Escuela de Ingeniería, UCN, Coquimbo (2023), 1–3

LABORATORIO NUMERO 07

Laboratorio 07 Regresion logistica

Daniel Olivares and Carlos Tapia

Abstract

La clasificación precisa de condiciones médicas mediante técnicas computacionales representa un avance significativo en el diagnóstico y tratamiento de enfermedades. En este laboratorio, se empleó la regresión logística, un método de aprendizaje automático, para diferenciar entre tumores benignos y malignos utilizando el conjunto de datos de cáncer de mama de Wisconsin. La implementación del modelo y la selección de hiperparámetros se realizaron en Python, utilizando bibliotecas como NumPy, Pandas y Scikit-learn. Los resultados destacaron la capacidad del modelo para clasificar con alta precisión, lo que demuestra su potencial en aplicaciones médicas diagnósticas.

Accurate classification of medical conditions using computational techniques represents a significant advancement in disease diagnosis and treatment. In this laboratory, logistic regression, a machine learning method, was employed to distinguish between benign and malignant tumors using the Wisconsin Breast Cancer dataset. The model implementation and hyperparameter selection were performed in Python, utilizing libraries such as NumPy, Pandas, and Scikit-learn. The results highlighted the model's ability to classify with high precision, demonstrating its potential in diagnostic medical applications.

Introducción

La capacidad de clasificar con precisión las condiciones médicas es un componente esencial en la toma de decisiones informadas dentro del campo de la medicina.Nos enfocamos en la aplicación de técnicas de regresión logística, una metodología de aprendizaje automático, para diferenciar entre tumores benignos y malignos.

A través del experimento, pusimos a prueba la hipótesis de que un modelo bien ajustado y entrenado con datos relevantes y preprocesados adecuadamente, sería capaz de clasificar con alta precisión las muestras proporcionadas. Además, se buscaron respuestas a las preguntas fundamentales relacionadas con la influencia de diferentes variables y cómo la selección de características puede impactar la efectividad del modelo.

Marco teórico

El marco teórico del presente laboratorio se centra en la regresión logística, un modelo estadístico que se aplica a problemas de clasificación binaria.

Regresión Logística

La regresión logística es una extensión del modelo lineal que utiliza una función logística para modelar una variable dependiente binaria. En el contexto de nuestro estudio, la variable dependiente es la presencia de tumores malignos (valor 1) o benignos (valor 0). Matemáticamente, el modelo se expresa como:

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}} \tag{1}$$

donde $P(Y = 1)$ es la probabilidad de que la variable de respuesta sea 1 (tumor maligno), e es la base del logaritmo natural, β_0 es el término de intersección, $\beta_1, \ldots, \beta_n$ son los coeficientes de las variables independientes $X_1, \ldots, X_n$.

Hiperparámetros en la Regresión Logística

Un hiperparámetro es un parámetro cuyo valor se establece antes del proceso de aprendizaje y se utiliza para controlar el proceso de aprendizaje. En la regresión logística, los hiperparámetros más relevantes incluyen:

- **La tasa de regularización** (C): Controla la complejidad del modelo, penalizando los valores grandes de los coeficientes para evitar el sobreajuste.
- **El tipo de penalización**: Determina si se utiliza L1 (lasso), L2 (ridge) o elastic net como término de regularización en la función de pérdida.
- **El número máximo de iteraciones** (*max_iter*): Establece el número de pasos que el algoritmo de optimización puede tomar para converger a los mejores coeficientes.
- **El algoritmo de optimización** (*solver*): Especifica el método utilizado para minimizar la función de pérdida.

Selección de Hiperparámetros

La selección de hiperparámetros es un paso crítico en la construcción de modelos de aprendizaje automático. En este laboratorio, se utilizó tanto la búsqueda en cuadrícula (*Grid Search*) como la búsqueda aleatoria (*Random Search*) para encontrar la combinación óptima de hiperparámetros que maximiza la precisión del modelo.

La **búsqueda en cuadrícula** implica probar exhaustivamente todas las combinaciones posibles de hiperparámetros en una cuadrícula predeterminada, mientras que la **búsqueda aleatoria** selecciona combinaciones al azar en un espacio de hiperparámetros definido. Ambas estrategias tienen como objetivo mejorar la capacidad predictiva y la eficiencia del modelo de regresión logística aplicado a la clasificación de tumores.

Materiales, procedimiento experimental y resultados

Para el desarrollo del laboratorio se utilizó el lenguaje de programación Python, versión 3.8. Se emplearon diversas bibliotecas especializadas para el análisis y modelado de datos, que incluyen:

- **NumPy**: Para la manipulación eficiente de arrays numéricos.
- **Pandas**: Proporciona estructuras de datos y herramientas de análisis.
- **Scikit-learn**: Utilizada para aplicar técnicas de aprendizaje automático como la regresión logística y para dividir los datos, normalizarlos, y realizar la selección de hiperparámetros.
- **Seaborn y Matplotlib**: Bibliotecas de visualización de datos para la creación de gráficos.

El conjunto de datos utilizado es el conocido *Breast Cancer Wisconsin (Diagnostic) Data Set*, que contiene mediciones de imágenes digitalizadas de masas mamarias y se utiliza para predecir si son benignas o malignas.

Procedimiento del Experimento

El experimento consistió en los siguientes pasos:

1. Carga y preprocesamiento del conjunto de datos utilizando Pandas.
2. División del conjunto de datos en subconjuntos de entrenamiento y prueba con una proporción de 80-20.
3. Normalización de los datos para garantizar una escala uniforme entre las características.
4. Entrenamiento del modelo de regresión logística con los datos normalizados.
5. Ajuste de hiperparámetros mediante búsqueda en cuadrícula y búsqueda aleatoria para mejorar el rendimiento del modelo.
6. Evaluación del modelo utilizando la matriz de confusión y el informe de clasificación.

Resultados

Los resultados obtenidos demostraron un alto nivel de precisión en la clasificación. La matriz de confusión reveló una cantidad mínima de falsos negativos y cero falsos positivos. El informe de clasificación proporcionó valores de precisión, recall y f1-score cercanos a 1, lo que indica un rendimiento excepcional del modelo. Las visualizaciones de las características importantes mostraron cuáles tenían más peso en la predicción del modelo, y las distribuciones de las características seleccionadas, como *radius_mean* y *texture_mean*, mostraron diferencias distintas entre las clases.

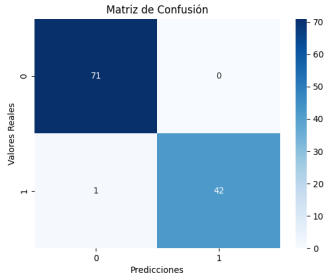

Figure 1. Matriz de confusión del modelo de regresión logística.

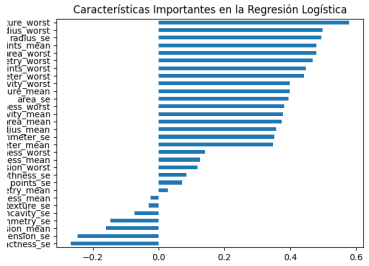

Figure 2. Características importantes determinadas por el modelo.

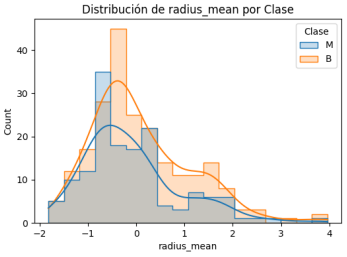

Figure 3. Distribución de la característica *radius_mean* por clase.

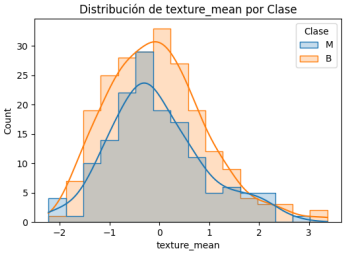

Figure 4. Distribución de la característica *texture_mean* por clase.

Discusión de resultados

Los resultados obtenidos en el laboratorio revelan una precisión significativa en la clasificación del conjunto de datos de cáncer de mama. La matriz de confusión (Figura 1) indica que el modelo fue capaz de identificar correctamente 71 de 71 tumores benignos y 42 de 43 tumores malignos. Esta alta tasa de verdaderos positivos y verdaderos negativos, con solo un falso negativo y ningún falso positivo, resalta la capacidad del

modelo para diferenciar entre las clases con una confiabilidad sobresaliente.

El informe de clasificación proporciona un desglose más detallado del rendimiento del modelo. Las medidas de precisión, recall y f1-score para ambas clases se acercan a 1, indicando que hay una armonía entre la sensibilidad y la especificidad del modelo. Estos valores sugieren que el modelo es igualmente bueno tanto para identificar la presencia de tumores malignos como para descartar la de tumores benignos.

La importancia de las características, como se muestra en la Figura 2, brinda una perspectiva sobre qué variables tienen más influencia en la predicción del modelo. Por ejemplo, las características que miden el peor radio y la peor textura (worst radius y worst texture) tienen los coeficientes más altos, lo que implica que juegan un papel crítico en la diferenciación de tumores malignos de los benignos. Estos hallazgos están en consonancia con la literatura médica que señala estas medidas como indicadores clave en la progresión y gravedad del cáncer de mama.

Las gráficas de distribución para *radius_mean* y *texture_mean* (Figuras 3 y 4) muestran una separación clara entre las clases, lo que indica que estas características son distintivas y contribuyen significativamente a la precisión del modelo. La superposición mínima entre las distribuciones de estas características sugiere que son factores predictivos robustos y confiables.

Este análisis subraya la importancia de la selección de hiperparámetros cuidadosa y considerada. La elección del hiperparámetro C y del tipo de penalización influye directamente en la capacidad del modelo para generalizar y evitar el sobreajuste. En este experimento, se encontró que un C de 0.1 con una penalización $l2$ y el uso de 'liblinear' como solver ofrecían el mejor rendimiento.

En resumen, los resultados respaldan la hipótesis de que un modelo de regresión logística, adecuadamente ajustado y entrenado con un conjunto de datos representativo, puede clasificar con éxito y alta precisión las muestras de cáncer de mama. Los datos analizados sugieren que el modelo no solo es efectivo sino también robusto, lo que es esencial en aplicaciones médicas donde la precisión es crítica.

Conclusiones

Los resultados obtenidos en este laboratorio validan la efectividad de la regresión logística para la clasificación de tumores como benignos o malignos. El análisis de la matriz de confusión y las métricas de rendimiento evidencian un modelo altamente preciso y confiable. Las características más influyentes identificadas corroboran con los indicadores conocidos en la literatura médica. La selección de hiperparámetros demostró ser un paso fundamental para optimizar el rendimiento del modelo. Estos hallazgos subrayan la importancia de la aplicación de técnicas de aprendizaje automático en el ámbito de la salud, especialmente en la detección temprana del cáncer de mama.