Aprendizagem 2024/25

Machine Learning Challenge

(ist1106992, ist1106154)

Discuss your findings on a (up to) 5 page document:

When describing our findings, we chose to split them in topics corresponding to the exercises to make it easier to understand our analysis. However, in the final part, we summarize the analysis and add a few extra information that didn't match any of the exercises but we found important to point out.

1.

After performing the logistic regression a few times and computing the average of the results, we obtained an accuracy of around 0.96.

This is a solid result, indicating that the model is effective in distinguishing between malignant and benign cases.

This suggests that the breast cancer data has features that are well-separated, even by a linear decision boundary. This means the dataset's feature space is structured in a way that a relatively simple model can classify the data accurately.

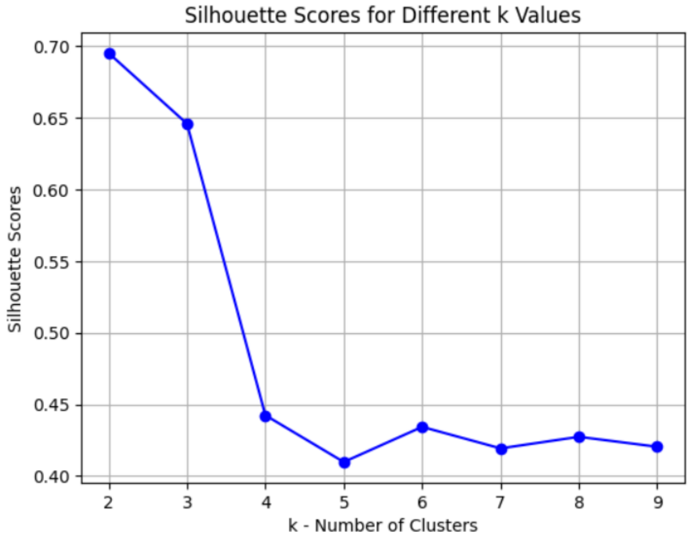
2.

Upon performing EM clustering, we obtained the silhouette scores for the different number of clusters ([0.6953546812827253, 0.6460072951798395, 0.44240066859293997, 0.40975176230240173, 0.4342973431576532, 0.4192675100180935, 0.4273557130544327, 0.4203979188373632] in order from 2 to 9 clusters).

We also know that overall silhouette score (average of the silhouettes of the individual points of the dataset) is used to evaluate the quality of clustering results. The individual silhouettes can range between -1 and 1 where lower values mean that the point has probably been assigned to the wrong cluster while higher values mean that the point matches nearly perfectly its neighbours (points of the same cluster) and very poorly points of different cluster, resulting in a better assignment of points to clusters. The overall silhouette of the model ranges between the same values as the previously explained silhouettes. For general interpretation, lower values suggest bad organization of points in clusters while values closer to 1 correspond to a lower intra-cluster distance (points in the same cluster are close to each other) and higher inter-cluster distance (clusters are far apart from each other).

Given this, when analysing the varying results, we concluded that the number of clusters impacts the model's quality of clustering. This happens because the number of clusters has an effect in cohesion and separation (with too few cluster the model has low separation and high cohesion - points within clusters are not very similar and clusters may overlap and with too many clusters can lead to overfitting - clusters become too small and lose their generality).

After this analysis, we concluded that there needed to be a balance in the model's cohesion and separation which are directly linked to its silhouette: the higher silhouette value belonged to k=2 (0.6953546812827253) which corresponds to the optimal number of clusters for this model and this data. This conclusion makes a lot of sense for this exercise because there are 2 possible outcomes (benign and malignant) and therefore 2 clusters to classify observations.

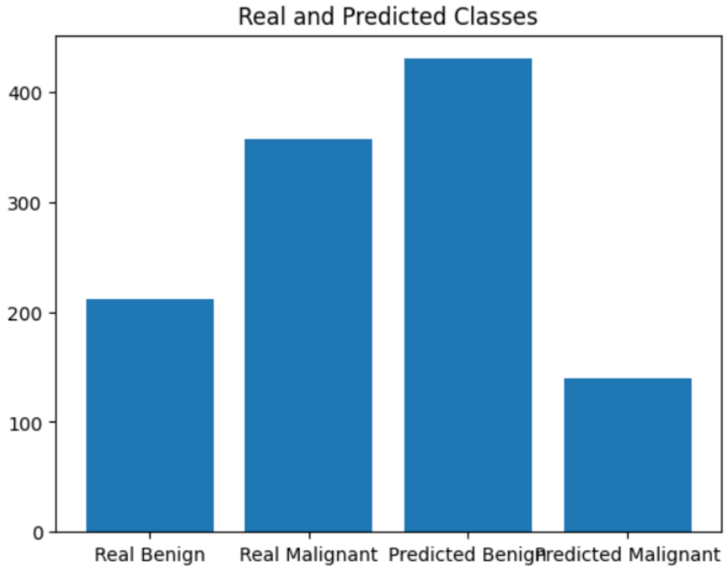


3.

Upon analysing the results of the probabilities, we assessed that for many observations there is a big difference between the probabilities for each cluster (for example [1.00000000e+000 1.99228071e-089] where the probability of belonging to the first cluster is almost 1 while the other cluster's probability is very close to 0), which means that the model assigns the points to cluster "confidently" because the clusters are fairly well-separated for those points.

Still, there are a few points where the probabilities are not as different (for example [2.06130572e-001 7.93869428e-001]). This means that for these points, the clusters are overlapped and the model isn't able to attribute the point to a cluster as precisely as for the others.

In conclusion, generally, this model is very confident in its attribution of points to clusters and if the clusters correlate well with the target labels (malignant and benign), then these probabilities can serve as features for a classifier.



4.

Logistic Regression applies the sigmoid activation function to the output of a linear regression to classify inputs in 1 of 2 categories. This works because the sigmoid returns a value between 0 and 1 and, with a well-defined threshold is possible to predict accurately which class the observations belong to.

Generally (for dataset with continuous outputs), as the number of clusters increases up to a certain point, clustering quality improves because the data is divided into finer partitions that capture more complex patterns. However, beyond this point, adding more clusters begins to decrease clustering quality due to over-segmentation: clusters become too small and fragmented, often capturing noise rather than meaningful patterns.

For this particular dataset (categorical), we assessed that the best number of clusters is the k that matches the number of possible outcomes (2). Given this, higher numbers of clusters would be detrimental to the model's performance (as we saw in exercise 2) and worsen the cluster evaluation. With this, we conclude that the number of clusters has an impact on the cluster evaluation.

Also, higher-quality clustering indicates that data points within each cluster are more similar to each other, which helps the logistic regression model to classify observations more accurately, leading to higher accuracy.

This suggests a relationship between cluster quality (as measured by cluster evaluation metrics) and the accuracy of the logistic regression model.

Given this, it's very important to choose the optimal number of clusters (which heavily depends on the dataset's properties) to generate the model. This allows better classification of observations and therefore better cluster evaluation proving that there is a relationship between the number of clusters, the cluster evaluation and the accuracy of the logistic regression model.

5.

For this exercise we obtained an accuracy of 0.6081871345029239, which, alone, indicates a moderate predictive performance for the RBF network. However, when using a naive approach (where the model predicts majority class for all the observations) we obtain a baseline accuracy of 0.6274165202108963 (probability of the class with majority). This is significant because the accuracy with RBF network is lower than the baseline (and the accuracy with logistic regression), meaning that this model underperforms and is currently not appropriate to evaluate the dataset.

Additionally, there are many factors that can affect the performance of the RBF model, such as clustering quality, feature representation and the overall model's capability.

In regards to the clustering quality, we disregarded slightly this factor because, as seen in the previous exercises the considered clusters are the best that could be obtained for this analysis. Given this, and as the RBF model uses the cluster's centres for its training, we concluded that this probably wasn't the most important factor to negatively affect the model. However, if the clustering had been suboptimal, it could have caused a worsening of the model's performance, especially because clustering adds an inherent layer of abstraction that may discard important details, leading to lower accuracy.

As for the feature representation, we could justify the model's poor performance with the fact that the model transforms data points into a new space based on their similarity to the cluster centres. As a consequence, if the test points are not well-matched to the training clusters, the RBF-transformed features may not capture relevant structure needed for accurate predictions.

When it comes to the overall model's capability, as it relies on a Ridge regression applied to the transformed data, the transformed features may not provide the needed information for the model to be able to distinguish between classes or, as this regression is linear, it may be too simplistic and limit the model's predictive capability.

Given this, we conclude that the obtained accuracy is below what is needed for a good machine learning model. This may be due to the cluster's characteristics or the type of regression used in the RBF network.

Conclusion

In summary, clustering, specifically EM clustering, reduces dimensionality by representing each data point as a probability distribution over clusters. For k=2, the clusters correspond reasonably well with the two target classes (malignant and benign), as supported by the high silhouette score of 0.6954.

The logistic regression performed well on the dataset, achieving an average accuracy of 96%, which demonstrates that the original feature space is highly suitable for linear classification without requiring dimensionality reduction.

The RBF network, used 2 cluster centres and achieved an accuracy of 60.82%, which is significantly lower than the logistic regression. This indicates that while clustering captures some data structure, it loses needed details for effective classification because it relies on intermediate transformations that can lose critical information, making the RBF network less robust, particularly when the clusters do not perfectly align with the target labels or when clusters overlap.

In addition, for this dataset (medical dataset), it's more important to achieve a higher recall instead of a higher precision because is more useful to categorize all malignant tumours as such (even if we classify a few benign as malignant) than to wrongly categorize some malignant tumours (like what happened in the plot from exercise 3). To avoid this type of issues, we could adjust the model's threshold in order to improve the classification process; this would happen because, for a threshold that benefits malignant classifications, for the same data, is more probable to classify correctly all malignant instances.