TÉCNICO LISBOA

Aprendizagem

LEIC IST-UL

RELATÓRIO - MERIT PRIZE CHALLENGE

Grupo 10:

Gabriel Ferreira 107030
Irell Zane 107161

2024/2025 – 1st Semester, P1

1 Logistic Regression Analysis

We conducted a comprehensive analysis of logistic regression performance using both raw and standardized data to evaluate the impact of preprocessing on model effectiveness. The standardization process involved scaling the features to have zero mean and unit variance, which helps prevent features with larger scales from dominating the model's learning process.

Table 1: Logistic Regression Performance Comparison

Preprocessing Method	Testing Accuracy	Training Accuracy
Without Scaling	97.66%	95.98%
Standardized Data	98.25%	98.74%

The results demonstrate that standardization led to improved performance across both training and testing datasets. The standardized model achieved a testing accuracy of 98.25%, representing a significant improvement over the non-standardized version. This improvement suggests that feature scaling is crucial for this dataset, likely due to the varying scales of the original features.

2 EM Clustering Analysis

Our implementation of Expectation-Maximization (EM) clustering utilized `scikit-learn`'s `GaussianMixture` model. We evaluated clustering quality using silhouette scores across different numbers of clusters ($k \in [2, 10]$) to determine the optimal cluster count for our dataset.

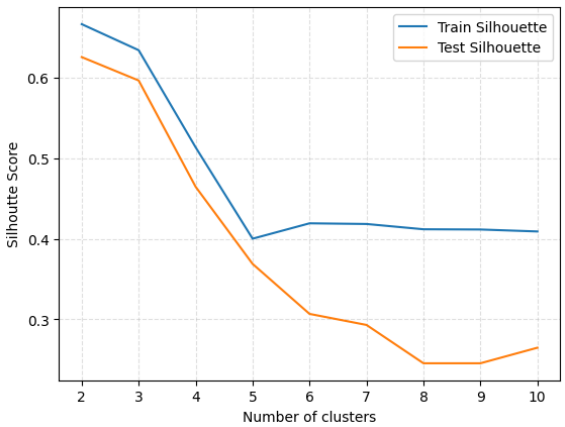

Figure 1: Silhouette scores vs. cluster count

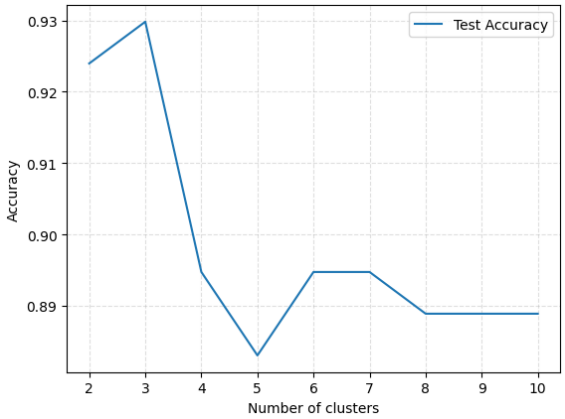

Figure 2: Logistic Reggression on the EM Probability Map

The silhouette analysis revealed several important trends, see Figure 1:

- A clear declining trend in clustering quality as the number of clusters increases;
- Growing disparity between training and testing silhouette scores with higher cluster counts, suggesting a loss in the generalization capacity of the clustering solution;
- Optimal clustering achieved with k=2, yielding silhouette scores of 0.626 (testing) and 0.667 (training);

4 Logistic Reggression on EM Probabilities Analysis

Upon analyzing Figure 2, the relationship between the number of clusters (k) and both performance metrics - the Silhouette scores and Logistic Regression accuracy on the mapped dataset - we can see that there is a notable relation: For low numbers of clusters, such as 2 and 3, both metrics show high values, and as the k increases, there is a sharp decline in both the silhouette score and the classification accuracy of the logistic model, which seem to follow the same trajectory, suggesting that a higher quality of clustering, as measured by the Silhouette score, is conducive to a better logistic regression model.

Using $k = 2$ or $k = 3$ the EM algorithm effectively serves as a dimensionality reduction technique, transforming the original high-dimensional feature space (30 features) into a lower dimensional feature space, while maintaining a high accuracy on the logistic regression model.

5 RBF Network Implementation

Our Radial Basis Function (RBF) Network implementation consisted of three key layers:

1. Input layer matching the dimension of observations;
2. Hidden layer corresponding to the number of clusters;
3. Output layer with a single neuron for binary classification;

Before training an RBF network we need to choose a Radial Basis Function. The most commonly used is the Gaussian and since we are working with EM Clustering it would not make sense to measure just the distance between the observation and the mean. We want the distance between the point of the observation and the distribution of cluster 1 and 2, the distance between a point and a distribution is the Mahalanobis distance.

The Gaussian Radial Basis Function employed in our network is thus defined as:

$$\phi_k(x) = exp(-\gamma \cdot (x - \mu)^T \Sigma_k^{-1} (x - \mu)) \tag{1}$$

Initial implementation with the standard[1] $\gamma = \frac{1}{2}$ yielded low accuracies:

- Train Accuracy: 62.56%;
- Test Accuracy: 63.15%;

This low accuracy was due to RBF values being in wildly different scales, most of them too small, see Figure 3. With so many of them being near zero, and some so small that their double precision floating point representation is zero, the model is unable to learn properly, and because of this, the model always made the same prediction regardless of input.

[1] As suggested by Wichert and Sa-Couto (2021, Chapter 10)

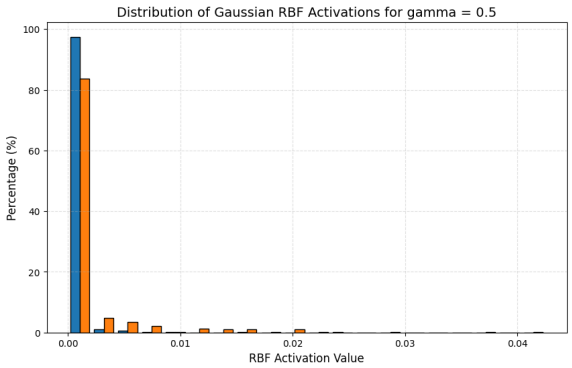

Figure 3: Hidden layer activation distribution ($\gamma = 0.5$)

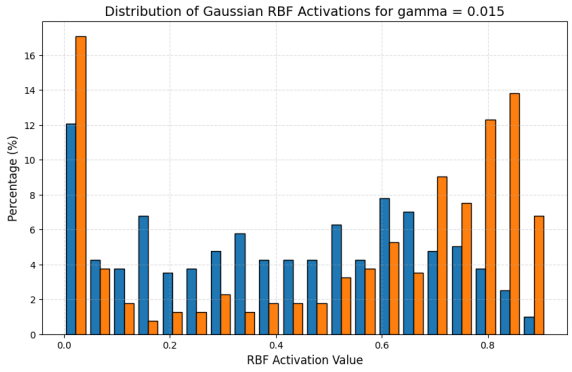

Figure 4: Hidden layer activation distribution ($\gamma = 0.015$)

To address this challenge we attempted to increase the scale by decreasing the gamma value, trying many different values, see Figure 5 and it seemed to be effective as a means to attenuate the scale differences, see Figure 4. We also considered that another good solution would be normalizing the results of the hidden layer, since this would guarantee that at least one neuron had a non near-zero value, meaning the regression model would be able to update its weights more effectively.

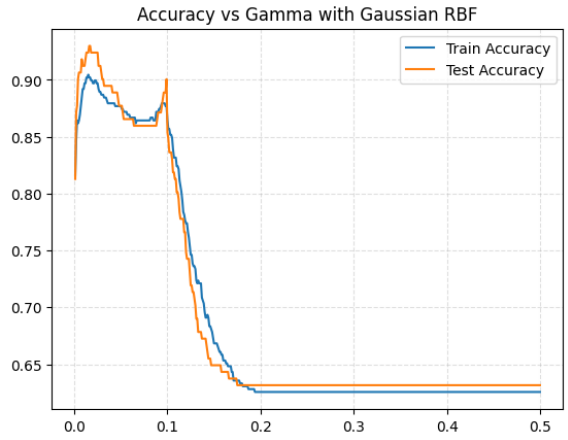

Figure 5: RBF Network gamma parameter tuning results

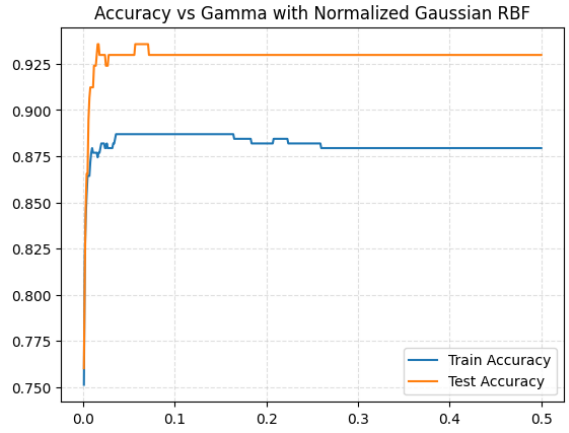

Figure 6: Normalized RBF Network gamma tuning results

The Normalized RBF (NRBF) implementation achieved several improvements:

- Increased testing accuracy from 92.40% ($\gamma = 0.015$) to 92.96% with $\gamma = 0.5$;
- Reduced sensitivity to gamma parameter selection, meaning that the standard γ of 0.5 can be chosen, making time costly hyperparameter tuning no longer a necessity;
- More stable and consistent performance across different configurations;
- Better distribution of activation values, leading to improved learning;

We also explored alternative basis functions. Here's the Linear RBF:

$$\phi_k(x) = \gamma \cdot (x - \mu)^T \Sigma_k^{-1} (x - \mu) \tag{2}$$

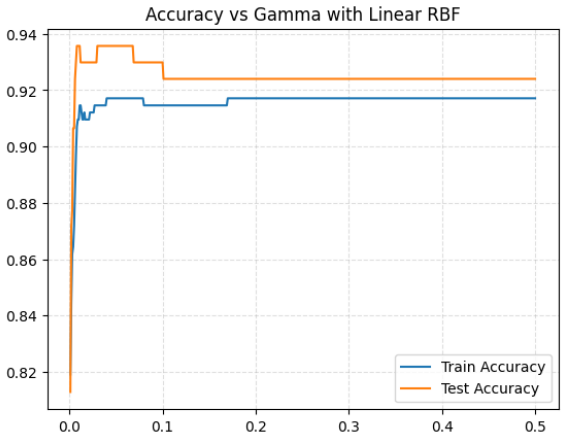

Figure 7: Linear RBF performance analysis

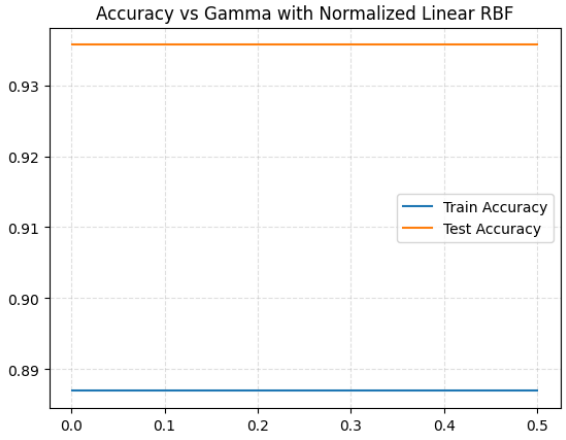

Figure 8: Normalized Linear RBF performance

The Linear RBF implementation showed promising results, outperforming the Gaussian RBF in terms of both accuracy and stability.

6 Comparative Analysis and Conclusions

Our comprehensive evaluation of different approaches revealed several key insights about both the models and the underlying data structure:

Table 2: Model Performance Summary

Model	Testing Accuracy
Logistic Regression (Standard Scaling)	98.25%
EM Probability Mapping ($k = 2$)	92.40%
Gaussian RBF Network (best γ parameter)	92.40%
Gaussian NRBF Network	92.96%
Linear RBF Network	92.40%
Linear NRBF Network	93.57%

Key findings from our analysis:

- Logistic regression's high accuracy (98.25%) suggests predominantly linear decision boundaries in the feature space
- EM clustering successfully reduced dimensionality while preserving most of the discriminative information
- The normalized RBF approach demonstrated robust performance across various parameter settings
- Preprocessing and normalization proved essential for optimal performance across all methods
- The dataset exhibits natural binary structure, aligning well with the classification objective

The practical implications of our findings include:

1. EM clustering with k=2 provides an effective approach for dimensionality reduction while maintaining good classification performance
2. Normalized RBF networks offer stable performance without requiring extensive parameter tuning
3. Feature standardization consistently improves model performance across different approaches
4. The choice between methods involves a trade-off between accuracy and computational complexity

References

[1] Luis Sa-Couto and Andreas Miroslaus Wichert, *Machine Learning - A Journey To Deep Learning: With Exercises And Answers*, International series of monographs on physics, World Scientific Pub Co Inc, 2021, ISBN: 9789811234057.