Merit Prize Challenge 2024/2025

Breast Cancer Dataset Analysis

Guilherme José
José Caldeira

November 2024

Contents

1 Introduction

Overview of the challenge, dataset, objectives, and the methodology followed.

1.1 Overview

In this task we aim to develop a model that can accurately classify a patient with or without cancer based on available medical data. This is often used to support healthcare professionals in order to enhance efficiency and enabling doctors to help a larger number of patients effectively.

1.2 Dataset description

The dataset consists of **30 numerical variables** and **1 binary variable**, the target. Patients are classified as having a **benign tumor** or a **malignant tumor**, target variable is **0** or **1**, respectively. The dataset has **no missing values**.

1.3 Logistic Regression

Logistic regression is a machine learning method used to make binary classification. It differs from linear regression in the activation function and the loss function to update the weights. While in linear regression there is no activation function, logistic regression uses the sigmoid function:

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

The expression of logistic regression is as it follows:

$$\hat{y} = \sigma(\mathbf{w} \cdot \mathbf{x} + b) = \frac{1}{1+e^{-(\mathbf{w} \cdot \mathbf{x} + b)}}$$

- $\mathbf{w}$: The weight vector.
- $\mathbf{x}$: The vector with each input variable
- b: The bias term.
- $\hat{y}$: The predicted probability of the positive class ($y = 1$).

To evaluate the model, we use cross validation as loss function:

$$\text{Loss} = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right]$$

1.4 EM Clustering

1.5 RBF Network

Radial Basis Function is a type of neural network that uses Radial Basis function as activation:

$$\phi(x, c) = \exp\left(-\gamma \|x - c\|^2\right)$$

- **x**: The input vector.
- **c**: The centroid vector.
- γ: Parameter to control the width of the RBF. Also represented by $\frac{1}{2\sigma^2}$.

RBF Networks have **3 layers**, one for the input, the middle layer, whose activation functino is RBF and the output layer. The middle layer has k neurons, being **k** the **number of centroids** of the model.

2 Logistic Regression on the original data

Logistic regression was applied to the raw dataset. These are the results:

- **Confusion matrix**:

	Predicted Positive	**Predicted Negative**
Actual Positive	107	1
Actual Negative	4	59

- **Accuracy**: 97.1%
- **Recall**: 99.1%
- **Specificity**: 93.7%

The accuracy score was excellent, this might be a confirmation that the dataset is complete and easily separable for this specific problem. Recall was even higher than accuracy, very close to 100%. This is an important point to consider, as prioritizing an increase in recall should be given higher importance in the model's objectives than increasing specificity.

3 EM Clustering Analysis

3.1 Clustering with Different k Values and Silhouette Evaluation

EM clustering was applied to the original data with k values ranging from 2 to 10. Then we computed the silhouette metric for each of the models. These were the results:

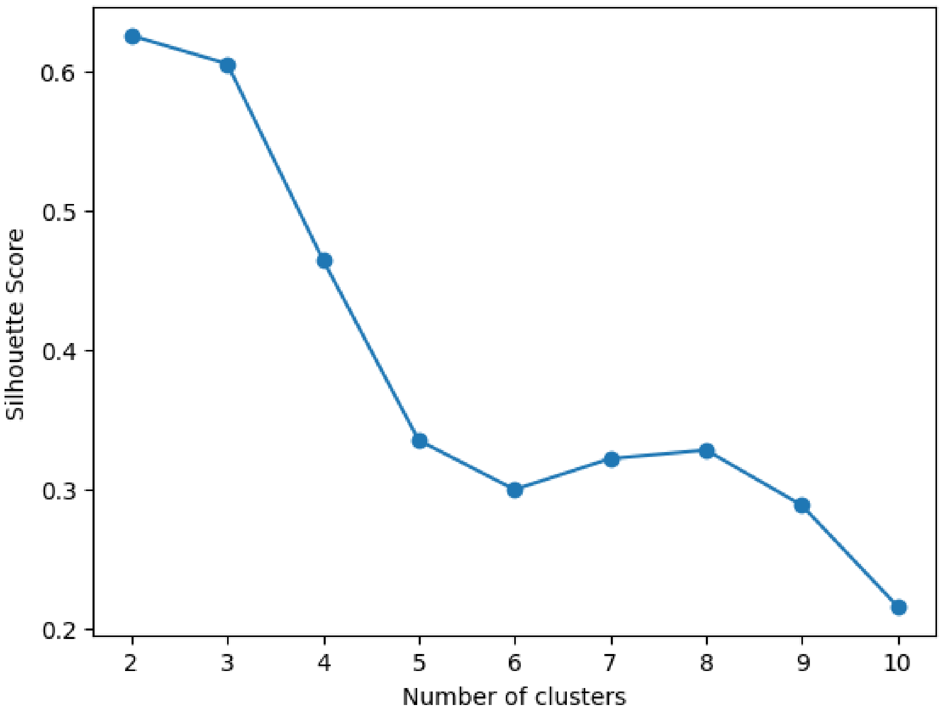

Figure 1: Silhoutte results across different k values.

The results match perfectly with the nature of the problem. The more compact clusters were the ones that were binary, just like the nature of the classification tasks, which had a score of **0.63**. This confirms that the data can be easily separable, as previously mentioned in Section 2.

3.2 Clustering Probabilities and Logistic Regression

We used the previous clustering models and for each of the models, we mapped the dataset into clustering probabilities, meaning each observation is a vector of probabilities for each cluster.

Later, we used that mapped dataset to feed a logistic regression. These were the results:

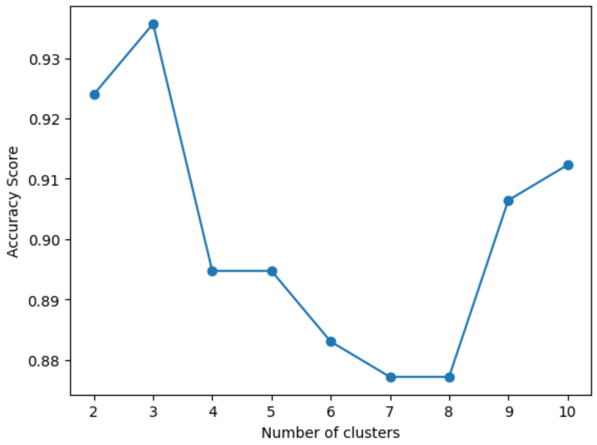

Figure 2: Accuracy results across different k values.

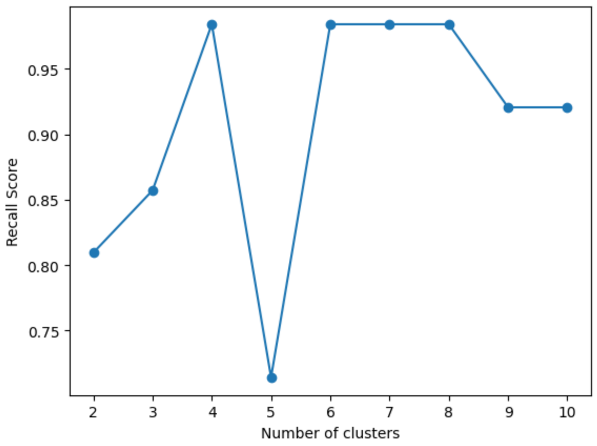

Figure 3: Recall across different k values.

The accuracy results show a slight increase from 2 to 3 in the x-axis (k values) and then decrease until k=8, after which it rises again. The first increase, even though silhouette is higher when k=2, 3 probability values might store more information than 2 and thus improve model accuracy. The decrease in accuracy is likely due to the poor silhouette score which makes the features (cluster probabilities) less informative. At the end there was a second increase in accuracy, this can be attributed to the increase in features outweight the poor silhouette scores.

The recall metric peaked when the accuracy was at its lowest, for $k \in \{4, 6, 7, 8\}$. This should be considered when choosing and evaluating models.

4 RBF Network Training

Training of the RBF network using the clustering with optimal k.

5 Discussion

Key findings, correlations, and overall insights from the analysis.

6 Conclusion

Summary of the challenge, results, and future directions.

References