Introducción a Inteligencia Artificial
Detección de SPAM

Gonzalo Gabriel Fernandez*, Roberto Enrique Castro Beltran†
Carrera de Especialización en Inteligencia Artificial
Universidad de Buenos Aires
Email: *fernandez.gfg@gmail.com, †electrobot73@gmail.com

Resumen

Uno de los problemas más comunes en la clasificación es la detección de correos electrónicos SPAM. Uno de los primeros modelos utilizados para abordar este problema fue el clasificador de Bayes ingenuo. La detección de SPAM es un problema persistente en el mundo digital, ya que los spammers continúan adaptando sus estrategias para eludir los filtros de correo no deseado. Además del clasificador de Bayes ingenuo, se han desarrollado y utilizado una variedad de técnicas más avanzadas en la detección de SPAM, que incluyen algoritmos de aprendizaje automático, redes neuronales y métodos basados en reglas.

I. Descripción del problema

En este trabajo práctico, utilizaremos un conjunto de datos que consta de 4601 observaciones de correos electrónicos, de los cuales 2788 son correos legítimos y 1813 son correos SPAM. Dado que el contenido de los correos electrónicos es un tipo de dato no estructurado, es necesario procesarlo de alguna manera. Para este conjunto de datos, ya se ha aplicado un procesamiento típico en el Procesamiento del Lenguaje Natural (NLP), que consiste en contar la frecuencia de palabras observadas en los correos.

El procesamiento de lenguaje natural (NLP) desempeña un papel fundamental en la detección de SPAM, ya que permite analizar el contenido de los correos electrónicos y extraer características relevantes para la clasificación. Además de contar la frecuencia de palabras, se pueden utilizar técnicas más sofisticadas, como la extracción de características semánticas y el análisis de sentimientos, para mejorar la precisión de los modelos de detección de SPAM.

En este proceso, se cuenta la cantidad de ocurrencias de cada palabra en los diferentes correos.

Con el fin de preservar la privacidad de los mensajes, la frecuencia de palabras se encuentra normalizada. El conjunto de datos está compuesto por 54 columnas de atributos que se denominan:

word_freq_XXXX: Donde XXXX es la palabra o símbolo. Los valores son enteros que van de 0 a 20k. Además, hay una columna adicional llamada spam, que es 1 si el correo es SPAM o 0 si no lo es.

Los clasificadores de Bayes ingenuos fueron los primeros filtros utilizados por las aplicaciones de correo electrónico, basados en este principio de palabras. La idea es que, partiendo de un dato a priori sobre la probabilidad de que un correo sea SPAM o no, ciertas palabras nos indicarán que la probabilidad a posteriori, dadas esas palabras, es más probable que el correo sea SPAM o no.

Para este trabajo práctico, se proporciona una notebook (ayuda.ipynb) con la lectura del conjunto de datos, la separación de los datos, entre otras ayudas para resolverlo.

Análisis del conjunto de datos

Las 10 palabras más encontradas en correos SPAM son:

- 'you'
- 'your'
- 'will'
- 'free'
- 'our'
- '!'
- 'all'
- 'mail'
- 'email'
- 'business'

Las 10 palabras más encontradas en correos no SPAM son:

- 'you'
- 'george'
- 'hp'
- 'will'
- 'your'
- 'hpl'
- 're'
- 'edu'
- 'address'
- 'meeting'

Las palabras en común son: 'you', 'will' y 'your'.

Clasificador de Bayes ingenuo y de regresión logística

Del conjunto total de datos se obtuvo un conjunto de entrenamiento y un conjunto de prueba (70 % y 30 % respectivamente).

Para clasificar los datos se utilizaron dos modelos: un clasificador de Bayes ingenuo y uno de regresión logística. Para ambos se utilizó la biblioteca *scikit learn*.

Se entrenó ambos modelos con el conjunto de datos para entrenamiento.

Análisis de tipo de error

En la figura 1 se puede observar las matrices de confusión para los dos clasificadores utilizados.

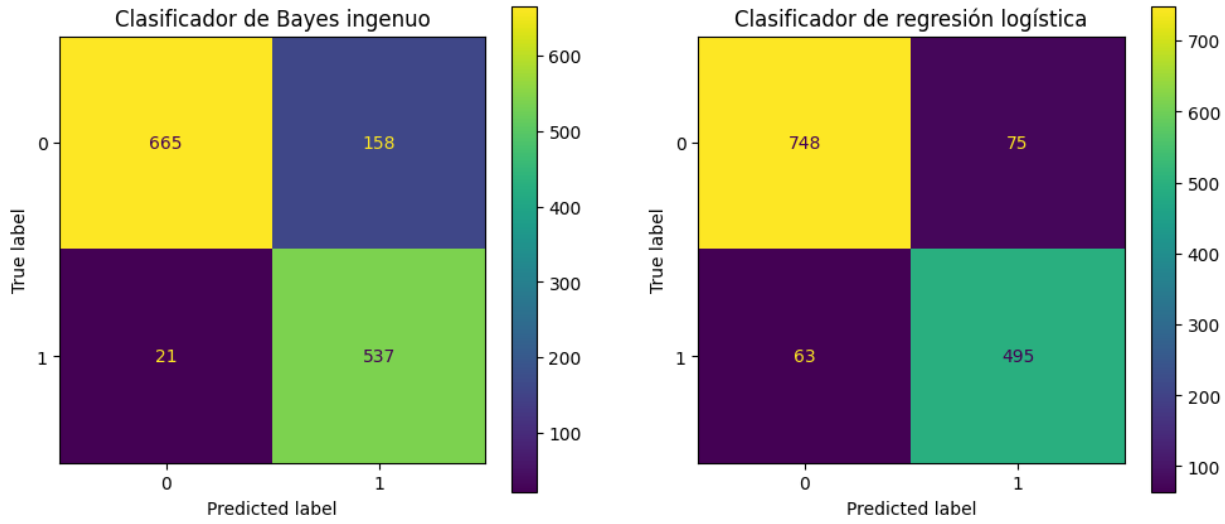

Figura 1. Matriz de confusión para modelos entrenados

De la figura 1 se observa que:

- El clasificador de Bayes ingenuo comete más errores de tipo I (falso positivo).
- El clasificador de regresión logística comete una ligera mayor cantidad de errores de tipo I.

Errores de tipo I y tipo II en detección de Spam:

- Error de tipo I (falso positivo):
  - Un correo legítimo es clasificado como spam.
  - Consecuencia: El usuario podría perder correos importantes que fueron enviados a la carpeta de spam.
- Error de tipo II (falso negativo):
  - Un correo spam es clasificado como legítimo.
  - Consecuencia: El usuario recibe correos no deseados en su bandeja de entrada, lo que puede ser molesto y potencialmente peligroso si contiene phishing o malware.

Para el problema de detección de spam, los errores de tipo I (falsos positivos) son más importantes porque implican la pérdida de correos legítimos, lo cual puede tener consecuencias significativas para el usuario.

Resultados obtenidos

En la tabla I se observa la precisión y la recuperación obtenida para cada clasificador utilizado.

	Precisión	Recuperación
Clasificador de Bayes ingenuo	0.77	0.96
Clasificador de regresión logística	0.86	0.88

Cuadro I
Resultados de precisión y recuperación para los modelos utilizados.

Las métricas evalúan lo siguiente:

- La precisión mide la proporción de verdaderos positivos sobre el total de predicciones positivas. Una alta precisión significa pocos falsos positivos.
- La recuperación mide la proporción de verdaderos positivos sobre el total de verdaderos positivos y falsos negativos. Una alta recuperación significa pocos falsos negativos.

Ambos modelos presentaron errores de tipo I (falsos positivos) en mayor proporción, sin embargo, el clasificador de Bayes ingenuo tuvo un mayor numero de errores de este tipo, es por ello que su precisión es menor comparado con la regresión logística.

Curva ROC

En la figura 2 se observa la curva ROC y el AUC (Área Bajo la Curva ROC) de ambos clasificadores.

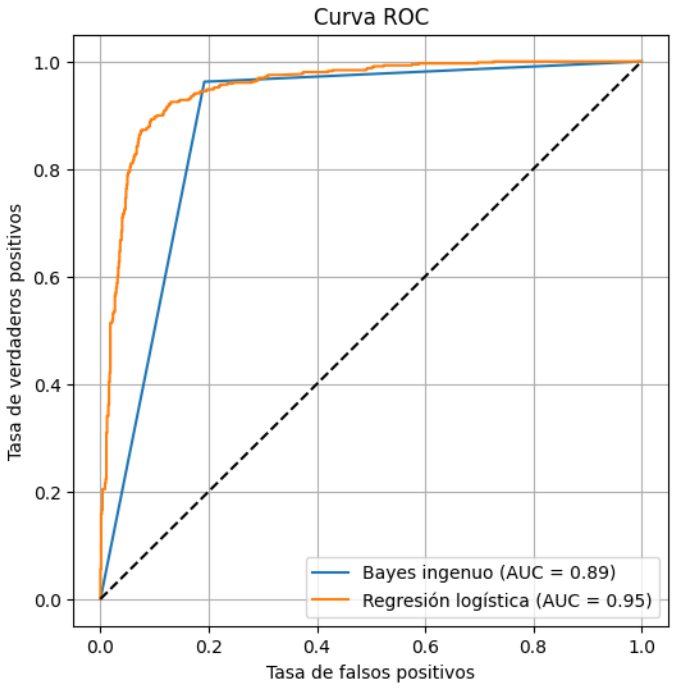

Figura 2. Curva ROC y AUC para los modelos utilizados.