Aprendizaje Automático II: Tarea 1

Sergio Soler Rocha
Universidad Nacional de Educación a Distancia

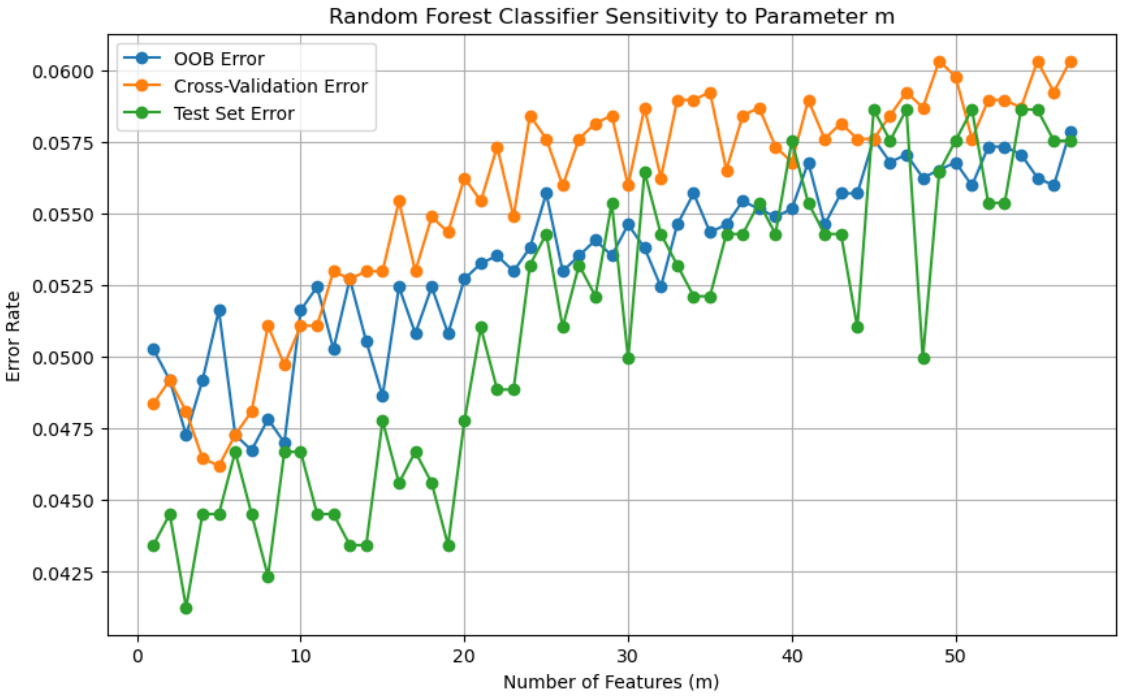

Figura 1: Sensibilidad del clasificador de bosque aleatorio al parámetro m.

1. Introducción

En este ejercicio, se llevará a cabo un análisis utilizando clasificadores de bosques aleatorios (random forest) sobre los datos de correos electrónicos SPAM proporcionados. El objetivo principal es explorar la sensibilidad del parámetro m en el rendimiento del modelo. Para ello, se ajustarán una serie de clasificadores de bosques aleatorios variando el valor de m, y se evaluará su desempeño utilizando tres métricas diferentes: el error out-of-bag (OOB, que es una medida de la precisión del modelo que se calcula utilizando muestras de datos que no se incluyeron en la construcción de cada árbol individual del bosque), el error de prueba y el error de validación cruzada.

Para visualizar la relación entre el parámetro m y el rendimiento del modelo, se generará una gráfica 1 que muestren el error OOB, el error de prueba (el conjunto de prueba ocupará el 20 % de los datos) y el error de validación cruzada (calculado utilizando 5 pliegues) en función de diferentes valores de m seleccionados(desde 1 hasta 57, que es el número máximo de características). Este análisis proporcionará información valiosa sobre cómo el ajuste del parámetro m afecta la capacidad predictiva del modelo de bosques aleatorios en la clasificación de correos electrónicos como SPAM o no SPAM.

2. Análisis de la gráfica

Al analizar la gráfica, se puede observar que, en términos generales, los tres errores estudiados muestran un comportamiento similar. Esto significa que existe una primera zona, aproximadamente entre m igual a 1 y m igual a 10, donde se encuentran los valores más bajos del error. A partir de ahí, se puede observar un crecimiento fluctuante del error a medida que se aumenta el valor de m.

El hecho de que los tres errores muestren una tendencia similar sugiere que el modelo se comporta de manera coherente en diferentes conjuntos de datos (entrenamiento, prueba y validación cruzada). Esto indica que el modelo es robusto y no está influenciado significativamente por la partición específica de los datos utilizados en cada conjunto. El error de validación cruzada es generalmente considerado como una estimación más confiable del rendimiento del modelo en datos no vistos, ya que utiliza múltiples particiones de los datos para evaluar el rendimiento. Dado que el error de validación cruzada sigue una tendencia similar al error OOB, significaría que el error OOB también puede ser una buena estimación del rendimiento del modelo en datos no vistos. La consistencia en la tendencia de los errores OOB y de validación cruzada indica que el modelo no está sobreajustando significativamente los datos de entrenamiento, ya que el error no muestra grandes discrepancias entre estos dos conjuntos de datos.

Otro punto importante a señalar es que el valor comúnmente recomendado para el parámetro m en problemas de clasificación es la raíz cuadrada del número de características [1]. En este caso al tener 57 características el valor redondeado sería 8. Analizando la gráfica, vemos que el error OOB, el error de prueba y el error de validación cruzada no necesariamente se minimizan en el punto donde m es igual a la raíz cuadrada del número de características. Sin embargo, se puede apreciar que hay una cierta estabilidad en los errores alrededor de ese valor y el error tiende a aumentar ligeramente a medida que m se aleja de él, lo que puede justificar que la elección de este valor podría ser una buena decisión.

3. Consideraciones sobre el uso de otros parámetros

Otros parámetros que se pueden considerar son:

- **Número de árboles (n_estimators)**: El número de árboles en el bosque aleatorio puede afectar significativamente el rendimiento del modelo. En el caso que nos ocupa hemos utilizado la cantidad de 100 árboles.
- **Profundidad máxima del árbol (max_depth)**: La profundidad máxima de los árboles en el bosque puede influir en la capacidad del modelo para capturar relaciones complejas en los datos. En nuestro caso se expanden hasta que todas las hojas sean puras o hasta que contengan menos de min_samples_split muestras.
- **Número mínimo de muestras requeridas para dividir un nodo interno (min_samples_split)**: Este parámetro controla la cantidad mínima de muestras requeridas para dividir un nodo interno. Utiliza el valor predeterminado, que es 2.
- **Número mínimo de muestras requeridas para estar en un nodo hoja (min_samples_leaf)**: Este parámetro controla la cantidad mínima de muestras requeridas para que un nodo sea considerado una hoja. Utilizamos el valor predeterminado, que es 1.

4. Eficiencia computacional del parámetro m

El parámetro m puede influir en la eficiencia computacional del algoritmo de bosques aleatorios al afectar el tiempo de entrenamiento, ya que, cuantas más características se consideren en cada división, más cálculos serán necesarios para evaluar todas las posibles divisiones y seleccionar la mejor. El uso de memoria, si el número de características es alto, se necesitará más memoria para almacenar los datos necesarios para calcular las divisiones de los nodos en los árboles. El riesgo de sobreajuste del modelo, Un valor bajo de m puede ayudar a reducir el sobreajuste del modelo, ya que limita la cantidad de características consideradas en cada división, lo que puede llevar a árboles más simples y menos propensos al sobreajuste.

En el caso que nos ocupa observamos una tendencia de aumento de tiempo de entrenamiento de forma lineal a medida que se aumenta el valor de m, Figura 2.

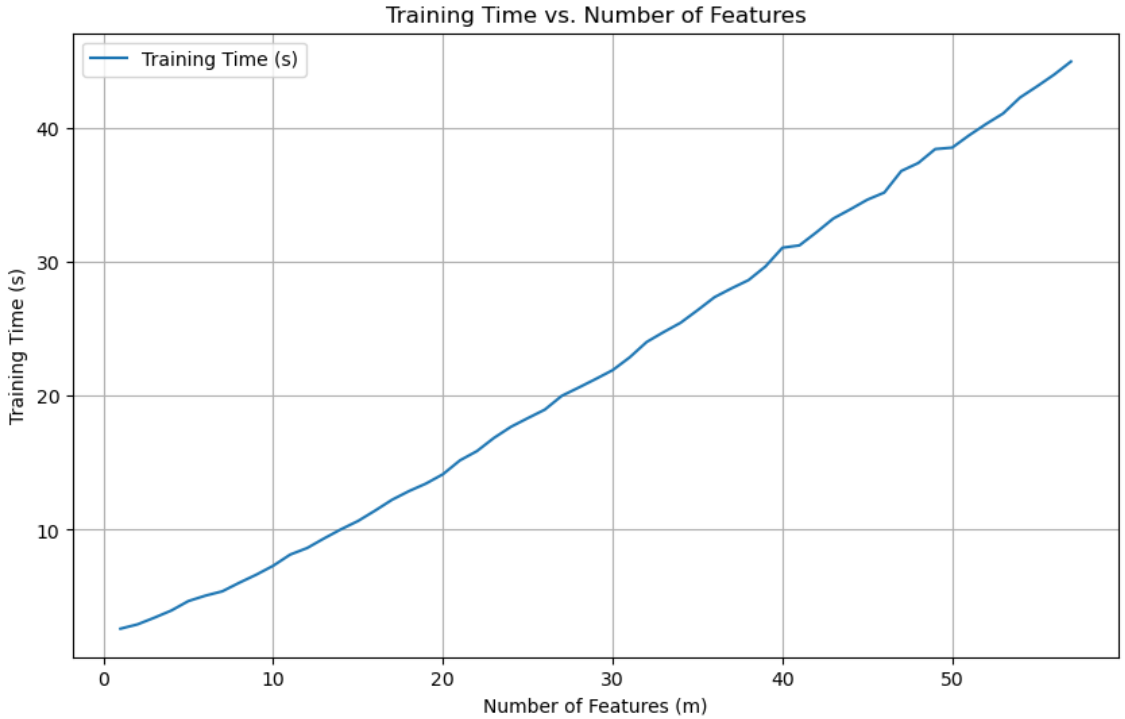

Figura 2: Tiempo de entrenamiento vs número máximo de características

Referencias

[1] Trevor Hastie, Robert Tibshirani, Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second Edition. Springer, Enero, 13, 2017.

[2] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Second Edition. Springer, Junio 21, 2023.

[3] *Random Forest: Bosque aleatorio. Definición y funcionamiento.* 25 de Ene, 2024. Disponible en: https://datascientest.com/es/random-forest-bosque-aleatorio-definicion-y-funcionamiento