Análise de Diferentes GMMs e o seu impacto numa RBFN

Vasco Pereira, 103368

1 - Introdução

Este relatório tem como objetivo explorar e avaliar o desempenho de diferentes modelos de machine learning aplicados ao dataset **Breast Cancer**, disponibilizado pela biblioteca **sklearn**. Este conjunto de dados contém 30 variáveis independentes que descrevem características celulares, como raio, textura, perímetro e simetria. O objetivo final é classificar tumores como benignos ou malignos com base nessas informações.Para garantir uma análise consistente, o dataset será dividido aleatoriamente em dois subconjuntos: **70% dos dados serão utilizados para o treino dos modelos**, enquanto os **30% restantes serão reservados para testes**. Os modelos usados incluem a **Regressão Logística**, que será utilizada como uma referência de desempenho, e uma **RBFN (Radial Basis Function Network)**. Além disso, serão avaliados dois tipos diferentes de covariância com os **Gaussian Mixture Models (GMMs)**: *full* e *tied*. Os GMMs serão utilizados para determinar o número ideal de clusters, que, posteriormente, servirão como neurônios na configuração da RBFN.

2 - Preparação do Dataset

Antes de tudo, o dataset será dividido em três subconjuntos distintos:

1. **Base Original**: O dataset sem quaisquer alterações, preservando os dados originais.

2. **Dataset Normalizado**: Onde os dados são normalizados, de forma a garantir que todos os componentes têm a mesma escala.

3. **Dataset com PCA**: A partir do dataset normalizado, será aplicado o algoritmo **PCA** (**Principal Component Analysis**) para reduzir a dimensionalidade do dataset, mantendo 95% da variância original.

Através desta separação é possível analisar se a normalização e a redução das dimensões resultam numa accuracy superior para este dataset específico.

3 - Regressão Logística

Ao aplicar o modelo de Regressão Logística no dataset os seguintes resultados foram obtidos:

Dataset	Accuracy
Base	96.5%
Normalizado	97.7%
PCA	97.7%

Como podemos observar na tabela, tanto o dataset normalizado como o dataset com PCA apresentaram um desempenho superior.

4 – Gaussian Mixture Models

O objetivo principal ao utilizar os **GMMs (Gaussian Mixture Models)** é determinar o número ideal de clusters que maximizam tanto o **valor da silhueta** quanto a **accuracy**. Posteriormente, iremos avaliar qual dos dois critérios resulta num maior desempenho ao ser usado como o número de neurónios de uma RBFN.

Além disso, como mencionado na introdução, serão explorados dois tipos diferentes de covariância: **full** e **tied**. Mas qual a diferença entre eles?

- **Full**: Cada cluster tem a sua própria matriz de covariância completa, o que fornece maior flexibilidade para modelar clusters com formas mais variadas.

- **Tied**: Todos os clusters partilham uma matriz de covariância geral, o que restringe todos os clusters a partilhar a mesma forma.

Esta análise permitirá avaliar não só o comportamento dos diferentes subconjuntos mas também como o tipo de covariância influencia o desempenho geral do modelo e sua integração na RBFN.

Ao aplicar dois GMMs (um para cada tipo de covariância) a cada dataset obteve-se os seguintes resultados para a covariância *full*:

Dataset	Número de clusters ideais para o valor de silhueta	Número de clusters ideais para a accuracy

Base	2	2
Normalizado	2	7
PCA	2	5

E os seguintes para a covariância *tied*:

Dataset	Número de clusters ideais para o valor de silhueta	Número de clusters ideais para a accuracy
Base	2	10
Normalizado	2	58
PCA	2	29

5 – RBFN

Uma RBFN é uma rede neural que utiliza funções de base radial. Neste caso, utiliza-se a função de base radial Gaussiana, que tem a seguinte fórmula:

$$e^{-\beta_i \cdot \|x - c_i\|^2}$$

Onde c_i é o centro do cluster i e β_i é igual a $\frac{1}{2 \cdot d_{mean}^2}$, sendo d_{mean} a distância média dos pontos do cluster i relativamente ao seu centro. Esta função será aplicada na *hidden layer* da RBFN seguido pela função sigmoid na *output layer*, de forma a classificar os tumores como benignos ou malignos.

Ao calcular o número de neurónios utilizando uma GMM com covariância *full* obteve-se os seguintes resultados:

Dataset	Accuracy com número de neurónios que maximizam o valor de silhueta	Accuracy com número de neurónios que maximizam o valor de accuracy
Base	88.9%	88.9%
Normalizado	93.0%	96.5%
PCA	88.9%	94.7%

De seguida, utilizou-se os resultados da GMM com covariância *tied*:

Dataset	Accuracy com número de neurónios que maximizam o valor de silhueta	Accuracy com número de neurónios que maximizam o valor de accuracy
Base	89.5%	90.6%
Normalizado	93.6%	96.5%

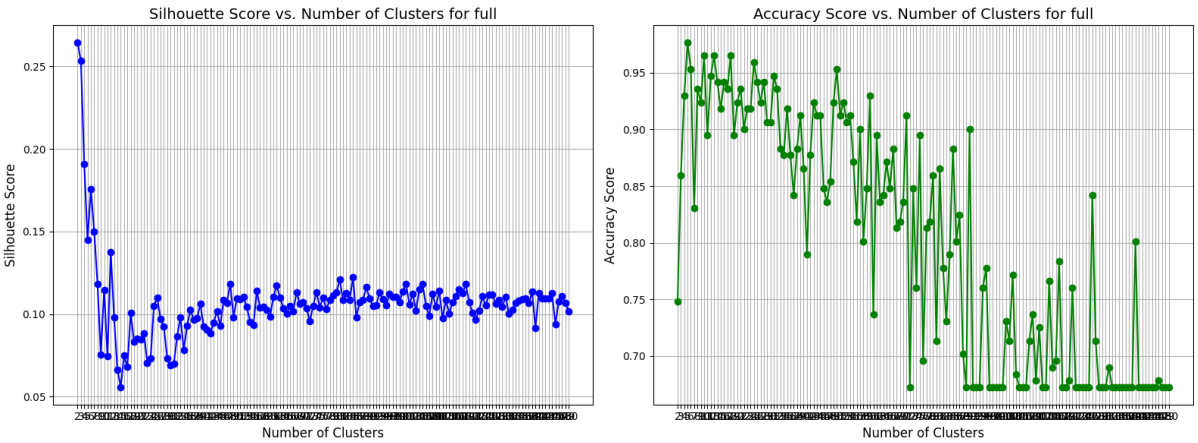
PCA	94.7%	98.8%

6 – Conclusão

Neste relatório, explorámos o impacto de diferentes abordagens na construção de uma RBFN, utilizando GMMs para determinar o número de neurónios ideais. Avaliámos dois tipos de covariância dos GMMs (*full e tied*) e analisámos o desempenho da RBFN com base no número de clusters que maximizam o valor de silhueta e a accuracy.

Os resultados indicam que a normalização e a redução dimensional com PCA são fundamentais para melhorar a performance do modelo, conforme observado na análise inicial com Regressão Logística como nos resultados da RBFN. Tanto o dataset normalizado quanto o dataset com PCA apresentaram accuracy superior em relação ao dataset original em todas as tabelas. Concluímos também que o número ótimo de clusters é **29 ao utilizar uma covâriancia *tied* e redução dimensional,** sendo este o único cenário cujo resultado da RBFN foi superior ao da Regressão Logística.

Ao comparar os diferentes tipos de covariância, concluímos que o modelo tied apresenta resultados consistentemente melhores que o modelo full, conforme ilustrado nas tabelas e figuras. Observou-se que a covariância full tem um desempenho inicial promissor, atingindo um pico de accuracy nos primeiros 15 clusters, mas decai gradualmente à medida que o número de clusters aumenta. Em contraste, a covariância tied mostrou-se mais estável e eficiente na representação das relações entre os dados.



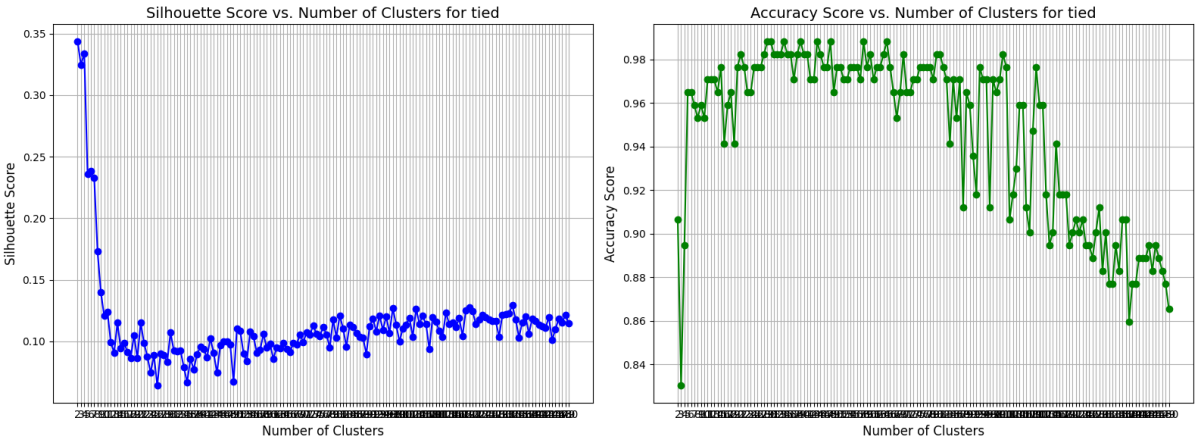


Figura 2: Comparação dos valor de silhueta e o número de clusters e do valor de accuracy e o número de clusters para o dataset reduzido com PCA e utilizando covariância tied

Das figuras, podemos concluir que, embora o valor de silhueta atinja seu máximo em dois clusters, a accuracy é completamente independente deste. O valor de silhueta avalia apenas a separação e coesão dos clusters, não refletindo diretamente o desempenho na tarefa de classificação. Assim, tanto as figuras quanto os resultados numéricos sugerem que a silhueta não é o critério mais confiável para determinar o número ideal de clusters em problemas de classificação.

Por último, também verificámos que, à medida que o número de clusters aumenta, tanto a accuracy quanto o valor da silhueta tendem a diminuir. Isso evidencia a importância de considerar um equilíbrio adequado no número de clusters para evitar overfitting e garantir um desempenho robusto do modelo.

Em resumo, este trabalho demonstrou a importância de combinar abordagens baseadas em normalização, redução dimensional e escolha adequada de parâmetros de clustering para otimizar o desempenho de modelos complexos como um RBFN.