Demostración Matemática para un Filtro de Spam en Español (Naive Bayes)

1 Conjunto de Datos de Entrenamiento

Mensajes de Spam:

- Gana dinero rápido
- Reclama tu premio

Mensajes de Ham:

- Reunión a las 3PM
- Actualización del proyecto necesaria

Mensaje de prueba: **Reclama tu dinero premio**

2 Construir el Vocabulario

El vocabulario es el conjunto de todas las palabras únicas en los mensajes:

- gana, dinero, rápido, reclama, tu, premio, reunión, a, las, 3PM, actualización, del, proyecto, necesaria

Tamaño del vocabulario (V) = 14

3 Contar las Palabras en Cada Clase

Conteo de palabras en Spam:

- gana: 1, dinero: 1, rápido: 1, reclama: 1, tu: 1, premio: 1

Total de palabras en spam = 6

Conteo de palabras en Ham:

- reunión: 1, a: 1, las: 1, 3PM: 1, actualización: 1, proyecto: 1, necesaria: 1

Total de palabras en ham = 7

4 Calcular las Priori (Priors)

$$P(\text{Spam}) = \frac{2}{4} = 0.5$$
$$P(\text{Ham}) = \frac{2}{4} = 0.5$$

5 Calcular las Probabilidades Condicionales con Suavizado de Laplace

$$P(\text{palabra—spam}) = \frac{\text{Conteo en spam} + 1}{\text{Total en spam} + V}$$
$$P(\text{palabra—ham}) = \frac{\text{Conteo en ham} + 1}{\text{Total en ham} + V}$$

Cálculos para el mensaje de prueba **Reclama tu dinero premio**:

Palabra	Conteo en Spam	P(palabra—spam)	P(palabra—ham)
reclama	1	$\frac{2}{20} = 0.1$	$\frac{1}{21} \approx 0.048$
tu	1	$\frac{2}{20} = 0.1$	$\frac{1}{21} \approx 0.048$
dinero	1	$\frac{2}{20} = 0.1$	$\frac{1}{21} \approx 0.048$
premio	1	$\frac{2}{20} = 0.1$	$\frac{1}{21} \approx 0.048$

Table 1: Probabilidades condicionales con suavizado de Laplace

6 Calcular las Probabilidades Conjuntas (Sin Logs)

$$P(\text{Spam—Mensaje}) = 0.5 \times (0.1)^4 = 0.5 \times 0.0001 = 0.00005$$
$$P(\text{Ham—Mensaje}) = 0.5 \times (0.048)^4 \approx 0.5 \times 0.000005 = 0.0000024$$

7 Calcular las Probabilidades Conjuntas (Con Logs)

$$\log P(\text{Spam—Mensaje}) = \log 0.5 + 4 \times \log 0.1 = -9.905$$
$$\log P(\text{Ham—Mensaje}) = \log 0.5 + 4 \times \log 0.048 = -12.833$$

8 Comparar los Resultados

- Spam (Log) = -9.905
- Ham (Log) = -12.833

El número menos negativo es más alto, por lo que el modelo clasifica este mensaje como spam.

9 Conclusión

El modelo clasifica correctamente el mensaje como spam porque las palabras "reclama", "tu", "dinero" y "premio" son más comunes en los mensajes de spam del conjunto de entrenamiento.