Proyecto final de matemática numérica

Cristina Hernández Fornaris a.uno@lab.matcom.uh.cu
Grupo D111

Alberto E Marichal Fonseca marichalalberto292@icloud.com
Grupo D111

Dalia Castro Valdes a.dos@lab.matcom.uh.cu
Grupo D111

Tutor(es):
Dr. Angela, León Mecías *MATCOM*
Lic. Rocío Ortíz Gancedo, *MATCOM*
Lic. Lázaro D González Martínez, *MATCOM*

Resumen

Este estudio se enfocó en el desarrollo de dos modelos predictivos para clasificar tumores mamarios como benignos o malignos. Para ello, se utilizó un dataset con 30 características de 569 personas, junto con su respectiva clasificación. Se aplicaron dos técnicas clave: la Descomposición en Valores Singulares (SVD) y la Regresión Logística (RL). El SVD permitió reducir la dimensionalidad de los datos, para identificar que los primeros 4 valores singulares concentraban la mayor parte de la información relevante. Posteriormente, el modelo de Regresión Logística, entrenado con los datos reducidos, mostró una alta exactitud del 94 % en la clasificación de los tumores. La combinación de estos métodos de aprendizaje automático y Álgebra Lineal demostró ser eficaz para optimizar el procesamiento de datos complejos y mejorar la precisión en la predicción de la naturaleza de los tumores. Los resultados subrayan el potencial de estas técnicas para avanzar en el diagnóstico del cáncer de mama y otros problemas médicos desafiantes. Este enfoque innovador puede contribuir significativamente a una detección temprana más efectiva y a un tratamiento más adecuado de esta enfermedad.

Abstract

This study focused on the development of two predictive models to classify breast tumors as benign or malignant. To do so, a dataset with 30 features from 569 people was used, along with their respective classification. Two key techniques were applied: Singular Value Decomposition (SVD) and Logistic Regression (RL). SVD allowed reducing the dimensionality of the data, identifying that the first 4 singular values concentrated most of the relevant information. Subsequently, the Logistic Regression model, trained with the reduced data, showed a high accuracy of 94 % in classifying tumors. The combination of these machine learning and linear algebra methods proved to be effective in optimizing the processing of complex data and improving the accuracy in predicting the nature of tumors. The results underline the potential of these techniques to advance the diagnosis of breast cancer and other challenging medical problems. This innovative approach can significantly contribute to more effective early detection and more appropriate treatment of this disease.

Palabras Clave: Regresión Logistica (RL), Descomposición en Valores Singulares (SVD)
Tema: Cáncer de mama

1. Introducción

El cáncer de mama se encuentra entre las enfermedades más comunes y complejas de la sociedad contemporánea y afecta a millones de personas en todo el mundo. Es el principal cáncer entre mujeres, lo que lo convierte en un importante problema de salud pública mundial. Así, diferenciar entre tumores benignos y malignos ha adquirido importancia en la detección temprana y el tratamiento eficaz.

Con los avances dados por la tecnología médica y la ciencia de datos en los últimos tiempos, se abren nuevas vías para una mejor detención y procedimientos de estadificación del cáncer de mama, sin los cuales de otro modo no habrían sido posibles.

El uso de algoritmos de aprendizaje autómatico combinados con análisis de big data traerá cambios drásticos con respecto a nuestra capacidad para predecir si los casos detectados probablemente sean benignos o malignos. En última instancia, esto revolucionará significativamente la forma en que abordamos estos casos.

En este estudio nos enfocamos en el desarrollo de dos modelos predictivos para clasificar tumores mamarios como benignos o malignos. Para ello utilizaremos un dataset que cuenta con 30 características diferentes, de este tipo de tumor, para 569 personas, junto con su respectiva clasificación en benigno (B) o maligno (M).

Aplicaremos dos técnicas de aprendizaje autómatico y Álgebra Lineal: la Descomposición en Valores Singulares (SVD) para reducir la dimensionalidad de los datos y la Regresión Logística (RL) como clasificador binario.

2. Datos

2.1 Características

Se cuenta con un dataset de kaggle[3] sobre el cáncer de mama con las siguientes características:

id: Representa un ID único de cada paciente.

diagnostico: Indica el tipo de cáncer. Esta propiedad puede tomar los valores "M" (Maligno) o "B" (Benigno).

Características del radio

radius mean: Media del radio de la célula

radius se: Error estándar del radio

radius worst: Peor radio

Características de la textura

texture mean: Media de la textura texture se: Error estándar de la textura texture worst: Peor textura

Características del perímetro

perimeter mean: Media del perímetro de la célula

perimeter se: Error estándar del perímetro

perimeter worst: Peor perímetro

Características de la suavidad

smoothness mean: Media de la suavidad de la superficie

smoothness se: Error estándar de la suavidad

smoothness worst: Peor suavidad

Características del área

area mean: Media del área de la célula

area se: Error estándar del área

area worst: Peor área

Características de la compacidad

compactness mean: Media de la compacidad

compactness se: Error estándar de la compacidad

compactness worst: Peor compacidad

Características de la concavidad

concavity mean: Media de la severidad de las concavidades

concavity se: Error estándar de la concavidad

concavity worst: Peor concavidad

Características de los puntos cóncavos

concave points mean: Media del número de puntos cóncavos

concave points se: Error estándar de los puntos cóncavos

concave points worst: Peor número de puntos cóncavos

Características de la simetría

symmetry mean: Media de la simetría

symmetry se: Error estándar de la simetría

symmetry worst: Peor simetría

Características de la dimensión fractal

fractal dimension mean: Media de la dimensión fractal

fractal dimension se: Error estándar de la dimensión fractal

fractal dimension worst: Peor dimensión fractal

3. Técnicas de Álgebra Lineal

3.1 *Regresión Logística*

La Regresión Logística es una técnica estadística utilizada para analizar y modelar la relación entre una variable dependiente categórica (generalmente binaria) y una o más variables independientes.

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

Dónde:

$$z = z(x, \beta) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Para buscar los parámetros $\beta_0, \beta_1 x_1, \ldots, \beta_n x_n$ el problema se reduce a buscar un $\beta*$ tal que:

$\sigma(\beta*) = \text{mín}_{\beta \in \mathbb{R}^{n+1}} \sigma(\beta)$

$$\nabla\sigma(\beta) = (\tfrac{\partial\sigma}{\partial\beta_0}, \tfrac{\partial\sigma}{\partial\beta_1}, \ldots, \tfrac{\partial\sigma}{\partial\beta_n})^T = \vec{0}_{\mathbb{R}^{n+1}}$$

que dan lugar al sistema de ecuaciones normales (SEN)

$\sum_{i=0}^{N} \{[s(x_i) - S(x_i, {\beta_0}^*, \ldots, {\beta_n}^*] \frac{\partial S}{\partial \beta_j}\} = 0$
$0 \leq j \leq n$

La solución del Sistema de Ecuaciones No Lineales es el vector $\beta*$ que constituye el único mínimo de σ y define la mejor función de aproximación mínimo cuadrática.

$$P(y = 1|X) = \sigma(\beta*)$$

Dónde $P(y = 1|X)$ es la probabilidad estimada de que la observación pertenezca a la clase positiva dadas sus características.

3.2 Descomposición en Valores Singulares (SVD)

El SVD permite mediante la descomposición de cualquier matriz A en el producto de tres matrices especiales, obtener otra cuyas dimensiones sean menor que la original y que mantenga la mayor cantidad posible de información relevante.

$$A = U\Sigma V^T$$

$U_{m\times m}$ Es una matriz ortogonal de vectores singulares izquierdos[1].

$\Sigma_{m\times n}$ Es una matriz diagonal de valores singulares no negativos y ordenados en forma decreciente. Cada valor singular indica la "importancia" o el peso de una dimensión en la matriz original. Los valores singulares más grandes representan dimensiones que capturan más varianza de los datos[1].

$V^T_{n\times n}$ Es una matriz ortogonal de dimensión $n \times n$ compuesta de autovectores de la matriz[1].

3.3 Modelo de Regresión Logística sin SVD

En el primer modelo se realizó directamente la Regresión Logística para clasificar el tumor. Se obtuvo resultados de 94 % de precisión, lo cual indica un alto índice de acierto.

	precision	recall	f1-score	support
0	0.92	0.96	0.94	47
1	0.97	0.94	0.95	67
macro avg	0.94	0.95	0.95	114
weighted avg	0.95	0.95	0.95	114
accuracy	-	-	0.95	114

Figura 1: Reportes del modelo de Regresión Logística ca sin SVD

3.4 Modelo de Regresión Logística con reducción SVD

En el segundo modelo para poder analizar el dataset de forma eficiente se redujo la dimensión de la matriz de datos. Se aplicó el SVD y se obtuvo las matrices laterales y un vector que contiene los valores singulares. Con los valores singulares se calculó la proporción de cada uno en relación con la suma total de todos estos y lo expresamos como porcentajes para entender cuanta "varianza" o "información" aporta cada componente en la descomposición, quedándonos la siguiente lista, donde se observa que en los primeros 4 valores es donde está la mayor cantidad de información.

[**87.986, 7.089, 2.516, 2.516, 1.586, 0.437, 0.163, 0.092, 0.041, 0.028, 0.020, 0.012, 0.006, 0.004, 0.003, 0.002, 0.001, 0.0013, 0.0013, 0.0009, 0.0008, 0.0006, 0.0005, 0.0004, 0.0003, 0.0002, 0.0002, 0.0001, 0.0001, 9.644645520411048e-05, 5.9235820488086806e-05**]

De esta manera, al probar reducir la dimensionalidad con estos valores identificamos que con k=3 y k=2 obteníamos los resultados más precisos al calcular la exactitud del modelo. De ahí que escogiéramos k=2, porque al disminuir el número de componentes se reduce la dimensionalidad de los datos, lo que puede mejorar la eficiencia en términos de tiempo de entrenamiento y predicción.

Por tanto se truncó la matriz con k=2 y se aproximó A para ser utilizada en el entrenamiento del modelo de regresión logística.

$$A \approx U_k \Sigma_k V^*$$

A continuación dividimos los datos reducidos en conjuntos de entrenamiento y prueba y al calcular el desempeño del modelo en datos no vistos obtuvimos un 94 % aproximadamente de exactitud.

	precision	recall	f1-score	support
0	0.96	0.91	0.93	47
1	0.94	0.97	0.96	67
macro avg	0.95	0.94	0.95	114
weighted avg	0.95	0.95	0.95	114
accuracy	-	-	0.95	114

Figura 2: Reportes del modelo de Regresión logística con SVD truncado en k=2

4. Conclusiones

Este estudio ha demostrado la eficacia de combinar técnicas avanzadas de aprendizaje automático y Álgebra Lineal en el análisis de datos médicos. Mediante la aplicación de SVD, logramos reducir la dimensionalidad de un conjunto de datos complejos con 30 características y 569 muestras, manteniendo la mayor parte de la información relevante. Posteriormente, el modelo de Regresión Logística, entrenado con estos datos reducidos, mostró una alta exactitud del 94 % en la clasificación de tumores, incluso igual que el modelo sin SVD. Esta combinación de métodos no solo optimiza el procesamiento de datos, sino que también mejora la precisión en la predicción de la naturaleza de los tumores. Los resultados subrayan el potencial de las técnicas de aprendizaje automático y análisis dimensional en la mejora de las herramientas de diagnóstico, proporcionando un avance significativo en la forma en que abordamos el cáncer de mama y otros problemas médicos complejos.

Referencias

[1] Luciano A Perez. *Descomposición SVD.*, 2015.

[2] Dimitrios Mitsotakis. B *Computational Mathematics An Introduction to Numerical Analysis and Scientific Computing with Python.*. 2023

[3] Datos. URL: https://www.kaggle.com/datasets/erdemtaha/can data.

[4] licencia. URL: https://creativecommons.org/licenses/by-nc-sa/4.0/bajoestalicencia.