

**Department of Electrical Engineering**

**University of Moratuwa**

**EN 3150 – Pattern Recognition**

**Learning from data and related challenges and classification**

Kularathna A. K. D. D. D. – 220332P

Date of Submission – 07/09/2025

This is submitted as a partial fulfilment for the module

EN 3150 Pattern Recognition

Department of Electronics and Telecommunication Engineering

University of Moratuwa

Table of Contents

[1. Linear Regression 3](#_Toc208098306)

[1.1. Why can the OLS line be “not aligned to the majority of data points”? 3](#_Toc208098307)

[1.2. Which weighting scheme gives a better line for inliers? Justify. 3](#_Toc208098308)

[1.3. Why is plain linear regression unsuitable for “which brain regions predict the task”? 6](#_Toc208098309)

[1.4. Next, the following two methods are being considered: 7](#_Toc208098310)

[1.5. Which method (LASSO or group LASSO) is more appropriate in this setting, and why? 7](#_Toc208098311)

[2. Logistic Regression 8](#_Toc208098312)

[2.1. Use the code given in listing 1 to load data. 8](#_Toc208098313)

[2.2. Now, use the code given in listing 2 to train a logistic regression model. Here, did you encounter any errors? If yes, what were they, and how would you go about resolving them? 8](#_Toc208098314)

[2.3. Why does the saga solver perform poorly? 11](#_Toc208098315)

[2.4. Now change the “saga” to “liblinear” and check the accuracy. 11](#_Toc208098316)

[2.5. Why does the "liblinear" solver perform better than "saga" solver? 13](#_Toc208098317)

[2.6. Explain why the model’s accuracy (with saga solver) varies with different random state values? 14](#_Toc208098318)

[2.7. Compare the performance of the "liblinear" and "saga" solvers with feature scaling. If there is a significant difference in the accuracy with and without feature scaling, what is the reason for that. You may use Standard Scaler available in sklearn library. 15](#_Toc208098319)

[2.8. Suppose you have a categorical feature with the categories ’red’, ’blue’, ’green’, ’blue’, ’green’. After encoding this feature using label encoding, you then apply a feature scaling method such as Standard Scaling or Min-Max Scaling. Is this approach correct? or not? What do you propose? 17](#_Toc208098320)

[3. Logistic regression First/Second-Order Methods 18](#_Toc208098321)

[3.1. Use the code given in listing 3 to generate data. Here, variable y and X are class labels and corresponding feature values, respectively. 18](#_Toc208098322)

[3.2. Implement batch Gradient descent to update the weights for the given dataset over20 iterations. State the method used to initialize the weights and reason for your selection. 19](#_Toc208098323)

[3.3. Specify the loss function you have used and state reason for your selection. 23](#_Toc208098324)

[3.4. Implement Newton’s method to update the weights for the given dataset over 20 iterations. 23](#_Toc208098325)

[3.5. Plot the loss with respect to number of iterations for batch Gradient descent and Newton methods in a single plot. Comment on your results. 26](#_Toc208098326)

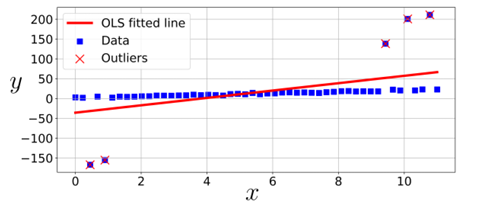
[3.6. Propose two approaches to decide number of iterations for Gradient descent and Newton’s method. 29](#_Toc208098327)

[3.7. Suppose the centers in listing 3 are changed to centers = [2, 2], [5, 1.5]]. Use batch Gradient descent to update the weights for this new configuration. Analyze the convergence behavior of the algorithm with this updated data and provide an explanation for convergence behavior. 31](#_Toc208098328)

[Reference 35](#_Toc208098329)

[Appendix 35](#_Toc208098330)

# Linear Regression

****

A black text on a white background

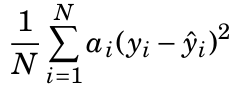
AI-generated content may be incorrect.**Ordinary Least Squares (OLS)**

## Why can the OLS line be “not aligned to the majority of data points”?

In the data we can see there are some outliers. Due to these outliers the residual becomes larger than the residual for an inlier. Since the OLS uses the square of the residual the impact from the outliers becomes very larger. Due to these effects of the outliers OLS fitted line is not aligned to majority of data points.

## Which weighting scheme gives a better line for inliers? Justify.

Let’s consider the modified loss function



Proposed scheme

* Scheme 1: For outliers ai = 0.01 and for inliers ai = 1
* Scheme 2: For outliers ai = 5 and for inliers ai = 1

Let’s check how these schemes affect the fitted lines using an approximate

Example

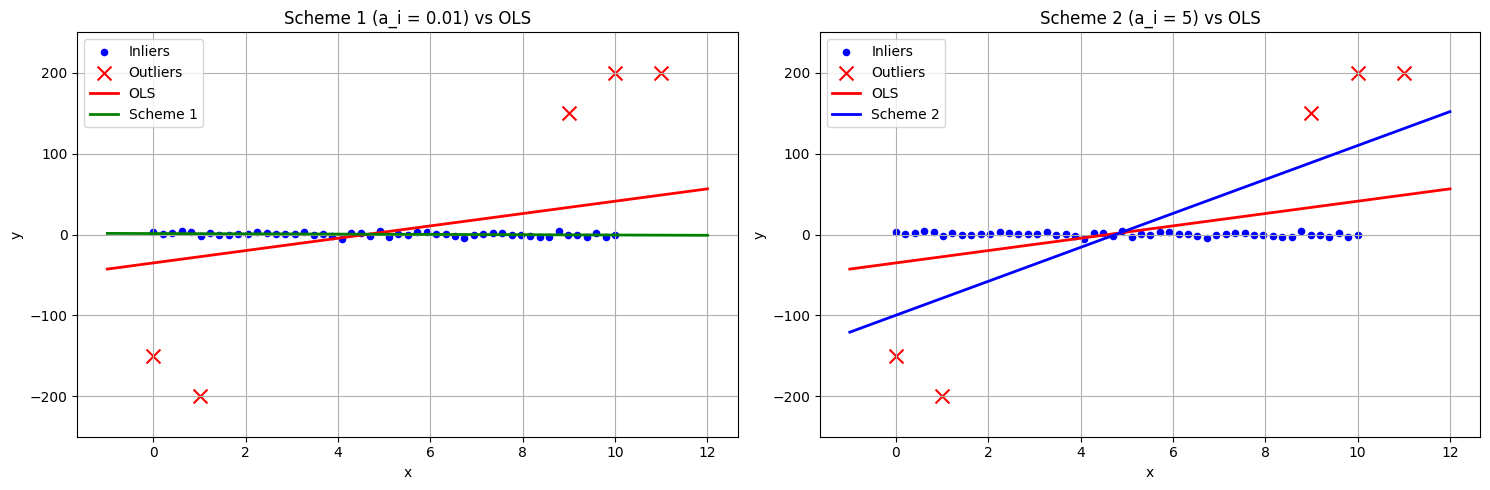
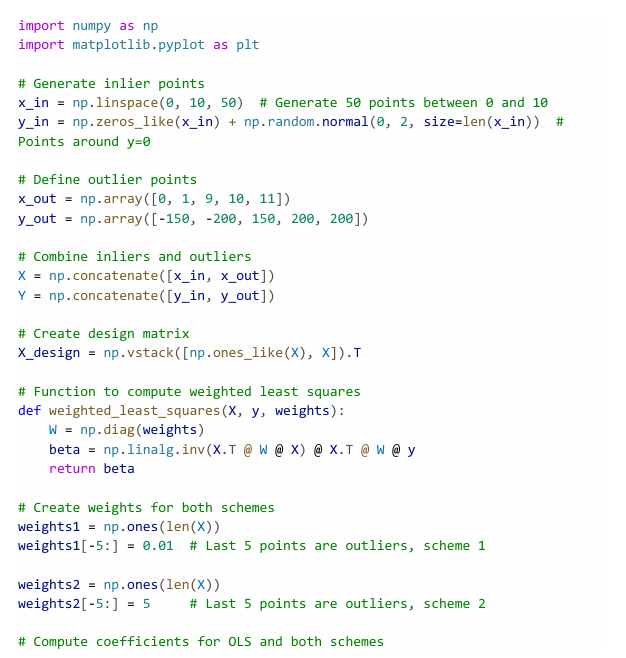
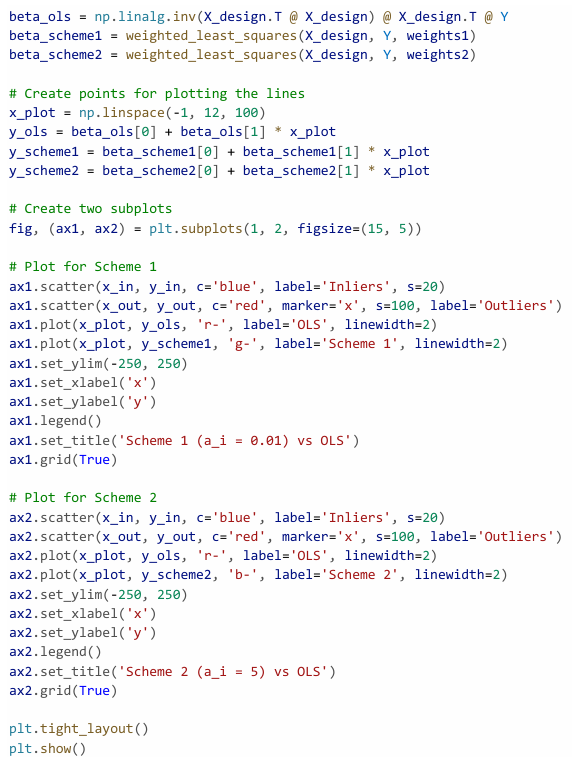


Figure .1

Code for the plots



Code .1

We can see that scheme 1 gives a better fitted line than scheme 2.

This is because it reduces weights the outliers reducing their impact therefore the fitted line aligns with most of the data points. However, scheme 2 does the exact opposite, it increases the weight of the outliers. This results in a greater deviation in the fitted line with actual inlier data points.

## Why is plain linear regression unsuitable for “which brain regions predict the task”?

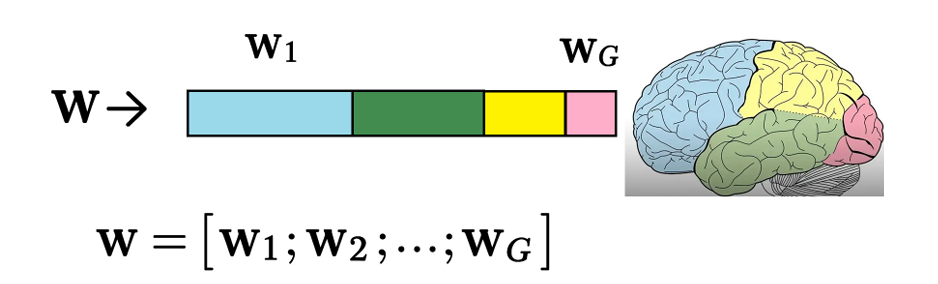


Figure 1.2

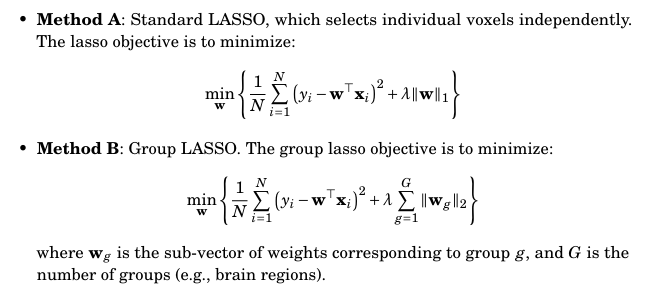
In brain image analysis (e.g., fMRI), the brain is divided into multiple regions as shown Figure 2, each consisting of many voxels (pixels). A researcher wants to identify which brain regions are most predictive of a specific cognitive task.

Each of the divided regions of the brain has many voxels, and we want to identify which brain regions are most predictive of a specific cognitive task.

For this task these are the issues that we encounter if we used plain Linear Regression.

* It has too many features but too few samples
  + In brain studies, we might have many voxels but only few people in the experiment. Linera regression needs more data than features to work properly. In our case since we have more features than subjects, the model can overfit. It will not be generalized.
* Features are highly correlated
  + Voxels inside the same brain region behave similarly. When predictors are almost copies of each other, the regression struggles to predict which voxel gets how much weight. Small changes in the data can make the weights change so much. This makes the results unstable and unreliable.
* No automatic feature selection
  + Ordinary regression tries to use all voxels at once. It doesn’t know how to ignore the irrelevant voxels. Due to this it is hard to know which voxels are important for predicting the task. Due to this reason, we will just end up with bunch of coefficients, many of which may be just noise.
* Ignores natural grouping of voxels into regions
  + Linear regression looks at voxels one by one, but it doesn’t consider the group structure. Therefore, we might get scattered voxels selected across the brain, which might not make sense biologically.

## Next, the following two methods are being considered:



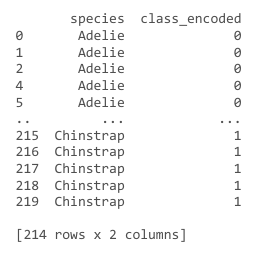
## Which method (LASSO or group LASSO) is more appropriate in this setting, and why?

* + Standard LASSO (Method A), this regression method adds an extra penalty to the model and because of that some coefficients were reduced up to zero. By doing this we can selects individual voxels. This is good for sparsity, but it tends to pick scattered voxels from correlated groups, which is hard to map back to the regions.
  + Group LASSO (Method B), Acts same as Standard LASSO but it respects group structure. What is special about Group LASSO is it adds a penalty that selects or drop entire regions.
  + Therefore, for this application Group LASSO is more suitable because it selects entire brain regions rather than scattered voxels. This makes the results more stable, easier to interpret.

# Logistic Regression

## Use the code given in listing 1 to load data.

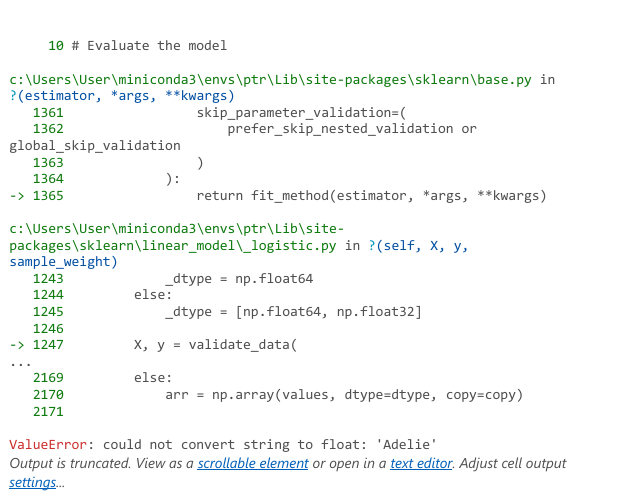
Listing 1’s code gives following output

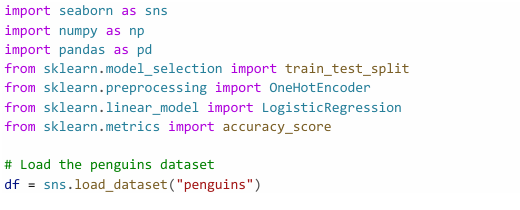


## Now, use the code given in listing 2 to train a logistic regression model. Here, did you encounter any errors? If yes, what were they, and how would you go about resolving them?

A screenshot of a computer code

AI-generated content may be incorrect.Yes, I encountered following error,



* + The error occurs because the logistic regression model can't handle categorical data. This happens because there are some string values in the columns like "species” and some other categorical features that need to be converted into numerical values.
  + As Solution we can use encoding to label those data. For this I used One-hot encoding. We can also use LebelEncoding too for this task, but One-hot encoding is better at handing these kinds of tasks.
  + Following code is the corrected one.

A screenshot of a computer code

AI-generated content may be incorrect.

A close-up of a computer code

AI-generated content may be incorrect.

Code 2. 1

A screenshot of a computer code

AI-generated content may be incorrect. This code gives following output

## Why does the saga solver perform poorly?

* + "saga" solver often needs feature scaling for proper convergence. Without scaling, the optimization problem is ill-conditioned. This gives following warning.
* ConvergenceWarning: The max\_iter was reached which means the coef\_ did not converge warnings.warn(
  + Default `max\_iter=100` This is usually too small. It should be set to a higher value.
  + These are the reasons why “saga” solver performs poorly.

## Now change the “saga” to “liblinear” and check the accuracy.

* + - Accuracy with liblinear solver
    - When used "liblinear”:
    - Accuracy = 1
    - When comparing to "saga”, "liblinear" shows perfect classification

A screenshot of a computer program

AI-generated content may be incorrect.

A computer screen shot of a code

AI-generated content may be incorrect.

Code 2.2

The code gives following output

A screenshot of a computer code

AI-generated content may be incorrect.

## Why does the "liblinear" solver perform better than "saga" solver?

According to the results

* + - saga: - Accuracy = 0.5813953488372093 with `ConvergenceWarning`
    - liblinear: - Accuracy = 1.0

Reasons behind these observations

1. Dataset size
   * + "liblinear" is designed for small to medium dense datasets like penguins.
     + However, "saga" is designed and optimized for very large datasets or with support to sparse high-dimensional problems. For small dense data, "saga" is often slower.
2. Optimization methods
   * + "liblinear" uses a coordinate descent/ dual optimization method, which is deterministic and converges reliably for binary classification.
     + But "saga" solver uses a stochastic gradient-based slover, which means it updates parameters using random mini-batches. On small data, this introduces noise and slow convergence.
3. Feature scaling
   * + "saga" is sensitive to the unscaled features because its step size depending on the feature magnitudes. Without scaling, optimization is ill-conditioned.
     + However liblinear is much less sensitive to feature scaling.
4. Convergence
   * + In our situation "saga" hit maximum iteration limit 100 without converging
     + However, "liblinear" converged properly. This led to perfect separation of the two species.

## Explain why the model’s accuracy (with saga solver) varies with different random state values?

1. Train/test split randomness

* "random\_state" is affecting how the dataset is split into training data and test data.
* With small dataset like penguins, changing the split can easily change which samples are in training vs test. Because the "saga" solver doesn’t converge well, even the small differences in the split lead to large changes in accuracy.

1. Stochastic optimization

* "saga" uses randomness in it for updating coefficients.
* Different random seeds lead to different optimization paths; this happens especially if the solver has not converged.
* Because it stops early, "saga" results depend more on that random path.

## Compare the performance of the "liblinear" and "saga" solvers with feature scaling. If there is a significant difference in the accuracy with and without feature scaling, what is the reason for that. You may use Standard Scaler available in sklearn library.

* With feature scaling
  + "saga" solvers accuracy improves significantly and converges becomes reliable
  + "liblinear" accuracy stays same because it is less sensitive to the feature scaling
* Reasons
  + "saga" is gradient-based stochastic solver, because of that it is very sensitive to feature magnitudes.
  + By scaling all features become comparable, hence improving optimization.
  + "liblinear" uses coordinate decent, it is less affected by scaling.
* Conclusion
  + Scaling makes "saga" better. but there is no significant difference for "liblinear".

A screen shot of a computer code

AI-generated content may be incorrect.



A computer code with text

AI-generated content may be incorrect.

Code 2.3

A screenshot of a computer code

AI-generated content may be incorrect. The code gives the following output

## Suppose you have a categorical feature with the categories ’red’, ’blue’, ’green’, ’blue’, ’green’. After encoding this feature using label encoding, you then apply a feature scaling method such as Standard Scaling or Min-Max Scaling. Is this approach correct? or not? What do you propose?

* This approach is incorrect.
* Here are the reasons why it is incorrect.

1. Artificial Ordering

* Since label encoder encode categories to numbers like red = 0, blue = 1, green = 2, The scaling of these values makes less sense because it assumes the distance between categories are meaningful.

1. Distance Problem

* After scaling, the distances between categories would be treated as meaningful
* For example, if red = 0 then blue = 1 and green = 2.
* Scaling would preserve these relative distances suggesting blue is between red and green.
  + - The proposed approach
      * + use **One-Hot Encoding**

A white grid with black text

AI-generated content may be incorrect.

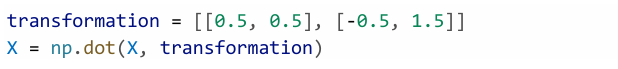
# Logistic regression First/Second-Order Methods

## Use the code given in listing 3 to generate data. Here, variable y and X are class labels and corresponding feature values, respectively.

Code used for this question

A screen shot of a computer code

AI-generated content may be incorrect.

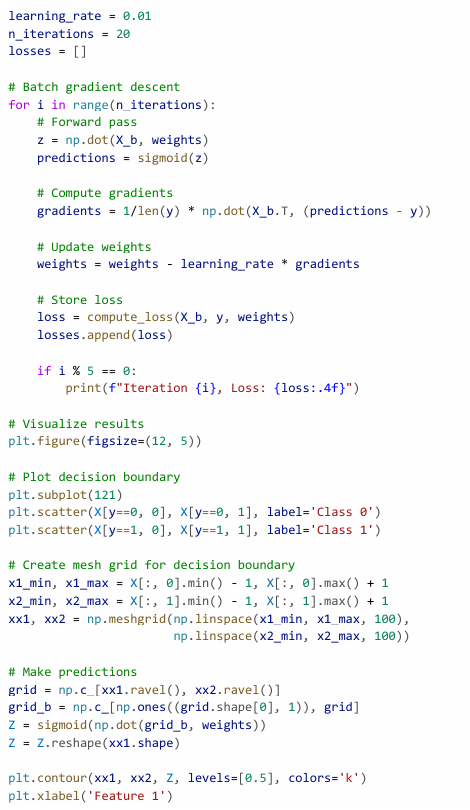


Code 3.1

## Implement batch Gradient descent to update the weights for the given dataset over20 iterations. State the method used to initialize the weights and reason for your selection.

A screenshot of a computer program

AI-generated content may be incorrect.Code implementation

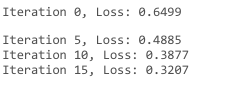


A screen shot of a computer program

AI-generated content may be incorrect.

Code 3.2

This code gives following output



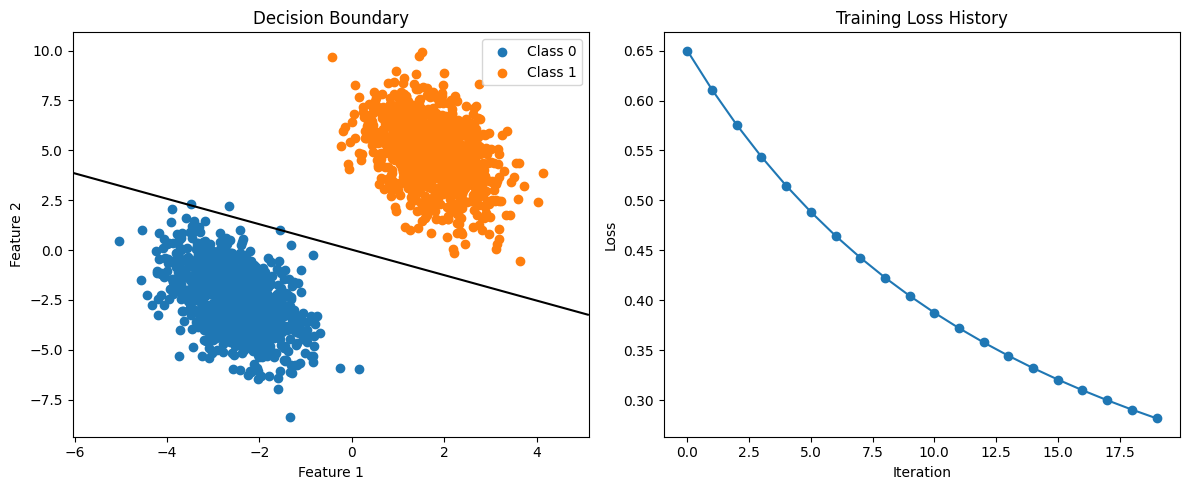


Figure 3.1

Final weights: [-0.0049357 0.15122723 0.23719972]

**Weight Initialization Method and Reasoning**

1. **Method Used**: Zero Initialization

  weights = np.zeros(X\_b.shape[1])  # 3 weights (bias, w1, w2)

1. **Reasons for Choosing Zero Initialization**:

* **Problem Type**: Binary classification with well-separated classes from “make\_blobs”
* **Convex Optimization**: Logistic regression is a convex problem, guaranteed to converge
* **Symmetry**: Equal starting point for both classes, no initial bias
* **Simplicity**: Deterministic initialization, reproducible results
* **Data Characteristics**: Classes are linearly separable, no need for complex initialization

1. **Why Other Methods Were Not Chosen**:

* **Random Initialization**: Not needed for convex problems
* **Small Random Values**: Would add unnecessary randomness
* **Xavier/Glorot**: Overkill for simple logistic regression
* **Constant Non-zero**: No advantage over zeros for this case

1. **Validation**:

* Loss curve shows proper convergence
* Decision boundary successfully separates classes
* No symmetry breaking issues encountered

## Specify the loss function you have used and state reason for your selection.

The loss function that is used here is the Binary Cross Entropy (BCE) Loss function. Here are the reasons for the selection

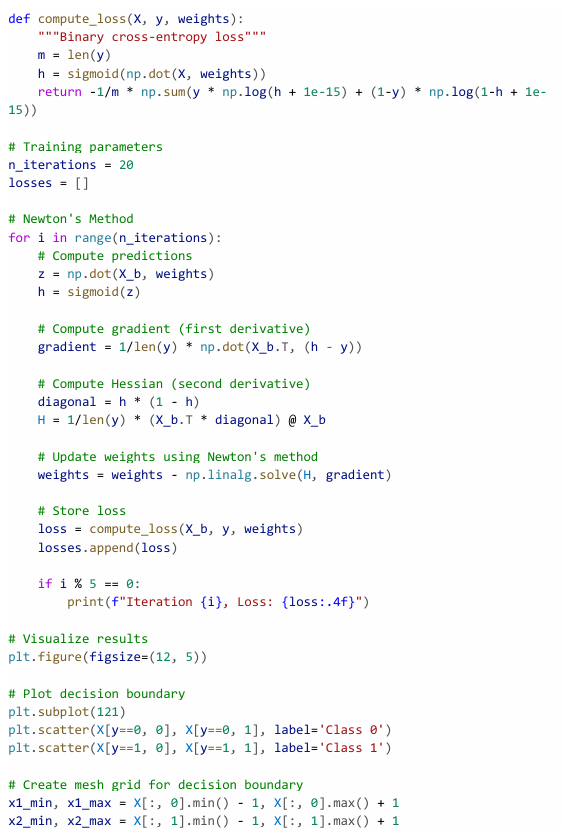
* Probabilistic Properties
* Output probabilities are between 0 and 1 through sigmoid function. And this is the common choice for binary classification problems.
* Mathematical Advantages
* It is a convex function that guarantees global minimum. And it is differentiable therefore can use gradient decent. Also, BCE penalizes confident wrong predictions heavily
* Compatibility
* The BCE pairs naturally with sigmoid activation. Because of that it provides proper probability estimates.
* Also efficiently works with batch gradient decent.

## Implement Newton’s method to update the weights for the given dataset over 20 iterations.

Code for the implementation

A screen shot of a computer code

AI-generated content may be incorrect.



A screenshot of a computer program

AI-generated content may be incorrect.

Code 3.3

Output of the following codeA number of numbers and symbols

AI-generated content may be incorrect.

A graph and diagram of a graph

AI-generated content may be incorrect.

Figure 3.2

Final weights: [-2.93259517 10.56207842 4.32766252]

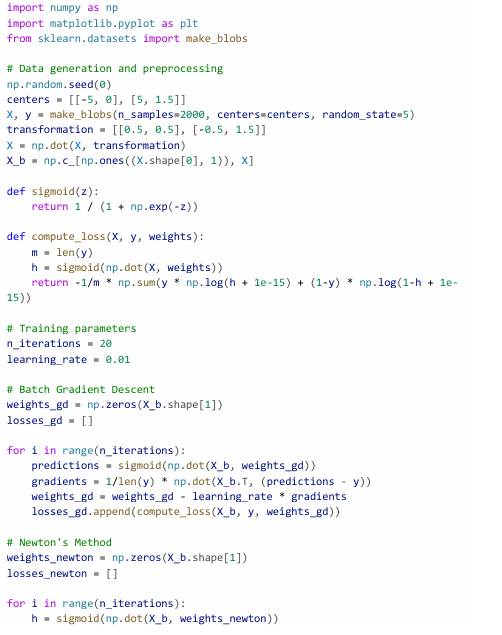
Final loss: 2.870387964766699e-09

## Plot the loss with respect to number of iterations for batch Gradient descent and Newton methods in a single plot. Comment on your results.

A graph with a blue line and red line

AI-generated content may be incorrect.

Figure 3.3

 Code used for this plot



Code 3.4

Output of this code

Final loss (GD): 0.2817968058115373

Final loss (Newton): 2.870387964766699e-09

**Observed behavior**

* Batch Gradient Decent gives a steadier linear but slow decrease in loss
* Newton's method loss has a very large decrease compared to the Batch gradient decent.

**Reasons**

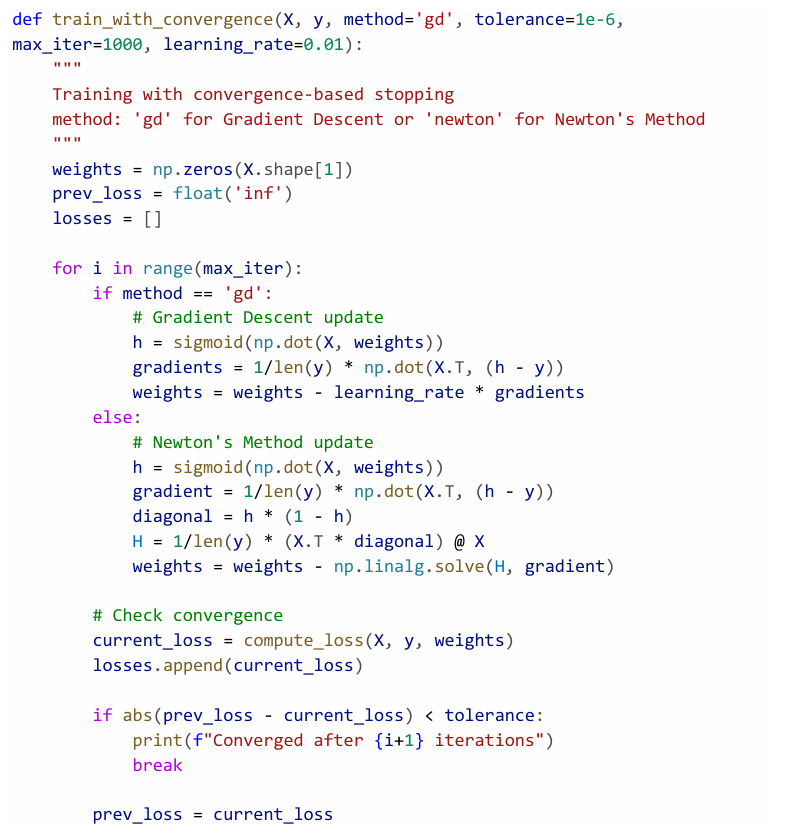
* The reason behind this is Newton's method uses second-order curvature information, giving a 2nd order convergence near the optimum.
* Overall, Newton's method is more efficient, but computational cost is very high for each iteration.

## Propose two approaches to decide number of iterations for Gradient descent and Newton’s method.

1. **Convergence-based Approach**

* Best for well-behaved, convex problems
* Monitors change in loss between iterations
* More suitable for Newton's method due to quadratic convergence
* Recommended tolerance: 1e-6 for Newton's, 1e-4 for GD

Code implementation



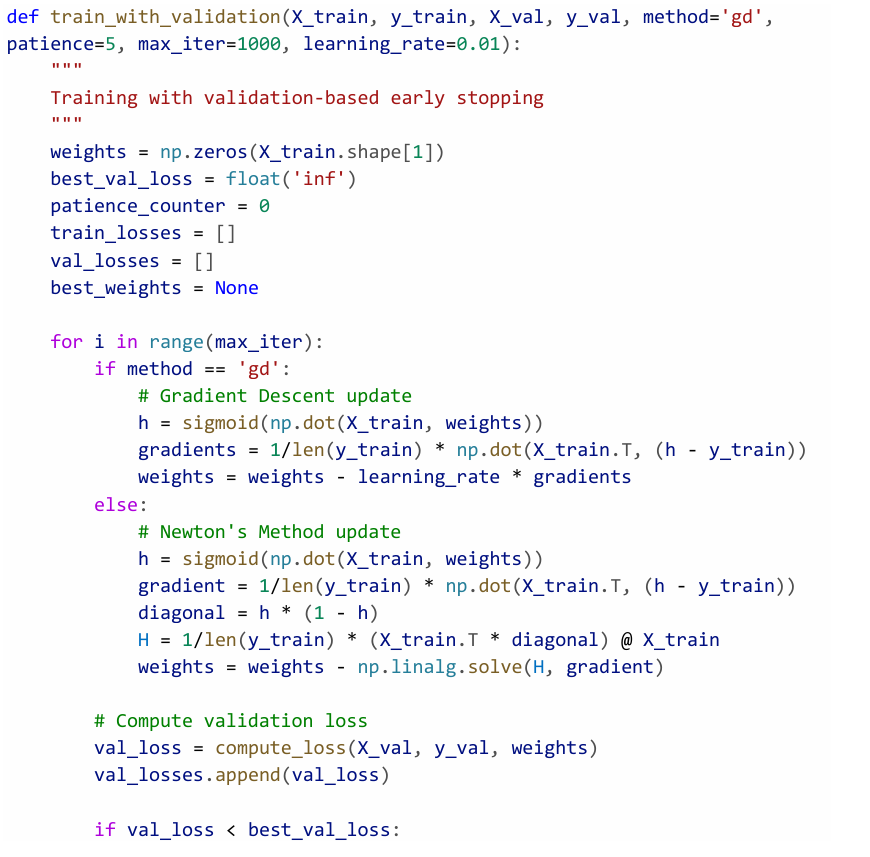
    return weights, losses

Code 3.5

1. **Validation-based Approach**

* Better for preventing overfitting
* More robust to noisy data
* Particularly useful for Gradient Descent
* Recommended patience: 5-10 iterations

Code implementation



A computer code with numbers and symbols

AI-generated content may be incorrect.

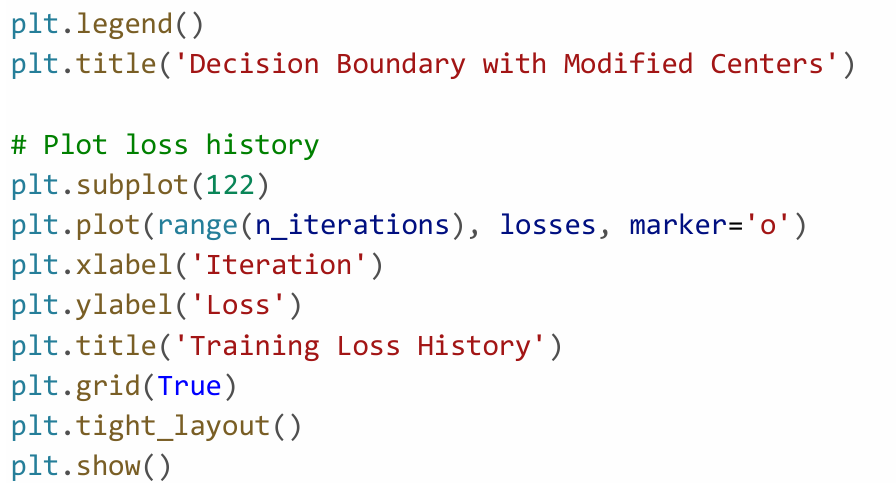
Code 3.6

## Suppose the centers in listing 3 are changed to centers = [2, 2], [5, 1.5]]. Use batch Gradient descent to update the weights for this new configuration. Analyze the convergence behavior of the algorithm with this updated data and provide an explanation for convergence behavior.

Code after making changes

A screenshot of a computer program

AI-generated content may be incorrect.



Code 3.7

A number of numbers and a loss

AI-generated content may be incorrect.Output of this code

A graph and diagram with numbers

AI-generated content may be incorrect.

Figure 3.4

Analysis of Convergence Behavior:

1. **Slower Convergence**

* The loss function shows slower convergence compared to the original centers
* Classes are closer together, making the decision boundary harder to find
* Requires more iterations to reach optimal separation

1. **Higher Final Loss**

* Final loss value is higher than with the original centers
* Classes have more overlap due to closer centers
* Perfect separation is more difficult to achieve

1. **Gradient Behavior**

* Gradients are smaller due to less distinct class separation
* Learning becomes more sensitive to learning rate
* Risk of getting stuck in suboptimal solutions increases

1. **Stability Issues**

* More sensitive to initialization
* May require careful tuning of learning rate
* Could benefit from adaptive learning rates

1. **Recommendations**

* Increase number of iterations
* Consider using adaptive learning rates
* May need to try different initializations
* Monitor for potential overfitting

# Reference

1. [***Lasso — scikit-learn 1.7.1 documentation***](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html)
2. [***LinearRegression — scikit-learn 1.7.1 documentation***](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html)
3. [***LogisticRegression — scikit-learn 1.7.1 documentation***](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html)
4. [***Cost Function and Gradient Descent | Sinhala***](https://www.youtube.com/watch?v=dGVdGJlhjCw&list=PL495mke12zYDHN9ONfcal1eQfo8VqmOgu&index=7)
5. [***Linear Regression | Supervised Learning Algorithm | Sinhala***](https://www.youtube.com/watch?v=yl1psqJf-3Q&list=PL495mke12zYDHN9ONfcal1eQfo8VqmOgu&index=5)
6. [***Logistic regression | Machine Learning Algorithm | Sinhala***](https://www.youtube.com/watch?v=WUeMWfim3GE&list=PL495mke12zYDHN9ONfcal1eQfo8VqmOgu&index=10)

# Appendix

**GitHub Repo link for the Assignment 02 -** [**Link**](https://github.com/DilshanKularathna/EN3150-Assignment02)