Convergence of the multi-class logistic regression dataset

In multi-class linear regression, each class is treated as a separate dimension in a continuous regression framework, with the goal of minimizing squared errors for each class label independently. By contrast, multi-class logistic regression optimizes the cross-entropy loss, which directly measures how well each feature combination separates the classes in probability space. Because these two objectives differ, one fitting continuous values versus one modeling class probabilities, their coefficient magnitudes and signs can vary. Consequently, the most influential features to distinguish each class can change, reflecting the different ways in which linear and logistic models weight and separate classes.

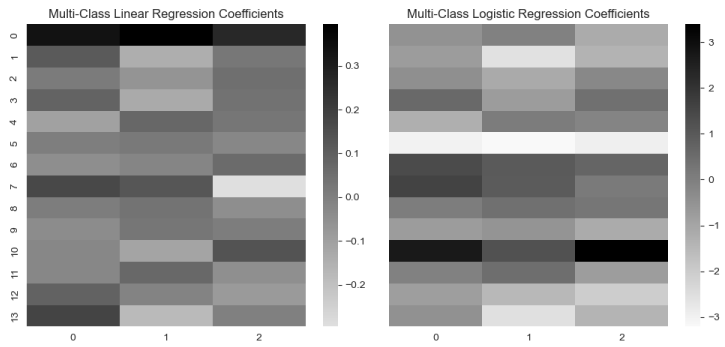

Figure 6: Multi-Class Linear Coefficients vs. Multi-Class Logistic Coefficients

3.2.3 Accuracy and AUROC evaluation

Same as before, we can evaluate the **multi-class linear regression** and **multi-class logistic regression** on the wine recognition test dataset by their respective accuracy for a certain threshold and their AUROC. As reference, we conduct the same experiment on the **KNN** and **DT** models.

Model (multi-class)	**Accuracy**	**Threshold**	**AUROC**		
			class 1	class 2	class 3
Linear Regression	99.63%	0.5	1.000	0.995	0.998
Logistic Regression	100%	0.5	1.00	1.00	1.00
K-Nearest Neighbors	94.44%	0.5	0.983	0.944	0.985
Decision Tree	88.89%	0.5	0.954	0.899	0.956

Table 4: Performance of models on multi-class classification on the wine recognition dataset

The **multi-class logistic regression** is once again outperforming other models on the prediction accuracy.

4 Discussion & Conclusion

In this assignment, we examined the performance of linear regression and logistic regression on classification tasks using the Breast Cancer Wisconsin and Wine Recognition datasets. Our findings confirm that logistic regression is better suited for classification tasks, as it explicitly models class probabilities and optimizes the cross-entropy loss, which directly minimizes misclassification errors. While linear regression surprisingly achieved a high AUROC, it lacks a clear decision boundary, making its classification performance unreliable compared to logistic regression.

A key observation was the difference in feature importance rankings between the two models. Linear regression assigns importance to features based on their variance, while logistic regression prioritizes features that maximize class separability. This distinction explains why some features, such as cell radius and texture in the Breast Cancer dataset, were important in both models, but logistic regression placed more emphasis on features directly linked to malignancy classification.

From a computational efficiency standpoint, linear regression is faster due to its closed-form solution, whereas logistic regression requires iterative optimization via gradient descent. However, the added computational complexity of logistic regression is justified by its superior predictive accuracy and ability to model decision boundaries effectively.

These results highlight several key takeaways: choosing the right loss function is critical for classification tasks, feature selection significantly impacts model performance, and logistic regression remains a strong baseline classifier due to its probabilistic nature and robustness across datasets. Future work could explore regularization techniques and nonlinear transformations to further enhance classification performance.

5 Statement and Contribution

Axel and Junghoon were responsible for implementing the Logistic Regression and Multi-Class Classification models, including data preprocessing, hyperparameter tuning, and performance evaluation. They also contributed to writing sections of the paper, focusing on the methodology and results. Melody focused on writing the other half of the paper, including the introduction, abstract, and discussion and conclusion sections.

6 References

Wolberg, W., Mangasarian, O., Street, N., & Street, W. (1993). Breast Cancer Wisconsin (Diagnostic) [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5DW2B.

Aeberhard, S. & Forina, M. (1992). Wine [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5PC7J.

Street, W.N., Wolberg, W.H., & Mangasarian, O.L. (1993). Nuclear feature extraction for breast tumor diagnosis. Electronic imaging.

7 Appendices

7.1 Appendix 1: pairwise correlation for the irrelevant features of the breast cancer dataset