Aprendizagem

LEIC IST-UL

RELATÓRIO - MERIT PRIZE CHALLENGE

Grupo 10:

Gabriel Ferreira	107030
Irell Zane	107161

2024/2025 – 1st Semester, P1

We began by loading the the required dataset, making a 70%-30% train-test split, and scaling the data using a `StandardScaler`.

1. *Perform logistic regression and indicate the accuracy.*

Testing Accuracy	Training Accuracy
98.25%	98.74%

2. *Perform EM clustering on the training data set with different number k of clusters. Evaluate the quality of the clusterings using Silhouette. Is the number of clusters correlated with the quality of clustering? Which is the optimal k?*

   We used `scikit-learn`'s `GaussianMixture` model to perform EM clustering on the breast cancer dataset features. An evaluation of the clustering quality was made using each solution's silhouette for numbers of clusters ranging from 2 to 10, see Figure 1.

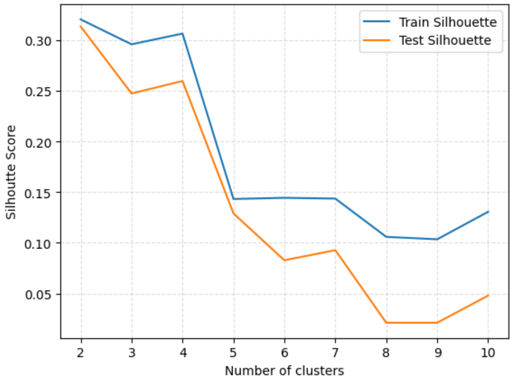


Figure 1: Silhouette of different EM clustering solutions.

   According to the Silhouette score, there is a trend of declining clustering quality as the number of clusters increases in the EM clustering solution. We can also see an increase in the difference in the Silhouette score of the classifications of the Testing, and Training samples, suggesting a negative trend in the generalization capacity as the number of clusters increases.

3. *Map the test set into probability values of the k-clusters. If you have a data point represented by a vector of dimension d, you will map it into a vector of dimension:* `prob=em_model.predict_proba(X)`

4. *Perform logistic regression on the mapped data set with the labels of the original test set. Indicate now the accuracy. Is there a relation between the number of clusters, the cluster evaluation and the accuracy of the logistic regression model?*

5. *Train an RBF network using the clustering with optimal k from 2).*

   An RBF Network would look like this:

   (a) Input layer with the size of the dimension of an observation

(b) Hidden layer with the size of the number of clusters

(c) Output layer with a single neuron

Before training an RBF network we need to choose an Radial Basis Function. A Radial Basis Function is any function that depends on the distance, the most commonly used is the Gaussian:

$$\phi_k(x) = \phi_k(d(x, c_k)) = exp(-\gamma \cdot d(x, c_k)^2) \tag{1}$$

Since we are dealing with EM Clusters, we want to measure 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:

$$d(x, c_k)^2 = (x - \mu)^T \Sigma_k^{-1} (x - \mu) \tag{2}$$

So our Radial Basis Function ends up like so:

$$\phi_k(x) = exp(-\gamma \cdot (x - \mu)^T \Sigma_k^{-1} (x - \mu)) \tag{3}$$

As in 'Machine Learning: A Journey to Deep Learning'[1], we thought the γ should be $\frac{1}{2}$; However when this was tried, the network would always make the same prediction regardless of the input. This was quite unexpected, we speculated that this was a float underflowing issue, arising from the negative exponent being too large resulting into loss of information. Thus we plotted the Accuracy against the Gamma to find tune this hyperparameter and settled on $\gamma = 0.016$, see Figure 2.

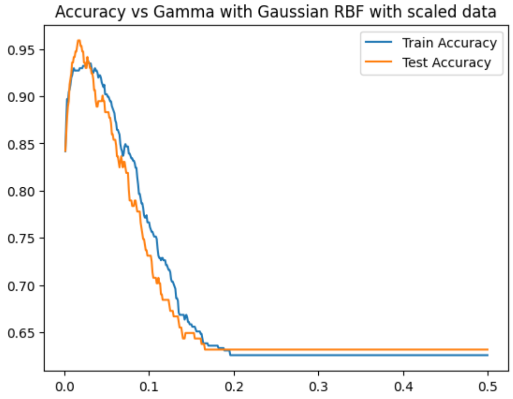
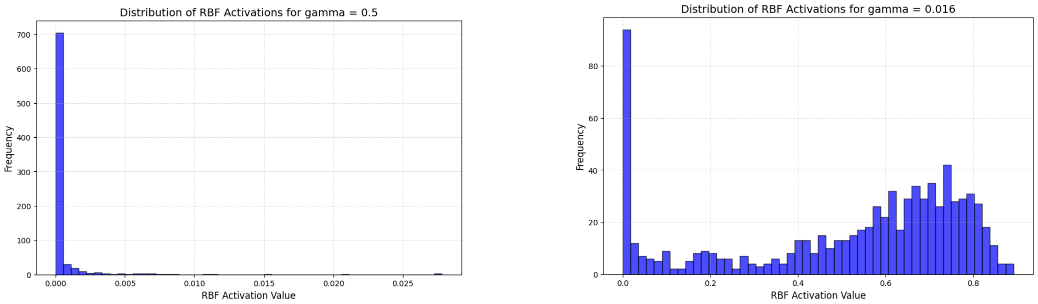


Figure 2: Gamma hyperparameter tuning

Comparing the distribution of the values of the hidden layers in both solutions, we can see that for $\gamma = 0.016$ there is a wider range of values, which have a less concentrated distribution, see Figure 3.



(a) $\gamma = 0.5$

(b) $\gamma = 0.016$

Figure 3: Hidden layer activation value distribution

6. *Discuss your findings.*

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.