ECSE551
Mini-project 1

Amirhossein Khademi,
Amin Fattahi,
Reza Ghasemi Pirbalouti

25 Feb 2024

Abstract

In this report, we tried to assess the performance of a Logistic Regression Binary classifier on two different data sets. The First data set was the data on Breast Cancer, and the second one was the data on the Ionosphere. The classifier was trained using gradient descent and the final models were chosen using 10-fold cross-validation. Also, all the algorithms have been implemented from scratch without using any of the Scikit-learn libraries. In the first data set, at the first step, we implemented the Logistic Regression classifier (Baseline) without any change in data, then we tried to implement three other improvement strategies including Normalization, adding higher-order features (HOF), and Principal Component Analysis (PCA) to improve the performance of the algorithm. In the second data set, after implementing the algorithm, we tried two different methods for improving the performance, including HOF, and PCA. We used the Mean Accuracy of the folds and the Elapsed Time as performance measures. For the First data set, HOF and the PCA significantly improve the accuracy(0.95, and 0.94 respectively) in comparison to baseline. Also, the Elapsed Time of the PCA (0.41 seconds) is approximately half all other methods. So, it seems that the PCA is the better method in comparison to other improvement methods in the first data set. For the second data set, the HOF Mean accuracy and the Elapsed Time are 0.91, and 0.8, respectively, which represent noticeable improvements compared to the baseline methods.

1 Introduction

In this project, we used the Logistic Regression Binary classifier to classify our data into two different classes. According to the project description, at first, we implemented data analysis and visualization on both data sets in section 2, then implemented the Logistic Regression Binary classifier from scratch in Python without using any of the Scikit-learn libraries. Then, we improved the accuracy and time through various improvement methods. For this aim, we implemented Normalization of the input data, adding higher-order features, and PCA. Also, we used K-fold cross-validation. The code, which is attached to this report, is class-implemented.

1.1 The Logistic Regression Binary classifier

In statistics, the Logistic Regression model is a popular statistical model used largely for classification. It means that given a set of observations, the Logistic Regression technique can help us classify them into two or more discrete classes. In Logistic Regression, one attempts to directly model the decision boundary between two binary class labels by estimating $P(y_i = 1|x_i)$ and therefore $P(y_i = 0|x_i)$ through the use of the Sigmoid function σ (Equation 1). The output of the Sigmoid function is interpreted as a probability (Equation 2 and 3) ranging between 0 and 1 while its argument corresponds to the log-odds ratio of $P(y = 1|x)$ over $P(y = 0|x)$ and is modeled as a weighted linear combination of the input data feature values (Figure 2).
In other words, we will have two probabilities for x_i, and the class with a higher probability

will be the final predicted class of x_i. To train the binary classifier, the logistic function's log-likelihood can be maximized by minimizing the associated Cross-Entropy Loss (Equation 4) on the training set using an optimization algorithm like Gradient Descent. Gradually optimizing the weight matrix will reduce the binary classifier's error by selecting the right search direction and learning rate (lr), which have to be defined in the algorithm as hyper-parameters [1]. Lastly, the trained model immediately determines the class label probabilities on the test set and applies the appropriate class label during classification. The ratio of correctly predicted class labels to the total number of test samples is then used to assess the final reported test accuracy.

1.2 Data sets, and important findings

The first data set is the Breast Cancer modifier data set, and the second one is The Ionosphere data set which will be discussed in the next part. Normalization, Adding Higher Order Features (HOF), and Principal Component Analysis (PCA) [3] have been used to enhance the performance of the classifier model. A brief explanation of these improvement methods can be found in the appendix, section 6.2.
According to the findings of this project, for the first data set the best approach was using PCA which achieved an accuracy of 95 percent with an Elapsed Time of 0.41 seconds, and for the second data set, the best approach was using HOF which achieved an accuracy of 92 percent, with an Elapsed Time of 0.8. Also, it is worth mentioning that given the wide range of the input values in the 1^{st} data set, the application of normalization techniques becomes necessary.

2 Data Sets

2.1 Overall

In this mini-project, the Logistic regression algorithm was developed for binary classification projects using two popular datasets: "Breast cancer modifier" and "Ionosphere modified". The Breast cancer modifier dataset is an essential tool in medical research, created to investigate and determine various factors that contribute to breast cancer through machine learning algorithms. This data set contains 569 samples of 30 features for each sample. In addition, in this dataset, m stands for Malignant, indicating a cancerous tumor and b stands for Benign, indicating a non-cancerous tumor. [4].
The Ionosphere data set which is used in the domain of atmospheric and radar signal processing, involves the radar returns from Ionosphere. This dataset contains 34 features relating to the characteristics of the radar signal returns. The classes stand for good radar returns (g) and bad radar returns (b) [2].

2.2 Feature analysis

In the case of the breast cancer modifier database, the features comprise continuous numbers as well as discrete numbers, whereas the ionosphere dataset exclusively comprises continuous numbers. Figure 1, which illustrates the visualization of the breast cancer modifier dataset, shows that the frequency of the features is approximately normally distributed.

Figure 1 shows that a significant number of features in the second dataset do not adhere to a Gaussian distribution.

Investigating the feature correlation matrix for both datasets(Figure 3 and 4) which are presented in the Appendix, reveals valuable insights into the interrelationships among the variables in the breast cancer dataset, such as the relation between features 1, 3 features 1, 4, etc. On the other hand, the are lower correlations between features in the ionosphere dataset.

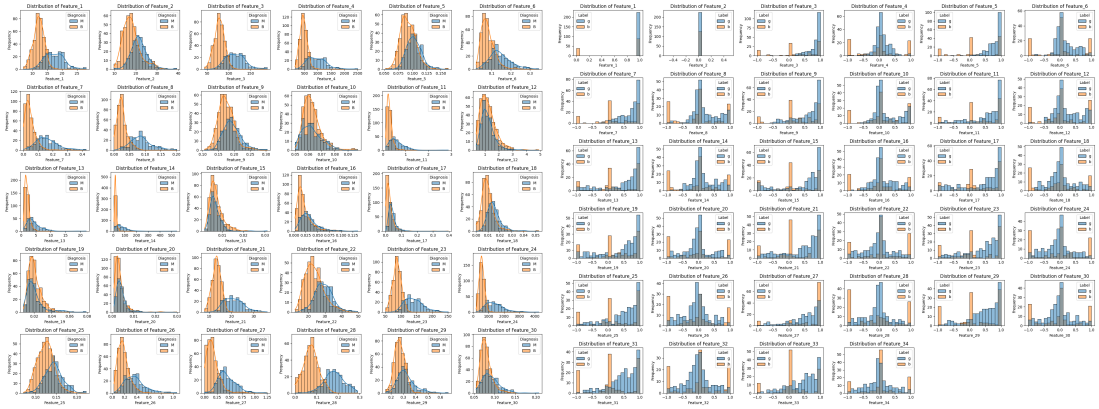

Figure 1: Features distribution. Left: Dataset 1, Right: Dataset 2

Eventually, to develop the performance and improve the accuracy of the model, new features including the second order of each feature were added to the dataset based on the requirements of the experiment.

In examining the datasets, it becomes evident that dataset 1 exhibits a diverse array of feature values spanning a wide range. Consequently, the application of normalization techniques becomes necessary to mitigate the potential amplification of gradients during the Sigmoid function's operation. Failure to address this could lead to complications in the Gradient Descent process, hindering the model's optimization. Conversely, dataset 2 showcases features confined within the interval of $[-1, 1]$, making normalization unnecessary. These distinctions in data characteristics necessitate tailored preprocessing approaches to ensure the effectiveness of subsequent analyses. The ensuing findings will be detailed in section 3.

3 Results

The results of applying the baseline method to the datasets are compared with those corresponding to the improvement strategies in this section. The results of the best approaches are outlined separately for each dataset in two sections.

To assess the impact of learning rates and the number of iterations on the model's accuracy in determining the optimal weight, various iterations and learning rates were examined in each model. It is needless to say, we have implemented cross-validation to find an optimal range of learning rates, as represented in Table 1, and 2. The convergence speed and accuracy of each experiment were then documented. In addition, the 10-fold cross-validation was utilized to choose the model for each experiment in both datasets.

3.1 Dataset 1

Four different strategies were applied to this dataset, including:

1. Baseline

In this experiment, the sole logistic regression model was used for learning from data.

2. Normalization

Normalization of the dataset was employed as a self-experiment alongside logistic regression to evaluate its impact on the results.

3. Normalization and addition of higher-order features (HOF)

In this experiment, in addition to normalizing the data, higher-order features were added to the dataset. The subset was generated by squaring each feature and appending it to the column

space, resulting in a subspace comprising 60 features. This subspace includes the squared values of the first 30 features. (Please refer to the equation 9)

4. Principal Component Analysis (PCA)

PCA was used as a strategy to identify the most important features and extract a new feature space. The sorted Eigenvalues are presented in the Appendix. To analyze it more clearly, Figure 6-a represents an elbow in Eigenvalues, such that we can conclude first 10 Eigenvalues contain the most important features/meanings. Thus, for dataset 1, we have performed a dimensional reduction using a PCA order 10.

Table 1 shows the results of four strategies used to perform and improve logistic regression on dataset 1. The table presents the results of accuracy and elapsed time required for each model considering a different learning rate. It is clear that, when compared to normalization, using higher-order features increased model accuracy by 95 percent. Meanwhile, as the PCA represents an optimal lower-dimensional space, the processing time, as expected, is roughly halved compared to alternative methods, while the accuracy is improved and comparable to the HOF.

	lr	1.00E-04	1.00E-05	1.00E-06
Baseline				
	Mean Accuracy	0.88	0.92	0.9
	Elapsed Time (seconds)	0.8	0.8	0.8
Normalized				
	Mean Accuracy	0.92	0.93	0.93
	Elapsed Time (seconds)	0.8	0.82	0.8
HOF				
	Mean Accuracy	0.95	0.95	0.95
	Elapsed Time (seconds)	0.83	0.85	0.79
PCA				
	Mean Accuracy	0.94	0.95	0.93
	Elapsed Time (seconds)	0.41	0.42	0.4

Table 1: Comparing the accuracy and the run time of different methods. (Dataset 1)

3.2 Dataset 2

According to the differences between this dataset and the previous one which was described in section 2, it is evident that different approaches may have better results on this dataset. The strategies that were used for this data set include:

1. Baseline method

In this experiment, the sole logistic regression model was used for learning from data.

2. Higher-order features addition (HOF)

The subset was generated by squaring each feature and appending it to the column space, resulting in a subspace comprising 68 column vectors (features). This subspace includes the squared values of the first 34 features. (Please refer to the equation 9)

3. Principal Component Analysis (PCA)

PCA was used as a strategy to identify the most important features. The sorted Eigenvalues are presented in the Appendix. In contrast to Dataset 1, the Eigenvalue curve of Dataset 2 does not tend to flatten out, and the Eigenvalues are not insignificant. This suggests that PCA may not be as effective for Dataset 2 compared to Dataset 1. This is what results in table 2 will confirm.

The results of each experiment are displayed in Table 2. The table includes the learning rates, elapsed time, and accuracy for each experiment to allow for a thorough comparison. The

results indicate that HOF enhancement significantly boosted the model's accuracy across various learning rates, whereas PCA enhancement is only limited to lower learning rates and while the processing time is lower.

	lr	1.0E+01	1.0E+00	1.0E-02	1.0E-03
Baseline					
	Mean Accuracy	0.87	0.88	0.82	0.69
	Elapsed Time (seconds)	0.78	0.76	0.69	0.71
HOF					
	Mean Accuracy	0.91	0.92	0.89	0.75
	Elapsed Time (seconds)	0.80	0.83	0.77	0.73
PCA					
	Mean Accuracy	0.86	0.86	0.89	0.84
	Elapsed Time (seconds)	0.38	0.39	0.39	0.37

Table 2: Comparing the accuracy and the run time of different methods. (Dataset 2)

4 Discussion and Conclusion

The baseline Logistic Regression Classifier model can classify inputs, if hyperparameters such as learning rate and number of iterations are set, appropriately. However, in cases such as dataset 1 which exhibits a diverse array of feature values, the application of normalization techniques becomes necessary to mitigate the potential amplification of gradients. To further improve the efficiency of the model, both in terms of time and accuracy, strategies such as PCA (which is a dimension reduction technique) and adding higher-order features can noticeably improve the efficiency of the model.
Furthermore, an increase in the learning rate was found to affect improving the model's accuracy on dataset 2. More precisely, increasing the learning rate led to a quicker convergence of the model towards an optimal global minimal error. The observed acceleration can be ascribed to the implementation of greater step sizes, which enable a more rapid advancement towards the loss function minimum.

5 Statement of Contribution

All tasks in the project were equally distributed among the three team members. Amirhossein, Reza, and Amin each contributed equally to developing the logistic regression model's structure, data processing, improvement strategies, and report preparation.

References

[1] Chao-Ying Joanne Peng, Kuk Lida Lee, and Gary M Ingersoll. An introduction to logistic regression analysis and reporting. *The journal of educational research*, 96(1):3–14, 2002.

[2] Wing S. Hutton L. Sigillito, V. and K. Baker. Ionosphere. UCI Machine Learning Repository, 1989. DOI: https://doi.org/10.24432/C5W01B.

[3] Gilbert Strang. *Linear algebra and learning from data*. SIAM, 2019.

[4] WIlliam Wolberg. Breast Cancer Wisconsin (Original). UCI Machine Learning Repository, 1992. DOI: https://doi.org/10.24432/C5HP4Z.

6 Appendix

6.1 Equations:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{1}$$

$$P(y_i = 1|x_i) = \sigma(W^T x_i) \tag{2}$$

$$P(y_i = 0|x_i) = 1 - \sigma(W^T x_i) \tag{3}$$

$$\text{Cross-Entropy} = -\frac{1}{N} \sum_{i=1}^{N} \left(y_i \log(\sigma(W^T x_i)) + (1 - y_i) \log(1 - \sigma(W^T x_i)) \right) \tag{4}$$

Where:

- N is the total number of samples in the dataset.
- y_i is the true label (0 or 1) of the i samples.
- W is the calculated weight matrix.

6.2 Explanation of improvements methods:

Normalization is a data preparation technique used to scale a dataset's numerical properties to a standard range without distorting the disparities in the ranges of values. Normalizing input data with wide ranges helps stabilize training by preventing saturation of activation functions like sigmoid, leading to faster convergence, better generalization, and improved gradient flow.

$$\text{mean} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{5}$$

$$\text{std} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2} \tag{6}$$

$$X_{\text{normalized}} = \frac{x_i - \text{mean}}{\text{std}}. \tag{7}$$

HOF involves creating new features from existing ones by raising them to higher powers or by multiplying different features together. This technique can help capture more complex relationships between variables in the data, potentially improving the model's ability to fit non-linear patterns and increasing its predictive power.

$$\text{Original Matrix:} \quad X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix} \tag{8}$$

$$\text{After adding higher-order features:} \quad X' = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} & x_{11}^2 \\ x_{21} & x_{22} & \dots & x_{2m} & x_{21}^2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm} & x_{n1}^2 \end{bmatrix}. \tag{9}$$

PCA is a method for reducing the dimensionality approach that converts high-dimensional datasets into a lower-dimensional space while retaining the majority of the variation in the data, hence simplifying their complexity.

$$\text{Original Data Matrix:} \quad X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix} \tag{10}$$

$$\text{After PCA:} \quad X' = \begin{bmatrix} x'_{11} & x'_{12} & \dots & x'_{1k} \\ x'_{21} & x'_{22} & \dots & x'_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x'_{n1} & x'_{n2} & \dots & x'_{nk} \end{bmatrix} \tag{11}$$

It involves: normalization, Eigen-decomposition, sorting and k^{th} order feature extraction.

6.3 Figures:

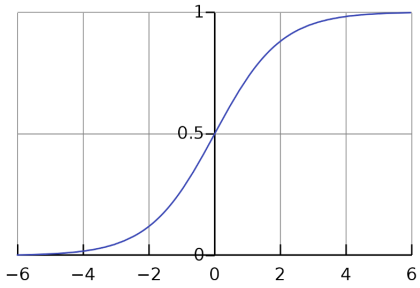

Figure 2: Sigmoid Function

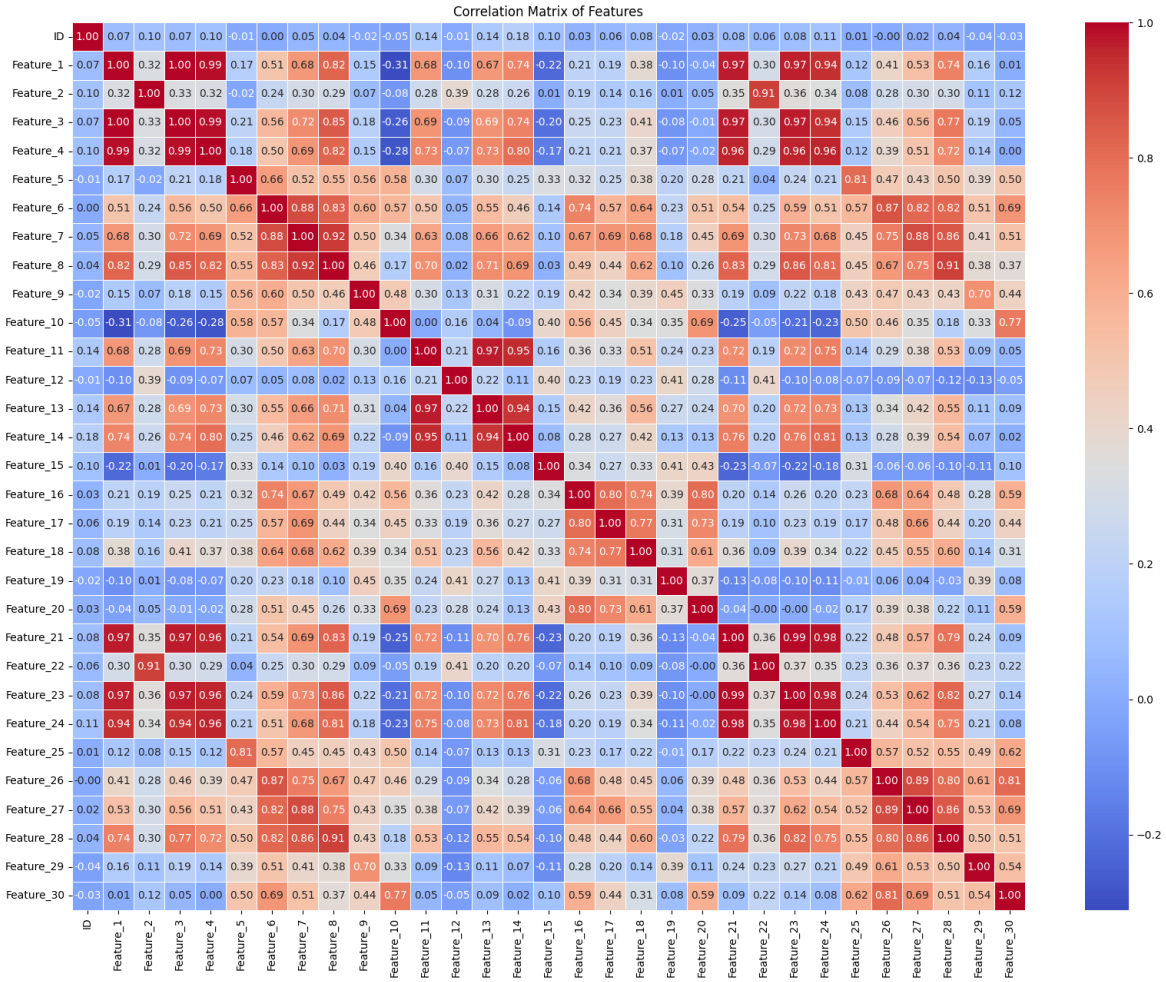

Figure 3: Features' correlation in dataset 1

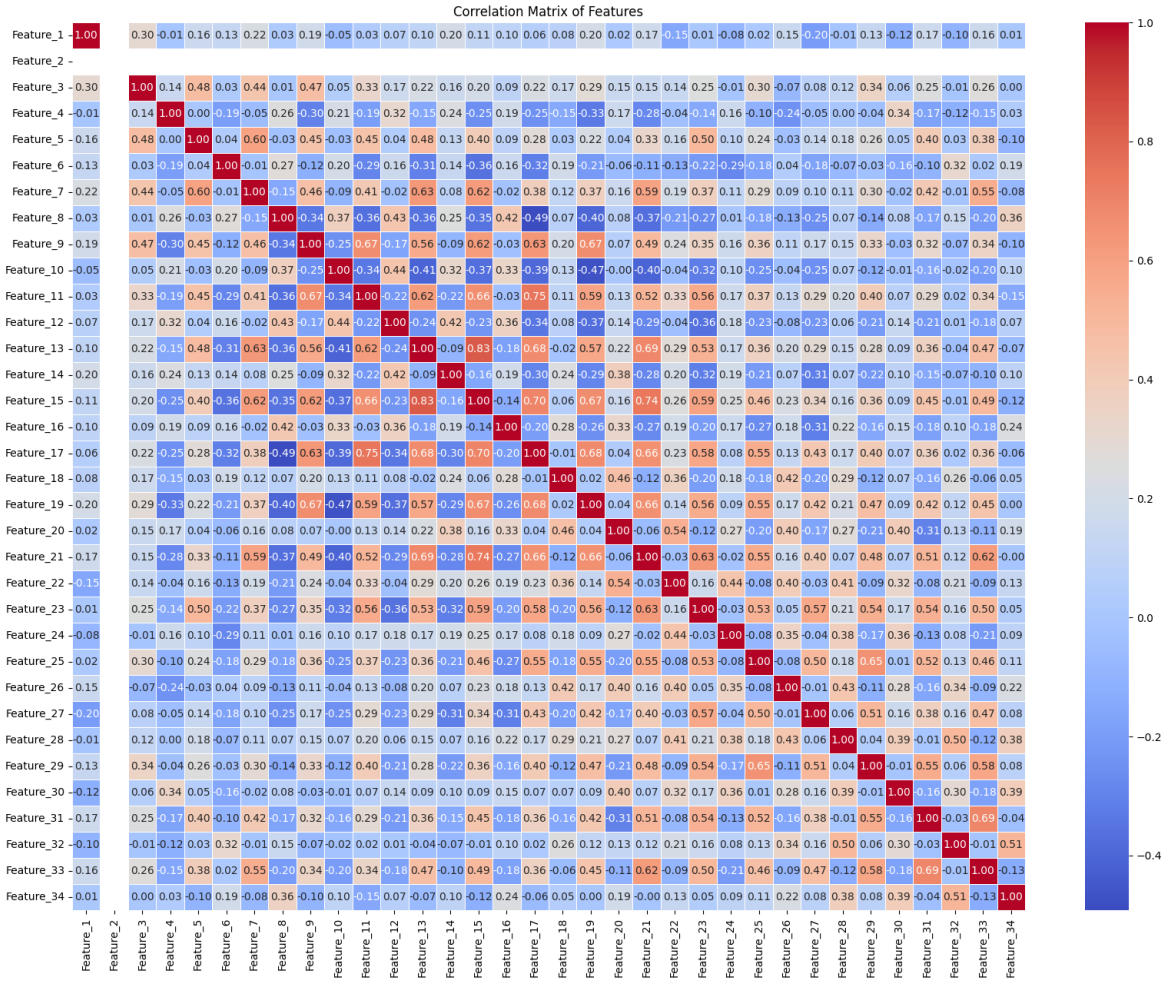

Figure 4: Features' correlation in dataset 2

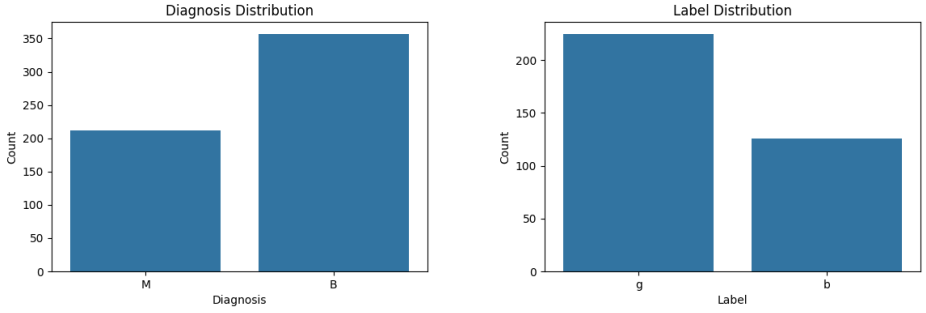

Figure 5: Class distribution. left: Dataset 1, right: Dataset 2

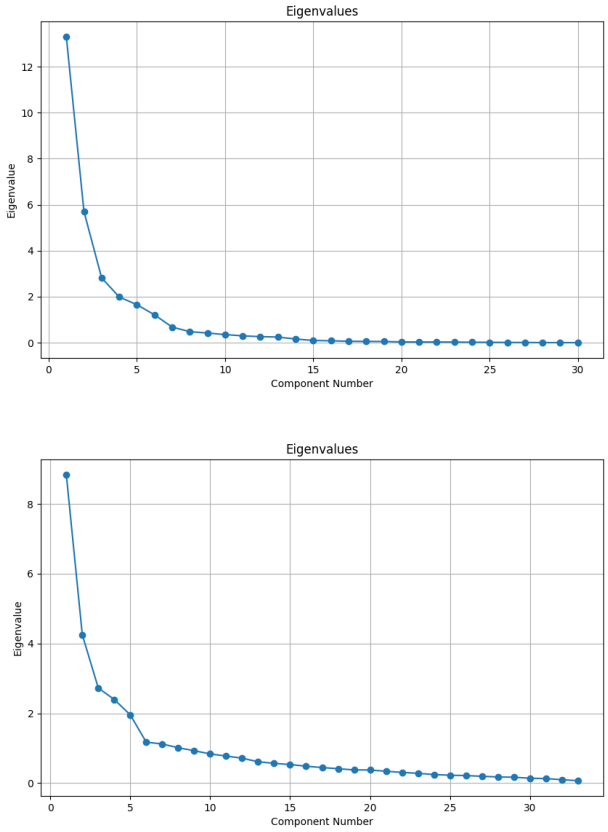

Figure 6: Eigenvalues - top: Dataset 1, bottom: Dataset 2