Breast Cancer Detection

Ricardo Caballeros ♠ Alfonso Moraga ♠ Celeste Batres♢ Julio Aragón♠
♠antonio.caballeros@galileo.edu ♢21000310@galileo.edu
♠celeste.batres@galileo.edu ♡ernesto.aragon@galileo.edu

Introducción

En este proyecto se evaluaron distintas arquitecturas de redes neuronales para clasificar tumores mamarios como benignos o malignos. A través de técnicas de feature engineering y ajustes en el diseño de los modelos, desarrollamos cuatro configuraciones distintas que ofrecieron distintos niveles de desempeño.

Descripción del dataset

Los features del dataset fueron extraidos de una imagen de un aspirado con aguja fina de una masa mamaria. Describen las características de los núcleos celulares presentes en la imagen. El dataset tiene un total de 32 atributos.

1) ID number
2) Diagnosis (M = maligno, B = benigno)
3) Radius
4) Texture
5) Perimeter
6) Area
7) Smoothness
8) Compactness
9) Concavity
10) Concave points
11) Symmetry
12) Fractal dimention

Los atributos del 3 al 12 incluyen tres métricas por característica: media (mean), error estándar (SE) y peor valor (worst). Por ejemplo, para el radio existen radius_mean, radius_se y radius_worst. El dataset contiene 357 casos benignos (B) y 212 malignos (M).

Metodología

Primero se realizó encoding de la variable diagnosis (valores "M"/"B") mediante mapeo de valores: M=0 (maligno), B=1 (benigno). Posteriormente, se aplicó oversampling a la clase minoritaria (malignos) para equilibrar el dataset, obteniendo 357 casos por clase.

Resultados

Se entrenaron cuatro modelos con arquitecturas de redes neuronales profundas utilizando diferentes funciones de activación (ReLU y Leaky ReLU), técnicas de regularización como dropout, normalización por lotes (batch normalization), y ajustes de hiperparámetros como learning rate o early stopping.

El objetivo principal fue identificar cuál arquitectura ofrecía el mejor rendimiento en la clasificación binaria de tumores mamarios como benignos o malignos.

Modelo	**Características**	
	Función Activación	**Capas**
Modelo 1	Leaky ReLU	3
Modelo 2	ReLU	3
Modelo 3	Leaky ReLU	8
Modelo 4	ReLU	2

Table 1: Comparación de arquitectura de modelos

La siguiente tabla resume el rendimiento de cada modelo en términos de precisión (accuracy) y pérdida (loss), tanto en los datos de entrenamiento como en los de prueba.

Modelo	**Accuracy**		**Loss**	
	Train	**Test**	**Train**	**Test**
Modelo 1	0.9820	0.9767	0.1475	0.1410
Modelo 2	0.9880	0.9860	0.0812	0.0770
Modelo 3	0.9860	0.9814	0.0605	0.0713
Modelo 4	0.9895	0.9720	0.1068	0.1063

Table 2: Comparación de rendimiento de modelos

Conclusiones

A partir del desarrollo de los modelos de redes neuronales para la detección de cáncer de mama, se pudo evidenciar que todos los modelos alcanzaron altos niveles de precisión, lo que sugiere que el conjunto de datos es apto para este tipo de tareas de clasificación binaria. Sin embargo, el segundo modelo, basado en la función de activación ReLU, demostró el mejor desempeño general al obtener una precisión del 98.60% en los datos de prueba y una pérdida mínima, sin caer en sobreajuste. Esto indica que una arquitectura más simple pero bien regularizada puede ser más efectiva que modelos excesivamente profundos.

Mejoras a futuro

Aunque los resultados obtenidos fueron satisfactorios, existen diversas oportunidades para mejorar los resultados. En primer lugar, se pueden explorar técnicas avanzadas de feature selection para reducir la dimensionalidad del dataset y mejorar la interpretabilidad del modelo.